Online health search via multi-dimensional information quality assessment based on deep language models

Zhang, Boya; Naderi, Nona; Mishra, Rahul; Teodoro, Douglas

How to cite

ZHANG, Boya et al. Online health search via multi-dimensional information quality assessment based on deep language models. In: JMIR AI, 2024, vol. 3, p. e42630. doi: 10.2196/42630

This publication URL: https://archive-ouverte.unige.ch/unige:178156
Publication DOI: 10.2196/42630

Original Paper

Online Health Search Via Multidimensional Information Quality Assessment Based on Deep Language Models: Algorithm Development and Validation

Boya Zhang[1], MSc; Nona Naderi[2], PhD; Rahul Mishra[1], PhD; Douglas Teodoro[1], PhD

[1]Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland

[2]Department of Computer Science, Université Paris-Saclay, Centre national de la recherche scientifique, Laboratoire Interdisciplinaire des Sciences du Numérique, Orsay, France

Corresponding Author:
Boya Zhang, MSc
Department of Radiology and Medical Informatics
University of Geneva
9 Chemin des Mines
Geneva, 1202
Switzerland
Phone: 41 782331908
Email: boya.zhang@unige.ch

Abstract

Background: Widespread misinformation in web resources can lead to serious implications for individuals seeking health advice. Despite that, information retrieval models are often focused only on the query-document relevance dimension to rank results.

Objective: We investigate a multidimensional information quality retrieval model based on deep learning to enhance the effectiveness of online health care information search results.

Methods: In this study, we simulated online health information search scenarios with a topic set of 32 different health-related inquiries and a corpus containing 1 billion web documents from the April 2019 snapshot of Common Crawl. Using state-of-the-art pretrained language models, we assessed the quality of the retrieved documents according to their usefulness, supportiveness, and credibility dimensions for a given search query on 6030 human-annotated, query-document pairs. We evaluated this approach using transfer learning and more specific domain adaptation techniques.

Results: In the transfer learning setting, the usefulness model provided the largest distinction between help- and harm-compatible documents, with a difference of +5.6%, leading to a majority of helpful documents in the top 10 retrieved. The supportiveness model achieved the best harm compatibility (+2.4%), while the combination of usefulness, supportiveness, and credibility models achieved the largest distinction between help- and harm-compatibility on helpful topics (+16.9%). In the domain adaptation setting, the linear combination of different models showed robust performance, with help-harm compatibility above +4.4% for all dimensions and going as high as +6.8%.

Conclusions: These results suggest that integrating automatic ranking models created for specific information quality dimensions can increase the effectiveness of health-related information retrieval. Thus, our approach could be used to enhance searches made by individuals seeking online health information.

(JMIR AI 2024;3:e42630) doi: 10.2196/42630

KEYWORDS

health misinformation; information retrieval; deep learning; language model; transfer learning; infodemic

Introduction

In today's digital age, individuals with diverse information needs, medical knowledge, and linguistic skills [1] turn to the web for health advice and to make treatment decisions [2]. The mixture of facts and rumors in online resources [3] makes it challenging for users to discern accurate content [4]. To provide high-quality resources and enable properly informed decision-making [5], information retrieval systems should

differentiate between accurate and misinforming content [6]. Nevertheless, search engines rank documents mainly by their relevance to the search query [7], neglecting several health information quality concerns. Moreover, despite attempts by some search engines to combat misinformation [8], they lack transparency in terms of the methodology used and performance evaluation.

Health misinformation is defined as health-related information that is inaccurate or misleading based on current scientific evidence [9,10]. Due to the lack of health literacy for nonprofessionals [11] and the rise of the infodemic phenomenon [12]—the rapid spread of both accurate and inaccurate information about a medical topic on the internet [13]—health misinformation has become increasingly prevalent online. Topics related to misinformation, such as "vaccine" or "the relationship between coronavirus and 5G" have gained scientific interest across social media platforms like Twitter and Instagram [14-16] and among various countries [17]. Thus, the development of new credibility-centered search methods and assessment measures is crucial to address the pressing challenges in health-related information retrieval [18].

In recent years, numerous approaches have been introduced in the literature to categorize and assess misinformation according to multiple dimensions. Hesse et al [19] proposed 7 dimensions of *truthfulness*, which include *correctness*, *neutrality*, *comprehensibility*, *precision*, *completeness*, *speaker trustworthiness*, and *informativeness*. On the other hand, van der Linden [20] categorized an infodemic into 3 key dimensions: *susceptibility*, *spread*, and *immunization*. Information retrieval shared tasks, such as the Text Retrieval Conference (TREC) and the Conference and Labs of the Evaluation Forum (CLEF), have also started evaluating quality-based systems for health corpora using multiple dimensions [21,22]. The CLEF eHealth Lab Series proposed a benchmark to evaluate models according to the *relevance*, *readability*, and *credibility* of the retrieved information [23]. The TREC Health Misinformation Track 2021 proposed further metrics of *usefulness*, *supportiveness*, and *credibility* [24]. These dimensions also appear in the TREC Health Misinformation Track 2019 as *relevancy*, *efficacy*, and *credibility*, respectively. Additionally, models by Solainayagi and Ponnusamy [25] and Li et al [26] incorporated similar dimensions, emphasizing source *reliability* and the *credibility* of statements. These metrics represent some of the initial efforts to quantitatively assess the effectiveness of information retrieval engines in sourcing high-quality information, marking a shift from the traditional query-document relevance paradigm [27,28]. Despite their variations, these information quality metrics focus on the following 3 main common topics: (1) *relevancy* (also called *usefulness* or *informativeness*) of the source to the search topic, (2) *correctness* (also called *supportiveness* or *efficacy*) of the information according to the search topic, and (3) *credibility* (also called *trustworthiness*) of the source.

Thanks to these open shared tasks, several significant methodologies have been developed to improve the search for higher-quality health information. Although classical bag-of-words–based methods outperform neural network approaches in detecting health-related misinformation when training data are limited [29], more advanced approaches are needed for web content. Specifically, research has proven the effectiveness of a hybrid approach that integrates classical handcrafted features with deep learning [18]. Further to this, multistage ranking systems [30,31], which couple the system with a label prediction model or use T5 [32] to rerank Okapi Best Match 25 (BM25) results, have been proposed. Particularly, Lima et al [30] considered the stance of the search query and engaged 2 assessors for an interactive search, integrating a continuous active learning method [33]. This approach sets a baseline of human effort in separating helpful from harmful web content. Despite their success, these models often do not take into account the different information quality aspects in their design.

In this study, we aimed to investigate the impact of multidimensional ranking on improving the quality of retrieved health-related information. Due to its coverage of the main information quality dimensions used in the scientific literature, we followed the empirical approach proposed in the TREC 2021 challenge, which considers *usefulness*, *supportiveness*, and *credibility* metrics, to propose a multidimensional ranking model. Using deep learning–based pretrained language models [34] through transfer learning and domain adaption approaches, we categorized the retrieved web resources according to different information quality dimensions. Specialized quality-oriented ranks obtained by reranking components were then fused [32] to provide the final ranked list. In contrast to prior studies, our approach relied on the automatic detection of harmful (or inaccurate) claims and used a multidimensional information quality model to boost helpful resources.

The main contributions of this work are 3-fold. We propose a multidimensional ranking model based on transfer learning and showed that it achieves state-of-the-art in automatic (ie, when the query stance is not provided) quality-centered ranking evaluations. We investigated our approach in 2 learning settings—transfer learning (ie, without query relevance judgments) and domain adaptation (ie, with query relevance judgments from a different corpus)—and demonstrated that they are capable of identifying more helpful documents than harmful ones, obtaining +5% and +7% help and harm compatibility scores, respectively. Last, we investigated how the combination of models specialized in different information dimensions impacts the quality of the results, and our analysis suggests that multidimensional aspects are crucial for extracting high-quality information, especially for unhelpful topics.

Methods

In this section, we introduce our search model based on multidimensional information quality aspects. We first describe the evaluation benchmark. We then detail the implementation methodology and describe our evaluation experiments using transfer learning and domain adaptation strategies.

TREC Health Misinformation Track 2021 Benchmark

Benchmark Data Set

To evaluate our approach, we used the TREC Health Misinformation Track 2021 benchmark [35] organized by the National Institute of Standards and Technology (NIST) [36].

The TREC Health Misinformation Track 2021 benchmark simulates web searches for specific health issues and interventions against a collection of English web documents [37]. For each topic, the benchmark annotates the quality of the retrieved web documents using a pooling approach, in which the top retrieved documents by systems participating in the challenge are evaluated according to their usefulness, correctness, and credibility and subsequently labeled as helpful or harmful. In this context, helpful documents are defined as those supportive of helpful treatments or that try to dissuade the reader from using unhelpful treatments, while harmful documents encourage the use of unhelpful treatments or dissuade the reader from using helpful treatments [24]. See Table S1 in Multimedia Appendix 1 for more detail on the annotation.

Health-Related Topics

A topic in the TREC Health Misinformation Track 2021 benchmark 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. NIST only provided assessments for 35 of the initial 50 topics. Among the assessed topics, 3 were further excluded due to the absence of harmful documents. Consequently, the benchmark consisted of 32 topics: 14 labeled as helpful and 18 labeled as unhelpful. For these queries, a total of 6030 query-document pairs were human-annotated according to different scales of usefulness, correctness, and credibility scores. A "helpful topic" refers to an intervention beneficial for treating a health issue, while an "unhelpful topic" indicates an ineffective intervention. The stance is supported by evidence from a credible source. Table 1 presents examples of the queries and descriptions of helpful and unhelpful topics.

Table 1. Examples of helpful and unhelpful topics with query and description.

Number	Query	Description	Stance
106	vitamin b12 sun exposure vitiligo	Can vitamin b12 and sun exposure together help treat vitiligo?	Helpful
102	tepid sponge bath reduce fever children	Is a tepid sponge bath a good way to reduce fever in children?	Unhelpful

Web Corpus

We used the Colossal Clean Crawled Corpus (C4), a collection of English-language web documents sourced from the public Common Crawl web scrape [38]. The corpus comprises 1 billion English documents from the April 2019 snapshot. To illustrate the contradictory nature of the web information within the corpus, in Table 2, we present 2 documents relevant to topic 102: "tepid sponge bath reduce fever in children." Although an article advises against the intervention ("Do Not Use Sponging to Reduce a Fever"), another article advises it could be a viable option ("Sponging is an option for high fevers").

Table 2. Examples of useful but contradictory documents for Topic 102: "Is a tepid sponge bath a good way to reduce fever in children?".

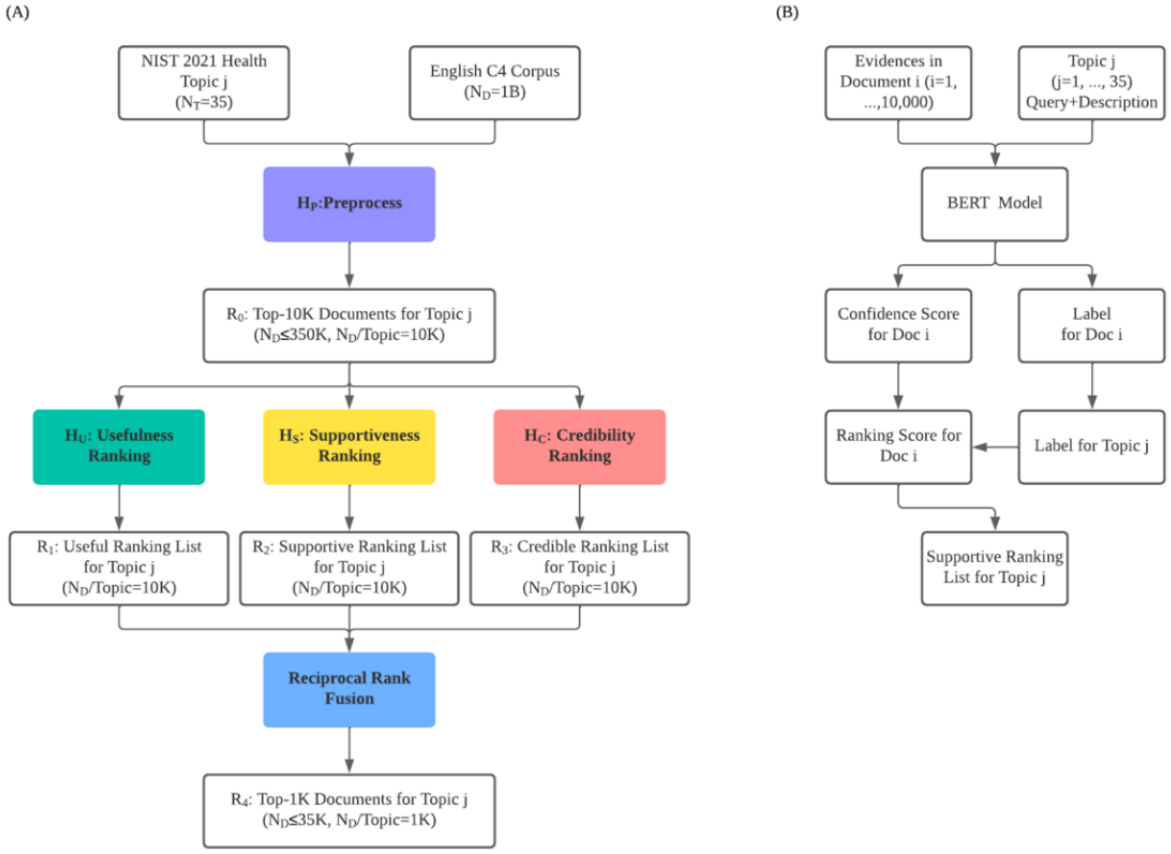
Article information	Article 1	Article 2
Doc ID	en.noclean.c4-train.07165-of-07168.96468	en.noclean.c4-train.00001-of-07168.126948
Time stamp	2019-04-25T18:00:17Z	2019-04-23T20:13:31Z
Text	[...] Do Not Use Sponging to Reduce a Fever. It is not recommended that you use sponging to reduce your child's fever. 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/

Quality-Based Multidimensional Ranking Conceptual Model

Phases

The quality-based multidimensional ranking model proposed in this work is presented in Figure 1A. The information retrieval process can be divided into 2 phases: *preprocessing* and *multidimensional ranking*. In the preprocessing phase, for a given topic j, N_D documents were retrieved based on their relevance (eg, using a BM25 model) [39]. In the multidimensional ranking phase, we further estimated the quality of the retrieved subset of documents according to the usefulness, supportiveness, and credibility dimensions. In the following sections, we describe the multidimensional ranking approach and its implementation using transfer learning and domain adaption. We then describe the preprocessing step, which can be performed based on sparse or dense retrieval engines.

Figure 1. Quality-based multidimensional ranking models: (A) general pipeline, (B) supportiveness model for the transfer learning approach. BERT: Bidirectional Encoder Representations from Transformers; C4: Colossal Clean Crawled Corpus; NIST: National Institute of Standards and Technology.



Multidimensional Ranking

To provide higher-quality documents at the top ranks, we proposed using a set of machine learning models trained to classify documents according to the usefulness, supportiveness, and credibility dimensions. For the initial rank list obtained in the preprocessing phase (see details in the following sections), the documents were reranked in parallel according to the following strategies for usefulness, supportiveness, and credibility.

Usefulness

The usefulness dimension is defined as *the extent to which the document contains information that a search user would find useful in answering the topic's question*. In this sense, it defines how pertinent a document is to a given topic. Thus, to compute the usefulness of retrieved documents, topic-document similarity models based on pretrained language models, such as Bidirectional Encoder Representations from Transformers (BERT)–base [40], mono-BERT-large [41], and ELECTRA [42], could be used. Given a topic-document pair, the language model infers a score that gives the level of similarity between the 2 input text passages. Although bag-of-words models, such as BM25, provide a strong baseline for usefulness, they do not consider word relations by learning context-sensitive representations as is the case with the pretrained language models, which are used to enhance the quality of the original ranking [28].

Supportiveness

The supportiveness dimension defines whether *the document supports or dissuades the use of the treatment in the topic's question*. Therefore, it defines the stance of the document on the health topic. In this dimension, documents are identified under 3 levels: (1) supportive (ie, the document supports the treatment), (2) dissuasive (ie, the document refutes the treatment), and (3) neutral (ie, the document does not contain enough information to make the decision) [35]. To compute the supportiveness of a document to a given query, the system should be optimized so that documents that are either supportive, if the topic is helpful, or dissuasive, if the topic is unhelpful, are boosted to the top of the ranking list, which means that correct documents are boosted and misinforming documents are downgraded.

Credibility

The credibility dimension defines *whether the document is considered credible by the assessor*, that is, how trustworthy the source document is. To compute this dimension, the content of the document itself could be used (eg, leveraging language features, such as readability [43]), which is assessable using the Simple Measure of Gobbledygook index [44]. Moreover, document metadata could be also used, such as incoming and outcoming links, which can be calculated with link analysis algorithms [45], and URL addresses considered to be trusted sources [46].

Transfer Learning Implementation

To implement the multidimensional ranking model in scenarios in which relevance judgments are not available, we proposed multiple (pretrained) models for each of the quality dimensions using transfer learning.

Usefulness

In this reranking step, we created an ensemble of pretrained language models—BERT-base, mono-BERT-large, and ELECTRA—all fine-tuned in the MS MARCO [47] data set. Each model then predicted the similarity between the topic and the initial list of retrieved documents. Their results were finally combined using reciprocal rank fusion (RRF) [32].

Supportiveness

In this reranking step (Figure 1B), we created an ensemble of claim-checking models—robustly optimized BERT approach (RoBERTa)–Large [48], BioMedRoBERTa-base [49], and SciBERT-base [50]—which were fine-tuned on the FEVER [51] and SciFact [52] data sets. Claim-checking models take a claim and a document as the information source and validate the veracity of the claim based on the document content [53]. Most claim-checking models assume that document content is ground truth. Since this is not valid in the case of web documents, we added a further classification step that evaluates the correctness of the retrieved documents. We used the top-*k* assignments [44] provided by the claim-checking models 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. A higher rank is then given to the correct supportive or dissuasive documents, a medium rank is given to the neutral documents, and a lower rank is given to the incorrect supportive or dissuasive documents. The rank lists obtained for each model were then combined using RRF.

Credibility

In this step, we implemented a random forest classifier trained on the Microsoft Credibility data set [54] with a set of credibility-related features, such as readability, openpage rank [45], and the number of cascading style sheets (CSS). The data set manually rated 1000 web pages with credibility scores between 1 (“very noncredible”) and 5 (“very credible”). We converted these scores for a binary classification setting—that is, scores of 4 and 5 were considered as 1 or *credible*, and scores of 1, 2, and 3 were considered as 0 or *noncredible*. For the readability score, we relied on the Simple Measure of Gobbledygook index [44], which estimates the years of education an average person needs to understand a piece of writing. Following Schwarz and Morris [54], we retrieved a web page’s PageRank and used it as a feature to train the classifier. We further used the number of CSS style definitions to estimate the effort for the design of a web page [55]. Last, a list of credible websites scrapped from the Health On the Net search engine [46] for the evaluated topics was combined with the baseline model to explore better performance. The result of the classifier was added to the unitary value of the Health On the Net credible sites [46].

Domain Adaptation Implementation

To implement the multidimensional ranking model in scenarios in which relevance judgments are available, we compared different pretrained language models—BERT, BioBERT [56], and BigBird [57]—for each of the quality dimensions using domain adaptation. In this case, each model was fine-tuned to predict the relevance judgment of a specific dimension (ie, usefulness, supportiveness, and credibility). Although the input size was limited to 512 tokens for the first 2 models, BigBird allows up to 4096 tokens.

We used the TREC 2019 Decision Track [33] benchmark data set to fine-tune our specific quality dimension models. The TREC 2019 Decision Track benchmark data set contains 51 topics evaluated across 3 dimensions: relevance, effectiveness, and credibility. Adhering to the experimental design set by [58], we mapped the 2019 and 2021 benchmarks as follows. The relevance dimension (2019) was mapped to usefulness (2021), with highly relevant documents translated as very useful and relevant documents as useful. The effectiveness dimension (2019) was mapped to supportiveness (2021), with effective labels reinterpreted as supportive and ineffective as dissuasive. The credibility dimension (2019) was directly mapped to credibility (2021) using the same labels.

The 2019 track uses the ClueWeb12-B13 [59,60] corpus, which contains 50 million pages. More details on the TREC 2019 Decision Track [33] benchmark are provided in Table S2 in Multimedia Appendix 1.

In the training phase, the language models received as input were the pair topic-document and a label for each dimension according to the 2019-2021 mapping strategy. At the inference time, given a topic-document pair from the TREC Health Misinformation Track 2021 benchmark, the model would infer its usefulness, supportiveness, or credibility based on the dimension on which it was trained.

Preprocessing or Ranking Phase

In the preprocessing step, which is initially executed to select a short list of candidate documents for the input query, a BM25 model was used. This step was performed using a bag-of-words model due to its efficiency. For the C4 snapshot collection, 2 indices were created, one using standard BM25 parameters and another fine-tuned using a collection of topics automatically generated (silver standard) from a set of 4985 indexed documents. For a given document, the silver topic was created based on the keyword2query [61] and doc2query [41] models to provide the query and description content, respectively. Using the silver topics and their respective documents, the BM25 parameters of the second index were then fine-tuned using grid search in a known-item search approach [62] (ie, for a given silver topic, the model should return in the top-1 the respective document used to generate it). The results of these 2 indices were fused using RRF.

Evaluation Metric

We followed the official TREC evaluation strategy and used the compatibility metric [46] 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 penalized, especially if they are relevant to the query content. In this context, 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) [63] similarity metric, which is top-weighted, with greater weight placed at higher ranks to address the indeterminate and incomplete nature of web search results [64]:

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

where the parameter p represents the searcher's patience or persistence and is set to 0.95 in our experiments and K is the search depth and is set to 1000 to bring pK-1 as close to 0 as possible. As shown in Equation 2, an additional normalization step was 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:

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

To ensure that helpful and harmful documents are treated differently, even if both might be relevant to the query content, the assessments were divided into “help compatibility” (help) and “harm compatibility” (harm) metrics. To evaluate the ability of the system to separate helpful from harmful information, the “harm compatibility” results were then subtracted from the “help compatibility” results, which were 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.

Experimental Setup

The BM25 indices were created using the Elasticsearch framework (version 8.6.0). The number of documents N_D retrieved per topic in the preprocessing step was set to 10,000 in our experiments. The pretrained language models were based on open-source checkpoints from the HuggingFace platform [65] and were implemented using the open-source PyTorch framework. The language models used for the usefulness dimension and their respective HuggingFace implemations were BERT base (Capreolus/bert-base-msmarco), BERT large (castorini/monobert-large-msmarco-finetune-only), and ELECTRA (Capreolus/electra-base-msmarco). The language models used for the supportiveness dimension were RoBERTa base (allenai/biomed_roberta_base), RoBERTa large (roberta-large), and SciBERT (allenai/scibert_scivocab_uncased). For the credibility dimension, we used the random forest algorithm of the scikit-learn library. In the domain adaptation setup, we partitioned the 2019 labeled data set into training and validation sets using an 80%:20% split ratio; the latter was used to select the best models. We then fine-tuned BioBERT (dmis-lab/biobert-base-cased-v1.1) with a batch size of 16, learning rate of 1^{-5}, and 20 epochs with early stopping set at 5 and utilizing the binary cross-entropy loss, which was optimized using the Adam optimizer. The BigBird model (google/bigbird-roberta-base) was fine-tuned with a batch size of 2, keeping all the other settings the same as the BioBERT model. All language models were fine-tuned using a single NVIDIA Tesla V100 graphics card with 32 GB of memory (see Multimedia Appendix 2 for more details). Results are reported using the compatibility and normalized discounted cumulative gain (nDCG) metrics. For reference, they were compared with the results of other participants of the official TREC Health Misinformation 2021 track, which have submitted runs for the automatic evaluation (ie, without using information about the topic stance). The code repository is available at [66].

Ethical Considerations

No human participants were involved in this research. All data used to build and evaluate the deep language models were publicly available and open aceess.

Results

Performance Results

In Table 3, we present the performance results of our quality-based retrieval models using the TREC Health Misinformation 2021 benchmark. Helpful compatibility (help) considers only helpful documents of the relevant judgment, while harmful compatibility (harm) considers only harmful documents and help-harm considers their compatibility difference (see Table S1 in Multmedia Appendix 1 for further detail). Additionally, we show the nDCG scores calculated using helpful (help) documents or harmful (harm) documents of the relevant judgment. The helpful_T, unhelpful_T, and all_T terms denote helpful topics, unhelpful topics, and all topics, respectively. H_U, H_S, and H_C rankings represent the combination of the preprocessing (H_P) results with the rerankings results for usefulness (H_U’), supportiveness (H_S’), and credibility (H_C’), respectively. For reference, we show our results compared with the models participating in the TREC Health Misinformation Track 2021: Pradeep et al [31] used the default BM25 ranker from Pyserini. Their reranking process incorporated a mix of mono and duo T5 models as well as Vera [67] on different topic fields. Abualsaud et al [68] created filtered collections that focus on filtering out nonmedical and unreliable documents, which were then used for retrieval with Anserini’s BM25. Schlicht et al [69] also used Pyserini’s BM25 ranker and Bio Sentence BERT to estimate usefulness and RoBERTa for credibility. The final score was a fusion of these individual rankings. Fernández-Pichel et al [70] used BM25 and RoBERTa for reranking and similarity assessment of the top 100 documents, trained an additional reliability classifier, and merged scores using CombSUM [71] or Borda Count. Bondarenko et al [72] used Anserini’s BM25 and PyGaggle’s MonoT5 for 2 baseline rankings, then reranked the top 20 from each using 3 argumentative axioms on seemingly argumentative queries.

Table 3. Performance results for the quality-based retrieval models.

Model	nDCG[a]		Compatibility				
	Help[b] ↑	Harm[c] ↓	Help ↑	Harm ↓	Help-harm ↑		
	all_T[d]	all_T	all_T	all_T	$helpful_T$[e]	$unhelpful_T$[f]	all
BM25[g] [39]	0.516	0.360	0.122	0.144	0.158	–0.162	–0.022
Pradeep et al [31]	0.602	0.378	0.195[h]	0.153	0.234[h]	–0.106	0.043
Abualsaud et al [68]	0.302	0.185[h]	0.164	0.123	0.179	–0.067	0.040
Schlicht et al [69]	0.438	0.309	0.121	0.103	0.157	–0.089	0.018
Fernández-Pichel et al [70]	0.603[h]	0.363	0.163	0.155	0.163	–0.113	0.008
Bondarenko et al [72]	0.266	0.226	0.129	0.144	0.150	–0.144	–0.015
Transfer learning							
H_U[i]	0.538[j]	0.324	0.142[j]	0.087[h]	0.156	–0.022[h]	0.056[h]
$H_U + H_S$[k]	0.477	0.315[j]	0.130	0.092	0.151	–0.049	0.038
$H_U + H_S + H_C$[l]	0.484	0.320	0.137	0.095	0.169[j]	–0.057	0.042
Domain adaptation							
H_U	0.510	0.327	0.128	0.100	0.146	–0.063	0.029
$H_U + H_S$	0.482	0.319	0.108	0.089	0.108	–0.050	0.019
$H_U + H_S + H_C$[l]	0.502	0.325	0.131	0.094	0.147	–0.048	0.037

[a]nDCG: normalized discounted cumulative gain.

[b]Help: results considering only helpful documents in the relevance judgment.

[c]Harm: results considering only harmful documents in the relevance judgment.

[d]all_T: all topics.

[e]$helpful_T$: helpful topics.

[f]$unhelpful_T$: unhelpful topics.

[g]BM25: Best Match 25.

[h]Best performance.

[i]H_U: usefulness model.

[j]Best performance among our models.

[k]H_S: supportiveness model.

[l]H_C: credibility model.

Our approach provides state-of-the-art results for automatic ranking systems in the transfer learning setting, with help-harm compatibility of +5.6%. This result was obtained with the usefulness model (H_U), which is the combination of preprocessing and usefulness reranking. It outperformed the default BM25 model [39] by 7% (P=.04) and the best automatic model from the TREC 2021 benchmark (Pradeep et al [31]) by 1%. In this case, although the help and harm compatibility metrics individually exhibited statistical significance (P=.02 and P=.01, respectively), the improvement in help-harm compatibility compared with the best automatic model was not statistically significant (P=.70). The usefulness model also stood out by achieving the best help and harm compatibility metrics among our models (14.2% and 8.7%, respectively; P=.50). Notice that, for the latter metric, the closest to 0, the better the performance. Interestingly, the usefulness model attained the highest nDCG score on help for all topics as well (P=.03). The combination of usefulness, supportiveness, and credibility models ($H_U + H_S + H_C$) provided the best help-harm (+16.9%) for helpful topics among our models (H_U: P=.40; $H_U + H_S$: P=.04).

Meanwhile, when calculating nDCG scores on harm, the combination of usefulness and supportiveness model ($H_U + H_S$) in the transfer learning and domain adaption settings outperformed the other model combinations (P=.50), indicating a different perspective of the best-performing model. Last, differently from what would be expected, in the domain adaption setting, the performance was poorer than the simpler transfer learning approach (2% decrease on average for the compatibility metric; P=.02). See Table S4 in Multimedia Appendix 3 for more information about using nDCG as a metric in a multidimensional evaluation.

Performance Stratification by Quality Dimension

In Table 4, we show the help, harm, and help-harm compatibility scores for the individual quality-based reranking models, which disregarded the preprocessing step (prime index). Additionally, we provide the nDCG scores for a more comprehensive view of the models' performance. H_P represents the preprocessing, and H_U', H_S', and H_C' stand for rerankings for usefulness, supportiveness, and credibility, respectively.

Table 4. Performance results for the individual ranking models.

Setting and model	nDCG[a]		Compatibility				
	Help[b] ↑	Harm[c] ↓	Help ↑	Harm ↓	Help-harm ↑		
	all_T[d]	all_T	all_T	all_T	$helpful_T$[e]	$unhelpful_T$[f]	all_T
H_P[g]	0.538[h]	0.341	0.126[h]	0.111	0.127[h]	–0.072	0.015
Transfer learning							
H_U'[i,j]	0.438	0.264	0.115	0.080	0.106	–0.020	0.036
H_S'[j,k]	0.140	0.102[h]	0.026	0.024	0.021	–0.013	0.002
H_C'[j,l]	0.131	0.113	0.031	0.035	0.033	–0.032	–0.003
Domain adaptation							
H_U'	0.436	0.277	0.077	0.038	0.099	–0.008	0.039[h]
H_S'	0.368	0.251	0.030	0.015[h]	0.030	0.003[h]	0.014
H_C'	0.443	0.296	0.079	0.064	0.104	–0.055	0.014

[a]nDCG: normalized discounted cumulative gain.

[b]Help: results considering only helpful documents in the relevance judgment.

[c]Harm: results considering only harmful documents in the relevance judgment.

[d]all_T: all topics.

[e]$helpful_T$: helpful topics.

[f]$unhelpful_T$: unhelpful topics.

[g]H_p: preprocess.

[h]Best performance.

[i]H_U': usefulness model.

[j]Unlike H_U, H_S, and H_C,H_U', H_S', and H_C' rankings are not combined with H_p.

[k]H_S': supportiveness model.

[l]H_C': credibility model.

In the transfer learning setting, the usefulness model (H_U') achieved the highest help-harm compatibility (+3.6%; P=.20). The preprocessing model gave the best help compatibility (+12.7%; H_U': P=.70; H_S' and H_C': P<.001). Additionally, the preprocessing model yielded the highest nDCG score for help (H_U': P=.10; H_S' and H_C': P<.001). On the other hand, the preprocessing model showed the highest harm compatibility (+11.1%; H_U': P=.33; H_S' and H_C': P<.01). The combination of the preprocessing and usefulness models (ie, H_U=+5.6%) improved the preprocessing model by 4.1% (from +1.5% to +5.6% on the help-harm compatibility; P=.06). For harm compatibility, the supportiveness model (H_S') achieved the best performance among the individual models (+2.4%; *Hp*: P<.001; *Hu*': P=.03; H_C': P=.34).

In the domain adaptation setting, the usefulness model (H_U') reached help-harm compatibility of +3.9%, similarly outperforming the other models (P=.32). The supportiveness model (H_S') achieved the best performance on harm compatibility (+1.5%; P=.07) and on help-harm compatibility for unhelpful topics (+0.3%; P=.50). Notice that +0.3% is the only positive help-harm compatibility for harmful topics throughout all the individual and combined models on both settings including the preprocessing step. Last, in the domain adaption setting, the performance of individual models was better than the simpler transfer learning approach (1% increase on average for the compatibility metric; P=.19).

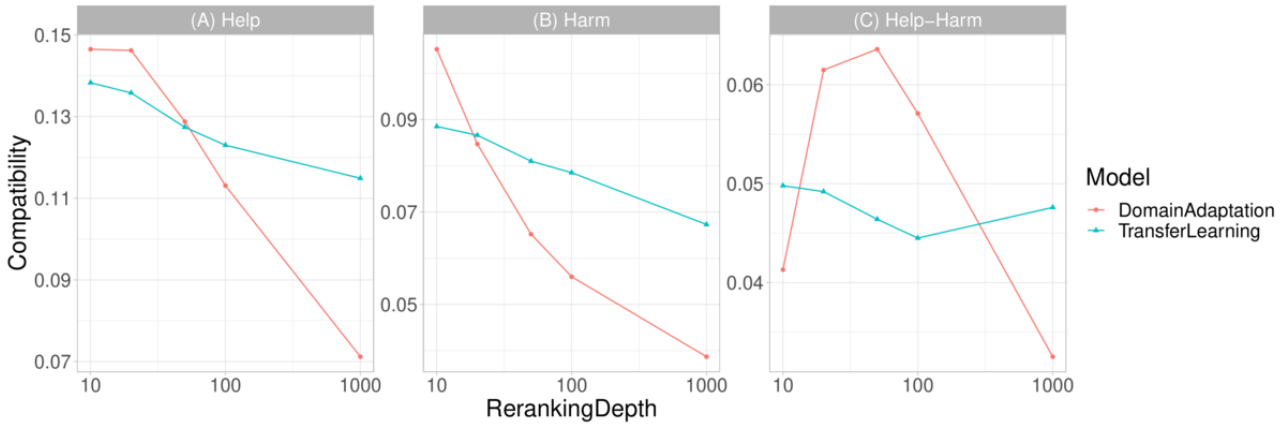
Reranking of the Top-N Documents

To further illustrate the effectiveness of the supportiveness and credibility dimensions, in Figure 2, we reranked only the top-n documents using the results of the usefulness model (H_U) as the basis. As we can see in Table 4, the overall effectiveness of the supportiveness (H_S') and credibility (H_C') models were considerably lower than that of the usefulness (H_U') model. The reason is that the relevance judgments were created using a

hierarchical approach: Only useful documents were further considered for supportiveness and credibility evaluations. As we reranked the documents in supportiveness and credibility dimensions without taking this hierarchy into account, their results might not be optimal. For example, low-ranking documents (ie, not useful) could have high credibility and, during the reranking process, could be boosted to the top ranks. Thus, we applied the supportiveness (H_S') and credibility (H_C') models to the usefulness model (H_U) results to rerank the top 10, 20, 50, 100, and 1000 documents, obtaining 2 new rankings, which were combined using RRF.

As the reranking depth increased from 10 to 1000, we observed a decrease in both help and harm compatibility. This suggests that both helpful and harmful documents were downgraded due to the inclusion of less useful but potentially supportive or credible documents. In the transfer learning setting, as the reranking depth increased, the help-harm compatibility decreased until the depth reached 100. Beyond this point, we observed a slight increase at the depth of 1000. In the domain adaptation setting, the help-harm compatibility increased above +6% when the reranking depth was between 20 and 50. This implies that, following the procedure of human annotation, by considering only the more useful documents, the supportiveness and credibility dimensions can help retrieve more helpful than harmful documents.

Figure 2. Compatibility performance for the top 10, 20, 50, 100, and 1000 reranking depths taking the results of usefulness as the basis.



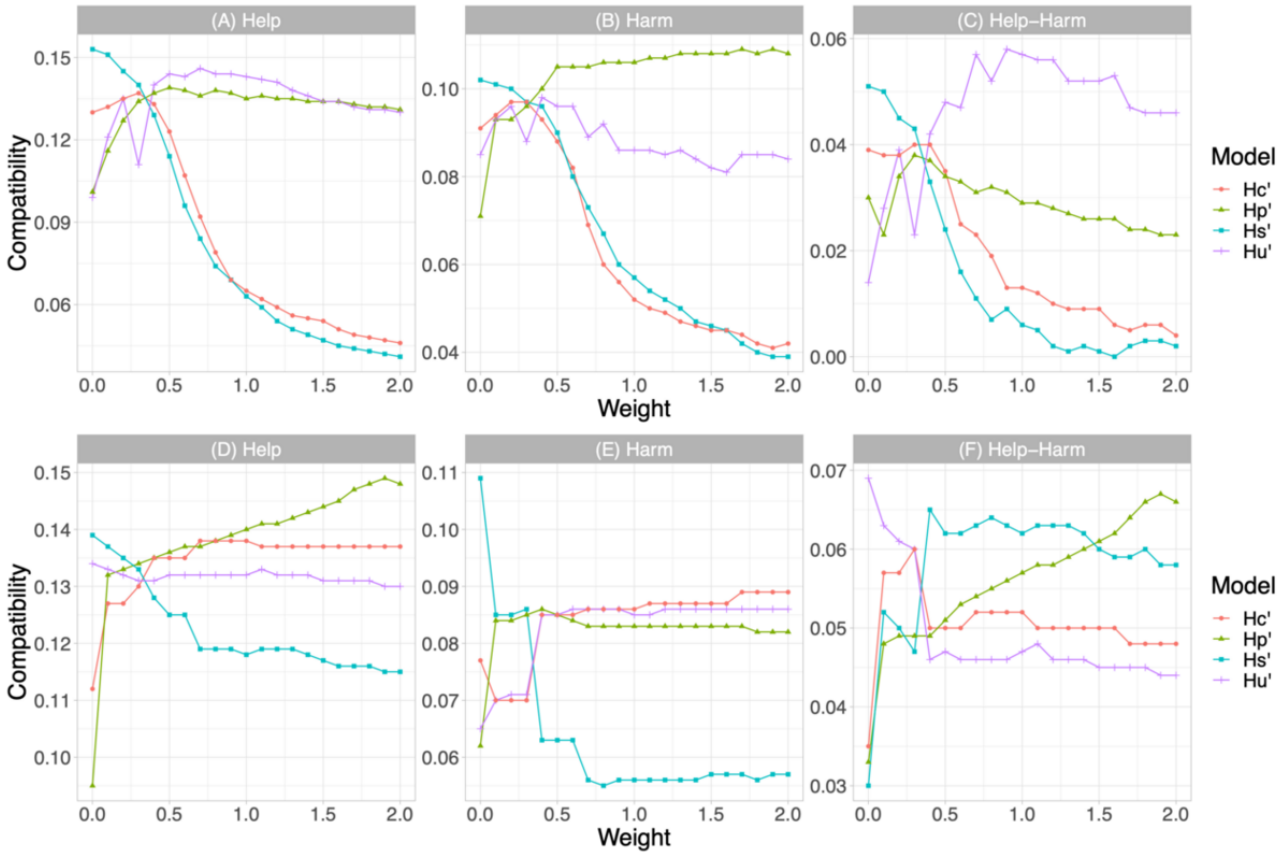
Quality Control

One of the advantages of the proposed multidimensional model is that we can optimize the results according to different quality metrics. In Figure 3, we show how the compatibility performance varies by changing the weight of the specific models (H_P, H_U', H_S', and H_C'). We normalized the score of the individual models to the unit and combined them linearly using a weight for 1 model between 0 and 2 while fixing the weight for the other 3 models at 0.33. For example, to see the influence of H_P in the final performance, we fixed the weights of H_U', H_S', and H_C' at 0.33 and varied the weight of H_P between 0 and 2. With weight 0, the reference model did not account for the final rank, while with weight 2, its impact was twice the sum of the other 3 models.

In the transfer learning setting, when we increased the weight of preprocessing and usefulness models, the help-harm compatibility increased to the best performance (+4.1% and +5.6%) then decreased slightly. For the supportiveness and credibility dimensions, the help-harm compatibility began to decrease once the weight was added. These results imply that the compatibility decreases with the weight addition regardless of whether it is helpful compatibility, harmful compatibility, or the difference between the 2.

In the domain adaptation setting, when we increased the weight of preprocessing, supportiveness, and credibility models individually, the help-harm compatibility increased then converged to +6.6%, +5.9%, and +4.8%, respectively. For the usefulness model, the help-harm compatibility decreased once the weight was added until it converged to +4.4%. It is worth noticing that, by combining the rankings linearly, the help-harm compatibility obtained from the domain adaptation setting may exceed the results we obtained when performing ranking combination with RRF (+3.7%), as well as the state-of-the-art result (+5.6%) in the transfer learning setting. The highest help-harm compatibility scores for each weighting combination were +6.6%, +6.8%, +6.5%, and +5.9% when varying the weights of H_P, H_U', H_S', and H_C', respectively.

Figure 3. Compatibility in the transfer learning approach (A-C) and compatibility in the domain adaptation approach (D-F), all with weights added to specific models.



Model Interpretation

To semantically explain the variation of help-harm compatibility, we set the search depth to 10. The help, harm, and help-harm compatibility of the 3 models are shown in Table 5. The help-harm compatibility was 1 when only helpful documents were retrieved in the top 10. Conversely, the help-harm compatibility was –1 when only harmful documents were retrieved in the top 10. A variation of 10% in the help or harm compatibility corresponded roughly to 1 helpful document exceeding the number of harmful documents retrieved in the top 10. Overall, the results show that retrieving relevant documents for health-related queries is hard, as, on average, only 1.5 of 10 documents were relevant (helpful or harmful) to the topic. In addition, we interpreted that the 3 models retrieved, on average, twice the number of helpful documents as harmful documents. Particularly, H_U had, on average, around 1 more helpful than harmful document in the top 10, of the 1.5 relevant documents retrieved. We also present the same analysis results for the domain adaptation setting, which also implies that, when the rankings were combined with RRF, the transfer learning approach outperformed the domain adaptation approach. See more details about the average compatibility for all the topics as the search depth K varied in Figure S1 in Multimedia Appendix 3.

Table 5. Help, harm, and help-harm compatibility with search depth set to 10 for the transfer learning setting and domain adaptation setting.

Setting and model	Help[a] ↑	Harm[b] ↓	Help-harm ↑
Transfer learning			
H_U[c]	0.112[d]	0.047[d]	0.065[d]
$H_U + H_S$[e]	0.088	0.050	0.038
$H_U + H_S + H_C$[f]	0.099	0.056	0.044
Domain adaptation			
H_U	0.094	0.060	0.034
$H_U + H_S$	0.074	0.070	0.003
$H_U + H_S + H_C$	0.087	0.076	0.011

[a]Help: results considering only helpful documents in the relevance judgment.

[b]Harm: results considering only harmful documents in the relevance judgment.

[c]H_U: usefulness model.

[d]Best performance.

[e]H_S: supportiveness model.

[f]H_C: credibility model.

Discussion

We propose a quality-based multidimensional ranking model to enhance the usefulness, supportiveness, and credibility of retrieved web resources for health-related queries. By adapting our approach in a transfer learning setting, we showed state-of-the-art results in the automatic quality ranking evaluation benchmark. We further explored the pipeline in a domain adaptation setting and showed that, in both settings, the proposed method can identify more helpful than harmful documents, as measured by +5% and +7% help-harm compatibility scores, respectively. By combining different reranking strategies, we showed that multidimensional aspects have a significant impact on retrieving high-quality information, particularly for unhelpful topics.

The quality of web documents is biased in terms of topic stance. For all models, helpful topics achieve higher help compatibility, while unhelpful topics achieve higher harm compatibility. The implication is that web documents centered around helpful topics are more likely to support the intervention and are helpful. On the other hand, web documents focusing on unhelpful topics present an equal chance of being supportive or dissuasive on the intervention and are helpful or harmful. Among other consequences, if web data are used to train large language models without meticulously crafted training examples using effective data set search methods [73], as the one proposed here, they are likely to further propagate health misinformation.

Automatic retrieval systems tend to find more helpful information on helpful topics with the information biased toward helpfulness and find more harmful information on unhelpful topics with the information slightly biased toward harmfulness. The help-harm compatibility ranged from +2.3% to +15.3% for helpful topics and from –5.7% to +0.2% for unhelpful topics. The difference shows that, for the improvement of quality-centered retrieval models, it is especially important to focus on unhelpful topics. Moreover, although specialized models might provide enhanced effectiveness, their combination is not straightforward. In our experiments, we showed that supportiveness and credibility models should be applied only in the top 20 to 50 retrieved documents to achieve optimal performance.

Finding the correct stance automatically is another key component of the automatic model. Automatic models show the ability to prioritize helpful documents, resulting in positive help-harm compatibility. However, they are still far from state-of-the-art manual models, with help-harm compatibility scores ranging from +20.8% [68] to +25.9% [31]. We acknowledge that the help-harm compatibility can improve significantly with the correct stance given. This information is nevertheless unavailable in standard search environments; thus, the scenario analyzed in this work is more adapted to real-world applications.

This work has certain limitations. In the domain adaptation setting, we simplified the task to consider 2 classes within each dimension for the classification due to the limited variety available in the labeled data set. Alternatively, we could add other classes from documents that have been retrieved. Moreover, the number of topics used to evaluate our models was limited (n=32), despite including 6030 human-annotated, query-document pairs, and thus reflects only a small portion of misinformation use cases.

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. We propose a multidimensional information quality ranking model that utilizes usefulness, supportiveness, and credibility to strengthen the factual reliability of health advice search results. Experiments conducted on publicly available data sets show that the proposed

model is promising, achieving state-of-the-art performance for automatic ranking in comparison with various baselines implemented on the TREC Health Misinformation 2021 benchmark. Thus, the proposed approach could be used to improve online health searches and provide quality-enhanced information for health information seekers. Future research could explore more granular classification models for each dimension, and a model simplification could provide an advantage for real-world implementations.

Acknowledgments

The study was funding by Innosuisse projects (funding numbers 55441.1 IP-ICT and 101.466 IP-ICT).

Data Availability

The data sets generated during and/or analyzed during this study are available in the Text Retrieval Conference (TREC) Health Misinformation Track repository [74] and GitLab repository [66].

Authors' Contributions

BZ, NN, and DT prepared the data, conceived and conducted the experiments, and analyzed the results. BZ, NN, and DT drafted the manuscript. All authors reviewed the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Additional Information on Benchmark Datasets.
[PDF File (Adobe PDF File), 22 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Fine-Tuning in the Domain Adaptation Setting.
[PDF File (Adobe PDF File), 69 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Supporting Experiment Results.
[PDF File (Adobe PDF File), 195 KB-Multimedia Appendix 3]

References

1. Goeuriot L, Jones GJF, Kelly L, Müller H, Zobel J. Medical information retrieval: introduction to the special issue. Inf Retrieval J. Jan 11, 2016;19(1-2):1-5. [doi: 10.1007/s10791-015-9277-8]
2. Chu JT, Wang MP, Shen C, Viswanath K, Lam TH, Chan SSC. How, when and why people seek health information online: qualitative study in Hong Kong. Interact J Med Res. Dec 12, 2017;6(2):e24. [FREE Full text] [doi: 10.2196/ijmr.7000] [Medline: 29233802]
3. Lee JJ, Kang K, Wang MP, Zhao SZ, Wong JYH, O'Connor S, et al. Associations between COVID-19 misinformation exposure and belief with COVID-19 knowledge and preventive behaviors: cross-sectional online study. J Med Internet Res. Nov 13, 2020;22(11):e22205. [FREE Full text] [doi: 10.2196/22205] [Medline: 33048825]
4. Ecker UKH, Lewandowsky S, Cook J, Schmid P, Fazio LK, Brashier N, et al. The psychological drivers of misinformation belief and its resistance to correction. Nat Rev Psychol. Jan 12, 2022;1(1):13-29. [doi: 10.1038/s44159-021-00006-y]
5. Krist AH, Tong ST, Aycock RA, Longo DR. Engaging patients in decision-making and behavior change to promote prevention. Stud Health Technol Inform. 2017;240:284-302. [FREE Full text] [Medline: 28972524]
6. Swire-Thompson B, Lazer D. Public health and online misinformation: challenges and recommendations. Annu Rev Public Health. Apr 02, 2020;41(1):433-451. [FREE Full text] [doi: 10.1146/annurev-publhealth-040119-094127] [Medline: 31874069]
7. Sundin O, Lewandowski D, Haider J. Whose relevance? Web search engines as multisided relevance machines. Asso for Info Science & Tech. Aug 21, 2021;73(5):637-642. [FREE Full text] [doi: 10.1002/asi.24570]
8. Sullivan D. How Google delivers reliable information in Search. Google. Sep 10, 2020. URL: https://blog.google/products/search/how-google-delivers-reliable-information-search/ [accessed 2024-04-18]
9. Di Sotto S, Viviani M. Health misinformation detection in the social web: an overview and a data science approach. Int J Environ Res Public Health. Feb 15, 2022;19(4):A. [FREE Full text] [doi: 10.3390/ijerph19042173] [Medline: 35206359]

10. Sylvia Chou W, Gaysynsky A, Cappella JN. Where we go from here: health misinformation on social media. Am J Public Health. Oct 2020;110(S3):S273-S275. [doi: 10.2105/ajph.2020.305905]
11. Kickbusch I. Health literacy: addressing the health and education divide. Health Promot Int. Sep 2001;16(3):289-297. [doi: 10.1093/heapro/16.3.289] [Medline: 11509466]
12. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: systematic review. J Med Internet Res. Jan 20, 2021;23(1):e17187. [FREE Full text] [doi: 10.2196/17187] [Medline: 33470931]
13. Eysenbach G. How to fight an infodemic: the four pillars of infodemic management. J Med Internet Res. Jun 29, 2020;22(6):e21820. [FREE Full text] [doi: 10.2196/21820] [Medline: 32589589]
14. Burki T. Vaccine misinformation and social media. The Lancet Digital Health. Oct 2019;1(6):e258-e259. [FREE Full text] [doi: 10.1016/s2589-7500(19)30136-0]
15. Lotto M, Sá Menezes T, Zakir Hussain I, Tsao S, Ahmad Butt Z, P Morita P, et al. Characterization of false or misleading fluoride content on Instagram: infodemiology study. J Med Internet Res. May 19, 2022;24(5):e37519. [FREE Full text] [doi: 10.2196/37519] [Medline: 35588055]
16. Mackey T, Purushothaman V, Haupt M, Nali M, Li J. Application of unsupervised machine learning to identify and characterise hydroxychloroquine misinformation on Twitter. The Lancet Digital Health. Feb 2021;3(2):e72-e75. [FREE Full text] [doi: 10.1016/s2589-7500(20)30318-6]
17. Nsoesie EO, Cesare N, Müller M, Ozonoff A. COVID-19 misinformation spread in eight countries: exponential growth modeling study. J Med Internet Res. Dec 15, 2020;22(12):e24425. [FREE Full text] [doi: 10.2196/24425] [Medline: 33264102]
18. Upadhyay R, Pasi G, Viviani M. Health misinformation detection in web content: a structural-, content-based, and context-aware approach based on Web2Vec. GoodIT '21: Proceedings of the Conference on Information Technology for Social Good. Sep 2021.:19-24. [doi: 10.1145/3462203.3475898]
19. Hesse BW, Nelson DE, Kreps GL, Croyle RT, Arora NK, Rimer BK, et al. Trust and sources of health information: the impact of the Internet and its implications for health care providers: findings from the first Health Information National Trends Survey. Arch Intern Med. 2005;165(22):2618-2624. [doi: 10.1001/archinte.165.22.2618] [Medline: 16344419]
20. van der Linden S. Misinformation: susceptibility, spread, and interventions to immunize the public. Nat Med. Mar 10, 2022;28(3):460-467. [doi: 10.1038/s41591-022-01713-6] [Medline: 35273402]
21. 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. ICTIR '17: Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval. Oct 2017.:209-216. [doi: 10.1145/3121050.3121074]
22. Upadhyay R, Pasi G, Viviani M. An overview on evaluation labs and open issues in health-related credible information retrieval. Proceedings of the 11th Italian Information Retrieval Workshop 2021. 2021.:1. [FREE Full text]
23. Suominen H, Kelly L, Goeuriot L, Krallinger M. CLEF eHealth Evaluation Lab 2020. Advances in Information Retrieval. 2020;12036:587-594. [doi: 10.1007/978-3-030-45442-5_76]
24. Clarke CLA, Maistro M, Smucker MD. Overview of the TREC 2021 Health Misinformation Track. NIST Special Publication: NIST SP 500-335: The Thirtieth Text REtrieval Conference (TREC 2021) Proceedings. 2022.:1. [FREE Full text]
25. Solainayagi P, Ponnusamy R. Trustworthy media news content retrieval from web using truth content discovery algorithm. Cognitive Systems Research. Aug 2019;56:26-35. [doi: 10.1016/j.cogsys.2019.01.002]
26. Li L, Qin B, Ren W, Liu T. Truth discovery with memory network. Tsinghua Science and Technology. Dec 2017;22(6):609-618. [doi: 10.23919/tst.2017.8195344]
27. Zhang E, Gupta N, Tang R, Han X, Pradeep R, Lu K, et al. Covidex: neural ranking models and keyword search infrastructure for the COVID-19 open research dataset. Proceedings of the First Workshop on Scholarly Document Processing. 2020.:31-41. [doi: 10.18653/v1/2020.sdp-1.5]
28. Teodoro D, Ferdowsi S, Borissov N, Kashani E, Vicente Alvarez D, Copara J, et al. Information retrieval in an infodemic: the case of COVID-19 publications. J Med Internet Res. Sep 17, 2021;23(9):e30161. [FREE Full text] [doi: 10.2196/30161] [Medline: 34375298]
29. Fernández-Pichel M, Losada DE, Pichel JC, Elsweiler D. Comparing TraditionalNeural Approaches for Detecting Health-Related Misinformation. In: Candan KS, editor. Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2021. Lecture Notes in Computer Science(), vol 12880. Cham, Switzerland. Springer International Publishing; 2021;78-90.
30. Lima LC, Wright DB, Augenstein I, Maistro M. University of Copenhagen participation in TREC Health Misinformation Track 2020. NIST Special Publication: NIST SP 1266: The Twenty-Ninth Text REtrieval Conference (TREC 2020) Proceedings. 2021.:1. [FREE Full text]
31. Pradeep R, Ma X, Nogueira R, Lin J. Vera: prediction techniques for reducing harmful misinformation in consumer health search. SIGIR '21: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. Jul 2021.:2066-2070. [doi: 10.1145/3404835.3463120]
32. Cormack GV, Clarke CLA, Buettcher S. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. 2009.:758-759. [doi: 10.1145/1571941.1572114]

33. Abualsaud M, Lioma C, Maistro M, Smucker M, Zuccon G. Overview of the TREC 2019 Decision Track. NIST Special Publication: SP 500-331: The Twenty-Eighth Text REtrieval Conference (TREC 2019) Proceedings. 2020.:1. [FREE Full text]
34. 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. NIST Special Publication: NIST SP 500-335: The Thirtieth Text REtrieval Conference (TREC 2021) Proceedings. 2022.:1. [FREE Full text]
35. Clarke CLA, Rizvi S, Smucker MD, Maistro M, Zuccon G. Overview of the TREC 2020 Health Misinformation Track. NIST Special Publication: NIST SP 1266: The Twenty-Ninth Text REtrieval Conference (TREC 2020) Proceedings. 2021.:1. [FREE Full text]
36. National Institute of Standards and Technology. URL: https://www.nist.gov/ [accessed 2024-04-18]
37. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research. 2020;21(140):1-67. [FREE Full text]
38. Common Crawl. URL: https://commoncrawl.org/ [accessed 2024-04-18]
39. Robertson S, Zaragoza H. The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval. Apr 2009;3(4):333-389. [doi: 10.1561/1500000019]
40. Li C, Yates A, MacAvaney S, He B, Sun Y. PARADE: Passage Representation Aggregation for Document Reranking. ACM Transactions on Information Systems. Sep 27, 2023;42(2):1-26. [doi: 10.1145/3600088]
41. Nogueira R, Yang W, Cho K, Lin J. Multi-stage document ranking with BERT. arXiv. Preprint posted online on October 31, 2019. [doi: 10.48550/arXiv.1910.14424 Focus to learn more]
42. Clark K, Luong MH, Le QV, Manning CD. Electra: Pre-training text encoders as discriminators rather than generators. arXiv. Preprint posted online on March 23, 2020. [doi: 10.48550/arXiv.2003.10555]
43. Zhou S, Jeong H, Green PA. How consistent are the best-known readability equations in estimating the readability of design standards? IEEE Trans. Profess. Commun. Mar 2017;60(1):97-111. [doi: 10.1109/tpc.2016.2635720]
44. 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. Jan 12, 2018;106(1):38-45. [FREE Full text] [doi: 10.5195/jmla.2018.262] [Medline: 29339932]
45. getPageRank. OpenPageRank. URL: https://www.domcop.com/openpagerank/documentation [accessed 2024-04-18]
46. Boyer C, Selby M, Scherrer J, Appel R. The Health On the Net Code of Conduct for medical and health websites. Comput Biol Med. Sep 1998;28(5):603-610. [FREE Full text] [doi: 10.1016/s0010-4825(98)00037-7] [Medline: 9861515]
47. Bajaj P, Campos D, Craswell N, Deng L, Gao J, Liu X, et al. MS MARCO: a human generated machine reading comprehension dataset. arXiv. Preprint posted online on October 31, 2018. [doi: 10.48550/arXiv.1611.09268]
48. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv. Preprint posted online on July 26, 2019. [doi: 10.48550/arXiv.1907.11692]
49. Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, et al. Don't stop pretraining: adapt language models to domains and tasks. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020.:8342-8360. [FREE Full text] [doi: 10.18653/v1/2020.acl-main.740]
50. 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.:3615-3620. [doi: 10.18653/v1/d19-1371]
51. Aly R, Guo Z, Schlichtkrull M, Thorne J, Vlachos A, Christodoulopoulos C, et al. The fact extraction and verification over unstructured and structured information (FEVEROUS) shared task. Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER). 2021.:1-13. [doi: 10.18653/v1/2021.fever-1.1]
52. Wadden D, Lin S, Lo K, Wang L, van Zuylen M, Cohan A, et al. Fact or fiction: verifying scientific claims. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020.:7534-7550. [doi: 10.18653/v1/2020.emnlp-main.609]
53. Stammbach D, Zhang B, Ash E. The choice of textual knowledge base in automated claim checking. Journal of Data and Information Quality. 2023;15(1):1-22. [doi: 10.1145/3561389]
54. Schwarz J, Morris M. Augmenting web pages and search results to support credibility assessment. CHI '11: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2011.:1245-1254. [doi: 10.1145/1978942.1979127]
55. Olteanu A, Peshterliev S, Liu X, Aberer K. Web credibility: Features exploration and credibility prediction. In: Serdyukov P, Braslavski P, Kuznetsov SO, Kamps J, Rüger S, Agichtein E, et al, editors. Advances in Information Retrieval. ECIR 2013. Lecture Notes in Computer Science, vol 7814. Berlin, Germany. Springer; 2013;557-568.
56. Lee J, Yoon W, Kim S, Kim D, Kim S, So C, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. Feb 15, 2020;36(4):1234-1240. [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
57. Zaheer M, Guruganesh G, Dubey K, Ainslie J, Alberti C, Ontanon S, et al. Big Bird: transformers for longer sequences. Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020). 2020. [FREE Full text]

58. Zhang D, Tahami AV, Abualsaud M, Smucker MD. Learning trustworthy web sources to derive correct answers and reduce health misinformation in search. SIGIR '22: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2022.:2099-2104. [doi: 10.1145/3477495.3531812]
59. The ClueWeb12 Dataset. The Lemur Project. URL: http://lemurproject.org/clueweb12/ [accessed 2024-04-18]
60. Zuccon G, Palotti J, Goeuriot L, Kelly L, Lupu M, Pecina P, et al. The IR Task at the CLEF eHealth evaluation lab 2016: User-centred health information retrieval. 2016. Presented at: CLEF 2016 - Conference and Labs of the Evaluation Forum; September 5-8, 2016;255-266; Évora, Portugal. URL: https://ceur-ws.org/Vol-1609/16090015.pdf
61. 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.:221-229. [doi: 10.18653/v1/K18-1022]
62. Ogilvie P, Callan J. Combining document representations for known-item search. SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. 2003.:143-150. [doi: 10.1145/860462.860463]
63. Webber W, Moffat A, Zobel J. A similarity measure for indefinite rankings. ACM Trans. Inf. Syst. Nov 23, 2010;28(4):1-38. [doi: 10.1145/1852102.1852106]
64. Clarke CLA, Smucker MD, Vtyurina A. Offline evaluation by maximum similarity to an ideal ranking. CIKM '20: Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 2020.:225-234. [doi: 10.1145/3340531.3411915]
65. Hugging Face. URL: https://huggingface.co [accessed 2024-04-18]
66. Zhang B, Naderi N, Mishra R, Teodoro D. Online health search via multi-dimensional information quality assessment based on deep language models. MedRxiv. Preprint posted online on January 11, 2024. [FREE Full text] [doi: 10.1101/2023.04.11.22281038]
67. 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. 2021.:94-103. [FREE Full text]
68. Abualsaud M, Chen IX, Ghajar K, Minh LNP, Smucker MD, Tahami AV, et al. UWaterlooMDS at the TREC 2021 Health Misinformation Track. NIST Special Publication 500-335: The Thirtieth Text REtrieval Conference Proceedings (TREC 2021). 2022.:1. [FREE Full text]
69. Schlicht I, Paula AD, Rosso P. UPV at TREC Health Misinformation Track 2021 ranking with SBERT and quality. NIST Special Publication 500-335: The Thirtieth Text REtrieval Conference Proceedings (TREC 2021). 2022.:1. [FREE Full text]
70. Fernández-Pichel M, Prada-Corral M, Losada DE, Pichel JC, Gamallo P. CiTIUS at the TREC 2021 Health Misinformation Track. NIST Special Publication 500-335: The Thirtieth Text REtrieval Conference Proceedings (TREC 2021). 2022.:1. [FREE Full text]
71. Belkin NJ, Kantor P, Fox EA, Shaw JA. Combining the evidence of multiple query representations for information retrieval. Information Processing & Management. May 1995;31(3):431-448. [doi: 10.1016/0306-4573(94)00057-A]
72. Bondarenko A, Fröbe M, Gohsen M, Günther S, Kiesel J, Schwerter J, et al. Webis at TREC 2021: Deep Learning, Health Misinformation, and Podcasts Tracks. NIST Special Publication 500-335: The Thirtieth Text REtrieval Conference Proceedings (TREC 2021). 2022.:1. [FREE Full text]
73. Teodoro D, Mottin L, Gobeill J, Gaudinat A, Vachon T, Ruch P. Improving average ranking precision in user searches for biomedical research datasets. Database (Oxford). Jan 01, 2017;2017:bax083. [FREE Full text] [doi: 10.1093/database/bax083] [Medline: 29220475]
74. 2021 Health Misinformation Track. TREC. 2022. URL: https://trec.nist.gov/data/misinfo2021.html [accessed 2024-04-18]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
BM25: Best Match 25
C4: Colossal Clean Crawled Corpus
CLEF: Conference and Labs of the Evaluation Forum
CSS: cascading style sheets
nDCG: normalized discounted cumulative gain
NIST: National Institute of Standards and Technology
RBO: rank-biased overlap
RoBERTa: robustly optimized BERT approach
RRF: reciprocal rank fusion
TREC: Text Retrieval Conference

Edited by B Malin; submitted 12.09.22; peer-reviewed by D Carvalho, D He, S Marchesin; comments to author 10.04.23; revised version received 12.07.23; accepted 15.01.24; published 02.05.24

Please cite as:
Zhang B, Naderi N, Mishra R, Teodoro D
Online Health Search Via Multidimensional Information Quality Assessment Based on Deep Language Models: Algorithm Development and Validation
JMIR AI 2024;3:e42630
URL: https://ai.jmir.org/2024/1/e42630
doi: 10.2196/42630
PMID: