

# An Overview on Evaluation Labs and Open Issues in Health-related Credible Information Retrieval

Discussion Paper

Rishabh Upadhyay, Gabriella Pasi and Marco Viviani

University of Milano-Bicocca – Department of Informatics, Systems, and Communication (DISCO)

Information and Knowledge Representation, Retrieval, and Reasoning (IKR3) Lab

Edificio U14, Viale Sarca, 336 – 20126 Milan, Italy – <https://ikr3.disco.unimib.it>

## Abstract

Faced with the problem of widespread health misinformation, in recent years credibility is being considered as one of the important dimensions for health-related information access and retrieval. To encourage research in this field, a couple of evaluation labs have been recently set up to provide large test collections, baselines, and evaluation metrics to interested researchers. The purpose of this article is to provide an overview of such evaluation labs, discussing their characteristics and open issues.

## Keywords

Health-related Information, Credibility, Consumer Health Search, Evaluation Labs.

## 1. Introduction

In the last years, the Web is increasingly being used to search for various health-related information, ranging from medical therapies and treatments to lifestyle and wellness. According to distinct surveys, a range of 60-70 percent of adults in the USA [1, 2], and around one in two EU citizens [3], look for health information online. Recently, as health-related searches on Google are getting so popular, the term "Dr. Google" has also been coined [4]. Furthermore, exchange health-related information on social media is also becoming a common practice [2]; Twitter, for example, is a widely used microblogging platform employed by both patients and healthcare professionals [5], for both consumer health search and advertisement purposes.

In this context, it is increasingly easy for people to run into *health misinformation* [6]. Although numerous attempts have been made to provide online users with genuine content in distinct domains, the development of automated approaches to tackle this issue in the health scenario is still in its infancy. Notwithstanding, relying on misinformation in such a context can be particularly harmful, especially for users without sufficient health literacy.


In this overview paper, our purpose is to outline and discuss the major issues related to health-related information credibility, and to present two evaluation labs that have been established in


---

IIR 2021 – 11th Italian Information Retrieval Workshop, September 13–15, 2021, Bari, Italy

✉ [rishabh.upadhyay@unimib.it](mailto:rishabh.upadhyay@unimib.it) (R. Upadhyay); [gabriella.pasi@unimib.it](mailto:gabriella.pasi@unimib.it) (G. Pasi); [marco.viviani@unimib.it](mailto:marco.viviani@unimib.it) (M. Viviani)

ORCID  0000-0001-9937-6494 (R. Upadhyay); 0000-0002-6080-8170 (G. Pasi); 0000-0002-2274-9050 (M. Viviani)

 © 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

recent years to the aim of accounting for the above issues when considering IR in the health domain.

## 2. Health-related Information Credibility and Evaluation Labs

The problem of the credibility of online information has been studied for at least a decade now, both in computer and data science [7]. While in social sciences *credibility* is understood as a subjective characteristic perceived by the information receiver [8], in computer science, and, therefore, in the development of automated solutions to verify the genuineness of information, it is necessary to produce an "objective" credibility assessment, by considering various characteristics (i.e., features) related to the contents, their authors and the social relations between users in the case of information spread through social platforms [7].<sup>1</sup> Such assessment can be used to produce either a *binary* (i.e., *credible* versus *non-credible*), *multi-class* (e.g., *credible*, *non-credible*, *non-judged*) or *ordinal classification* (e.g., *non-credible*, *partially credible*, *credible*, *highly credible*) of the information, or even a credibility-based *ranking*.

In recent years, some works have tried to consider the credibility of information as an aspect of *relevance* in various Information Retrieval tasks [9, 10], including *consumer health search* (CHS). In this context, some evaluation initiatives based on the Cranfield paradigm have been set up to allow researchers to test the ability of their systems to account for credibility in relevance assessment. Below, we discuss the major evaluation labs that addresses the above-mentioned issue. Although in evaluation initiatives such as FIRE and NTCIR, information credibility has somewhat been considered (in particular in the UrduFake [11] and Lab-PoliInfo [12] labs), only in TREC and CLEF a couple of recent labs have considered the health domain.

### 2.1. TREC

The *Text REtrieval Conference* (TREC) has included the *Health Misinformation Track* in 2019.<sup>2</sup> Participants in this track are required, among other sub-tasks, to develop systems that "return relevant and credible information that will help searchers make correct decisions" in the health domain. Depending on the Track edition, other criteria have been considered beyond credibility, which are briefly illustrated in the following sections.

#### 2.1.1. Data Collections

The 2019 Health Misinformation Track used the ClueWeb12-B13 dataset as a corpus.<sup>3</sup> Such a corpus is constituted by English Web pages collected in 2012 related to various health issues, and containing both correct and incorrect information, and of varying credibility and quality. The 2020 Track used a dataset provided by Common Crawl, in particular related to different news collected in the first four months of 2020.<sup>4</sup> On such dataset, 74 COVID-19 related topics have

---

<sup>1</sup>The problem is made even more complex by the fact that some related but not totally overlapping terms are used in the literature in addition to credibility, including *veracity*, *trustworthiness*, *reliability*, etc. This article is not intended to disambiguate the use of these terms, but this is an issue that will certainly need to be further addressed.

<sup>2</sup><https://trec-health-misinfo.github.io/2019.html>

<sup>3</sup><https://lemurproject.org/clueweb12/>

<sup>4</sup><https://commoncrawl.org/2016/10/news-dataset-available/>

been selected, and Web pages filtered accordingly. In the current 2021 edition, the "noclean" version of the C4 dataset used by Google has been employed.<sup>5</sup>

### 2.1.2. Human Assessment

In the different editions of the Health Misinformation Track, documents have been labeled by human assessors with respect to distinct criteria: *relevance*, *efficacy*, and *credibility* in 2019, *usefulness*, *correctness*, and *credibility* in 2020, and *usefulness*, *credibility*, and *supportiveness* in 2021. Efficacy concerns the presence in the document of "correct" information regarding the topic's treatment. This is similar to the correctness criterion employed in 2020. Both efficacy and correctness have been assessed on a three-point scale, including a "non-judged" label. Supportiveness is intended as the ability of the document to support or dissuade the use of the treatment in the topic's question. This criterion has been assessed on a three-point scale, including a neutral value. It is important to note that in all three editions, documents judged as non-relevant (or non-useful) have not been further assessed with respect to additional criteria.

As regards the criterion that interests us most, namely credibility, it had been assessed on a three-point scale (including a "non-judged" label) in 2019, and on a binary scale in the two last editions. Human assessors were asked to provide a credibility label based, among others, on the following aspects: the amount of expertise, authoritativeness, and trustworthiness of the document, the indication of an author or an institute that published the Web document and their credentials, the presence of citations to trustworthy/credible sources, the style of writing (well written or poorly written), the purpose for which the document is written (to provide information or for advertising purposes). In each edition, around 20,000 labeled documents with over 50 topics have been provided.

### 2.1.3. Baselines and Evaluation Metrics

In both the 2019 and 2020 Health Misinformation Track, baselines based on the BM25 retrieval model implemented by employing the Anserini toolkit,<sup>6</sup> with default parameters, have been employed. Both the baselines and the submitted runs have been evaluated with respect to the following measures, proposed in [13], to take into account the different criteria considered: *Normalized Local Rank Error* (NLRE): the rank positions of documents are compared pairwise checking for "errors", which are defined as misplacement of documents, i.e., relevant or credible documents placed after non-relevant or non-credible documents; *Normalized Weighted Cumulative Score* (nWCS): a single label out of the multiple criteria is generated, and the standard nDCG measure is computed; *Convex Aggregating Measure* (CAM): each criterion is considered separately, and either AP or nDCG with respect to the ranking obtained with the single criterion is computed. Finally, the average AP or nDCG value is computed. In 2020, runs have been also evaluated in terms of "traditional" evaluation measures, i.e., nDCG@*k* and MAP, to compare measures accounting for relevance only and those that account for usefulness, credibility, and correctness.

---

<sup>5</sup><https://huggingface.co/datasets/allenai/c4>

<sup>6</sup><https://github.com/castorini/anserini>

## 2.2. CLEF

The *Conference and Labs of the Evaluation Forum* (CLEF) has included, starting from 2018, tasks related to the automatic identification and verification of claims in social media, with the *CheckThat! Lab* [14]; however, it is only since 2020 that the concept of credibility has been considered in the context of CHS in the eHealth Lab,<sup>7</sup> which also sees us as co-organizers. The aim is to assess the ability of systems to retrieve documents that are relevant, readable, and credible; in addition, there is a sub-task that is specifically dedicated to credibility prediction.

### 2.2.1. Data Collections

For both the 2020 and 2021 eHealth editions, Web pages has been collected by repeatedly submitting a set of CLEF eHealth 2018 queries to the Microsoft Bing APIs,<sup>8</sup> over a period of a few weeks. The list of obtained Web documents have been further augmented to add other reliable and unreliable Web pages, and this augmentation was based on the Web sites previously compiled by health institutions and agencies. Additionally, in the 2021 edition, social media content from Reddit and Twitter has been also considered. Such content has been gathered with respect to 150 health-related topics. Queries have been manually generated from such topics, and used to filter posts and tweets from Reddit and Twitter, respectively. A Reddit document consists of a so-called "submission", i.e. a post that have a title and a description, in which a question is generally made, and a "comment", i.e., a "reply" to the submission, whereas for Twitter, a single tweet and related metadata constitutes the document.

### 2.2.2. Human Assessment

In CLEF eHealth, documents have been labeled with respect to three criteria, i.e., (topical) *relevance*, *readability* (or *understandability*), and *credibility*. Relevance and readability have been assessed on a three-point scale, i.e., *non-relevant/readable*, *partially relevant/readable*, *relevant/readable*. Regarding credibility, it was considered useful to introduce a fourth label, namely "not able to judge", given the peculiarity of this criterion.

In particular, in assessing the credibility of Web pages and social content, human assessors have been required to consider the availability of trustworthiness indicators of the source (e.g., expertise, Web reputation, etc.), the syntactic and semantic characteristics of the content (e.g., the writing style), the emotions that the text seeks to evoke, the presence of verifiable facts and assertions (e.g., by the presence of citations or external links), the analysis of the social relationships of the author of a post (in the case of social content).

### 2.2.3. Baselines and Evaluation Metrics

In CLEF eHealth 2020, organizers have developed baseline methods based on the Okapi BM25 retrieval model and query expansion optimized via reinforcement learning. The query expansion model has been pre-trained using the TREC-CAR, Jeopardy, and Microsoft Academic datasets from [15], and the expanded queries employed as the input of the BM25 model. In CLEF eHealth

---

<sup>7</sup><https://clefehealth.imag.fr/>

<sup>8</sup>[https://github.com/CLEFeHealth/CLEFeHealth2018IRtask/blob/master/clef2018\\_queries\\_task2\\_task3.txt](https://github.com/CLEFeHealth/CLEFeHealth2018IRtask/blob/master/clef2018_queries_task2_task3.txt)

2021, six baselines systems based on the Okapi BM25, Dirichlet Language Model (DirichletLM), and Term Frequency times Inverse Document Frequency (TF×IDF) retrieval models with default parameters have been provided. Further details can be found in [16].

The following evaluation metrics have been used to assess the baselines and the submitted runs: MAP, BPref, nDCG, uRBP and cRBP, to evaluate the the systems with respect to the ranking produced by considering three criteria, and Accuracy, F1-score, and AUC, to assess the goodness of the (binary) classification of documents with respect to credibility only. Regarding the uRBP (*understandability Rank Biased Precision*) and cRBP (*credibility Rank Biased Precision*) metrics, they serve the purpose to account for the contribution of understandability and credibility in the ranking produced by the retrieval models. uRBP has been introduced in [17] while cRBP has been employed for the first time (based on uRBP), in the 2020 edition of CLEF eHealth [18].

### 3. Discussion and Open Issues

Access to credible health-related information is one of the most challenging aspects of current research in Information Retrieval. The development of evaluation initiatives that take this into consideration, as outlined in this article, is undoubtedly promising, however some aspects need to be further considered.

It is first necessary to consider that the health domain is characterized by the presence of medical experts, and that health-related content can be marked by the use of a very specific language. It is also necessary to consider that health-related information is disseminated both in the form of Web pages and short texts (i.e., social content), and so it is necessary to consider the problem of what constitutes a single unit of retrievable information. Finally, there is the problem of evaluating the effectiveness of a retrieval system in considering credibility over other relevance criteria.

We have seen that current evaluation labs attempt to consider some of these issues. However, in the future, it would be necessary to act in the following directions: (*i*) provide different scenarios regarding content published by experts (for informational purposes) versus content published for example in virtual communities (in a context of opinion exchange) (a first attempt was made in CLEF eHealth 2020); (*ii*) better consider that evaluation by human assessors may be different than evaluating Web pages and synthetic social content (a first attempt was made in CLEF eHealth 2021); (*iii*) identify well, with respect to both Web pages and social content, what content is being assessed (e.g., in a Web page there may be several sections with different credibility, whereas the credibility of social media posts may be considered individually or with respect to the thread containing it); (*iv*) develop new credibility-oriented assessment measures (the work of [13] and the metrics employed in both the TREC Health Misinformation Track and CLEF eHealth constitute an important first step in this direction).

### Acknowledgments

This work is supported by the EU Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie Grant Agreement No 860721 – DoSSIER: “Domain Specific Systems for Information Extraction and Retrieval”.

## References

- [1] K. Purcell, et al., Understanding the participatory news consumer, *Pew Internet and American Life Project 1* (2010) 19–21.
- [2] S. Fox, et al., *The social life of health information*, California Healthcare Foundation, 2011.
- [3] AA.VV., ICT usage in households and by individuals (isoc\_i). Reference Metadata in Euro SDMX Metadata Structure (ESMS), Technical Report, EUROSTAT, 2021. URL: <https://ec.europa.eu/eurostat/web/products-eurostat-news/-/edn-20210406-1>.
- [4] M. L. Millenson, J. L. Baldwin, L. Zipperer, H. Singh, Beyond Dr. Google: the evidence on consumer-facing digital tools for diagnosis, *Diagnosis 5* (2018) 95–105.
- [5] M. L. Antheunis, et al., Patients’ and health professionals’ use of social media in health care: motives, barriers and expectations, *Patient education and counseling 92* (2013) 426–431.
- [6] W.-Y. S. Chou, et al., Addressing health-related misinformation on social media, *Jama 320* (2018) 2417–2418.
- [7] M. Viviani, G. Pasi, Credibility in social media: opinions, news, and health information—a survey, *WIREs Data Mining and Knowledge Discovery 7* (2017) e1209.
- [8] M. Metzger, A. Flanagin, Online health information credibility, *Encyclopedia of Health Communication*. Thousand Oaks, CA: SAGE (2011) 976–978.
- [9] W. Weerkamp, M. de Rijke, Credibility-inspired ranking for blog post retrieval, *Information retrieval 15* (2012) 243–277.
- [10] D. G. P. Putri, et al., Social search and task-related relevance dimensions in microblogging sites, in: *International Conference on Social Informatics*, Springer, 2020, pp. 297–311.
- [11] M. Amjad, et al., UrduFake@FIRE2020: Shared Track on Fake News Identification in Urdu, in: *Forum for Information Retrieval Evaluation*, 2020, pp. 37–40.
- [12] Y. Kimura, et al., Overview of the NTCIR-14 QA Lab-PoliInfo Task, in: *14th NTCIR Conference on Evaluation of Information Access Technologies*, 2019, pp. 121–140.
- [13] C. Lioma, et al., Evaluation measures for relevance and credibility in ranked lists, in: *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, 2017, pp. 91–98.
- [14] A. Barrón-Cedeno, et al., Overview of CheckThat! 2020: Automatic identification and verification of claims in social media, in: *International Conference of the Cross-Language Evaluation Forum for European Languages*, Springer, 2020, pp. 215–236.
- [15] R. Nogueira, K. Cho, Task-oriented query reformulation with reinforcement learning, *arXiv preprint arXiv:1704.04572* (2017).
- [16] L. Goeriot, et al., Consumer Health Search at CLEF eHealth 2021, in: *CLEF (Working Notes)*, 2021.
- [17] G. Zuccon, Understandability biased evaluation for information retrieval, in: *European Conference on Information Retrieval*, Springer, 2016, pp. 280–292.
- [18] L. Goeriot, et al., Overview of the CLEF eHealth 2020 Task 2: Consumer Health Search with Ad Hoc and Spoken Queries, in: *CLEF (Working Notes)*, 2020.