Consumer Health Search at CLEF eHealth 2021

Lorraine Goeuriot[1], Hanna Suominen[2,3,4], Gabriella Pasi[5], Elias Bassani[5,6], Nicola Brew-Sam[2], Gabriela González-Sáez[1], Liadh Kelly[7], Philippe Mulhem[1], Sandaru Seneviratne[2], Rishabh Upadhyay[5], Marco Viviani[5] and Chenchen Xu[2,3]

[1]*Université Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000 Grenoble France*
[2]*The Australian National University, Canberra, ACT, Australia*
[3]*Data61/Commonwealth Scientific and Industrial Research Organisation, Canberra, ACT, Australia*
[4]*University of Turku, Turku, Finland*
[5]*University of Milano-Bicocca, DISCo, Italy*
[6]*Consorzio per il Trasferimento Tecnologico - C2T, Milan, Italy*
[7]*Maynooth University, Ireland*

Abstract

This paper details materials, methods, results, and analyses of the Consumer Health Search Task of the CLEF eHealth 2021 Evaluation Lab. This task investigates the effectiveness of information retrieval (IR) approaches in providing access to medical information to laypeople. For this a *TREC*-style evaluation methodology was applied: a shared collection of documents and queries is distributed, participants' runs received, relevance assessments generated, and participants' submissions evaluated. The task generated a new representative web corpus including web pages acquired from a 2021 CommonCrawl and social media content from Twitter and Reddit, along with a new collection of 55 manually generated layperson medical queries and their respective credibility, understandability, and topicality assessments for returned documents. This year's task focused on three subtask: (i) *ad-hoc* IR, (ii) weakly supervised IR, and (iii) document credibility prediction. In total, 15 runs were submitted to the three subtasks: eight addressed the ad-hoc IR task, three the weakly supervised IR challenge, and 4 the document credibility prediction challenge. As in previous years, the organizers have made data and tools associated with the task available for future research and development.

Keywords

Dimensions of Relevance, eHealth, Evaluation, Health Records, Medical Informatics, Information Storage and Retrieval, Self-Diagnosis, Test-set Generation

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

✉ lorraine.goeuriot@univ-grenoble-alpes.fr (L. Goeuriot); hanna.suominen@anu.edu.au (H. Suominen); gabriella.pasi@unimib.it (G. Pasi); elias.assani@unimib.it (E. Bassani); nicola.brew-sam@anu.edu.au (N. Brew-Sam); gabriela-nicole.gonzalez-saez@univ-grenoble-alpes.fr (G. González-Sáez); liadh.kelly@mu.ie (L. Kelly); philippe.mulhem@univ-grenoble-alpes.fr (P. Mulhem); sandaru.seneviratne@anu.edu.au (S. Seneviratne); r.upadhyay@campus.unimib.it (R. Upadhyay); marco.viviani@unimib.it (M. Viviani); chenchen.xu@anu.edu.au (C. Xu)

🌐 http://mrim.imag.fr/User/lorraine.goeuriot/ (L. Goeuriot); https://researchers.anu.edu.au/researchers/suominen-h (H. Suominen); https://www.unimib.it/gabriella-pasi (G. Pasi); https://researchers.anu.edu.au/researchers/brew-sam-n (N. Brew-Sam); https://www.maynoothuniversity.ie/people/liadh-kelly (L. Kelly); https://ikr3.disco.unimib.it/people/marco-viviani/ (M. Viviani)

iD 0000-0001-7491-1980 (L. Goeuriot); 0000-0002-4195-164 (H. Suominen); 0000-0002-6080-8170 (G. Pasi); 0000-0001-7544-4499 (N. Brew-Sam); 0000-0003-1131-5238 (L. Kelly); 0000-0002-2274-9050 (M. Viviani)

CEUR Workshop Proceedings (CEUR-WS.org)

1. Introduction

In today's information overloaded society, using a web search engine to find information related to health and medicine that is credible, easy to understand, and relevant to a given information need is increasingly difficult, thereby hindering patient and public involvement in healthcare [1]. These problems are referred to as credibility, understandability, and topicality dimensions of relevance in information retrieval (IR) [2, 3, 4, 5, 6]. The CLEF eHealth lab (https://clefehealth.imag.fr/) of the Conference and Labs of the Evaluation Forum (CLEF, formerly known as Cross-Language Evaluation Forum, http://www.clef-initiative.eu/) is a research initiative that aims at providing datasets and gathering researchers working on information extraction, management and retrieval tasks in the medical comain. Since its establishment in 2012, CLEF eHealth has included eight IR tasks on CHS with more more than twenty subtasks in total [7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17]. Its usage scenario is to ease and support laypeople, policymakers, and healthcare professionals in understanding, accessing, and authoring eHealth information in a multilingual setting.

In 2021, CLEF eHealth initiative has organised a CHS task with the following three IR subtasks:

1. Adhoc IR,
2. Weakly Supervised IR, and
3. Document Credibility Prediction.

The task has challenged researchers, scientists, engineers, analysts, and graduate students to develop better IR systems to support creating web-based search tools that return webpages that are better suited to laypeople's information needs from perspectives of

1. information topicality (i.e., how relevant are the contents to the search topic),
2. information understandability (i.e., how easily can a layperson understand the contents), and
3. information credibility (i.e., should the contents be trusted).

As a continuation of the previous CLEF eHealth IR tasks that ran in 2013–2018, and 2020 [2, 18, 19, 20, 21, 22, 23, 24], the 2021 CHS task has embraced the Text REtrieval Conference (TREC) -style evaluation process. Namely, it has designed, developed, and deployed a shared collection of documents and queries, called for the contribution of runs from participants, and conducted the subsequent formation of relevance assessments and evaluation of the participants' submissions.

The main contributions of the CLEF eHealth 2021 task on CHS are as follows:

1. generating a novel representative web corpus,
2. collecting layperson medical queries,
3. attracting new submissions from participants,
4. contributing to IR evaluation metrics relevant to the three dimensions of document relevance, and
5. evaluating IR systems (i.e., runs submitted by the task participants or from organizers' baseline systems) on newly conducted assessments.

The remainder of this paper is structured as follows: First, Section 2 details the task. Then, Section 3 introduces the document collection, topics, baselines, pooling strategy, and evaluation metrics. After this, Section 4 presents the participants and their approaches while Section 5 addresses their results. Finally, Section 6 concludes the paper.

2. Description of the Tasks

In this section, we provide a description of the three subtasks offered in this year's CHS task, namely: Subtask 1 on ad-hoc IR; Subtask 2 on weakly supervised IR; and Subtask 3 on document credibility prediction.

2.1. Subtask 1: Adhoc Information Retrieval

Similar to previous years of the CHS task, this was a standard ad-hoc IR task. A document collection was provided to task participants along with realistic layperson medical information need use cases, both described in Section 3. The purpose of the task was to evaluate IR systems' abilities to provide users with credible, understandable, and topical documents. As such, participating teams submitted their runs, which were pooled together with baseline runs and manual relevance assessments conducted, also described in Section 3. These systems' performance was assessed on multiple dimensions of relevance — credibility, understandability, and topicality. Evaluation metrics employed for this task are described in Section 3.6.

2.2. Subtask 2: Weakly Supervised Information Retrieval

This task aimed to evaluate the ability of machine learning-based ad-hoc IR models, trained with weak supervision, to retrieve relevant documents in the health domain. In order to train neural models to address the search task, a large collection of real-world health related queries extracted from commercial search engine query logs and synthetic (weak) relevance scores computed with a competitive IR system were considered. The submissions were evaluated against the same test set as Subtask 1 submissions, in order to allow a full comparison of traditional vs neural approaches. Details on the methods and materials associated with this task are described in Section 3.

2.3. Subtask 3: Document Credibility Prediction

The purpose of this task was the automatic assessment of the credibility of information that is disseminated online, through the Web and social media. Using the dataset related to Subtask 1, this task aimed at comparing approaches estimating documents credibility. The ground truth for this classification task is based on the credibility labels assigned to documents during the relevance assessment process. Evaluation of the runs includes classical classification metrics and investigates the ability of the participants systems to perform well on both web documents and social media content.

3. Materials and Methods

In this section, we will describe the materials and methods used in the CHS task of the CLEF eHealth evaluation lab 2021. After introducing our new document collection and novel topics, we will describe our baseline systems and pooling methodology. Finally, we will address our relevance assessments and evaluation metrics.

3.1. Document Collection

The 2021 CLEF eHealth Consumer Health Search document collection consisted of two separate crawls of documents: web documents acquired from the CommonCrawl and social media documents composed by Reddit and Twitter submissions.

First, we acquired web pages from the CommonCrawl. We extracted an initial list of websites from the 2018 CHS task of CLEF eHealth. This list was built by submitting a set of medical queries to the Microsoft Bing Application Programming Interfaces (through the Azure Cognitive Services) repeatedly over a period of a few weeks, and acquiring the uniform resource Locators (URL) of the retrieved results [22, 13]. We included the domains of the acquired URLs in the 2021 list, except some domains that were excluded for decency reasons. After this, similarly to 2018, we augmented the list by including a number of known reliable and unreliable health websites, domains, and social media contents of ranging reliability levels. Finally, we further extended the list by including websites that were highly relevant for the task queries to finalize our list of 600 domains. In summary, we introduced thirteen new domains in 2021 compared to the 2018 collection, and then the newly crawled domains from the latest CommonCrawl 2021-04 (https://commoncrawl.org/2021/02/january-2021-crawl-archive-now-available/).

Second, we complemented the collection with social media documents from Reddit and Twitter. As the first step, we selected a list of 150 health topics related to various health conditions. Then, we generated search queries manually from those topics and submitted them to Reddit to retrieve posts and comments. After this, we applied the same process on Twitter to get related tweets from the platform.

Please note that for the purposes of the 2021 CHS task, we defined a social media document as a text obtained by a single interaction. This implied that

- on Reddit, a document is composed by a post, one comment of the post, and associated meta-information and
- on Twitter, a document is a single tweet with its associated meta-information.

3.2. Topics

The set of topics of CLEF eHealth IR task aimed at being representative of laypersons' medical information needs in various scenarios. This year, the set of topics was collected from two sources as follows:

- The first part was based on our insights from consulting laypeople with lived experience of multiple sclerosis (MS) or diabetes to motivate, validate, and refine the search scenarios; we, as experts in IR and CHS tasks, captured these layperson-informed insights as the scenarios.
- The second part was based on use cases from Reddit health forums; we extracted and manually selected a list of topics from Google trends to best fit each use case.

As a result, we had a new set of 55 queries in English for authoring realistic search scenarios. To describe the scenarios, we enriched each query manually by labels in English either to characterize the search intent (for manually created queries) or to capture the submission text (for social medial queries) (Table 3.2).

Scenario	Query	Narrative
8	*best apps daily activity exercise diabetes*	*I'm a 15 year old with diabetes. I'm planning to join the school hiking club. What are the best apps to track my daily activity and exercises?*
57	*multiple sclerosis stages phases*	*I read that MS develops with several stages and includes phases of relapse. I want to know more about how this disease develops over time.*
126	*birth control suppression antral follicle count*	*So I've been on birth control taking only active pills for months now and I went in to a doctor's office to inquire about egg freezing. She seemed optimistic that my ultrasound would be very reassuring but instead she came away from the ultrasound deeply concerned. I did an AMH test and it came back 0.3 which was consistent with what was seen in the ultrasound. I'm terrified. Like...suicidal terrified. I don't have any underlying conditions and while I'm not young I'm not old either. I don't smoke and I'm not terribly overweight. I don't have insulin resistance and my reproductive organs looks good other than my ovaries. I'd been taking the combo pill to skip periods for months now. I don't even take the brown pills. When I first heard that you can skip periods and only have 2-3 a year when taking brown pills I started on them. The first time I tried to have a period on the brown pills nothing happened. No period. This was 2 years ago. I freaked out but they said I just respond strongly to the pills. So I went off of the pills completely and had 2-3 normal periods starting about 2 months later. Could the birth control be suppressing my antral follicle count and amh this much? Please help.*

Table 1
Illustration of Some of Our Queries and Narratives

Subtasks 1, 2, and 3 used these 55 queries with 5 released for training and 50 reserved for testing; the test topics contained a balanced sample of the manually constructed and automatically extracted search scenarios.

3.3. Baseline Systems

With respect to both Subtask 1 and Subtask 2, we provided six baseline systems. These systems applied the Okapi BM25 (BM for Best Matching), Dirichlet Language Model (DirichletLM), and Term Frequency times Inverse Document Frequency (TF×IDF) algorithms with default parameters. Each of them was implemented without and with pseudo relevance feedback (RF) using default parameters (DFR Bo1 model [25] on three documents, selecting ten terms). This resulted in the $3 \times 2 = 6$ baseline systems. The systems are implemented using Terrier version 5.4 [26] as following:

```
terrier batchretrieve -t topics.txt -w [TF_IDF|DirichletLM|BM25] [-q|].
```

Regarding Subtask 3, where it is necessary to evaluate the effectiveness of the approaches with respect to assessing the credibility of information, it was decided to approach the problem as a binary classification (identification of credible versus non-credible information). For this reason, simple baselines were developed based on supervised classifiers that act on a set of features that can be extracted from the documents under consideration. Dealing with documents that are both Web pages and social content, it was necessary to consider, in baseline development, only those features that are directly extractable from the text of the documents. For social content, it would also be possible to consider other metadata related to the social network of the authors of the posts, but for consistency with the evaluation of Web pages, such metadata have not been considered.

The employed supervised classifiers are based on Support Vector Machines (SVM), Random Forests (RF), and Logistic Regression (LR), i.e., the machine learning techniques that have proven to be more effective for binary credibility assessment in the literature [27]. Such baselines act on the following linguistic features: the TF×IDF text representation, the word embedding text representation obtained by Word2vec pre-trained on Google News,[1] and, given the health-related scenario considered, the word embedding representation obtained by Word2vec pre-trained on both PubMed and Medical Information Mart for Intensive Care III (MIMIC-III).[2] Python and the `scikit-learn` library [28] have been employed for implementing the machine learning algorithms and for producing the TF×IDF text representation. Furthermore, we have considered threshold tuning to get the optimal cut-off for the runs, based on the Youden's index [29].

3.4. Pooling Methodology

Similar to the 2016, 2017, 2018, and 2020 pools, we created the pool using the rank-biased precision (RBP)-based Method A (Summing contributions) [30] in which documents are weighted according to their overall contribution to the effectiveness evaluation as provided by the RBP formula (with $p = 0.8$, following a study published in 2007 on RBP [31]). This strategy, called RBPA, has been proven more efficient than traditional fixed-depth or stratified pooling to evaluate systems under fixed assessment budget constraints [32], as it was the case for this task. All participants' runs were considered on the document's pool, along with six baselines provided by the organizers. In order to guarantee the judgements of the documents of the participants' runs, half of the pool was composed by their documents and half from documents of the baselines' runs which resulted in 250 documents per query in the pool.

3.5. Relevance Assessment

The credibility, understandability, and topicality assessments were performed by 26 volunteers (19 women and 7 men) in May–June 2021. Of these assessors, 16 were from Australia, 4 from Italy, 3 from France, 2 from Ireland, and 1 from Finland. We recruited, trained, and supervised them by using bespoke written materials from April to June 2021. The recruitment took place via email and on social media, using both our existing contacts and snowballing.

We implemented these assessments online by expanding and customising the Relevation! tool for relevance assessments [33] to capture the three dimensions of document relevance, and their scale (see [17, Figures 4–6] for illustrations of the online assessment environment).

[1] https://github.com/mmihaltz/word2vec-GoogleNews-vectors
[2] https://github.com/ncbi-nlp/BioSentVec

We associated every query with 250 documents to be assessed with respect to their credibility, understandability, and topicality. Initially, we allocated each assessor with 2 queries for their assessment and then revised these allocations based on their individual needs and availability. In the end, every assessor completed 1–4 queries.

Ethical approval (2021/013) was obtained for all aspects of this assessment study involving human participants as assessors from the Human Research Ethics Committee of the Australian National University (ANU). Each study participant provided informed consent.

3.6. Evaluation Metrics

We considered evaluation measures that allowed to evaluate both:

1. the effectiveness of the systems with respect to the ranking produced by taking into account the three criteria considered, and
2. the accuracy of the (binary) classification of documents with respect to credibility (Subtask 3).

Specifically, with regard to the first aspect, this included the following performance evaluation measures: Mean Average Precision (MAP), preference-based BPref metric, normalized Discounted Cumulative Gain (nDCG), understandability-based variant of Rank Biased Precision (uRBP), and credibility-ranked Rank Biased Precision (cRBP) [6].

With respect to the second aspect, we referred to classical measures to assess the goodness of a classifier, such as Accuracy and the Area under the Receiver Operating Characteristic (ROC) Curve (AUC) [34]. In particular, since Subtask 3 is also devoted to assess credibility in relation to information needs (topics), also the average of a Topic-based Credibility Precision w.r.t. each topic, namely CP(q), has been considered.

A brief explanation of the three measures that need more detail, namely uRBP, cRBP, and CP(q) is provided below.

3.6.1. Understandability-Ranked Biased Precision

The uRBP measure evaluates IR systems by taking into account both topicality and understandability dimensions of relevance. In particular, the function for calculating uRBP was [35]:

$$\text{uRBP} = (1-\rho)\sum_{k=1}^{K}\rho^{k-1}r(d@k)\cdot u(d@k), \tag{1}$$

where $r(d@k)$ is the relevance of the document d at position k, $u(d@k)$ is the understandability value of the document d at position k, and the persistent parameter ρ models the user desire to examine every answer, which was set to 0.50, 0.80, and 0.95 to obtain three version of uRBP, according to different user behaviors.

3.6.2. Credibility-Ranked Biased Precision

In CLEF eHealth 2020, we have adapted uRBP to credibility, obtaining the so-called credibility-ranked biased precision (cRBP) measure [6]. In this case the function for calculating cRBP was the same used to calculate uRBP, replacing $u(d@k)$ by the credibility value of the document d at position k, $c(d@k)$:

$$\text{cRBP} = (1-\rho)\sum_{k=1}^{K}\rho^{k-1} r(d@k)\cdot c(d@k). \tag{2}$$

As in uRBP, the parameter ρ was set to three values, from an impatient user (0.50) to more persistent users (0.80 and 0.95).

3.6.3. Topic-based Credibility Precision

The precision in retrieving credible documents can be calculated over the top-k documents in the ranking, for each topic (query) q, as follows:

$$\text{CP}(q) = \frac{\#credible\_retrieved\_docs\_top\_k(q)}{\#retrieved\_docs\_top\_k(q)}.$$

4. Participants and Approaches

In 2021, 43 teams registered for the task on the web site and two teams submitted runs to the subtasks. We provided the registered participants with the crawler code and the domain list of the crawl for both the web documents and social media documents. They also had access to indexes built by organizers from the document collection on demand. Participants' submissions were due by 8 May 2021.

Of the four run submissions, two were to Subtask 1 on Adhoc IR, one was to Subtask 2 on Weakly Supervised IR, and one was to Subtask 3 on Document Credibility Prediction. The teams were from two countries (i.e., China and Italy) in two continents (i.e., Asia and Europe). In Subtask 1, the submissions were by

1. a 4-member team from the School of Computer Science, Zhongyuan University of Technology (ZUT) in Zhengzhou, China [36] and
2. a 2-member team from the Information Management Systems (IMS) Research Group, University of Padova (UniPd), Padova, Italy [37].

In Subtasks 2 and 3, the submissions were by the leader of this IMS UniPd team [37] — a regular participant in our previous CHS tasks.

The two teams submitted the following types of document ranking approaches as runs to the Adhoc IR subtask (Table 4): Team ZUT used a learning-to-rank approach in all its four submitted runs, but with four different machine learning algorithms to train their models [36]. In contrast, Team UniPd founded their four submissions on a renown Python framework for IR called PyTerrier as variants of Reciprocal Rank Fusion with the provided Terrier index [37].

The UniPd team submitted closely related approaches to the other two subtasks (Table 4). Namely, they also submitted their aforementioned four approaches to the Weekly Supervised IR subtask and two of them and another two of their variants to the Document Credibility Prediction subtask [37].

Subtasks	Team	Run	Approach	Algorithmic Details
1	ZUT	i	Learning to Rank	LM
1	ZUT	ii	Learning to Rank	MR
1	ZUT	iii	Learning to Rank	RFs
1	ZUT	iv	Learning to Rank	RB
1–3	UniPd	a	RRF using PyTerrier with the provided Terrier index	BM25, QLM, & DFR
1 & 2	UniPd	b	RRF using PyTerrier with the provided Terrier index	BM25, QLM, & DFR using the RM3 relevance language model for pseudo RF
1–3	UniPd	c	RRF using PyTerrier with the provided Terrier index	BM25, QLM, & DFR on manual variants of the query
1 & 2	UniPd	d	RRF using PyTerrier with the provided Terrier index	BM25, QLM, & DFR on manual variants using RM3 pseudo RF
3	UniPd	e	RRF using PyTerrier with the provided Terrier index	UniPd's runs a & b above (i.e., the ones without manual variants of the query) merged with min-max normalization
3	UniPd	f	RRF using PyTerrier with the provided Terrier index	UniPd's runs c & d above (i.e., the ones with manual variants of the query) merged with min-max normalization

Table 2

Summary of Runs Submitted by Participating Teams. Acronyms: BM — Best Match, DFR — divergence from randomness, LM — LambdaMART, MR — MART, RB — RankBoost, RF — relevance feedback, RFs — random forests, RRF — Reciprocal Rank Fusion, QLM — query likelihood model

5. Results

In 2021, the CLEF eHealth CHS task generated a new representative Web corpus including Web pages acquired from a 2021 CommonCrawl, and social media content from Twitter and Reddit. A new collection of 55 manually generated layperson medical queries was also created, along with their respective credibility, understandability, and topicality assessments for returned documents. In total, 15 runs were submitted to the three subtasks on adhoc IR, weakly supervised IR, and document credibility prediction, respectively.

5.1. Coverage of Relevance Assessments

A total of $12,500$ assessments were made on $11,357$ documents: $7,400$ Web documents, and $3,957$ social media documents. Figure 1 shows the number of social media and Web documents that were assessed for each query. The bottom part shows which queries were created from discussion with patients (*Expert queries*), and queries created from discussions on social media (*Social media queries*). We can see that Web documents took a bigger part of the pool of documents. Nevertheless, the proportion of social media documents was bigger for queries based on social media. A special case is presented in queries 22, 63, and 116, where the pool

	Social Media			Web		
# of documents	Highly	Somewhat	Not	Highly	Somewhat	Not
Topical	925	1, 046	1, 555	2, 175	2, 540	4, 259
Understandable	1, 160	1, 766	600	4, 758	3, 014	1, 202
Credible	12	1, 427	1, 967	4, 552	3, 123	753

Table 3
Number of topical (a.k.a relevant), understandable, and credible documents in social media and Web documents.

of documents was composed by Web documents only. This means that the submitted runs retrieved only Web documents.

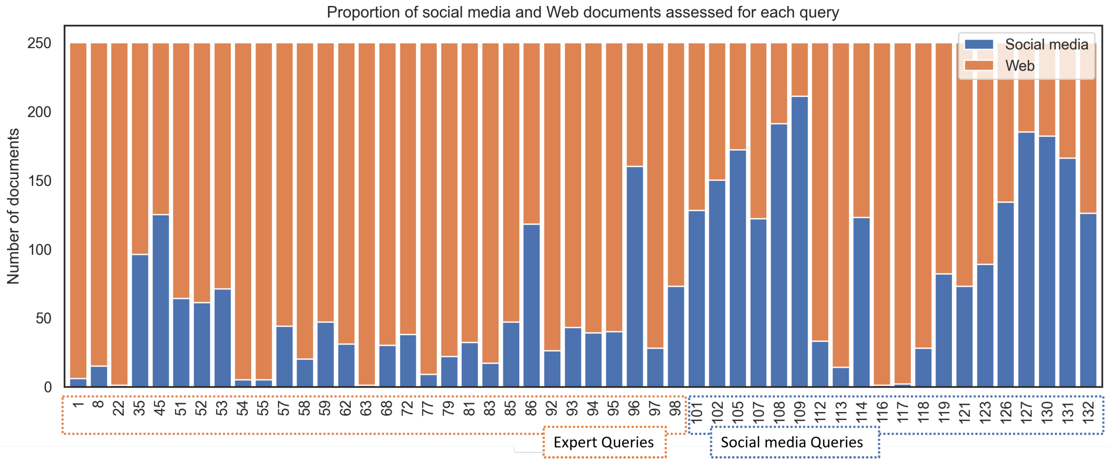

Figure 1: Proportion of social media and Web documents per each query, orange represents the proportion of Web documents and blue social media. The bottom line shows queries based on experts discussions with patients, and queries based on reddit posts.

While the distribution of the assessments on topicality and understandability dimensions were similar on social media and Web documents, the credibility assessments presented a big difference (Table 5.1), with only 12 documents (less than 1% of social media assessed documents) were assessed as highly credible in the social media set, in contrast to the 4, 552 (54% of Web assessed documents) highly credible Web documents. Finally, an important part of social media documents were assessed as not credible. Figure 2 shows the relationship between the three dimensions of relevance; its diagonal exemplifies the distribution of each dimension. Each plot exposes the number of documents evaluated as “highly”, “somewhat”, or “not”, with respect to two dimensions of relevance (one in each axis). By example, in social media assessments the Figure 2 shows for the not credible documents that a fraction is not topical relevant, a set of documents is somewhat topical relevant, and a small part of the not credible is anyways highly topically relevant.

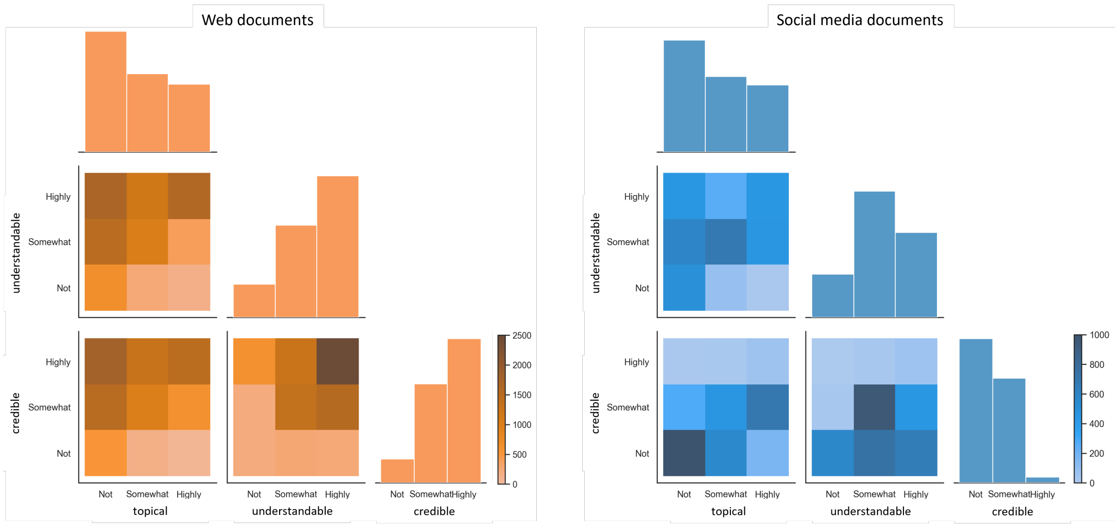

Figure 2: Relationship between pairs of relevance dimensions, and distribution of assessments for each type of document.

5.2. Subtask 1: Adhoc Information Retrieval

In this section we present the results for the Adhoc IR subtask where the systems were evaluated in different dimensions of relevance. For topicality, we evaluated the systems using MAP, BPref, and NDCG@10 performances metrics. For readability, we made use of uRBP that considers topicality and understandability of the documents. Finally, in order to include credibility in our evaluation, we measured systems' cRBP performance that considers topicality and credibility of the document.

Table 5.2 presents the ranking of participant systems and organizers baselines, with respect to three metrics of topical relevance: MAP, BPref, and NDCG@10. The team achieving the highest results was UniPd. Their top run, `original_rm3_rrf` used Reciprocal Rank Fusion with BM25, QLM, DFR approaches using pseudo relevance feedback with 10 documents and 10 terms (query weight 0.5). It achieved 0.43 MAP and 0.51 BPref. The best system of ZUT team was their run3, which used learning to rank techniques with a model trained using Random Forests. ZUT run3 was the best systems of the team on the three metrics with 0.4 MAP, 0.47 BPref, and 0.66 NDCG@10. For BPref and NDCG@10, the organizers baseline using TF×IDF with query expansion obtained higher results.

Table 5.2 compares the results of our baseline systems with respect to last year's Adhoc IR task results. In the case of MAP and BPref metrics, all the systems had better performance on 2021 test collection; for NDCG@10, DirichletLM with and without relevance feedback had a better performance on 2020 test collection. Despite this clear improvement from 2020 to 2021, the ranking of systems has changed in all the metrics. Namely, in 2020, the best MAP and BPref performance was achieve by DirichletLM, and TF×IDF with respect to NDCG@10 metric. In contrast, in this year's task, the best baseline MAP, Bpref, and NDCG@10 was achieved by TF×IDF with relevance feedback.

Rank	Team - run	MAP	Team - run	BPref	Team - run	NDCG@10
1	UniPd original_rm3_rrf	0.431	Baseline terrier_TF×IDF_qe	0.511	Baseline terrier_TF×IDF_qe	0.654
2	UniPd simplified_rm3_rrf	0.431	UniPd original_rm3_rrf	0.508	Baseline terrier_TF×IDF	0.646
3	ZUT run3_clef2021_task2	0.409	UniPd simplified_rm3_rrf	0.508	Baseline terrier_BM25	0.636
4	Baseline terrier_TF×IDF_qe	0.397	Baseline terrier_BM25_qe	0.499	Baseline terrier_BM25_qe	0.635
5	Baseline terrier_BM25_qe	0.390	Baseline terrier_TF×IDF	0.474	UniPd original_rrf	0.619
6	ZUT run4_clef2021_task2	0.373	Baseline terrier_DirichletLM	0.472	ZUT run3_clef2021_task2	0.615
7	Baseline terrier_DirichletLM	0.369	Baseline terrier_BM25	0.471	UniPd original_rm3_rrf	0.614
8	Baseline terrier_TF×IDF	0.366	ZUT run3_clef2021_task2	0.469	Baseline terrier_DirichletLM	0.595
9	Baseline terrier_BM25	0.364	ZUT run4_clef2021_task2	0.447	ZUT run4_clef2021_task2	0.565
10	ZUT run2_clef2021_task2	0.338	ZUT run1_clef2021_task2	0.441	Baseline terrier_DirichletLM_qe	0.536
11	ZUT run1_clef2021_task2	0.338	ZUT run2_clef2021_task2	0.428	ZUT run1_clef2021_task2	0.526
12	Baseline terrier_DirichletLM_qe	0.242	Baseline terrier_DirichletLM_qe	0.369	ZUT run2_clef2021_task2	0.482
13	UniPd original_rrf	0.236	UniPd original_rrf	0.300	UniPd simplified_rm3_rrf	0.478
14	UniPd simplified_rrf	0.215	UniPd simplified_rrf	0.298	UniPd simplified_rrf	0.459

Table 4
Ad-hoc task ranking of systems

	MAP		BPref		NDCG@10	
year	2020	2021	2020	2021	2020	2021
terrier_BM25	0.2627	0.3641	0.3964	0.4707	0.5919	0.6364
terrier_BM25_qe	0.2453	0.3903	0.3784	0.4994	0.5698	0.6352
terrier_DirichletLM	**0.2706**	0.3694	**0.416**	0.4724	0.6054	0.5952
terrier_DirichletLM_qe	0.1453	0.2423	0.2719	0.3691	0.5521	0.5362
terrier_TF×IDF	0.2613	0.3663	0.3958	0.4744	**0.6292**	0.6464
terrier_TF×IDF_qe	0.25	**0.3974**	0.3802	**0.5106**	0.608	**0.6535**

Table 5
MAP, BPref, and NDCG@10 2020 and 2021 Adhoc CHS task on baseline systems. Best system in bold

The results for all participants for readability and credibility evaluation is shown in Table 5.2. For readability evaluation the best run was the baseline system TF×IDF with relevance feedback. The best participant system was ZUT's run3 that is the same system that presented the best performance in terms of topicality as well. For credibility evaluation, the best run was the baseline system DirichletLM while the best participant system was UniPd's `original_rm3_rrf` — the best system of the team in topicality evaluation.

Rank	Team - run	rRBP	Team - run	cRBP
1	Baseline terrier_TF×IDF_qe	0.523	Baseline terrier_DirichletLM	0.458
2	Baseline terrier_TF×IDF	0.509	UniPd original_rm3_rrf	0.452
3	Baseline terrier_BM25_qe	0.507	Baseline terrier_TF×IDF_qe	0.450
4	Baseline terrier_BM25	0.501	Baseline terrier_BM25_qe	0.432
5	ZUT run3_clef2021_task2	0.494	Baseline terrier_DirichletLM_qe	0.429
6	UniPd original_rrf	0.491	UniPd original_rrf	0.427
7	UniPd original_rm3_rrf	0.475	Baseline terrier_TF×IDF	0.418
8	Baseline terrier_DirichletLM	0.463	ZUT run3_clef2021_task2	0.414
9	ZUT run4_clef2021_task2	0.450	ZUT run4_clef2021_task2	0.409
10	Baseline terrier_DirichletLM_qe	0.408	Baseline terrier_BM25	0.406
11	ZUT run1_clef2021_task2	0.408	UniPd simplified_rm3_rrf	0.331
12	UniPd simplified_rm3_rrf	0.369	ZUT run2_clef2021_task2	0.330
13	UniPd simplified_rrf	0.359	ZUT run1_clef2021_task2	0.319
14	ZUT run2_clef2021_task2	0.349	UniPd simplified_rrf	0.317

Table 6
Runs ranked for readability relevance dimension (rRBP) and credibility (cRBP).

5.3. Subtask 2: Weakly Supervised Information Retrieval

The purpose of the task was to evaluate systems based on machine learning, trained on the weakly supervised dataset. UniPD submitted their runs to Subtask 1 for this subtask, which are not trained on the weakly supervised dataset. Their submission provides a kind of baseline for a system that does not use any information. Since no other team submitted runs to this task, we cannot provide any evaluation.

5.4. Subtask 3: Document Credibility Prediction

The UniPd team, to assess document credibility, reused the runs computed in Subtask 1 and grouped them in order to produce a single score for each document. Their simple hypothesis was that documents that have a higher score across different search engines are also more credible. Their approach did not consider any additional information about the provenance of the document.

5.4.1. Credibility Assessment as a Binary Classification Problem

The results illustrated in this section were obtained by performing a binary credibility assessment using the CLEF 2020 eHealth dataset for training the baselines and testing on a subset of the CLEF 2021 eHealth data, that is, those that were employed for both runs `subtask1_ims_original` and `subtask1_ims_simplified` submitted by the UniPd team. In this case, the topics against which the documents were retrieved are not taken into account. The results of classifying documents according to their credibility using both simple baselines (illustrated in Section 3.3) and considered runs are illustrated in Table 7.

The results obtained from both baselines and runs were evaluated by means of classical measures in the context of document classification, that is, the AUC and Accuracy (as discussed in Section 3.6).

Model	AUC	Accuracy
Baseline SVM_orig_tf_idf	56.89%	63.38%
Baseline RF_orig_tf_idf	47.37%	70.94%
Baseline LR_orig_tf_idf	65.12%	64.98%
Baseline SVM_orig_w2v_google	59.81%	64.26%
Baseline RF_orig_w2v_google	51.82%	71.11%
Baseline LR_orig_w2v_google	67.70%	65.47%
Baseline SVM_orig_w2v_bio	65.30%	64.33%
Baseline RF_orig_w2v_bio	56.10%	76.06%
Baseline LR_orig_w2v_bio	**69.50%**	65.80%
Baseline SVM_simp_tf_idf	57.31%	63.56%
Baseline RF_simp_tf_idf	54.40%	74.34%
Baseline LR_simp_tf_idf	61.43%	63.19%
Baseline SVM_simp_w2v_google	60.24%	65.54%
Baseline RF_simp_w2v_google	54.43%	75.64%
Baseline LR_simp_w2v_google	62.35%	64.87%
Baseline SVM_simp_w2v_bio	63.97%	64.66%
Baseline RF_simp_w2v_bio	56.26%	**78.14%**
Baseline LR_simp_w2v_bio	67.87%	67.45%
Run subtask1_ims_original	**62.33%**	**48.37%**
Run subtask1_ims_simplified	51.45%	45.11%

Table 7
AUC and Accuracy values for baselines and runs when credibility assessment is treated as a binary classification problem.

As it can be observed from the table, having considered only linguistic features, the classification results obtained were not particularly significant to the considered problem, neither for the runs nor for the baselines considered. The idea discussed by the members of the UniPd group, that documents having a higher score across different search engines are also more credible, did not appear to be supported by these results, but nevertheless merited future investigation, in particular in relation to the results reported in the next section. Regarding the baselines, it was possible to observe that, when supervised classifiers employ the word embedding text representation obtained by Word2vec pre-trained on biomedical datasets, such as Pubmed and MIMIC-III, we obtained best results as compared to their counterparts employing text representation features based on TF×IDF and Word2vec pre-trained on the general-purpose Google News dataset. However, it must be considered that, in general, the Accuracy values in particular could be influenced by the fact that the dataset considered was strongly unbalanced, being made up of about 20% of documents labeled as credible and 80% labeled as non-credible. Probably, this should also raise the need to deepen an analysis about the way in which human assessors evaluate the documents proposed to them.

5.4.2. Topic-based Credibility Assessment

Regarding the problem of assessing the credibility of the documents retrieved with respect to the topics considered, we provide below the results relative to the runs sent by UniPD, namely `subtask2_ims_original` and `subtask2_ims_simplified`. In this case, for each topic considered, the precision in retrieving credible documents with respect to the topic was evaluated based on the value of CP(q) calculated as shown in Section 3.6.3.

In Table 8, we report the average CP(q) values with respect to the set of topics considered. Specifically, the top-100 and top-200 documents retrieved with respect to each topic were taken into account in computing the CP(q) metric.

Model	No of top-k documents	Average CP(q)
Run subtask2_ims_original	$k = 100$	0.6767
Run subtask2_ims_simplified	$k = 100$	0.5221
Run subtask2_ims_original	$k = 200$	0.5205
Run subtask2_ims_simplified	$k = 200$	0.3433

Table 8
Average CP(q) values when credibility assessment is performed by considering documents retrieved with respect to specific topics.

As it can be observed from the table, with respect to the evaluations carried out in the previous case related to binary credibility assessment (i.e., Section 5.4.1), it seems that the methods applied in the two submitted runs that consider credibility associated with documents retrieved with respect to specific topics, actually managed to find credible information in the first top-k positions, in particular as k decreases. In general, the most effective method seems to be the one implemented in the run `subtask2_ims_original`; however, from specific observations with respect to the results submitted by the participants (which are omitted in this paper), we can state that the second method, the one implemented in `subtask2_ims_simplified`, proved to be particularly efficient for some topics (even more than the method implemented in `subtask2_ims_original`) and much less so with respect to others, leading to a lower average CP(q) value. It is therefore worth investigating this aspect more in the future.

6. Conclusions

This paper has described methods, results and analysis of the Consumer Health Search (CHS) challenge at the CLEF eHealth 2021 Evaluation Lab. The task considered the problem of lay people searching for medical information related to their health condition on the web. In particular, across web pages and social media content. The task included three subtasks on ad hoc IR, weakly supervised IR, and document credibility prediction. 15 runs were submitted to these tasks.

The CLEF eHealth CHS challenge, first ran at the inaugural CLEF eHealth lab in 2013, and has offered since then TREC-style IR challenges with medical datasets consisting of large medical web and lay person query collections. As a by-product of this evaluation exercise, the task contributes to the research community a collection with associated assessments and evaluation framework that can be used to evaluate the effectiveness of retrieval methods for health information seeking on the web. Queries, assessments, and participants' runs for the 2021 CHS challenge are publicly available at https://github.com/CLEFeHealth/CHS-2021 and previous years' CHS collections are available at https://github.com/CLEFeHealth/.

Acknowledgments

Thanks: The CHS task of the CLEF eHealth 2021 evaluation lab has been supported in part by the CLEF Initiative. It has also been supported in part by the Our Health in Our Hands (OHIOH) initiative of the Australian National University (ANU), as well as the ANU School of Computing, ANU Research School of Population Health, and Data61/Commonwealth Scientific and Industrial Research Organisation. OHIOH is a strategic initiative of the ANU which aims to transform health care by developing new personalised health technologies and solutions in collaboration with patients, clinicians, and health care providers. Moreover, the task has been supported in part by the bi-lateral Kodicare (Knowledge Delta based improvement and continuous evaluation of retrieval engines) project funded by the French ANR (ANR-19-CE23-0029) and Austrian FWF. Finally, the task has been supported in part 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". We are also thankful to the people involved in the query creation and relevance assessment exercises. Last but not least, we gratefully acknowledge the participating teams' hard work. We thank them for their submissions and interest in the task.

Author Contribution Statement: With equal contribution, Task 2 was led by LG, GP, and HS, and organized by EB, NB-S, GG-S, LK, PM, GP, SS, HS, RU, MV, and CX.

References

[1] S. H. Soroya, A. Farooq, K. Mahmood, J. Isoaho, S. e Zara, From information seeking to information avoidance: Understanding the health information behavior during a global health crisis, Information Processing and Management 58 (2021) 102440.

[2] L. Goeuriot, G. J. Jones, L. Kelly, J. Leveling, A. Hanbury, H. Müller, S. Salantera, H. Suominen, G. Zuccon, ShARe/CLEF eHealth Evaluation Lab 2013, Task 3: Information retrieval to address patients' questions when reading clinical reports, CLEF 2013 Online Working Notes 8138 (2013).

[3] H. Suominen, L. Kelly, L. Goeuriot, Scholarly influence of the Conference and Labs of the Evaluation Forum eHealth Initiative: Review and bibliometric study of the 2012 to 2017 outcomes, JMIR Research Protocols 7 (2018) e10961.

[4] J. Palotti, G. Zuccon, A. Hanbury, Consumer health search on the web: Study of web page understandability and its integration in ranking algorithms, J Med Internet Res 21 (2019) e10986.

[5] H. Suominen, L. Kelly, L. Goeuriot, The scholarly impact and strategic intent of CLEF eHealth Labs from 2012 to 2017, in: N. Ferro, C. Peters (Eds.), Information Retrieval Evaluation in a Changing World: Lessons Learned from 20 Years of CLEF, Springer International Publishing, Cham, 2019, pp. 333–363.

[6] L. Goeuriot, Z. Liu, G. Pasi, G. G. Saez, M. Viviani, C. Xu, Overview of the CLEF eHealth 2020 task 2: consumer health search with ad hoc and spoken queries, in: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings, 2020.

[7] H. Suominen (Ed.), The Proceedings of the CLEFeHealth2012 — the CLEF 2012 Workshop

on Cross-Language Evaluation of Methods, Applications, and Resources for eHealth Document Analysis, NICTA, 2012.

[8] H. Suominen, S. Salanterä, S. Velupillai, W. W. Chapman, G. Savova, N. Elhadad, S. Pradhan, B. R. South, D. L. Mowery, G. J. Jones, J. Leveling, L. Kelly, L. Goeuriot, D. Martinez, G. Zuccon, Overview of the ShARe/CLEF eHealth Evaluation Lab 2013, in: Information Access Evaluation. Multilinguality, Multimodality, and Visualization, Springer Berlin Heidelberg, 2013, pp. 212–231.

[9] L. Kelly, L. Goeuriot, H. Suominen, T. Schreck, G. Leroy, D. L. Mowery, S. Velupillai, W. Chapman, D. Martinez, G. Zuccon, J. Palotti, Overview of the ShARe/CLEF eHealth Evaluation Lab 2014, in: Information Access Evaluation. Multilinguality, Multimodality, and Visualization, Springer Berlin Heidelberg, 2014, pp. 172–191.

[10] L. Goeuriot, L. Kelly, H. Suominen, L. Hanlen, A. Névéol, C. Grouin, J. Palotti, G. Zuccon, Overview of the CLEF eHealth Evaluation Lab 2015, in: Information Access Evaluation. Multilinguality, Multimodality, and Visualization, Springer Berlin Heidelberg, 2015.

[11] L. Kelly, L. Goeuriot, H. Suominen, A. Névéol, J. Palotti, G. Zuccon, Overview of the CLEF eHealth Evaluation Lab 2016, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer Berlin Heidelberg, 2016, pp. 255–266.

[12] L. Goeuriot, L. Kelly, H. Suominen, A. Névéol, A. Robert, E. Kanoulas, R. Spijker, J. Palotti, G. Zuccon, CLEF 2017 eHealth Evaluation Lab overview, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer Berlin Heidelberg, 2017, pp. 291–303.

[13] H. Suominen, L. Kelly, L. Goeuriot, A. Névéol, L. Ramadier, A. Robert, E. Kanoulas, R. Spijker, L. Azzopardi, D. Li, Jimmy, J. Palotti, G. Zuccon, Overview of the CLEF eHealth Evaluation Lab 2018, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer Berlin Heidelberg, 2018, pp. 286–301.

[14] L. Kelly, H. Suominen, L. Goeuriot, M. Neves, E. Kanoulas, D. Li, L. Azzopardi, R. Spijker, G. Zuccon, H. Scells, J. Palotti, Overview of the CLEF eHealth Evaluation Lab 2019, in: F. Crestani, M. Braschler, J. Savoy, A. Rauber, H. Müller, D. E. Losada, G. Heinatz Bürki, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2019, pp. 322–339.

[15] L. Goeuriot, H. Suominen, L. Kelly, A. Miranda-Escalada, M. Krallinger, Z. Liu, G. Pasi, G. Gonzalez Saez, M. Viviani, C. Xu, Overview of the CLEF eHealth evaluation lab 2020, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, H. Joho, C. Lioma, C. Eickhoff, A. Névéol, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2020, pp. 255–271.

[16] L. Goeuriot, H. Suominen, L. Kelly, L. A. Alemany, N. Brew-Sam, V. Cotik, D. Filippo, G. G. Saez, F. Luque, P. Mulhem, G. Pasi, R. Roller, S. Seneviratne, J. Vivaldi, M. Viviani, C. Xu, CLEF eHealth 2021 Evaluation Lab, in: Advances in Information Retrieval — 43st European Conference on IR Research, Springer, Heidelberg, Germany, 2021.

[17] H. Suominen, L. Goeuriot, L. Kelly, L. Alonso Alemany, E. Bassani, N. Brew-Sam, V. Cotik, D. Filippo, G. Gonzalez-Saez, F. Luque, P. Mulhem, G. Pasi, R. Roller, S. Seneviratne, R. Upadhyay, J. Vivaldi, M. Viviani, C. Xu, Overview of the CLEF eHealth evaluation lab 2021, in: CLEF 2021 - 11th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS), Springer, Heidelberg, Germany, 2021.

[18] L. Goeuriot, L. Kelly, W. Lee, J. Palotti, P. Pecina, G. Zuccon, A. Hanbury, H. M. Gareth J.F. Jones, ShARe/CLEF eHealth Evaluation Lab 2014, Task 3: User-centred health information retrieval, in: CLEF 2014 Evaluation Labs and Workshop: Online Working Notes, Sheffield, UK, 2014.

[19] J. Palotti, G. Zuccon, L. Goeuriot, L. Kelly, A. Hanburyn, G. J. Jones, M. Lupu, P. Pecina, CLEF eHealth Evaluation Lab 2015, Task 2: Retrieving Information about Medical Symptoms, in: CLEF 2015 Online Working Notes, CEUR-WS, 2015.

[20] G. Zuccon, J. Palotti, L. Goeuriot, L. Kelly, M. Lupu, P. Pecina, H. Mueller, J. Budaher, A. Deacon, The IR Task at the CLEF eHealth Evaluation Lab 2016: User-centred Health Information Retrieval, in: CLEF 2016 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS, 2016.

[21] J. Palotti, G. Zuccon, Jimmy, P. Pecina, M. Lupu, L. Goeuriot, L. Kelly, A. Hanbury, CLEF 2017 Task Overview: The IR Task at the eHealth Evaluation Lab, in: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum, CEUR Workshop Proceedings, 2017.

[22] . Jimmy, G. Zuccon, J. Palotti, Overview of the CLEF 2018 consumer health search task, in: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum, CEUR Workshop Proceedings, 2018.

[23] L. Goeuriot, G. J. Jones, L. Kelly, J. Leveling, M. Lupu, J. Palotti, G. Zuccon, An Analysis of Evaluation Campaigns in ad-hoc Medical Information Retrieval: CLEF eHealth 2013 and 2014, Springer Information Retrieval Journal (2018).

[24] L. Goeuriot, H. Suominen, L. Kelly, Z. Liu, G. Pasi, G. S. Gonzales, M. Viviani, C. Xu, Overview of the CLEF eHealth 2020 task 2: Consumer health search with ad hoc and spoken queries, in: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum, CEUR Workshop Proceedings, 2020.

[25] G. Amati, Probabilistic Models for Information Retrieval Based on Divergence from Randomness, Ph.D. thesis, Glasgow University, Glasgow, the UK, 2003.

[26] I. Ounis, C. Lioma, C. Macdonald, V. Plachouras, Research directions in terrier: a search engine for advanced retrieval on the web, CEPIS Upgrade Journal 8 (2007).

[27] M. Viviani, G. Pasi, Credibility in social media: opinions, news, and health information—a survey, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 7 (2017) e1209.

[28] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, G. Varoquaux, API design for machine learning software: experiences from the scikit-learn project, in: ECML PKDD Workshop: Languages for Data Mining and Machine Learning, 2013, pp. 108–122.

[29] W. J. Youden, Index for rating diagnostic tests, Cancer 3 (1950) 32–35.

[30] A. Moffat, J. Zobel, Rank-biased precision for measurement of retrieval effectiveness, ACM Transactions on Information Systems 27 (2008) 2:1–2:27.

[31] L. A. Park, Y. Zhang, On the distribution of user persistence for rank-biased precision, in: Proceedings of the 12th Australasian document computing symposium, 2007, pp. 17–24.

[32] A. Lipani, J. Palotti, M. Lupu, F. Piroi, G. Zuccon, A. Hanbury, Fixed-cost pooling strategies based on ir evaluation measures, in: European Conference on Information Retrieval, Springer, 2017, pp. 357–368.

[33] B. Koopman, G. Zuccon, Relevation!: an open source system for information retrieval relevance assessment, in: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, ACM, 2014, pp. 1243–1244.

[34] H. Suominen, S. Pyysalo, M. Hiissa, F. Ginter, S. Liu, D. Marghescu, T. Pahikkala, B. Back, H. Karsten, T. Salakoski, Performance evaluation measures for text mining, in: M. Song, Y. Wu (Eds.), Handbook of Research on Text and Web Mining Technologies, IGI Global, Hershey, Pennsylvania, USA, 2008, pp. 724–747.

[35] G. Zuccon, Understandability biased evaluation for information retrieval, in: Advances in Information Retrieval, 2016, pp. 280–292.

[36] H. Yang, X. Liu, B. Zheng, G. Yang, Learning to rank for Consumer Health Search, in: CLEF 2021 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS, September 2021.

[37] G. Di Nunzio, F. Vezzani, IMS-UNIPD @ CLEF eHealth Task 2: Reciprocal Ranking Fusion in CHS, in: CLEF 2021 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS, September 2021.