An Unsupervised Approach to Genuine Health Information Retrieval Based on Scientific Evidence

Rishabh Upadhyay, Gabriella Pasi, and Marco Viviani(✉)

Department of Informatics, Systems, and Communication (DISCo), Information and Knowledge Representation, Retrieval, and Reasoning (IKR3) Lab, Edificio ABACUS, University of Milano-Bicocca, Viale Sarca, 336, 20126 Milan, Italy
{rishabh.upadhyay,gabriella.pasi,marco.viviani}@unimib.it
https://ikr3.disco.unimib.it/

Abstract. In contemporary society, more and more people refer to information they find online to meet their information needs. In some domains, such as health information, this phenomenon has been particularly on the rise in recent years. On the one hand, this could have a positive impact in increasing people's so-called health literacy, which would benefit the health of the individual and the community as a whole. On the other hand, with a significant amount of health misinformation circulating online, people and society could face very serious consequences. In this context, the purpose of this article is to investigate a solution that can help online users find health information that is relevant to their information needs, while at the same time being genuine. To do so, in the process of retrieval of estimated relevant information, the genuineness of the information itself is taken into consideration, which is evaluated by referring to scientific articles that can support the claims made in the online health information considered. With respect to the literature, the proposed solution is fully unsupervised and does not require any human intervention. It is experimentally evaluated on a publicly accessible dataset as part of the TREC 2020 Health Misinformation Track.

Keywords: Health misinformation · Information Retrieval · Information genuineness · Multidimensional relevance · Health literacy

1 Introduction

It has now been demonstrated how, especially in recent years, people increasingly rely on information they find online about various tasks and contexts [21]. One may want to find news about specific events, to retrieve people's opinions with respect to certain products or services, to seek information about diseases, symptoms, and treatments, etc. In the latter area, we refer specifically to *Consumer Health Search* (CHS), which indicates search conducted by laypersons looking for health advice online [32]. As early as 2013, the *Pew Research Center*,

R. Chbeir et al. (Eds.): WISE 2022, LNCS 13724, pp. 119–135, 2022.
https://doi.org/10.1007/978-3-031-20891-1_10

a "nonpartisan *fact tank* informing the public about issues, attitudes and trends shaping the world",[1] was highlighting how a large proportion of the U.S. population searched and consulted health information online, even going so far as to exclude the figure of the doctor when making decisions with respect to their own health [14]. By means of a recent *Eurostat* survey,[2] it was shown that also in Europe online health information seeking has been steadily increasing over the years, especially among young people.

Such a phenomenon must take into account the problem that a great deal of information disseminated online actually turns out to be *false* or *misleading* [35]; this, especially in the case of an area as sensitive as health, can have even very serious repercussions both at the individual level and at the level of society as a whole; let us think, for example, of the proposition of "miracle cures" for cancer or other diseases, or the amount of non-genuine information that has occurred with respect to COVID-19 in recent years [3,12]. In most cases, laypeople are unable to discern genuine health information from non-genuine one, because of their insufficient level of *health literacy* [5,30]; this latter concept was included in the glossary of the *World Health Organization* (WHO) in 1988, and indicates "the ability of a citizen to obtain, process, and understand basic health information in order to make informed choices" [16]. At the same time, clinical experts cannot take charge of evaluating every single piece of information that appears online, because of the volume and speed with which it is constantly generated. This is why it is necessary to develop automated solutions or tools that can support non-expert users in avoiding behaviors that are harmful to their health when they come into contact with *misinformation*.

In this article, we refer to this latter concept which, in the literature, we believe to be the most general one compared to that of *disinformation*, which often refers to false or misleading information that is generated *on purpose* [36]. In fact, we do not aim at estimating the purpose for which non-genuine information is disseminated, but we aim at investigating a solution that can somehow help users limit access to *health misinformation*, intended as "a health-related claim of fact that is currently false due to a lack of scientific evidence" [7]. To do this, we propose the development of a *retrieval model* that takes into account both the *topical relevance* of health-related content with respect to user queries and the *genuineness* of the information itself, by comparing such content with what is reported in scientific articles, which we consider as a reputed source of scientific evidence. In this work, information genuineness is also considered a query-dependent dimension of relevance, and it is computed in a totally unsupervised manner, requiring no human intervention w.r.t. the definition of indicators of information genuineness or the formal definition of knowledge bases. The proposed model is evaluated against data made publicly accessible as part of the *Health Misinformation Track* at TREC 2020, and against a baseline and various experimental model configurations that demonstrate its potential effectiveness.

[1] https://www.pewresearch.org/.

[2] https://ec.europa.eu/eurostat/web/products-eurostat-news/-/edn-20220406-1 (accessed on May 25, 2022).

2 Related Work

The scientific community has rather recently begun to propose solutions to the growing health misinformation circulating online. A good portion of them started from works that addressed the problem of misinformation detection in general (e.g., fake news detection, opinion spam detection, etc.), and used their solutions applied to the health domain. These are mainly works that treat misinformation detection as a *binary classification* task [35]. Other works have addressed the problem by attempting to take more account of the point characteristics of the health domain and developing ad hoc solutions. Again, these are mainly supervised solutions acting on both health-related content in the form of Web pages [17,33] and social media content [2,28,38].

However, since health information can be somewhat verified against the presence of experts, prior knowledge, and/or quality content, a recent research direction is investigating how to automatically use such different forms of "scientific evidence" in the context of health misinformation detection. In this context, some work has involved *experts* and, in general, *human assessors* in evaluating health information genuineness. For example, DISCERN [6] is a brief questionnaire, developed within the DISCERN Project,[3] which aims at providing users with a possible reliable way of assessing the quality of written information on treatment choices for a health problem. The *Health on the Net foundation* (HON) [4] has issued a code of conduct and quality label for medical sites by considering different attributes such as: disclosure of authorship, sources, updating of information, disclosure of editorial and publicity policy, as well as confidentiality.[4] The approach known as HC-COVID [18] focuses on COVID-related health misinformation detection by employing a crowdsourcing-based knowledge graph, used as a source of evidence, built by leveraging the collaborative efforts of expert and non-expert crowd workers. The drawbacks characterizing these approaches are mainly related to the high level of human intervention needed, e.g., to manually assign quality indicators to each new piece of content, to recruit expert and non-expert crowd workers, to guarantee the quality of annotations, etc.

Other work has formalized evidence-based health-related concepts into *knowledge bases* (through the use of ontologies or knowledge graphs) to compare online health-related claims against such knowledge. For example, in *MedFact* [28], the authors develop an algorithm for checking social media post based on the so-called *Evidence-Based Medicine* [27], i.e., integrating individual clinical expertise with the best available external clinical evidence from systematic research and trusted medical information sources such as the *Turning Research Into Practice* (TRIP) database.[5] In another pretty recent model named DETER-

[3] http://www.discern.org.uk/.

[4] According to HON, "the HONcode is not an award system, nor does it intend to rate the quality of the information provided by a Web site. It only defines a set of rules to: (*i*) hold Web site developers to basic ethical standards in the presentation of information; (*ii*) help make sure readers always know the source and the purpose of the data they are reading". https://www.hon.ch/HONcode/.

[5] https://www.tripdatabase.com/home.

RENT [11], the authors focus on explainable healthcare misinformation detection by leveraging a medical knowledge graph named *Knowlife* [13], built on top of medical content extracted from *PubMed*,[6] and other health portals such as *Mayo Clinic*,[7] *RxList*,[8] the *Wikipedia Medicine Portal*,[9] and *MedlinePlus*.[10] The potential disadvantages of these latter approaches lie mainly in the complexity of the knowledge base formalization, which is difficult to build automatically, and subject to constant updating issues.

All the above-mentioned works do not address the problem as an Information Retrieval task. Only some recent works are starting to be proposed in this area, to produce a *ranking* of search results within an *Information Retrieval System* (IRS) while also taking into consideration as a relevance criterion that of information genuineness. Among them, Vera [23] is a solution that identifies *harmful* and *helpful* documents by considering a multi-stage ranking architecture. Specifically, the top-ranked topically relevant documents – retrieved by means of the BM25 retrieval model [26] – are re-ranked by using the mono-T5 and duo-T5 retrieval models [24], by exploiting the passages with the highest probability of being relevant within the documents; subsequently, a label prediction model is trained using the TREC 2019 Decision (Medical Misinformation) Track data [1] to consider information genuineness (referred as *credibility* in the paper) and to re-rank again documents based also on this criterion. In [29] the authors consider, beyond topical relevance computed using BM25, another relevance dimension related to information genuineness, i.e., *information quality*. In this work, quality estimation is performed by training a multi-label classifier that returns a probabilistic score for ten quality criteria considered (e.g., Does the story adequately quantify the benefits of the intervention? Does the story establish the availability of the treatment/test/product/procedure?, etc.).[11] Specifically, a RoBERTa-based model is trained on the *Health News Review* dataset presented in [39], labeled with respect to the above-mentioned quality criteria. Once distinct rankings are obtained on the basis of topical relevance and information quality scores, they are merged by means of *Reciprocal Rank Fusion* [9]. The work described in [25], in the context of social search, uses the *query likelihood model* [10] to calculate topical relevance, a *Multi Criteria Decision Making* (MCDM) approach to calculate information genuineness [34], and a simple linear combination to obtain the final relevance value. The works just illustrated suffer from: (*i*) the need to have labeled datasets available to calculate information genuineness scores [23,29], which can be unavailable or characterized by bias related to domain dependence or to choices made during the labeling process (first and foremost the subjectivity of human assessors), and (*ii*) the need for human intervention in defining the computational model of information genuineness [25].

[6] https://pubmed.ncbi.nlm.nih.gov/.
[7] https://www.mayoclinic.org/.
[8] https://www.rxlist.com/.
[9] https://en.wikipedia.org/wiki/Portal:Medicine.
[10] https://medlineplus.gov/.
[11] https://www.healthnewsreview.org/about-us/review-criteria/.

3 Considering Information Genuineness Based on Scientific Evidence in Health Information Retrieval

The proposed solution, which aims to address the various problems associated with the approaches presented in the previous section, is based on the development of a *retrieval model* capable of considering both *topical relevance* and *information genuineness* in providing access to health-related content. The model focuses, in particular, on the idea of calculating the second criterion on the basis of comparing health claims in distinct health documents and medical journal articles, which are considered reliable sources of scientific evidence for a given query. In this way, we obtain two *query-dependent* relevance scores related to each distinct criterion, which are combined through a suitable *aggregation strategy* for obtaining the final *Retrieval Status Value* (RSV), based on which the estimated relevant documents are ranked. Neither human intervention, nor complex knowledge bases, nor labeled datasets are needed for this purpose. The architecture of the proposed model is illustrated in Fig. 1.

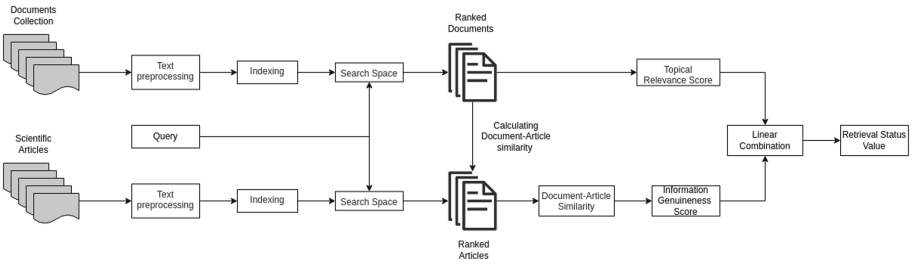


Fig. 1. The proposed retrieval model, considering both topical relevance and information genuineness (based on scientific evidence in the form of medical journal articles).

3.1 Computing Topical Relevance

Topical relevance constitutes the core relevance dimension in any IRS, and assesses how well the content of a document topically meets the information needs of users, which are usually expressed by means of a query [10]. There are several approaches in literature to estimate topical relevance, one of the most effective is still Okapi BM25 [26], which is a lexical-based unsupervised model, a strong baseline for distinct IR tasks, based on a probabilistic interpretation of how terms contribute to the relevance of a document and uses easily computed statistical properties such as functions of term frequencies, document frequencies, and document lengths. Using BM25, the *topical relevance score* of a document d with respect to a query q, denoted as $trs(d, q)$, is calculated as follows:

$$trs(d,q) = \sum_{t \in q,d} \log \frac{N - df(t) + 0.5}{df(t) + 0.5} \cdot \frac{tf(t,d) \cdot (k_1 + 1)}{tf(t,d) + k_1 \cdot (1 - b + b\frac{l_d}{L})} \quad (1)$$

The left part of the equation allows to compute the *inverse document frequency* of a term with respect to the entire document collection; specifically, N denotes the total number of documents in the collection, and $df(t)$ refers to the *document frequency* for the term t, i.e., the number of documents in which t appears. In the second part, $tf(t,d)$ denotes the *term frequency*, i.e., the number of times the term t appears in the document d. Since document collections usually are constituted by documents with different lengths, length normalization is performed in the denominator; specifically, l_d refers to length of the document d, L refers to the average document length, while k_1 (a positive tuning parameter that calibrates the document term frequency scaling) and b (determines the document length scaling) are internal BM25 parameters.

3.2 Computing Information Genuineness

Various approaches have been proposed in the literature to evaluate *information genuineness*,[12] whether health-related or not, whether applied to IR or not. As illustrated in Sect. 2, most of them need either human intervention, or hand-built knowledge bases, or datasets labeled for the purpose.

Without using any of these solutions, in the proposed approach we initially indexed open-source articles extracted from reputed medical journals,[13] such as the *Journal of the American Medical Association* (JAMA),[14] and *eLife*,[15] considered as sources of trustworthy scientific evidence. From these articles, we employed BM25 to retrieve topically relevant ones by considering as queries those extracted from the dataset employed in this work for evaluation purposes, illustrated in Sect. 4. Each retrieved journal article was compared with each retrieved document for the considered query, by using *cosine similarity*. To represent both documents and journal articles, we used two BERT-based textual representation models, one pre-trained on MSMarco,[16] and the other on the *Pubmed* and *PubMed Central* (PMC) datasets.[17] We obtained, this way, dense vector representations based on chunks of 512 tokens, along with a sliding-window of 500 words (to keep context of the past passage) on the whole document. For the top-n retrieved documents and the top-k retrieved journal articles,[18] we obtained an $n \times k$ *similarity matrix*, where rows represent the documents, columns the journal articles, and each cell of the matrix contains the similarity score between the document and the journal article, as shown in Fig. 2.

[12] Although there are numerous terms that have been used in the literature, to refer to this dimension of relevance (e.g., *credibility*, *veracity*, *truthfulness*, etc.), in this and other works we prefer to use the concept of *genuineness* as an abstract term that can grasp various aspects of the above concepts.

[13] https://openmd.com/guide/finding-credible-medical-sources.

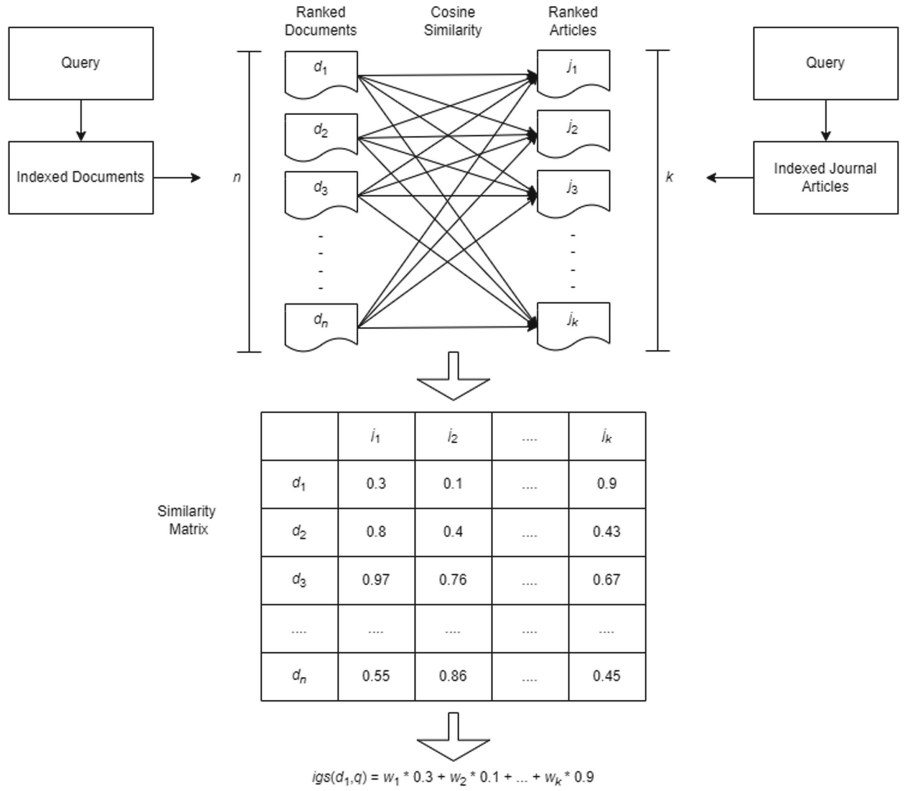
[14] https://jamanetwork.com/.

[15] https://elifesciences.org/.

[16] https://huggingface.co/sentence-transformers/msmarco-distilbert-base-v4.

[17] https://github.com/dmis-lab/biobert.

[18] Where $k << n$, to keep the focus on document retrieval and consider only the most relevant journal articles.



	j_1	j_2	....	j_k
d_1	0.3	0.1	....	0.9
d_2	0.8	0.4	....	0.43
d_3	0.97	0.76	....	0.67
....	....	....	....	....
d_n	0.55	0.86	....	0.45

Fig. 2. Information genuineness score calculation. q denotes the query that is used to retrieve both documents and journal articles.

The *information genuineness score* for each document d with respect to a query q, denoted as $igs(d, q)$, is obtained by *linearly combining* the similarity scores among d and the top-k journal articles j_i that were estimated relevant for the same query for which d was retrieved, by considering distinct weights proportional to the positions in the *ranking* of the retrieved journal articles. Formally:

$$igs(d, q) = w_1 * \cos(d, j_1) + w_2 * \cos(d, j_2) + \ldots + w_k * \cos(d, j_k) \quad (2)$$

In Eq. 2, $w_1, w_2, \ldots, w_k$ denote the weights assigned to each similarity score, such that $\sum w_i = 1$ and $w_i \geq w_{i+1}$ $(1 \leq i \leq k-1)$. This second condition serves to consider the position in the rank in which the journal articles were positioned with respect to the similarity to the documents retrieved (i.e., the higher the position, the higher the weight). The way in which the w_i weights are actually assigned, for evaluation purposes, is illustrated in detail in Sect. 4.3.

3.3 Computing the Retrieval Status Value

Once the two relevance dimension scores have been obtained, both of which in this case are dependent on the query formulated by the user, it was necessary to aggregate them to obtain the *Retrieval Status Value*, denoted as $\text{RSV}(d, q)$, which represents the final relevance score of a document with respect to a query given topical relevance and information genuineness. In this case, we also opted for a *linear combination* among the scores. Formally:

$$\text{RSV}(d, q) = w_{trs} * trs(d, q) + w_{igs} * igs(d, q) \tag{3}$$

In Eq. 3, w_{trs} denotes the weight assigned to the topical relevance score, and w_{igs} denotes the weight assigned to the information genuineness score. Also in this case, each weight w_{**s} is actually assigned, for evaluation purposes, as illustrated in Sect. 4.3. In the same section, the solution adopted to normalize the two relevance dimension scores in the same numerical range is also explained, since they are calculated in different ranges.

4 Evaluation Framework and Results

This section describes the experimental evaluation framework that was set up to assess the effectiveness of the retrieval model presented in this article. A BM25 baseline and several model configurations are evaluated on a public dataset and by means of suitable evaluation metrics.[19] The purpose of this experimental evaluation is to punctually assess the effectiveness of such configurations of the proposed approach in using external reputed sources (medical journal articles) to consider information genuineness as a query-dependent dimension of relevance, compared to the simple baseline chosen that uses topical relevance alone. Comparison with IR solutions that consider information genuineness (or related concepts) in the context of supervised or requiring human intervention solutions (see Sect. 2) is currently under study.

4.1 The TREC Health Misinformation Track Dataset

The TREC *Health Misinformation Track* fosters research on retrieval methods that promote reliable and correct information over misinformation for health-related decision-making tasks.[20] In this work, we used a subset of the dataset provided by the Track in its 2020 edition [8]. The original dataset is constituted by *CommonCrawl news*,[21] sampled from January, 1st 2020 to April 30th, 2020, which contains health-related news articles from all over the world. For our experiments, given the large volume of the original dataset, we selected 219,245

[19] Materials and documentation for reproducing the experiments are available at the following link: https://github.com/ikr3-lab/misinformation-wise2022.

[20] https://trec-health-misinfo.github.io/.

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

English news related to COVID-19. The dataset has a fixed structure, organized into *topics*. Each topic includes a *title*, a *description*, which reformulates the title as a question, a *yes/no answer*, which is the actual answer to the description field based on the provided evidence, and a *narrative*, which describes helpful and harmful documents in relation to the given topic. For example, for the topic title field: 'ibuprofen COVID-19', the value of the other attributes in the dataset are, for the description: 'Can ibuprofen worsen COVID-19?', for the yes/no answer: 'no', and for the narrative: 'Ibuprofen is an anti-inflammatory drug used to reduce fever and treat pain or inflammation'.

The considered dataset also consists of an *evaluation set* of 5,340 labeled data. The data is labeled with respect to *usefulness*, *answer*, and *credibility*. Usefulness corresponds to topical relevance, answer indicates if the document provides an answer to the query contained in the description field, and credibility is the concept that, in the document collection, is used to indicate information genuineness. In this work, we just considered as labels usefulness and credibility. Both of them are provided on a binary scale, i.e., useful or non-useful, and credible or non-credible.

4.2 Evaluation Metrics

The TREC Health Misinformation Track, in addition to provide publicly available data, also provides an evaluation tool in which standard IR evaluation measures are implemented, especially when referring to multiple dimensions of relevance, which we have therefore also used for our experiments. The measures considered in this work are *Average Precision* (AP) and *Normalized Discounted Cumulative Gain* [15] on the first 10 results (NDCG@10), both computed by means of the MM *evaluation framework* for multidimensional relevance estimation [22], and two different implementations of the *Convex Aggregating Measure* (CAM) [20], one based on *Mean Average Precision* (MAP) [15], and the other on NDCG@n, for a distinct number of n results retrieved. Specifically:

- The MM framework for multidimensional relevance evaluation allows to incorporate distinct relevance criteria in the assessment of the effectiveness of an IRS along with topical relevance. In such a framework, firstly the evaluation results for each dimension of relevance are calculated separately using distinct evaluation measures. Taking inspiration from the measures used by the TREC Decision Track 2019 [1], we considered both *Average Precision* and NDCG@10. Finally, these scores are combined into a measure using the *weighted harmonic mean*. As the weighted harmonic mean is particularly sensitive to a single lower-than-average value, thus it will reward systems that are consistently more effective across all relevance dimensions.
- The *Convex Aggregating Measure* (CAM) is defined as the *convex sum* of the distinct evaluation results computed with respect to each relevance dimension considered. Formally:

$$\mathrm{CAM}(r) = \lambda_{rel} M_{rel}(r) + \lambda_{cred} M_{cre}(r) \tag{4}$$

where r denotes the number of documents, and M_{rel}, and M_{cre} denote the relevance (i.e., topical relevance), and credibility (i.e., information genuineness) evaluation measures considered. In our work, we applied both Mean Average Precision and NDCG@n. In the equation, λ denotes a weight to assign more importance to one of the two relevance dimensions, under the condition that $\lambda_{rel} + \lambda_{cred} = 1$. For our evaluation, we set the value of λ for each dimension to 0.5, as performed in the TREC 2020 Health Misinformation Track.

4.3 Implementation Technical Details

This section provides some technical details related to the implementation of the proposed solution, regarding indexing and other basic IR operations, the assignment of weights for the calculation of information genuineness and RSV, and the normalization of topical relevance and information genuineness scores into a single numerical range.

Basic IR Operations. To index documents, compute topical relevance scores, and implementing BM25-based retrieval models, we employed the implementation of BM25 provided in *PyTerrier*,[22] with default parameters. To retrieve documents, we used the description of the topic in the considered TREC 2020 dataset as the query. The same procedure was also adopted to find the journal articles related to the query considered, which are used to calculate the information genuineness score, as illustrated in Sect. 3.2.

Assignment of Weights. For assigning both w_i and w_{**s} weights, different solutions can be adopted. One can choose these values heuristically, or on the basis of greedy strategies, or on other ad hoc models. It is not the purpose of this article to determine which solution is best. Exclusively for evaluation purposes, it was decided to employ the weight assignment solution mentioned in [37] in the Information Retrieval field for computing the w_i weights; with this solution, ten queries were randomly selected, and a *grid search* strategy was performed with distinct weights to assess the best results, in computing both $trs(d, q)$ and $igs(d, q)$, in terms of CAM_{MAP}.[23] This latter is the official metric used in TREC 2020 Health Misinformation Track [8], fully described in Sect. 4.1. Regarding the calculation of the w_{**s} weights, it was decided to heuristically consider the case where equal importance is given to both dimensions of relevance, i.e., $w_{trs} = w_{igs} = 0.5$, the case where topical relevance is considered more important, i.e., $w_{trs} = 0.6$ and $w_{igs} = 0.4$, and the case where information genuineness is considered more important, i.e., $w_{trs} = 0.4$ and $w_{igs} = 0.6$.

[22] PyTerrier is a Python-based retrieval framework for simple and complex information retrieval (IR) pipelines by making use of Terrier IR platform for basic document indexing and retrieval. https://github.com/terrier-org/pyterrier.

[23] This choice does not impact the general unsupervised nature of the solution proposed in this article. It is only the simplest, least expensive, and already used in the IR literature solution to be able to provide an initial experimental evaluation.

Normalization of Relevance Dimension Scores. Topical relevance scores are computed using Eq. 1, which does not return values in a predetermined numerical range. In contrast, information genuineness scores, obtained by Eq. 2, take values in the range [0,1]. Hence, to make sure that both relevance scores are in the same numerical range, we normalized topical relevance scores using the *min-max normalization* [19]. Formally:

$$trs'(d,t) = \frac{trs(d,q) - \min_{trs(q)}}{\max_{trs(q)} - \min_{trs(q)}} \tag{5}$$

where $trs'(d,q)$ and $trs(d,q)$ are the normalized and original topical relevance scores of a document d for a query q; $\min_{trs(q)}$ and $\max_{trs(q)}$ are the minimum and maximum topical relevance scores for all documents retrieved for q.

4.4 Results and Discussion

This section illustrates and discusses the results obtained by considering a simple BM25 retrieval model as a baseline, with respect to different model configurations proposed in this paper, against the considered evaluation metrics. In particular, the differences in model configurations aim to assess the impact of different textual representation of documents and journal articles (i.e., BERT vs BioBERT) and the relevance dimensions considered (we recall that distinct weights were heuristically attributed to topical relevance and information genuineness). The results are illustrated in Tables 1, 2, and 3, where:

- **BM25** denotes the baseline model, based on just topical relevance;
- **Model (1)** denotes the proposed retrieval model in which we compute topical relevance scores using BM25, and information genuineness scores by considering documents and top-10 journal articles as scientific evidence, both represented as BERT embeddings. In this model, the RSV is calculated by linearly combining (Eq. 3) topical relevance and information genuineness scores with equal weights (i.e., $w_{trs} = 0.5$ and $w_{igs} = 0.5$);
- **Model (2)** denotes the proposed retrieval model which differs from the previous one only in that RSV is calculated by assigning different weights to different relevance dimensions in the linear combination, specifically 0.6 to topical relevance, and 0.4 to information genuineness;
- **Model (3)** differs from Model (2) because, in the linear combination, a weight equal to 0.4 is assigned to topical relevance, and to 0.6 to information genuineness;
- **Model (4)** differs from Model (1) because documents and top-10 journal articles are represented as BioBERT embeddings;
- **Model (5)** differs from Model (2) because documents and top-10 journal articles are represented as BioBERT embeddings;
- **Model (6)** differs from Model (3) because documents and top-10 journal articles are represented as BioBERT embeddings.

Table 1. Experimental results obtained by using the MM evaluation framework in terms of Average Precision (AP) ad NDCG@10. Evaluations are performed by considering the same number of top-k journal articles, i.e., $k = 10$, as scientific evidence. Statistically significant results ($p < 0.05$ using the t-test [31]) are denoted with $*$.

Model	w_{trs}	w_{igs}	AP	NDCG@10	Embeddings
BM25	–	–	0.461	0.8601	–
Model (1)	0.5	0.5	0.469	0.8676	BERT
Model (2)	0.6	0.4	0.474	0.8701	BERT
Model (3)	0.4	0.6	0.476	0.8747	BERT
Model (4)	0.5	0.5	0.479	0.8785*	BioBERT
Model (5)	0.6	0.4	0.481*	0.8813*	BioBERT
Model (6)	0.4	0.6	0.493*	0.8951*	BioBERT

Table 1, which summarizes the effectiveness results of model configurations in terms of both AP and NDCG@10, shows that joint consideration of two relevance criteria in computing the final RSV improves system performance compared with topical relevance alone (BM25), regardless of the specific model configuration selected. Again for all the proposed models, analysis of the results suggests that assigning a higher weight to information genuineness leads to slightly higher performance. Finally, we can observe that Models (4), (5), and (6) are the best performing, all of which are based on the BioBERT representation. This suggests how taking into account the medical scientific vocabulary within (semantic- and context-aware) textual representation can actually improve health search results even in terms of information genuineness.

The greater effectiveness of models that are based on the BioBERT representation compared with the BERT representation, almost under each model configuration, also emerges from Table 2, which summarizes results in terms of both CAM_{MAP} and $\text{CAM}_{\text{NDCG@n}}$. This observation remains valid for whatever the number of documents retrieved from the system (in this case we chose to consider $n = 5$, 10, 15 and 20 retrieved documents). Also with respect to these measures, it can be observed that assigning higher weights to information genuineness produces slightly better results. The results in terms of these measures are even more significant because of the very nature of the two metrics, which explicitly combine assessments related to the two distinct dimensions of relevance.

Finally, to test the effectiveness of both textual representations as the number of articles taken as scientific evidence increases, i.e., for $k = 5$, 10, and 15, we kept fixed the number of retrieved documents on which the assessments were made (specifically, $n = 20$), and employed both Model (3) and Model (6), which are the ones who provided the best results in Table 2 for the BERT and BioBERT representations. From Table 3, summarizing these results in terms of both CAM_{MAP} and $\text{CAM}_{\text{NDCG@20}}$, we can observe that increasing the number of journal articles taken into account as scientific evidence actually contributes positively to the improved results obtained. Also in this case, the superiority of

Table 2. Experimental results in terms of Convex Aggregating Measure (CAM), w.r.t. both Mean Average Precision (MAP) and NDCG@n, for the top-n documents (# n docs) considered in different runs. The number of top-k journal articles considered as scientific evidence (# k j.arts) is fixed, i.e., $k = 10$. Statistically significant results.

Model	# n docs	w_{trs}	w_{igs}	# k j.arts	CAM_{MAP}	$\text{CAM}_{\text{NDCG@}n}$	Embeddings
BM25		1	–	–	0.0631	0.1435	–
Model (1)		0.5	0.5	10	0.0641	0.1434	BERT
Model (2)		0.6	0.4	10	0.0685	0.1475	BERT
Model (3)	5	0.4	0.6	10	0.0697	0.1495	BERT
Model (4)		0.5	0.5	10	0.0701	0.1487	BioBERT
Model (5)		0.6	0.4	10	0.0721	0.1500	BioBERT
Model (6)		0.4	0.6	10	**0.0894**	**0.1688**	BioBERT
BM25		1	–	–	0.1047	0.2052	–
Model (1)		0.5	0.5	10	0.1073	0.2057	BERT
Model (2)		0.6	0.4	10	0.1085	0.2084	BERT
Model (3)	10	0.6	0.4	10	0.1145	0.2151	BERT
Model (4)		0.5	0.5	10	0.1124	0.2112	BioBERT
Model (5)		0.6	0.4	10	0.1177	0.2161	BioBERT
Model (6)		0.4	0.6	10	**0.1249**	**0.2299**	BioBERT
BM25		1	–	–	0.0631	0.1435	–
Model (1)		0.5	0.5	10	0.1399	0.249	BERT
Model (2)		0.6	0.4	10	0.1435	0.2535	BERT
Model (3)	15	0.4	0.6	10	0.1485	0.2552	BERT
Model (4)		0.5	0.5	10	0.1489	0.2541	BioBERT
Model (5)		0.6	0.4	10	0.1507	0.259	BioBERT
Model (6)		0.4	0.6	10	**0.1597**	**0.2702**	BioBERT
BM25		1	–	–	0.1676	0.285	–
Model (1)		0.5	0.5	10	0.1649	0.2845	BERT
Model (2)		0.6	0.4	10	0.1726	0.2905	BERT
Model (3)	20	0.4	0.6	10	0.1797	0.2945	BERT
Model (4)		0.5	0.5	10	0.1753	0.2902	BioBERT
Model (5)		0.6	0.4	10	0.1783	0.2948	BioBERT
Model (6)		0.4	0.6	10	**0.1978**	**0.3102**	BioBERT

the model based on the BioBERT representation is confirmed, regardless of the number of journal articles considered.

Table 3. Comparison of Model (3) and Model (6) by considering the same number, i.e., $n = 20$, of retrieved documents and a different number of top-k journal articles (# k j.arts), as scientific evidence. Statistically significant results.

Model	# k j.arts	CAM_{MAP}	$\text{CAM}_{\text{NDCG@20}}$	Embedding
Model (3)	5	0.1698	0.285	BERT
Model (6)		**0.1787**	**0.2953**	BioBERT
Model (3)	10	0.1797	0.2945	BERT
Model (6)		**0.1978**	**0.3102**	BioBERT
Model (3)	15	0.1810	0.2912	BERT
Model (6)		**0.1975**	**0.3109**	BioBERT

5 Conclusions and Further Research

In the context of the spread of increasingly health misinformation online, in this article we addressed the problem of how to provide online users with topically relevant yet genuine information by proposing a retrieval model that considers scientific evidence in the form of reputed medical international journal articles in calculating so-called information genuineness.

Unlike other approaches that have been presented in the literature, which rely on the use of experts, or manually constructed knowledge bases, or labeled datasets and supervised approaches to assess the genuineness of information in retrieval models, in this article we have attempted to give a simple yet effective unsupervised solution to compare health content circulating online directly with the content of scientific articles, thereby succeeding in providing an automatic and non-time-consuming solution.

The results obtained showed that this approach is indeed effective when considering together topical relevance (calculated by state-of-the-art methods) and information genuineness calculated as in the proposed method. In particular, it can be seen that if the documents in the collection and the articles taken as scientific evidence are represented by embeddings related to the domain under consideration, the proposed solution is even more effective.

As for future developments to consider, there is first of all the comparison with other literature baselines in IR, both in terms of effectiveness and efficiency. It will also be necessary to further study the impact of individual relevance dimensions on the final results, as in this article we have only begun this investigation by heuristically testing a few configurations addressing this issue (which, in any case, have made it possible to observe that information genuineness can have a non-negligible impact in calculating the best Retrieval Status Value). Automated methods will also have to be considered to build knowledge bases that can actually exploit more semantic information than simply comparing textual representations between documents and reference articles.

Acknowledgment. This work was supported by the EU Horizon 2020 ITN/ETN on Domain Specific Systems for Information Extraction and Retrieval (H2020-EU.1.3.1., ID: 860721).

References

1. Abualsaud, M., et al.: Overview of the TREC 2019 Decision Track (2020)
2. Bal, R., et al.: Analysing the extent of misinformation in cancer related tweets. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 14, pp. 924–928 (2020)
3. Barua, Z., et al.: Effects of misinformation on COVID-19 individual responses and recommendations for resilience of disastrous consequences of misinformation. Prog. Disaster Sci. **8**, 100119 (2020)
4. Boyer, C., Selby, M., Appel, R.: The health on the net code of conduct for medical and health websites. In: Proceedings of the 9th World Congress on Medical Informatics, vol. 2, pp. 1163–1166. IOS Press (1998)
5. Chang, Y.S., Zhang, Y., Gwizdka, J.: The effects of information source and eHealth literacy on consumer health information credibility evaluation behavior. Comput. Hum. Behav. **115**, 106629 (2021)
6. Charnock, D., et al.: DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J. Epidemiol. Commun. Health **53**(2), 105–111 (1999)
7. Chou, W.Y.S., Oh, A., Klein, W.M.: Addressing health-related misinformation on social media. JAMA **320**(23), 2417–2418 (2018)
8. Clarke, C.L.A., et al.: Overview of the TREC 2020 Health Misinformation Track (2020). https://trec.nist.gov/pubs/trec29/papers/OVERVIEW.HM.pdf
9. Cormack, G.V., Clarke, C.L., Buettcher, S.: Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in IR, pp. 758–759 (2009)
10. Croft, W.B., Metzler, D., Strohman, T.: Search Engines: Information Retrieval in Practice, vol. 520. Addison-Wesley Reading (2010)
11. Cui, L., et al.: DETERRENT: knowledge guided graph attention network for detecting healthcare misinformation. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 492–502 (2020)
12. Enders, A.M., et al.: The different forms of COVID-19 misinformation and their consequences. Harv. Kennedy Sch. Misinf. Rev. (2020)
13. Ernst, P., et al.: KnowLife: a knowledge graph for health and life sciences. In: 2014 IEEE 30th International Conference on Data Engineering, pp. 1254–1257. IEEE (2014)
14. Fox, S., Duggan, M.: Health online 2013. Health **2013**, 1–55 (2013)
15. Järvelin, K., Kekäläinen, J.: IR evaluation methods for retrieving highly relevant documents. In: ACM SIGIR Forum, vol. 51, pp. 243–250. ACM, New York (2017)
16. Kickbusch, I.S.: Health literacy: addressing the health and education divide. Health Promot. Int. **16**(3), 289–297 (2001)
17. Kinkead, L., Allam, A., Krauthammer, M.: AutoDiscern: rating the quality of online health information with hierarchical encoder attention-based neural networks. BMC Med. Inform. Decis. Mak. **20**(1), 1–13 (2020)

18. Kou, Z., Shang, L., Zhang, Y., Wang, D.: HC-COVID: a hierarchical crowdsource knowledge graph approach to explainable COVID-19 misinformation detection. In: Proceedings of the ACM on Human-Computer Interaction, vol. 6, no. GROUP, pp. 1–25 (2022)
19. Lee, J.H.: Analyses of multiple evidence combination. In: Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in IR, pp. 267–276 (1997)
20. Lioma, C., Simonsen, J.G., Larsen, B.: Evaluation measures for relevance and credibility in ranked lists. In: Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval, pp. 91–98 (2017)
21. Metzger, M.J., Flanagin, A.J.: Psychological approaches to credibility assessment online. In: The Handbook of the Psychology of Communication Technology, pp. 445–466 (2015)
22. Palotti, J., Zuccon, G., Hanbury, A.: MM: a new framework for multidimensional evaluation of search engines. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, pp. 1699–1702 (2018)
23. Pradeep, R., Ma, X., Nogueira, R., Lin, J.: Vera: prediction techniques for reducing harmful misinformation in consumer health search. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in IR, pp. 2066–2070 (2021)
24. Pradeep, R., Nogueira, R., Lin, J.: The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models. arXiv preprint arXiv:2101.05667 (2021)
25. Putri, Divi Galih Prasetyo., Viviani, Marco, Pasi, Gabriella: Social search and task-related relevance dimensions in microblogging sites. In: Aref, S., et al. (eds.) SocInfo 2020. LNCS, vol. 12467, pp. 297–311. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60975-7_22
26. Robertson, S., Zaragoza, H.: The Probabilistic Relevance Framework: BM25 and Beyond. Now Publishers Inc. (2009)
27. Sackett, D.L.: Evidence-based medicine. In: Seminars in Perinatology, vol. 21, pp. 3–5. Elsevier (1997)
28. Samuel, Hamman, Zaïane, Osmar: MedFact: towards improving veracity of medical information in social media using applied machine learning. In: Bagheri, Ebrahim, Cheung, Jackie C. K.. (eds.) Canadian AI 2018. LNCS (LNAI), vol. 10832, pp. 108–120. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-89656-4_9
29. Schlicht, I.B., de Paula, A.F.M., Rosso, P.: UPV at TREC Health Misinformation Track 2021 Ranking with SBERT and Quality Estimators. arXiv preprint arXiv:2112.06080 (2021)
30. Schulz, P.J., Nakamoto, K.: The perils of misinformation: when health literacy goes awry. Nat. Rev. Nephrol. 1–2 (2022)
31. Smucker, M.D., Allan, J., Carterette, B.: A comparison of statistical significance tests for information retrieval evaluation. In: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, pp. 623–632 (2007)
32. Suominen, H., et al.: Overview of the CLEF eHealth evaluation lab 2021. In: Candan, K.S., et al. (eds.) CLEF 2021. LNCS, vol. 12880, pp. 308–323. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-85251-1_21
33. Upadhyay, R., Pasi, G., Viviani, M.: Health misinformation detection in web content: a structural-, content-based, and context-aware approach based on Web2Vec. In: Proceedings of the Conference on Information Technology for Social Good, pp. 19–24 (2021)

34. Viviani, Marco, Pasi, Gabriella: A multi-criteria decision making approach for the assessment of information credibility in social media. In: Petrosino, Alfredo, Loia, Vincenzo, Pedrycz, Witold (eds.) WILF 2016. LNCS (LNAI), vol. 10147, pp. 197–207. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-52962-2_17
35. Viviani, M., Pasi, G.: Credibility in social media: opinions, news, and health information-a survey. Wiley Interdiscip. Rev.: Data Mining Knowl. Discov. **7**(5), e1209 (2017)
36. Wardle, C., et al.: Thinking about 'information disorder': formats of misinformation, disinformation, and mal-information. In: Ireton, C., Posetti, J. (eds.) Journalism,'Fake News' & Disinformation, pp. 43–54. UNESCO, Paris (2018)
37. Wu, S., et al.: Assigning appropriate weights for the linear combination data fusion method in information retrieval. Inf. Process. Manag. **45**(4), 413–426 (2009)
38. Zhao, Y., Da, J., Yan, J.: Detecting health misinformation in online health communities: incorporating behavioral features into machine learning based approaches. Inf. Process. Manag. **58**(1), 102390 (2021)
39. Zuo, C., Zhang, Q., Banerjee, R.: An empirical assessment of the qualitative aspects of misinformation in health news. In: Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda, pp. 76–81 (2021)