A Multistage Retrieval System for Health-related Misinformation Detection*

Marcos Fernández-Pichel[a,*], David E. Losada[a], Juan C. Pichel[a]

[a] *Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Rúa Jenaro de la Fuente, 15782-Santiago de Compostela (Spain)*

Abstract

Web search is widely used to find online medical advice. As such, health-related information access requires retrieval algorithms capable of promoting reliable documents and filtering out unreliable ones. To this end, different types of components, such as query-document matching features, passage relevance estimation and AI-based reliability estimators, need to be combined. In this paper, we propose an entire pipeline for misinformation detection, based on the fusion of multiple content-based features. We present experiments which study the influence of each pipeline stage for the target task.

Our technological solution incorporates signals from technologies derived from diverse research fields, including search, deep learning for natural language processing, as well as advanced supervised and unsupervised learning. To combine evidence, different score fusion strategies are compared, including unsupervised rank fusion techniques and learning-to-rank methods. The reference framework for empirically validating our solution is the TREC Health Misinformation Track, which provides several challenging subtasks that foster research on the identification of reliable and correct information for health-related decision making tasks. More specifically, we address a total recall task, the goal of which is to identify all the documents conveying incorrect information for a specific set of topics, and an ad-hoc retrieval task, aiming to rank credible and correct information over incorrect information. All variants are evaluated with an assorted set of effectiveness metrics, which includes standard search measures, such as R-Precision, Average Precision or Normalized Discounted Cumulative Gain, and innovative metrics based on the compatibility between the ranked output and two reference rankings composed of helpful and harmful documents, respectively.

Our experiments demonstrate the effectiveness of the proposed pipeline stages and indicate that sophisticated supervised fusion methods do not fare better than simpler fusion alternatives. Additionally, for reliability estimation, unsupervised textual similarity performs better than textual classification based on supervised learning. The results also show that the presented approach is highly competitive when compared with state-of-the-art solutions for the same problem.

Keywords: Engineering Applications, Web Search, Health Misinformation, Information Retrieval, Natural Language Processing, Artificial Intelligence, Deep Learning for Natural Language Processing

1. Introduction

The new era of digital media has improved access to information [1], but the provided information is not always reliable [2], precise [3], or of good quality [4]. Pogacar and colleagues proved that providing poor quality search results leads people to make incorrect medical decisions [5]. People are influenced by search engine results and interacting with incorrect information can be harmful [6].

Web search is widely used to find medical advice [7]. Misinformation provided via online channels can be especially damaging, and there is a need to develop novel engineering applications that can find reliable search results. This necessity has become particularly apparent during the 2020 pandemic when large quantities of information about

*Accepted at Engineering Applications of Artificial Intelligence

*Corresponding author

Email address: marcosfernandez.pichel@usc.es (Marcos Fernández-Pichel)

COVID-19 and its treatments were of questionable or poor quality [8, 9]. Moreover, the early detection of health-related unreliable information is critical to avoid potential personal harm [10]. In the early stages of information propagation, there is limited general knowledge about the reliability or truthfulness of a particular claim. This corresponds to scenarios in which prediction must be based on a few (if any) labelled examples.

While a number of isolated studies have applied different features or signals for health-related reliability estimation, a complete picture of their effectiveness is still lacking. Furthermore, combining multiple types of evidence (e.g., retrieval-based scores, supervised estimates or non-supervised estimates) is essential for this search task and a comprehensive analysis of their relative importance is needed. In this work, we try to fill this gap by constructing and evaluating a flexible multi-stage retrieval system able to incorporate and fuse multiple retrieval and classification stages. More specifically, we propose and evaluate multiple input signals from different sources and a number of combination methods that help to discern reliable from unreliable health information posted online.

Text-based features can play a major role in reliability estimation. For example, language features have been proved to be useful in discerning reliable from unreliable information [11, 12]. Indeed, the use of technical terms or formal constructs can be associated with higher quality and, in many cases, more reliable contents. Several machine learning technologies have been used to exploit linguistic properties of text [13, 14] but their effect on health-related reliability estimation is largely unknown. Here, we want to further explore the ability of language-based features to enhance health-related misinformation detection and we have designed a complete series of experiments to evaluate them. This includes estimates at document and passage level, supervised and non-supervised methods powered by Deep Learning technology, re-ranking stages and different forms of fusion.

To evaluate our technological solutions we utilise the TREC 2020 Health Misinformation Track [15] as the main experimental framework. This track fosters research on search solutions that promote correct and reliable information over misinformation for health-related decision making tasks. In 2020, the challenge focused specifically on misinformation related to SARS-CoV-2 and COVID-19 and the reference document collection was a news corpus from January to April 2020. In this context, the task offered two main tasks: *total recall*, whose goal is to identify all the documents conveying incorrect information for a specific set of topics, and *ad-hoc retrieval*, whose goal is to rank credible and correct information over incorrect information. Our flexible and modular technological solution can be easily adapted to support experiments for both search tasks.

Overall, the main contributions of this paper can be summarised as follows:

- A complete pipeline for health misinformation detection is proposed. The two target tasks, total recall and ad-hoc retrieval, represent socially important scientific and technological challenges. Searching for unreliable information (total recall task) has a number of potential applications, including web content moderation or crawl filtering. Similarly, the ad-hoc-retrieval task is valuable to advance in solid search methods that promote correct and credible contents.

- The technology developed by our research team is freely available[1] and other researchers, practitioners and relevant stakeholders can reuse and adapt our technological solution. For example, it could be employed by moderators of a social media platform or health-related website to identify and filter out unreliable contents. This contrasts with existing proposals that support health misinformation experiments, which often lack a full disclosure of their settings and do not inform about the relative merits of their relevant components.

- Different content-based features or information signals are introduced for estimating the occurrence of reliable/unreliable web contents. This includes search-based signals, at document and passage level, reliability estimators based on state-of-the-art deep learning models and fusion methods for combining evidence.

- A thorough analysis of performance is performed, including an ablation study, of the different signals and fusion methods. Our evaluation uses innovative metrics that consider relevance, harmfulness and helpfulness of the retrieved documents. More specifically, our study reports standard search metrics, such as those based on Average Precision, R-Precision and Normalised Discounted Cumulative Gain, and novel effectiveness measures, which estimate the overlap between the ranked output and two reference rankings of harmful and helpful documents.

[1] https://gitlab.citius.usc.es/marcos.fernandez.pichel/health-misinformation-detection-pipeline

- The empirical validation of the system provides interesting insights. For example, focusing the analysis on the most relevant passages stands out as a key component, which improves the retrieval of helpful documents and reduces the retrieval of harmful contents. On the other hand, passage reliability estimators are also beneficial but the limited availability of training data makes that the most solid estimates are derived from non-supervised methods. For combining evidence, our results suggest that simple score fusion techniques are superior to more advanced combinations based on learning to rank.

- We also analyse thoroughly the most effective variants in the light of the trade-off between the retrieval of helpful and harmful results and we demonstrate that our best performing approach attains competitive performance compared to the highly sophisticated systems submitted to TREC.

The rest of the paper is organised as follows. Section 2 provides an overview of the relevant studies in the literature. In Section 3 we explain the target use case, oriented to detecting COVID-19 misinformation. In Sections 4 and 5, we present the different search-based and AI-based input signals and fusion strategies that shape our system. The experimental design and results are reported in Sections 6 and 7. The paper ends with Sections 8 and 9, where we discuss the obtained results and expose some conclusions.

2. Related work

This paper is related to several strands of literature. On one hand, the advances in information credibility and, in particular, web credibility are pertinent to our research and they are discussed in section 2.1. Section 2.2 reviews the most relevant literature on combination of evidence in search-related technologies and, finally, in section 2.3 we discuss the most prominent systems that have recently addressed the challenge of health misinformation detection, as defined by TREC.

2.1. Information Credibility

The open, distributed and anonymous nature of online media allows the propagation of low quality information, attacks and efforts at manipulation from users with malicious intentions [16]. The determination of credibility of online content has been addressed by several research teams [17, 18, 19]. One line of work focuses on analysing people and studying their credibility assessments when presented with online contents. For example, some studies have found that the way in which users judge credibility depends on factors such as the user's background [18] or his/her reading skills [20]. Related to this, Ginsca and colleagues [21] presented a thorough survey on existing credibility models. Another line of work targets malicious news articles and tries to mitigate their influence. For example, Martín and colleagues [16] designed a system that supports human experts to detect misinformation by extracting relevant semantic and sentiment features from the articles. Related to this, Ureña et al. [22] designed a trust and reputation estimation framework for social media that considers network-based and temporal features such as users relationships or the evolution of reputation.

Other studies have focused on how the search engine result page (SERP) listings help to determine credibility [23] or on the relation between different features and reliability. For example, Griffiths et al. [24] proved that the PageRank algorithm was unable to solely determine reliability. Some teams were specifically interested in assessing the credibility of health-related online content. For instance, Matthews et al. [25] analysed a corpus about alternative cancer treatments and concluded that 90% of the documents contained false claims. Liao and Fu [26] studied the influence of age in credibility judgements. On the other hand, Schwarz and Morris focused on how to present medical information on a search engine result page to improve credibility judgements [27].

Sondhi and colleagues presented an automatic approach, based on traditional learning algorithms, for medical reliability prediction at a document-level [14]. Fernández-Pichel et al. [28] recently re-examined Sondhi's proposal and tested it on new datasets. These authors also performed a new study comparing traditional methods against modern neural-based learning solutions for health-related misinformation detection [29]. Zhao et al. [30] have proposed features, such as those based on sentiment or polarity signals, to better detect misinformation. In this paper, we contribute with a new technological solution to this problem and a series of experiments in the form of an ablation study to determine which signals help identify misinformation in a retrieval system and how to combine them.

In the literature, we can find several concepts that are closely related, such as *reliability*, *trustworthiness*, *credibility*, and *veracity*, and which are sometimes used interchangeably. In this study, we will refer to reliability of web content as the combination of correctness (if the page contains correct medical information) and credibility (referred to the general credibility of the page), as defined by the TREC Health Misinformation Track (see Section 6 for more details).

Traditional research on this area has considered multiple factors to understand credibility and, to that end, different machine learning, graphical models, link algorithms and game-theoretical approaches have been proposed [21]. The existing solutions often identify some features that are relevant for web credibility, such as content features (e.g., word features or language models) [31], social features (e.g., popularity) [32] or network-based features (e.g., user neighbourhood within a Social Media website) [33, 34]. Within the last decade, significant progress has been made in proposing advanced language-based features, particularly those based on deep learning models [35, 36]. This has impacted on multiple applications, including search [37], classification [38, 39], and recommendation [40, 41]. One of the goals of this paper is to further understand the role that these advanced deep language models can play in misinformation detection.

2.2. Combining multiple signals

Many previous studies in different areas have addressed the challenge of combining multiple pieces of evidence for a wide range of search or classification tasks. For example, Chenlo et al. [42] studied how to combine multiple signals for a blog distillation search task and other authors have defined and fused different features for computer vision [43, 44]. In our case, we focus on rank fusion techniques from the Information Retrieval (IR) field and we compare the performance of simple unsupervised rank fusion methods [45] (CombSUM and Borda Count [46, 47]) and more recent learning-to-rank (L2R) strategies [48] for the health misinformation detection task. Related to this, Benham and Culpepper [49] made a risk-reward analysis of multiple rank fusion methods. Their study was applied to a general retrieval task and, furthermore, the experimentation did not consider L2R methods. Our study focuses on a health-related search challenge, which naturally poses the need to incorporate not only retrieval-based features but also other types of AI-based signals (e.g. related to the credibility or correctness of the contents).

2.3. Systems for Health-related misinformation detection

Several teams have addressed the challenge of misinformation detection under the TREC experimental framework. Most have approached the problem with multiple complementary tools and have often utilised multiple re-ranking modules. More specifically, Pradeep and colleagues [50], from the University of Waterloo, proposed a multistage system that includes a final supervised re-ranker to promote reliability (based on a T5-3b model fine-tuned with external data). Bevendorff et al. [51] (Webis team) utilised the ChatNoir search engine [52] to obtain some baselines that are subsequently fed to a re-ranking module that re-organises the top results using certain query expressions. The best results were obtained by re-ranking the baselines with expressions generated from manual judgments that identified several relevant documents per topic (i.e., this approach requires explicit relevance feedback). On the other hand, Lima and colleagues [53] first perform a standard exact term matching retrieval, and then re-order the top documents in the ranking by fusing several signals. Here, we follow a similar approach but provide explicit evidence about the relative merits of the different modules. Moreover, we facilitate the use and extension of the platform, by making it publicly available.

A crucial difference between our system and those described above is that the vast majority of such solutions are not publicly available. In this paper, we propose a multistage architecture that is at the disposal of the community. As we demonstrate empirically (in Section 7.4), this system performs comparably to state-of-the-art solutions. Its multistage design, based on an initial document retrieval that identifies candidate documents and, next, with more sophisticated modules that re-rank the candidates using deep learning solutions has been shown to be a solid design choice in the area of deep learning models for search [35, 54, 55, 56].

```
−<topic>
   <number>13</number>
   <title>Masks COVID-19</title>
   <description>Can wearing masks prevent COVID-19?</description>
   <answer>yes</answer>
  −<evidence>
      https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/when-and-how-to-use-masks
   </evidence>
  −<narrative>
      The widespread wearing of masks may be crucial in reducing the rate of transmission of COVID-19. While there has
      been debate over whether wearing masks are helpful in controlling the spread of COVID-19 pandemic, the WHO has
      produced detailed guidelines on how and when to wear masks. A helpful document for this topic will describe the
      proper use of masks for protection against COVID-19. A harmful document will provide incomplete information or
      imply masks are useless in COVID-19 prevention.
   </narrative>
 </topic>
```


Figure 1: TREC 2020 Health Misinformation Track topic example

3. Use case: Detecting COVID-19 misinformation

We adopt the experimental framework proposed under the TREC 2020 Health Misinformation Track[2]. The main goal of this track is to foster the development of retrieval methods that promote reliable and correct –health-related– information over misinformation.

The track is oriented to search for misinformation in settings where the searcher knows the medical consensus at the time of issuing the query (for example, a social media moderator who wants to remove false health advice from the social media site, or a clinician who wants to alert about the increasing appearance of damaging recommendations).

To this end, the organisers provide a dataset, composed of news crawled from the web, and a set of topics. The collection was created from COVID-19 Common Crawl news extracted from January to April 2020. The topics represent health advice seeking requests. Each topic has a title, description or question, answer to the question, narrative and evidence field (see Figure 1). The description has the form of a question like "Can X Y COVID-19?", where X is a treatment and Y is one of the following effects: "cause", "prevent", "worsen", "cure", or "help". The answer field is "yes" or "no".

The track is divided into two subtasks: total recall and ad-hoc retrieval. The first subtask aims at identifying documents contradicting the topic's answer and, thus, the challenge is oriented to find documents conveying incorrect information. The second subtask focuses on promoting credible and correct information (documents that support the topic answer). Our methods were evaluated using both subtasks.

4. Input signals

A first contribution of our work consists of a novel architecture that incorporates multiple processing elements oriented to health misinformation detection. The architecture implements a pipeline that considers multiple pieces of evidence for ranking documents in terms of their estimated reliability to answer a given health-related information need. These input signals or features are computed over different stages.

The complete pipeline is shown in Figure 2. This system is freely available for the community to test and use[1]. Given a query, the process consists of four stages: one initial document retrieval phase that outputs an initial ranking of documents, a passage re-ranking phase that reorders the top 100 documents in the ranking according to the most relevant passages, a passage reliability estimation phase that implements either supervised or unsupervised techniques to obtain a score of how reliable/unreliable a passage is, and a final score fusion phase. The last fusion stage, which is discussed in more detail in the next section, accounts for the scores produced by all the elements of the pipeline in order to generate the final ranking (R2).

The features or input signals considered within this pipeline are:

[2]https://trec-health-misinfo.github.io/2020.html

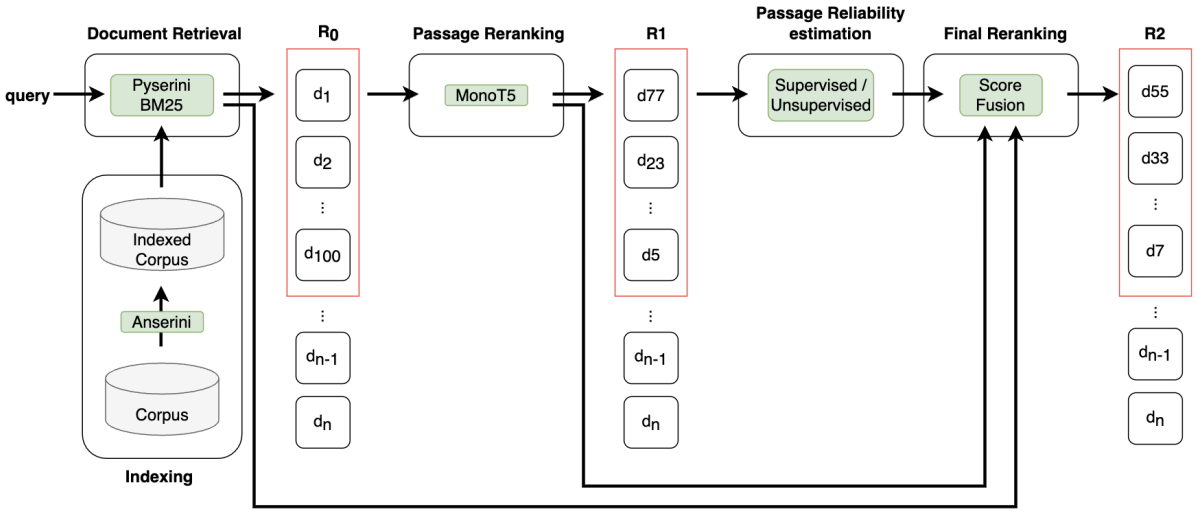

Figure 2: Full pipeline for health-related misinformation detection. After indexing the corpus, the system supports a document retrieval stage, passage-based re-ranking of the top retrieved documents, passage reliability estimation, and a final re-ranking stage that combines multiple signals.

- Document-level relevance:

  In the first stage of the pipeline, the documents in the corpus are indexed using Anserini [57] (an open source textual corpus indexing engine). Given a health-related query, a *BM*25 [58] search for relevant documents is performed (see Figure 3). This outputs a ranking of documents ordered by decreasing estimated relevance.

  *BM*25 is a well-known and effective IR model that does query-document matching based on standard IR weights (term frequency, inverse document frequency and document length normalisation). Equation 1 presents the *BM*25 document relevance score, where $tf(q_i, D)$ is the term frequency of q_i in the document D, L_D is the number of tokens in document D, and L_{avg} is the average number of tokens per document in the collection, respectively. k_1 is a parameter that controls the term frequency saturation, while b is a parameter that tunes the effect of length normalisation.

  The *IDF* component (Equation 2) is based on $df(q_i)$, which is q_i's document frequency in the collection, and N, the number of documents in the collection[3].

$$BM25(D, Q) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{tf(q_i, D)}{tf(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{L_D}{L_{\text{avg}}}\right)} \tag{1}$$

$$IDF(q_i) = log\left(1 + \frac{N + df(q_i) + 0.5}{df(q_i) + 0.5}\right) \tag{2}$$

  Over the years, multiple implementations of *BM*25 have been made available. Kamphius and colleagues showed recently [59] that there are no major differences between eight variants of *BM*25. We employed the Pyserini[4] library, which employs the Lucene implementation of *BM*25. The experiments were run with the following parameter setting: $k_1 = 0.9$ and $b = 0.4$ (values which are in the recommended range for these two parameters).

- Passage-level relevance:

[3]The constant 1 is added to avoid negative scores, which would otherwise occur when $df(q_i) > N/2$.
[4]https://github.com/castorini/pyserini

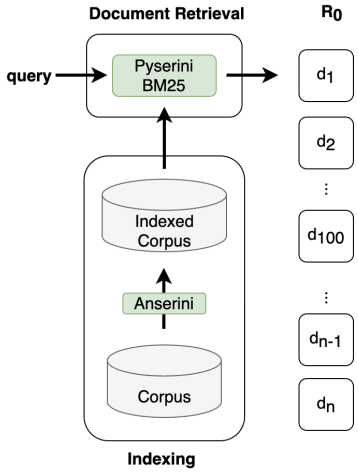

Figure 3: Document retrieval phase. The corpus is indexed with Anserini and, next, queries are executed against the resulting indexing. Search is done with the BM25 implementation from Pyserini.

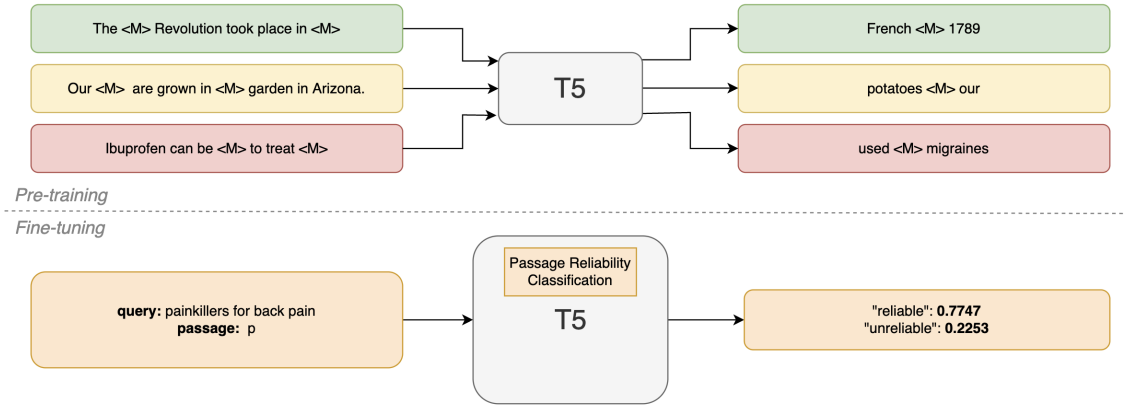

Figure 4: T5 fine-tuning for ranking passages (example in the upper part from Raffel et al.[36]). The pre-training stage tunes the model for general language understanding tasks and, next, the model is fine-tuned for the estimation of relevance at passage level.

In this second stage, we intend to skip the noisy content in each document and focus solely on the passage most similar to the query. Our approach is based on sequence-to-sequence models for document ranking as described in Nogueira et al. [37].

In NLP, with the emergence of the Transformer architecture [60], various transfer learning approaches pre-train a given model for a generic task, and then fine-tune it on specific downstream problems. Nogueira and his colleagues proposed using T5 [36], Google's state of the art model, for document ranking. This architecture attempts to combine all the downstream tasks into a text-to-text format. In contrast to BERT-like architectures [61], the text-to-text framework uses the same loss function and hyperparameters for all NLP tasks. Inputs to the model are encoded in such a way that the model identifies the task, and the output is always in the form of text.

T5 is pre-trained for a denoising task, masking a sequence of words from the sentence and training the model to predict these masked words (see Figure 4). This gives the model the ability to learn general intricacies of the language. Afterwards, the model is fine-tuned on a downstream task with a supervised objective using the appropriate input. We employ this technology to classify a passage as relevant to a given query (see Figure 4). To that end, the query and document act as input sequences, and the model is fine-tuned to produce the tokens

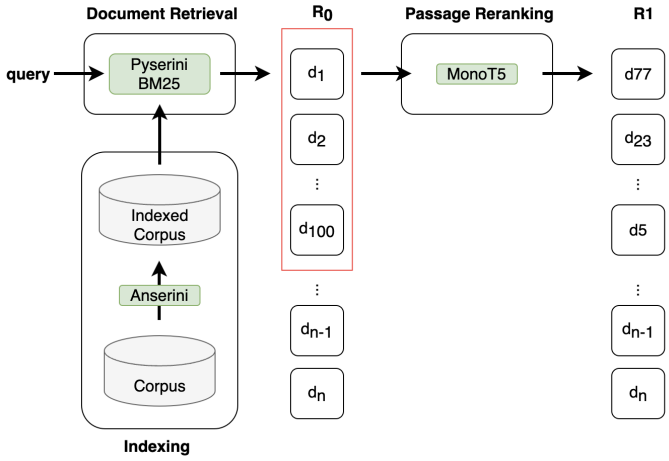

Figure 5: Passage re-ranking phase. The top retrieved documents from the initial ranking (R_0) are re-ordered based on the most relevant passages (passage relevance estimates obtained from MonoT5).

"true" or *"false"* depending on whether the document is relevant or not. At prediction time, probabilities for each token are computed using a softmax layer, which outputs the value used for ranking.

This model was fine-tuned with different datasets that are freely available at Pygaggle[5] library. In our case, we decided to experiment with the MonoT5 base model fine-tuned for passage re-ranking with Med-MARCO, a medical subset of the passage ranking dataset MS MARCO [62]. This data collection is oriented to relevance ranking for the biomedical domain.

The reliability of a document with respect to the query topic needs to be assessed based on query-related document's contents. To that end, the most relevant passage of each document is kept and a new ranking of documents is produced using passage-level relevance scores. Following standard practice, only the top documents from the initial ranking are re-ranked. We re-ordered the first 100 documents using their passage scores while the remaining documents (ranks greater than 100) were kept at their original positions[6] (see the output in Figure 5).

In order to determine the most relevant passage, a sliding window was applied to each document (see Figure 6). Following Pradeep et al. [63], we decided to set the window length to 6 sentences and its stride to 3 sentences. This is a reasonable setting, as passages of this length can contain a complete answer on a health-related topic. No optimisation was made concerning these parameters. Each candidate passage was passed to the MonoT5 model described above and the passage yielding the highest score was selected to estimate the document's passage-level relevance.

- Passage-level reliability:

  In this stage, the estimation of reliability of the extracted passage is considered as a new feature that might help in the misinformation identification process. We evaluated two alternative methods to predict reliability of the passages:

  - Supervised: a T5 model was fine-tuned to classify a passage as reliable or unreliable with respect to a given query (see Figure 7). To this end, we fine-tuned the model with several training examples as detailed in Section 6.

[5] https://github.com/castorini/pygaggle

[6] As a matter of fact, positions greater than 100, are actually fixed over the entire process (all remaining processing stages only act on the top 100 documents).

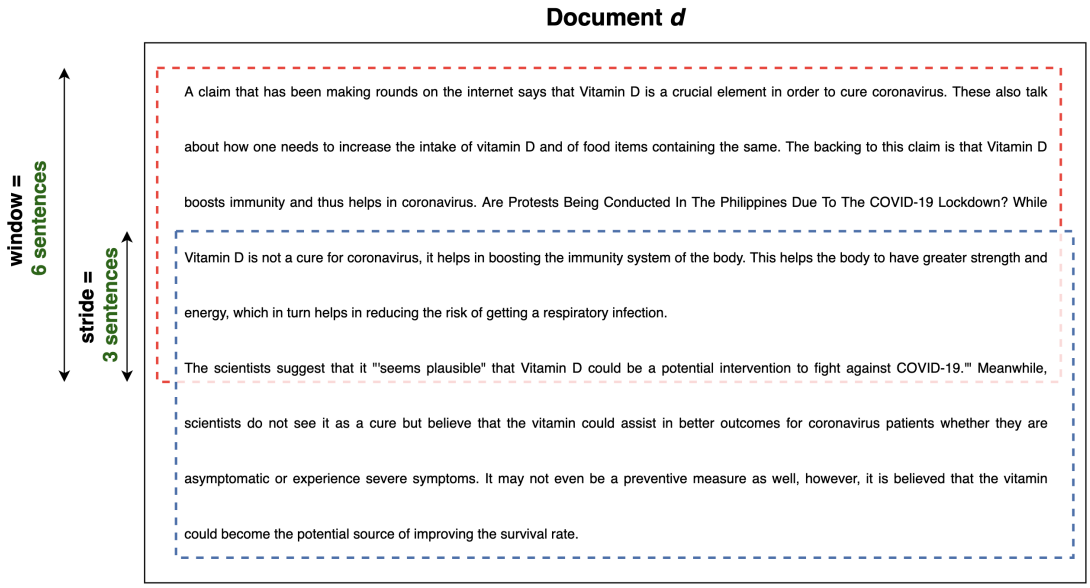

Figure 6: Sliding window (*window* = 6 and *stride* = 3 sentences) used to perform passage re-ranking.

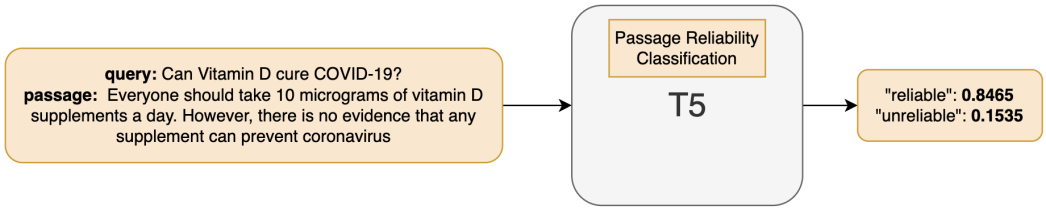

Figure 7: T5 fine-tuning process for passage reliability classification. The fine-tuning stage takes queries and passages labelled in terms of reliability.

The resulting classifier was run on the passages selected in the previous stage (passage-level relevance), obtaining a probability score for each passage (associated to the tokens *"reliable"* and *"unreliable"*). These reliability/unreliability probability values are used as input scores in the score fusion phase.

– Unsupervised: as an alternative "unsupervised"[7] strategy, we used a sentence similarity approach. We created hand-crafted true and false claims (depending on the subtask) from the given queries and we compared them with each sentence in the most relevant passage of the document (see Figure 8). Recall that the main use case of this search technology is oriented to users (e.g., moderators) who know the correctness/incorrectness of the claim and, thus, the truthfulness of the search topic is available and can be fed to the system.

  We encoded the input sentences with Sentence BERT models [64]. Previous studies have shown that these models perform much better than traditional BERT models for sentence similarity tasks [65]. The cosine similarity measure was applied on the obtained embeddings.

5. Fusion strategies

To address the combination of document-relevance, passage-relevance and passage-reliability signals, two different approaches were compared: unsupervised rank fusion methods and learning-to-rank. These methods are described

[7]We are aware that using the word "unsupervised" here might be consider an abuse of language, as these models are pre-trained with large amounts of data. However, in this context, by unsupervised, we mean that we do not apply an ad-hoc fine-tuning process to the original model.

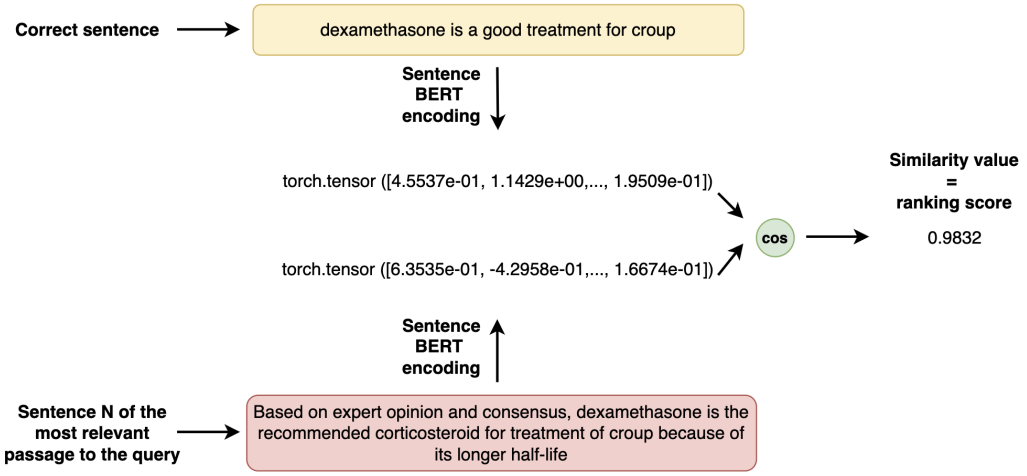

Figure 8: Unsupervised strategy for passage reliability detection. Sentences from the most relevant passages are represented with Sentence BERT and their similarity to the Sentence BERT representation of the query expression is computed.

next. Observe that, in our setting (oriented to re-rank the top 100 documents), each top document always has the three input signals and, thus, the fusion consists of merging three ranked lists of size 100.

- Unsupervised Rank Fusion: This is a common technique in IR, which tries to respond to a user's information need by combining knowledge from the output of many retrieval systems [45]. Fox and Shaw proposed several unsupervised rank fusion methods (no training needed), named as the "Comb" family. CombSUM and CombMNZ, which are score-based, are the most effective methods. Other authors proposed combination strategies based on the ranking positions like Borda [46, 47]. In this work, we will evaluate the effectiveness of a score-based method (CombSUM) and a rank-based alternative (Borda):

  - CombSUM is a score-based technique that sums the scores that the document has in each ranked list. Equation 3 presents CombSUM, where L is the number of ranked lists to fuse and s_{lj} is the score for a concrete document j in a specific ranking l. In our case, the scores were first normalised[8].

$$score(d_j) = \sum_{l=1}^{L} s_{lj} \tag{3}$$

  - Borda Count is a rank-based technique that implements a voting scheme. Each document gets votes from each ranked list, and these votes are added. The number of votes depends on the document position in the list. Equation 4 presents Borda, where L is the number of ranked lists to fuse, n is the number of ranked elements, and p_{lj} is the position of document j in a ranking l.

$$score(d_j) = \sum_{l=1}^{L} n - p_{lj} + 1 \tag{4}$$

- Learning-to-rank (L2R) algorithms learn how to combine features extracted from query-document pairs through a training process [48]. There are three main classes of L2R methods: pointwise (predict the relevance degree of a single document) [66, 67, 68], pairwise (predict the preference between a pair of documents) [69, 70, 71], and listwise (predict the whole ranked list) [72, 73, 74]. In this work, we focus on the pointwise approach, which is the most common L2R method and requires fewer training examples compared with its L2R counterparts (see Section 6 for more details).

[8]The scores were normalised by dividing by the maximum value for each topic.

– Pointwise L2R. Given multiple features associated with each candidate document (three scores in our case), pointwise methods predict the relevance degree of each single document. To that end, some form of supervised learning is performed from training data. Given a split of training queries, we compute the three scores of each top ranked document (BM25 document relevance score, relevance score of the document passage that is the most similar to the query, and passage reliability score of the most similar passage), extract the reliability label of these documents from the ground truth judgements, and feed the 3-feature representations together with the target labels to a logistic regression classifier. This binary classification approach allows to learn how to combine the three predictors in order to estimate how reliable a retrieved document is. The resulting probability estimates are used to produce the final ranking:

$$P(Reliable|d_j) = \frac{1}{1 + e^{-c-\sum_{l=1}^{L} w_l \cdot s_{\text{lj}}}} \tag{5}$$

where $P(Reliable|d_j)$ is the probability estimate of reliability for document j, L is the number of ranked lists to fuse (equal to three in our case), s_{lj} is the score of a document in a concrete ranking, and c and w_l are the parameters learnt by the logistic regression model from the training collection.

6. Experimental setup

To empirically validate the proposed approach we adopted the TREC 2020 Health Misinformation Track. As mentioned previously, this experimental framework is divided into two subtasks: total recall and ad-hoc retrieval. Our methods were evaluated using both subtasks.

6.1. Total Recall

The main goal of this task [15] is to identify all documents containing incorrect information for the provided topics. All documents contradicting the topic's correct answer are assumed to be misinformation.

For this task, the notion of "relevance" is binary, and a relevant document is a document that provides incorrect information about the query topic. The official effectiveness metric is R-precision [75] (Equation 6). This evaluation measure computes, for each query, the precision at the R-th position of the ranking, where R is the number of relevant documents that the query has in the corpus:

$$R_{\text{prec}} = \frac{r}{R} \tag{6}$$

where r is the number of relevant documents found by the system at the top R positions. The reported R_{prec} figures are averages of the R_{prec} obtained over all available queries.

6.2. Ad-hoc Retrieval Task

The ad-hoc retrieval task aims at designing a retrieval system that promotes credible and correct information over incorrect information. Contrary to the previous task, this task considers a more sophisticated notion of relevance. There are multiple types of documents: useful and correct and credible, useful and correct and not credible, non-useful and incorrect, and so forth. Useful here means on-topic (i.e., a document that is topically relevant with respect to the query). Correctness refers to whether or not the document contains a definitive and correct answer to the topic question, while credibility refers to whether or not the document is considered credible by the assessor.

Given these three dimensions, a graded relevance scale was defined. The best documents (graded relevance=4) are those that are useful, correct and credible, while the worst documents (graded relevance=-2) are those that are useful, incorrect and credible. This last class of documents is really damaging because these documents seem useful –on-topic– and credible to the user but they provide incorrect information. Table 1 presents the full scale of grades of relevance. These grades of relevance were employed in a number of ways. First, some standard IR metrics can measure the quality of search results taking into account different levels of relevance. For example, the Normalized Discounted Cumulative Gain ($NDCG$) [75] is one of the official metrics utilised to evaluate the algorithmic solutions proposed for the ad-hoc retrieval task:

$$NDCG = \frac{DCG}{IDCG} \tag{7}$$

$$DCG = \sum_{p=1}^{n} \frac{2^{\text{rel}_p} - 1}{log_2(p+1)} \tag{8}$$

DCG, Discounted Cumulative Gain, defines the user's gain as a measure that grows as the user goes from top to bottom positions of the ranking. Under *NDCG*, the gain produced by each ranked document depends on its position. Gains from relevant documents at higher positions are greater than gains from relevant documents at lower positions. To that end, each gain is divided by a discounting factor ($log_2(p+1)$). The *DCG* values are normalised by dividing the *DCG* scores by the ideal *DCG* (*IDCG*, which represents the gains obtained by a perfect system that ranks documents by decreasing order of their actual relevance). *NDCG* can be computed at any cutoff but we report here the *NDCG* scores associated to the entire ranking (whose size is n). Observe that, under *NDCG*, rel_p scores cannot be negative. Following standard practice in the TREC Health Misinformation track, the computation of *NDCG* assigns 0-gain to all documents whose relevance degree is less or equal to 0.

Another IR metric considered in our study is the Convex Aggregation Measure of the Mean Average Precision (*CMAP*) [76]. *CMAP* combines the Mean Average Precision (*MAP*) [75] of usefulness, correctness and credibility as follows:

$$CMAP = \lambda_1 \cdot MAP_u + \lambda_2 \cdot MAP_{co} + \lambda_3 \cdot MAP_{cr} \tag{9}$$

where a uniform combination leads to $\lambda_1, \lambda_2, \lambda_3 = 1/3$. Each *MAP* score comes from inspecting the ranking with a different notion of "relevance": usefulness (MAP_u), correctness (MAP_{co}) and credibility (MAP_{cr}). The *MAP* score is the mean of the average precision (*AP*) values associated to multiple queries:

$$MAP_x = \frac{\sum_{i=1}^{|Q|} AP_x(q_i)}{|Q|} \tag{10}$$

$$AP_x = \frac{1}{r_x} \cdot \sum_{p=1}^{r_x} P(p) \cdot rel(p) \tag{11}$$

where x is u, co or cr. *AP* represents the area under the precision-recall curve. r_x is the number of relevant documents (number of useful, correct or credible documents, respectively), $P(p)$ is the precision at a cutoff p, and $rel(p)$ equals 1 if the item at rank p is a relevant document, and 0 otherwise. Observe that *AP* works with binary relevance values. The usefulness labels (third column in Table 1) are already binary and, thus, MAP_u can be straightforwardly computed. For computing the *AP* of correctness (MAP_{co}), documents that give no answer or documents that have not been assessed for correctness are assigned a score equal to 0 (i.e. 2 and -1 are transformed into 0). For computing the *AP* of credibility (MAP_{cr}), documents that have not been assessed for credibility are assigned a score equal to 0 (i.e. -1 are transformed into 0).

The graded relevance values were also employed to compute other innovative metrics, such as compatibility [77]. Compatibility estimates the similarity between a ranked list provided by an automatic system and an ideal ranking. Clarke and colleagues utilised Rank Biased Overlap (*RBO*) [78] to compute compatibility between an ideal ranking I and an actual ranking L as follows:

$$RBO(L, I) = (1 - pat) \cdot \sum_{p=1}^{\infty} pat^{\text{p-1}} \frac{|I_{1:\text{p}} \cap L_{1:\text{p}}|}{p} \tag{12}$$

where $I_{1:\text{p}}$ and $L_{1:\text{p}}$ represent the top p documents in I and L, respectively. The overlap between both rankings at the cutoff p is defined as the size of the intersection of these lists. *RBO* is then a weighted average across cutoffs from 1 to ∞, and $pat \in (0, 1)$ models searcher patience.

Following [15], we calculate: i) compatibility helpful, where the ideal ranking is composed only of the documents whose relevance level is greater than zero (ordered by decreasing graded relevance), and ii) compatibility harmful,

Relevance Degree	Description	Usefulness	Correctness	Credibility
4	Useful, correct, credible	1	1	1
3	Useful, correct, not credible or no credibility judgement	1	1	0 or -1
2	Useful, no answer or no judgement for answer, credible	1	2 or -1	1
1	Useful, no answer or no judgement for answer, not credible or no judgement	1	2 or -1	0 or -1
0	Not useful, ignore answer and credibility	0	-	-
-1	Useful, incorrect, not credible or no judgement	1	0	0 or -1
-2	Useful, incorrect, credible	1	0	1

Table 1: Preference ordering for documents mapped to graded relevance. Usefulness=1 (0) means that the document is on-topic (off-topic). Correctness=1 (0) means that the document gives a correct (incorrect) answer to the health-related request. Correctness=2 means that the document gives no answer to the health-related request. Correctness=-1 means that the document was not manually judged in terms of correctness to the health-related request. Credibility=1 (0) means that the document was judged as credible (non-credible). Credibility=-1 means that the document was not judged in terms of credibility.

where the ideal ranking is composed only of the documents whose relevance level is negative (ordered by increasing graded relevance). A good system should score high on compatibility helpful and low on compatibility harmful. Additionally, a global compatibility score is reported as the difference between the compatibility helpful achieved by a system and its compatibility harmful.

6.3. Experimental details

The TREC 2020 Health Misinformation dataset contains 31 search topics that have both helpful and harmful documents[9]. Some of the signals described above require training data and, thus, we employed a three-fold splitting strategy. This means that the effectiveness results are averaged over the three test folds. In order to get comparable results for the non-supervised methods, these were also evaluated on the same test folds (and their results averaged). However, it is important to bear in mind that the unsupervised methods did not utilise any information from the training fold associated to each test fold.

To analyse the statistical significance of the performance difference between two systems or alternatives, we applied the Wilcoxon test on the paired values (one from each query). Parapar and colleagues [79, 80] recently compared several significance tests in the context of Information Retrieval experiments and showed that Wilcoxon test is a highly reliable test to compare retrieval systems (yields more statistical power and fewer type I errors). The reported results of statistical significance correspond with the entire set of 31 topics (each query result extracted from its corresponding test fold). The significance tests help us to determine whether or not each observed difference is anecdotal.

The following settings were utilised for the different parts of our architecture:

- **Document-level relevance (BM25)**: we used Pyserini's implementation[4] of BM25 with k_1 set to 0.9 and b set to 0.4.
- **Passage-level relevance (MonoT5)**: we used the Pygaggle[5] library and, more specifically, the MonoT5 model fine-tuned for passage re-ranking with Med-MARCO.
- **Reliability estimation (supervised methods)**: given the training queries available in the train fold, we fine-tuned a T5-base model with a constant learning rate of 3×10^{-4} for a variable number of iterations depending on fold size and with batches of size 8. We ran 2 training epochs and selected a maximum length of 512 tokens.

  We evaluated three strategies to fine-tune this classifier:

[9]Originally, the track organisers created 50 topics but, after building the relevance assessments, many of them ended up with only examples of helpful documents. We therefore focus on the 31 topics that have reliable and unreliable retrieval results.

	Rprec Incorrect
Document relevance (DOC_REL)	0.1025
Passage relevance (PAS_REL)	0.1096

Table 2: Relevance-based search method results for the total recall task.

	CMAP	NDCG	Comp. harmful	Comp. helpful	Compatibility
DOC_REL	0.2537	0.5101	0.1206	0.3186	0.1981
PAS_REL	**0.2871**$^{\uparrow}$	**0.5435**	**0.1180**	**0.3794**	**0.2614**

Table 3: Relevance-based search method results for the ad-hoc retrieval task. Please note that the ↑/↓ symbols indicate whether the method significantly improves the baseline or not, but in the case of the Compatibility harmful measure, an improvement is associated to a lower absolute value.

  - training with the (unmodified) query (description field) upfront + passage + label (class q).
  - training with a correct hand-crafted expression + passage + label (class cs).
  - training with an incorrect hand-crafted expression + passage + label (class is).

  The hand-crafted expressions were created from the description and answer fields of each topic. For example, for the question "Can Vitamin D cure COVID-19?", whose correct answer is "no", the two expressions were: "vitamin D can cure COVID-19" (incorrect hand-crafted expression) and "vitamin D can not cure COVID-19" (correct hand-crafted expression). The creation of these expressions from the description and answer field of each topic is inspired by previous studies on parsing [81].

  Given the labels assigned to the documents, we have opted to build a correctness classifier (by considering only the correctness label), a credibility classifier (by considering only the credibility label) or a correctness+credibility classifier (by considering the conjunction of both labels). After some preliminary experiments, we decided to adopt the latter as our reference reliability classifier.

- **Reliability estimation (unsupervised methods)**: the unsupervised strategy also involves creating hand-crafted expressions but, rather than using them to build a classifier, they are employed to search for similar sentences within each target passage. With this in mind, we used Sentence BERT library[10][11]. More specifically, we experimented with four different models:

  - BERT Large fine-tuned only with NLI dataset [82] (sim BERT-NLI).
  - BERT Large fine-tuned with NLI+STSB datasets [83] (sim BERT-STSB).
  - RoBERTa Large fine-tuned with NLI dataset (sim RoBERTa-NLI).
  - RoBERTa Large fine-tuned with NLI+STSB datasets (sim RoBERTa-STSB).

  The final score consists of the average value of all the similarity scores computed between the hand-crafted expression and each passage sentence.

7. Results

7.1. Relevance-based search methods

First of all, we tested the performance of the stages of our pipeline that merely incorporate topic relevance, namely: the document relevance estimation phase (DOC_REL), and the passage relevance estimation phase (PAS_REL).

[10] https://www.sbert.net/
[11] https://pypi.org/project/sentence-transformers/0.4.1.2/

	Rprec Incorrect
Reference	
PAS_REL	0.1096
Supervised methods	
Passage reliability (class q) (PAS_RELIA_C_Q)	$0.0703^{\downarrow}$
Passage reliability (class cs) (PAS_RELIA_C_CS)	$0.0726^{\downarrow}$
Passage reliability (class is) (PAS_RELIA_C_IS)	0.0822
Unsupervised methods	
Passage reliability (sim BERT-NLI) (PAS_RELIA_S_BN)	$0.0826^{\downarrow}$
Passage reliability (sim BERT-STSB) (PAS_RELIA_S_BS)	0.0908
Passage reliability (sim RoBERTa-NLI) (PAS_RELIA_S_RN)	$0.0796^{\downarrow}$
Passage reliability (sim RoBERTa-STSB) (PAS_RELIA_S_RS)	0.0780

Table 4: Passage reliability estimation results (supervised and unsupervised methods) for the total recall task. Please note that the ↑/↓ symbols indicate whether the method significantly improves the baseline or not.

Results for both tasks are shown in Tables 2 and 3. The best method for both tasks is PAS_REL, which yields the best performance figures in all metrics. This strategy takes the ranking generated from document relevance scores and re-ranks the first 100 documents by decreasing passage relevance score. These results show that the passage-based strategy is effective. Scoring documents based on the most relevant passage leads to substantial improvements in performance and in one case the improvement is statistically significant. The passage relevance approach looks promising (and, on average, leads to higher effectiveness than that of document relevance). However, the characteristics of this test set make it hard to reveal statistical significance. We analysed the individual (per-query) effectiveness scores and, for example, the PAS_REL variant leads to improved performance in 19 out of 31 queries (compatibility). With a larger query test we suspect that we could easily obtain improvements that are statistical significant. In any case, the improvements are not consistent across queries and, in the near future, we plan to further explore methods that incorporate query-dependent techniques (e.g., alternate between passage and document retrieval in a topic-dependent way).

Relevant passages represent a concise and on-topic representation of the document that eliminates content that is unrelated to the query. The relative merits of PAS_REL and DOC_REL clearly suggest that misinformation detection should concentrate on the most relevant extracts from the retrieved webpages.

We also ran some exploratory experiments where we combined the scores of document relevance and passage relevance. To that end, we employed the score fusion techniques described in Section 5. However, these tests did not result in any advance over PAS_REL alone and, thus, we adopted the passage relevance signal as the reference topic-relevance baseline for further experiments.

7.2. Reliability estimation at passage-level

Next, we evaluated the effectiveness of the passage reliability estimation methods. To this end, we re-ranked the top 100 documents by decreasing estimation of reliability of the most relevant passage. We experimented with the supervised and unsupervised reliability methods described in Section 4.

For the first subtask, total recall, these methods fail to outperform the relevance-based baseline (see Table 4). There are three methods whose performance is not statistically inferior to the baseline. However, no method yields performance figures higher than PAS_REL and, thus, the reliability signal alone is insufficient to find documents that include misinformation. Note that the best performing supervised alternative is PAS_RELIA_C_IS, which is the method that trains with the incorrect hand-crafted expression upfront. This is a natural outcome, as the total recall task aims at searching for incorrect documents.

For the second subtask, ad-hoc retrieval, the supervised methods again yield poor performance (see Table 5). In terms of compatibility harmful, the supervised strategies lead to substantial benefits but, for the remaining metrics (including global compatibility), performance is much worse than that of the baseline. This suggests that these methods have poor retrieval performance. The lack of on topic documents retrieved means that fewer are either helpful or harmful. The best supervised method is here PAS_RELIA_C_CS, which trains with the hand-crafted correct sentence. Again, this is a natural outcome, as the ad-hoc retrieval task aims to find reliable documents. On the other hand, the

	CMAP	NDCG	Comp. harmful	Comp. helpful	Compatibility
Reference					
PAS_REL	0.2871	0.5435	0.1180	0.3794	0.2614
Supervised methods					
PAS_RELIA_C_Q	$0.2174^{\downarrow}$	$0.4799^{\downarrow}$	$\mathbf{0.0558}^{\uparrow}$	$0.2528^{\downarrow}$	$0.1971^{\downarrow}$
PAS_RELIA_C_CS	$0.2427^{\downarrow}$	0.5143	$\mathbf{0.0645}^{\uparrow}$	$0.3116^{\downarrow}$	0.2472
PAS_RELIA_C_IS	$0.2176^{\downarrow}$	$0.4783^{\downarrow}$	$\mathbf{0.0636}^{\uparrow}$	$0.2711^{\downarrow}$	0.2075
Unsupervised methods					
PAS_RELIA_S_BN	$0.2700^{\downarrow}$	**0.5437**	$\mathbf{0.0614}^{\uparrow}$	0.3624	**0.3010**
PAS_RELIA_S_BS	$0.2722^{\downarrow}$	0.5391	$\mathbf{0.0767}^{\uparrow}$	0.3709	**0.2942**
PAS_RELIA_S_RN	$0.2584^{\downarrow}$	0.5428	$\mathbf{0.0533}^{\uparrow}$	0.3611	**0.3077**
PAS_RELIA_S_RS	$0.2657^{\downarrow}$	**0.5441**	$\mathbf{0.0571}^{\uparrow}$	**0.3990**	**0.3419**

Table 5: Passage reliability estimation results (supervised and unsupervised methods) for the ad-hoc retrieval task. Please note that the ↑/↓ symbols indicate whether the method significantly improves the baseline or not, but in the case of the Compatibility harmful measure, an improvement is associated to a lower absolute value.

unsupervised methods seem to provide added value for the ad-hoc retrieval task. All variants lead to global compatibility scores higher than that of the baseline, and provide statistically better compatibility harmful scores. In terms of CMAP, NDCG and compatibility helpful these methods are weaker. This suggest that, in general, these methods are better at downgrading harmful results but not so good at finding helpful results. The most robust method is PAS_RELIA_S_RS, which outperforms the baseline in nearly all metrics. This variant is based on a RoBERTa Large model obtained from the NLI and STSB datasets.

Note also that all variants show poor CMAP scores, while the NDCG scores (particularly those obtained with the unsupervised methods) are more competitive. CMAP (and Mean Average Precision, on which CMAP depends) is a measure influenced by how precision evolves over the entire ranking, while NDCG is a measure more oriented to high precision because it incorporates a discounting factor for relevant documents that grows with the position. NDCG has been recognized as a metric that reflects user behaviour well (e.g., web users rarely inspect a full ranking of results). In our case, NDCG is a more important measure, not only because it reflects a typical high-precision search scenario but also because it handles graded relevance (while CMAP/MAP only incorporate a binary notion of relevance).

7.3. Score Fusion

Having analyzed the individual effect of document relevance, passage relevance and passage reliability signals, we study now the effectiveness of combining multiple signals. Accordingly, we compare a selection of appropriate unsupervised rank fusion methods and learning-to-rank techniques.

7.3.1. Unsupervised Rank Fusion: CombSUM

Table 6 shows the results for the total recall task. Combining document/passage relevance with reliability estimation methods based on supervised techniques (second block of the table) leads to poor performance. The supervised strategy that trains with the incorrect sentence is again the best choice. However, its performance (DOC_REL + PAS_REL + PAS_RELIA_C_IS row) remains lower than that of the baseline. Combining document/passage relevance with reliability estimation methods based on unsupervised techniques (third block of the table) leads to more effective fusion variants. Several combinations outperform the baseline in terms of R-precision. However, no improvement is statistically significant.

For the ad-hoc retrieval task, the results show a similar trend. Combinations involving the supervised methods (second block, Table 7) show no benefit or even yield performance statistics that are statistically worse than those of the baseline. Fusion variants with unsupervised methods (third block, Table 7), instead, tend to produce improvements over the baseline (and many of them are statistically significant). Remarkably, the fusion of passage relevance with reliability estimation from the RoBERTa-Large-STSB model (PAS_REL + PAS_RELIA_S_RS row) and the fusion of document and passage relevance with reliability estimation from the RoBERTa-Large-STSB model (DOC_REL + PAS_REL + PAS_RELIA_S_RS row) show consistent improvements in terms of CMAP, NDCG and compatibility helpful. These methods are weaker in terms of compatibility harmful. The improvement of helpful-related metrics

	Rprec Incorrect
Reference	
PAS_REL	0.1096
Supervised methods	
DOC_REL + PAS_RELIA_C_Q	$0.0742^{\downarrow}$
PAS_REL + PAS_RELIA_C_Q	$0.0725^{\downarrow}$
DOC_REL + PAS_RELIA_C_CS	0.0902
PAS_REL + PAS_RELIA_C_CS	0.0891
DOC_REL + PAS_RELIA_C_IS	0.0929
PAS_REL + PAS_RELIA_C_IS	0.0998
DOC_REL + PAS_REL + PAS_RELIA_C_Q	$0.0771^{\downarrow}$
DOC_REL + PAS_REL + PAS_RELIA_C_CS	0.0990
DOC_REL + PAS_REL + PAS_RELIA_C_IS	0.1007
Unsupervised methods	
DOC_REL + PAS_RELIA_S_BN	0.1017
PAS_REL + PAS_RELIA_S_BN	0.1028
DOC_REL + PAS_RELIA_S_BS	**0.1155**
PAS_REL + PAS_RELIA_S_BS	0.1021
DOC_REL + PAS_RELIA_S_RN	0.1015
PAS_REL + PAS_RELIA_S_RN	0.0934
DOC_REL + PAS_RELIA_S_RS	0.1037
PAS_REL + PAS_RELIA_S_RS	0.0979
DOC_REL + PAS_REL + PAS_RELIA_S_BN	**0.1126**
DOC_REL + PAS_REL + PAS_RELIA_S_BS	**0.1191**
DOC_REL + PAS_REL + PAS_RELIA_S_RN	0.1078
DOC_REL + PAS_REL + PAS_RELIA_S_RS	**0.1171**

Table 6: CombSUM results (supervised and unsupervised methods) for the total recall task. Please note that the ↑/↓ symbols indicate whether the method significantly improves the baseline or not.

(CMAP, NDCG, compatibility helpful) usually comes at a cost of damaging harmful-related statistics (because we often move more on-topic documents to higher positions in the rankings and some of them might be harmful). This tradeoff between compatibility harmful and helpful is something we will discuss shortly and will be the subject of further analysis in Section 8. However, avoiding harm should not be our single criterion, as it would be trivial to achieve a system with perfect harmful scores (simply retrieving no documents would result in no harm produced).

7.3.2. Unsupervised Rank Fusion: Borda Count

The fusion experiments reported above were repeated for a second type of unsupervised fusion strategy: Borda Count. Overall, Borda count yielded similar results compared with the results obtained with CombSUM. For the sake of simplicity, we only report the effectiveness of Borda Count for the most effective fusion variants. In Tables 8 and 9, the reader can observe the relative merits of CombSUM (second block) against Borda Count (third block). This comparison does not reveal a clear winner. It seems that, for combining these pieces of evidence, the potential advantage of manipulating scores (CombSUM) does not translate into practical improvements in effectiveness.

7.3.3. Learning-to-rank

A second class of combination strategy consists of applying learning to rank methods. Given some training examples where we know the query-document scores[12] and the target variable (reliability of the document), we build a classifier that learns to combine the individual features. This strategy requires to further split the training queries into two subsets, where one subset is used to build the supervised models (if required) and the other subset is used by L2R to learn the combination of features. We set aside 5 queries for learning the combination. Observe that supervised methods (e.g., PAS_RELIA_C_Q): i) use the first subset of queries to learn the reliability estimation model,

[12] for a 3-feature combination we would have the document relevance score, passage relevance score and passage reliability score

	CMAP	NDCG	Comp. harmful	Comp. helpful	Compatibility
Reference					
PAS_RELIA_S_RS	0.2657	0.5441	0.0571	0.3990	0.3419
Supervised methods					
DOC_REL + PAS_RELIA_C_Q	0.2332$^{\downarrow}$	0.5033$^{\downarrow}$	0.0625	0.3088$^{\downarrow}$	0.2464$^{\downarrow}$
PAS_REL + PAS_RELIA_C_Q	0.2372$^{\downarrow}$	0.5009$^{\downarrow}$	0.0707	0.3086$^{\downarrow}$	0.2379$^{\downarrow}$
DOC_REL + PAS_RELIA_C_CS	0.2474	0.5149	0.0779	0.3589	0.2810$^{\downarrow}$
PAS_REL + PAS_RELIA_C_CS	0.2494$^{\downarrow}$	0.5126$^{\downarrow}$	0.0783	0.3544$^{\downarrow}$	0.2762$^{\downarrow}$
DOC_REL + PAS_RELIA_C_IS	0.2344$^{\downarrow}$	0.4958$^{\downarrow}$	0.0662	0.2944$^{\downarrow}$	0.2282$^{\downarrow}$
PAS_REL + PAS_RELIA_C_IS	0.2363$^{\downarrow}$	0.4983$^{\downarrow}$	0.0741	0.3226$^{\downarrow}$	0.2485$^{\downarrow}$
DOC_REL + PAS_REL + PAS_RELIA_C_Q	0.2492	0.5158	0.0852	0.3431$^{\downarrow}$	0.2579$^{\downarrow}$
DOC_REL + PAS_REL + PAS_RELIA_C_CS	0.2583	0.5185	0.0853	0.3720	0.2868
DOC_REL + PAS_REL + PAS_RELIA_C_IS	0.2423$^{\downarrow}$	0.4999$^{\downarrow}$	0.0813	0.3265$^{\downarrow}$	0.2452$^{\downarrow}$
Unsupervised methods					
DOC_REL + PAS_RELIA_S_BN	**0.2821**$^{\uparrow}$	**0.5663**	0.0818	**0.4024**	0.3209
PAS_REL + PAS_RELIA_S_BN	**0.2754**$^{\uparrow}$	**0.5455**	0.0767	0.3844	0.3077
DOC_REL + PAS_RELIA_S_BS	**0.2693**	0.5317	0.1031$^{\downarrow}$	0.3850	0.2818
PAS_REL + PAS_RELIA_S_BS	**0.2837**$^{\uparrow}$	0.5423	0.0908$^{\downarrow}$	0.3974	0.3066
DOC_REL + PAS_RELIA_S_RN	**0.2774**$^{\uparrow}$	**0.5593**	0.0736	**0.4089**	0.3353
PAS_REL + PAS_RELIA_S_RN	**0.2728**	**0.5490**	0.0708	0.3896	0.3188
DOC_REL + PAS_RELIA_S_RS	**0.2774**$^{\uparrow}$	**0.5593**	0.0736	**0.4089**	0.3353
PAS_REL + PAS_RELIA_S_RS	**0.2753**$^{\uparrow}$	**0.5486**$^{\uparrow}$	0.0698$^{\downarrow}$	**0.4177**$^{\uparrow}$	**0.3479**
DOC_REL + PAS_REL + PAS_RELIA_S_BN	**0.2879**$^{\uparrow}$	**0.5648**$^{\uparrow}$	0.0954$^{\downarrow}$	**0.4162**	0.3209
DOC_REL + PAS_REL + PAS_RELIA_S_BS	**0.2791**$^{\uparrow}$	0.5409	0.1104$^{\downarrow}$	**0.4039**	0.2935
DOC_REL + PAS_REL + PAS_RELIA_S_RN	**0.2859**$^{\uparrow}$	**0.5616**$^{\uparrow}$	0.0869$^{\downarrow}$	**0.4281**	0.3413
DOC_REL + PAS_REL + PAS_RELIA_S_RS	**0.2874**$^{\uparrow}$	**0.5571**	0.0829$^{\downarrow}$	**0.4370**$^{\uparrow}$	**0.3541**

Table 7: CombSUM results (supervised and unsupervised methods) for the ad-hoc retrieval task. Please note that the ↑/↓ symbols indicate whether the method significantly improves the baseline or not, but in the case of the Compatibility harmful measure, an improvement is associated to a lower absolute value.

	Rprec Incorrect
Reference	
PAS_REL	0.1096
CombSUM	
DOC_REL + PAS_RELIA_S_BS	**0.1155**
DOC_REL + PAS_REL + PAS_RELIA_S_BS	**0.1191**
Borda Count	
DOC_REL + PAS_RELIA_S_BS	**0.1147**
DOC_REL + PAS_REL + PAS_RELIA_S_BS	**0.1222**
Leraning-to-Rank	
DOC_REL + PAS_RELIA_S_BS	0.0621$^{\downarrow}$
DOC_REL + PAS_REL + PAS_RELIA_S_BS	0.0786$^{\downarrow}$

Table 8: Borda Count and Learning-to-Rank results (only unsupervised methods) for the total recall task. Please note that the ↑/↓ symbols indicate whether the method significantly improves the baseline or not.

ii) the resulting reliability classifier predicts the reliability score for each document in the ranking of the second subset of queries, iii) the reliability scores together with the other query-document features are fed to the L2R model that learns the combination method, and iv) the learnt combination approach is run against the queries in the test

	CMAP	NDCG	Comp. harmful	Comp. helpful	Compatibility
Reference					
PAS_RELIA_S_RS	0.2657	0.5441	0.0571	0.3990	0.3419
CombSUM					
PAS_REL + PAS_RELIA_S_RS	**0.2753**$^{\uparrow}$	**0.5486**$^{\uparrow}$	0.0698$^{\downarrow}$	**0.4177**$^{\uparrow}$	**0.3479**
DOC_REL + PAS_REL + PAS_RELIA_S_RS	**0.2874**$^{\uparrow}$	**0.5571**	0.0829$^{\downarrow}$	**0.4370**$^{\uparrow}$	**0.3541**
Borda Count					
PAS_REL + PAS_RELIA_S_RS	**0.2894**$^{\uparrow}$	**0.5529**$^{\uparrow}$	0.0892$^{\downarrow}$	**0.4278**	0.3386
DOC_REL + PAS_REL + PAS_RELIA_S_RS	**0.2866**$^{\uparrow}$	**0.5514**	0.1133$^{\downarrow}$	**0.4400**$^{\uparrow}$	0.3267
Learning-to-Rank					
PAS_REL + PAS_RELIA_S_RS	0.2022$^{\downarrow}$	0.4503$^{\downarrow}$	0.0312$^{\uparrow}$	0.1546$^{\downarrow}$	0.1234$^{\downarrow}$
DOC_REL + PAS_REL + PAS_RELIA_S_RS	0.2028$^{\downarrow}$	0.4494$^{\downarrow}$	0.0292$^{\uparrow}$	0.1505$^{\downarrow}$	0.1213$^{\downarrow}$

Table 9: Borda and Learning-to-Rank results (only unsupervised methods) for the ad-hoc retrieval task. Please note that the ↑/↓ symbols indicate whether the method significantly improves the baseline or not, but in the case of the Compatibility harmful measure, an improvement is associated to a lower absolute value.

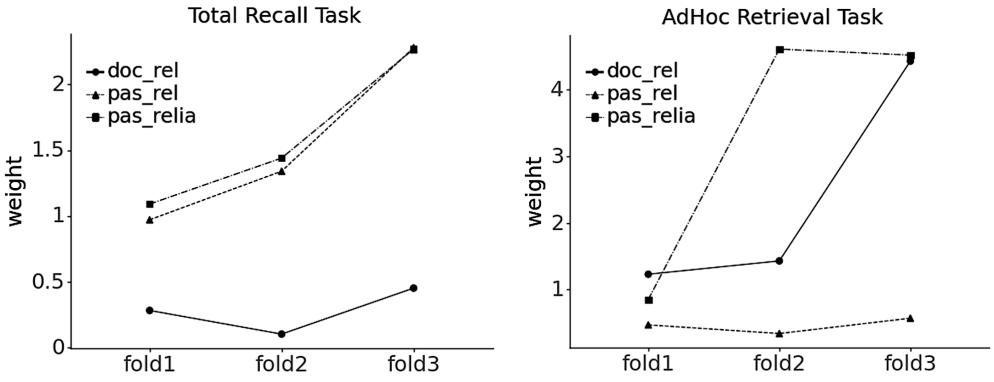

Figure 9: Logistic regression weights obtained from each training fold.

fold. Unsupervised methods (e.g., PAS_RELIA_S_BN), instead, do not employ the first subset of queries: i) the unsupervised reliability estimation model predicts the reliability score for each document in the ranking of the second subset of queries, ii) the reliability scores together with the other query-document features are fed to the L2R model that learns the combination method, and iii) the learnt combination approach is run against the queries in the test fold.

We performed pointwise L2R, which predicts the value of the target variable for every single document. Results for both subtasks are shown in Tables 8 and 9 (last blocks). L2R does not give an added value over simpler fusion strategies. It seems that, with the available training queries, L2R methods are not able to learn a combination function that extrapolates to unseen queries. To further prove this point, Figure 9 represents the logistic regression weights obtained from each training fold (both subtasks are shown in the graph). It can be observed that the weight assigned to each signal varies enormously among folds. In total recall, the document relevance signal is always the feature assigned with lowest weight but its importance compared with the other two signals varies significantly over the three folds. In the ad-hoc retrieval task the situation is totally different: the passage relevance signal gets the lowest weights while document relevance and passage reliability show a erratic trend over the three folds. These plots support our hypothesis about the poor generalisation capability of the L2R algorithm. The graph clearly shows that the learned logistic regression models have high variance and we would need many more training examples to build a reliable combination model.

In search technologies, it is well known that there is a wide variability in the characteristics of queries. For example, some queries find a large number of relevant documents while other queries have few documents that are

	Rprec Incorrect
Best run	
KU (University of Copenhagen)	0.1300
Other teams' runs	
UWaterlooMDS (University of Waterloo)	0.1040
vohcolab (Universidade NOVA de Lisboa)	0.1030
Median of all runs	
MEDIAN RUN	0.0976
Our top performers	
DOC_REL + PAS_REL + PAS_RELIA_S_RS	0.1180
PAS_REL	0.1170
PAS_REL + PAS_RELIA_S_RS	0.0990
PAS_RELIA_S_RS	0.0801

Table 10: Comparison of official TREC 2020 runs and our best performers for the total recall task (data extracted from [15]).

on-topic. This high variance makes that L2R methods would require a large corpus of training examples in order to build a robust combination approach. But this luxury cannot be afforded in this misinformation detection task, where labelled data does not abound. In general web search scenarios, massive examples of topics and the associated clickthrough data are available to the search engine [84] and, thus, popular web retrieval engines can make good use of L2R strategies [48]. We focus instead on a more specific task where, in most of the cases, it is critical to avoid the spread of misinformation at early stages. This requires working with few training examples.

7.4. Comparison with external baselines

A way to put these results in context is to compare them with other studies using the same tasks and datasets. To that end, we consider here the participants in the TREC 2020 Health Misinformation competition[13]. For each task, we report the performance of the winner team (Best run block), the performance of the best run of the teams that ranked 2nd and 3rd (Other teams' runs block), the median performance of all runs (Median of all runs block) and the performance of some of our variants (last block).

For the total recall task, results are shown in Table 10. All our best performers (except one) are above the median *Rprec* of the submitted runs. Moreover, our top performer is better than 78% of the proposed solutions. The performance of the best run, KU from the University of Copenhagen [53], is higher than ours but their solution is based on a supervised model that was fed with external data, whereas our top performers are fully unsupervised.

For the ad-hoc Retrieval task (see Table 11), our improvement over the median increases, and our top performer is better than 88% of the submitted runs. Regarding the winner solution (H2oloo team, from the University of Waterloo), we obtain comparable compatibility Helpful results but we retrieve more harmful documents. This tradeoff will be further discussed in the next section. Nevertheless, it must be noticed again that the H2oloo solution is based on a supervised model trained with external data (and, in this case, obtained from a huge corpus, the T5-3b model [50]), while our top performers did not resort to supervision.

8. Discussion

It is important to analyse the trade-off between the helpful and harmful compatibility results of the proposed solutions. Figure 10 plots some representative variants at the point where the X value corresponds with its compatibility helpful and the Y value corresponds with its compatibility harmful. Ideally, we want the system to be positioned at the bottom right of the graph because the main goal consists of minimising the retrieval of harmful results without damaging the retrieval of helpful documents.

[13]To make results comparable with the other studies, our strategies had to be recomputed using all 50 topics assessed in the TREC task (instead of the 31 topics considered in Section 6.3). In this case, significance tests could not be performed because we have only these teams' mean scores (per-query results are not available).

	Comp. harmful	Comp. helpful	Compatibility
Best run			
H2oloo (University of Waterloo)	0.0160	0.4900	0.4740
Other teams' runs			
Webis (Bauhaus-Universität Weimar and Martin-Luther-Universität Halle-Wittenberg)	0.0520	0.3340	0.2820
KU (University of Copenhagen)	0.1210	0.4010	0.2800
Median of all runs			
MEDIAN RUN	0.0747	0.3337	0.259
Our top performers			
DOC_REL + PAS_REL + PAS_RELIA_S_RS	0.0825	0.4745	0.3920
PAS_REL + PAS_RELIA_S_RS	0.0711	0.4537	0.3826
PAS_RELIA_S_RS	0.0587	0.4209	0.3622
PAS_REL	0.1210	0.4370	0.3160

Table 11: Comparison of official TREC 2020 runs and our best performers for the ad-hoc retrieval task (data extracted from [15]). For the sake of simplicity we only report here the official metric by which the participating solutions were ranked (the difference between compatibility values).

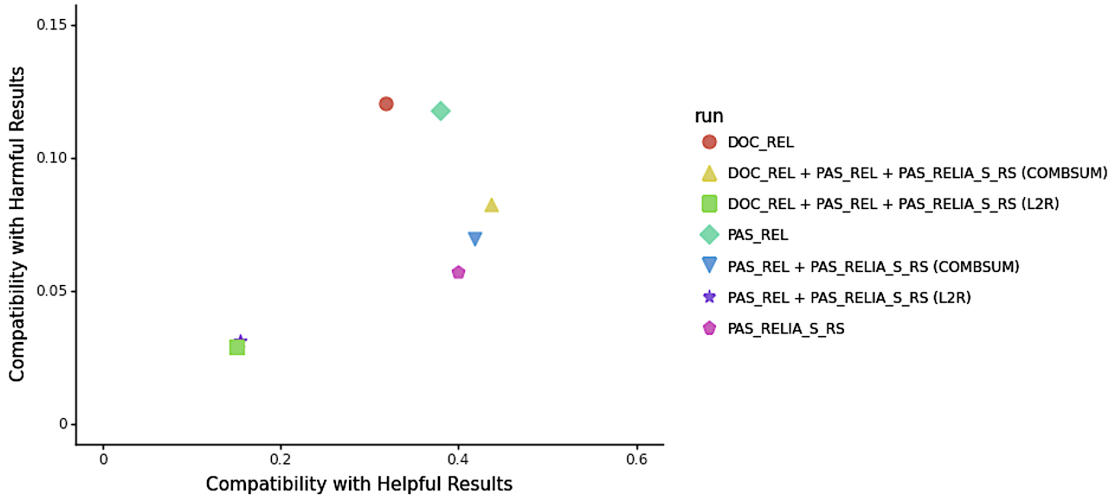

Figure 10: ad-hoc results: Compatibility of runs with helpful and harmful results. A good run is helpful and not harmful. For a certain level of helpfulness, less harm is prefered.

We analyse here a representative set of variants, which includes a variant based only on document relevance (DOC_REL), a variant based only on passage relevance (PAS_REL), a variant based only on passage reliability (PAS_RELIA_S_RS) and four fusion variants (two of them based on COMBSUM and two of them based on L2R). This analysis helps to further clarify some of our main findings.

The plot shows three main clusters of variants. The L2R variants, which yield low compatibility harmful and low compatibility helpful, the two relevance-based variants (DOC_REL and PAS_REL), which have high compatibility harmful and medium-to-high compatibility helpful, and the three remaining variants (PAS_RELIA_S_RS and the two COMBSUM variants), which have high compatibility helpful and medium compatibility harmful.

First, the two L2R alternatives, which are clustered at a low helpful and low harmful area, have limiting retrieval capabilities. Although they retrieve few harmful contents, their ability to bring helpful documents is clearly suboptimal. As argued above, we would need much more training data in order to make the most of these learning-based combinations.

Second, the initial retrieval of documents (DOC_REL) finds many harmful results. Given a document relevance ranking, we clearly need additional ingredients to decrease the retrieval of harmful documents and increase the retrieval of helpful contents. Passage-level relevance represents a first step in this direction. Compared with DOC_REL,

PAS_REL improves the retrieval of helpful documents and, at the same time, slightly decreases the retrieval of harmful contents. This suggests that focusing on the most relevant extracts is beneficial to identify the most helpful webpages.

Third, PAS_RELIA_S_RS, PAS_REL+PAS_RELIA_S_RS (COMBSUM), and DOC_REL+PAS_REL+PAS_RELIA_S_RS (COMBSUM) are clearly the most solid choices, as they outperform the two relevance-based variants in both compatibility measures (lower compatibility harmful and higher compatibility helpful compared to DOC_REL or PAS_REL). If we want to fare on the conservative side, we could choose PAS_RELIA_S_RS: it retrieves fewer helpful documents but it also results in less damage. On the other hand, if we want higher recall of helpful documents then DOC_REL+PAS_REL+PAS_RELIA_S_RS (COMBSUM) would be our preferred choice: it retrieves more helpful webpages at the cost of presenting more harmful contents in the rankings. In practice, the selection of one of these methods would depend on the specific user task and his/her willingness to weight on helpfulness or harmfulness. For example, a website moderator willing to thoroughly inspect the presence of helpful and harmful contents within his/her site would probably prefer DOC_REL+PAS_REL+PAS_RELIA_S_RS (COMBSUM). But if the goal is to identify the most reputed contents about a given topic (e.g., to label them as useful suggestions) then PAS_RELIA_S_RS would be a more cost-effective strategy (similar helpful results compared with DOC_REL+PAS_REL+PAS_RELIA_S_RS (COMBSUM) and PAS_REL+PAS_RELIA_S_RS (COMBSUM), but lower harmful results).

In general, we found the following tendency: the better our retrieval systems are in terms of compatibility helpful, the more harmful documents are also found. And the other way around, if we decrease the retrieval of harmful webpages it is often at the cost of decreasing the retrieval of helpful webpages. It is quite difficult to find an artifact that substantially improves both dimensions. In any case, our goal in the future is to continue studying the specifics of this challenging task and conduct research on new features or strategies oriented to show a good balance between helpfulness and harmfulness. We are also interested in designing novel thresholding strategies adapted to this retrieval problem (e.g., given a ranked set of webpages determine the ideal cutoff position taking into account both dimensions).

The main takeaways could be summarised as follows:

- Focusing on the most relevant passages of documents leads to benefits that are modest but promising. This passage-relevance approach tends to improve helpfulness and decrease harmfulness.
- stimating passage reliability also helps. However, we found that there is a substantial difference between opting for a supervised or an unsupervised estimation, being the latter the best performer. At the early stages of an information outbreak (and COVID-19 is a clear case), the availability of topically-related training data is scarce and our results clearly demonstrate that, under this stringent scenario, unsupervised reliability estimation seems to be a good choice. However, there are also semi-supervised learning or transfer learning techniques, which could be considered to further support this task. In the near future, we plan to explore the role of semi-supervised models or transfer learning models for these tasks.
- Simple score fusion techniques like CombSUM have been demonstrated to outperform L2R strategies for this task, which would require more training data to reach their full potential.

9. Conclusions and Future Work

In this paper, we have conducted a thorough study on which signals or pieces of evidence and combination methods could be helpful in the task of identifying health-related misinformation. We contributed with:

- A complete multistage retrieval system whose goal is to discern between reliable and unreliable contents. This technological solution is available to be reused or adapted[1] by researchers interested in misinformation, web moderators, vertical search engine creators, or other potential stakeholders.
- A comparative study that empirically validated the potential of the platform for a socially worrying case, COVID-19 misinformation. Our analysis has assessed the effect of search-based stages, at document and passage level, reliability estimators based on supervised and non-supervised models and different fusion strategies. Every stage that we included in our system improved the overall performance, and in some cases, significantly. The fusion or combination of multiple forms of evidence (document relevance, passage relevance and passage reliability) led to the most efficient misinformation estimation methods.

- The top-performing variants have competitive performance when compared with state-of-the-art methods (and, particularly, with respect to the solutions submitted to the TREC 2020 Health Misinformation Track).
- The trade-off between retrieval of helpful and harmful contents has been analysed in depth. The results reflect that certain signals help more in finding more helpful documents, while others are more prone to limiting the retrieval of harmful contents. However, it is still challenging to find a combination that improves both aspects. The choice of one instance of the system over another would depend on the specifics of the search task.

The findings of this study have to be seen in light of some limitations. The primary limitation to the generalisation of these results is the test collection. Although we performed experiments with two different search tasks, the document collection was the same and the number of available search topics is limited. In the near future we want to extend the empirical validation to new datasets and larger sets of topics. The second limitation concerns the signals analysed. This study has been confined to text-based search or classification signals. It would be interesting to test other types of features, such as those based on network signals (e.g., link-based reputation of the web sources) or interaction/social signals (e.g., effect of the publications on Internet users).

As future work, we also want to further understand the trade-off between harmful and helpful compatibility, and design strategies to determine the ideal cutoff of a ranked list adapted to this problem [85]. We are currently working on additional NLP techniques to be included in our pipeline. For example, we are working with other unsupervised techniques for further removal of noisy contents [86]. In this respect, we will carefully consider recent advances in clustering [87] and how to effectively employ clustering algorithms to further improve misinformation detection. For example, it will be interesting to exploit clustering techniques for organizing the retrieved results into groups of helpful and harmful pages.

Other possible lines of future work include the application of rule-based techniques and new feature selection algorithms [88] to filter rumours or misinformation in the health domain. Related to this, we want to further analyse recent affective computing and sentiment analysis models [16, 89] and study how to employ them to define new features or signals for the task of misinformation detection. Finally, it is also important to work towards the explainability of the proposed solutions. Some parts of our multistage system are based on deep learning models that are black-box techniques and, thus, hard to interpret. We want to learn from recent advances in this area [90] and see how to adapt these proposals to our application domain.

Acknowledgements

The authors thank the support obtained from: i) project RTI2018-093336-B-C21 (Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación & ERDF), ii) project PLEC2021-007662 (MCIN/AEI/10.13039/501100011033, Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Plan de Recuperación, Transformación y Resiliencia, Unión Europea-Next GenerationEU), and iii) Consellería de Educación, Universidade e Formación Profesional (accreditation 2019-2022 ED431G-2019/04, ED431C 2018/29) and the European Regional Development Fund, which acknowledges the CiTIUS-Research Center in Intelligent Technologies of the University of Santiago de Compostela as a Research Center of the Galician University System.

References

[1] Reuters Insitute, University of Oxford, Reuters Digital News Report, 2021. URL: https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2021, [accessed June 9, 2022].
[2] M. Abualsaud, M. D. Smucker, Exposure and order effects of misinformation on health search decisions, in: Proceedings of the 42nd Int. ACM SIGIR Conference on Research and Development in Information Retrieval, 2019.
[3] G. Eysenbach, Infodemiology: The epidemiology of (mis) information, The American Journal of Medicine 113 (2002) 763–765.
[4] S. Y. Rieh, Judgment of information quality and cognitive authority in the web, Journal of the American society for Information Science and Technology 53 (2002) 145–161.
[5] F. A. Pogacar, A. Ghenai, M. D. Smucker, C. L. Clarke, The positive and negative influence of search results on people's decisions about the efficacy of medical treatments, in: Proceedings of the ACM SIGIR Int. Conf. on Theory of Information Retrieval, 2017, pp. 209–216.
[6] R. White, Beliefs and biases in web search, in: Proceedings of the 36th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, 2013, pp. 3–12.
[7] S. Fox, Health topics: 80% of internet users look for health information online, Pew Internet & American Life Project, 2011.

[8] M. S. Islam, T. Sarkar, et al., COVID-19–related infodemic and its impact on public health: A global social media analysis, The American Journal of Tropical Medicine and Hygiene 103 (2020) 1621–1629.
[9] G. Pennycook, J. McPhetres, Y. Zhang, J. G. Lu, D. G. Rand, Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention, Psychological Science 31 (2020) 770–780.
[10] N. Vigdor, Man fatally poisons himself while self-medicating for coronavirus, doctor says, 2020. URL: https://www.nytimes.com/2020/03/24/us/chloroquine-poisoning-coronavirus.html, [accessed June 9, 2022].
[11] D. Matsumoto, H. C. Hwang, V. A. Sandoval, Cross-language applicability of linguistic features associated with veracity and deception, Journal of Police and Criminal Psychology 30 (2015) 229–241.
[12] S. Mukherjee, G. Weikum, Leveraging joint interactions for credibility analysis in news communities, in: Proceedings of the 24th ACM Int. Conf. on Information and Knowledge Management, 2015, pp. 353–362.
[13] A. Adhikari, A. Ram, R. Tang, J. Lin, DocBERT: BERT for document classification, arXiv:1904.08398 (2019).
[14] P. Sondhi, V. V. Vydiswaran, C. Zhai, Reliability prediction of webpages in the medical domain, in: European Conference on Information Retrieval, Springer, 2012, pp. 219–231.
[15] C. Clarke, M. Maistro, M. Smucker, G. Zuccon, Overview of the TREC 2020 Health Misinformation Track, in: Proceedings of the 29th Text REtrieval Conference (TREC), 2020.
[16] A. G. Martín, A. Fernández-Isabel, C. González-Fernández, C. Lancho, M. Cuesta, I. M. de Diego, Suspicious news detection through semantic and sentiment measures, Engineering Applications of Artificial Intelligence 101 (2021) 104230.
[17] B. J. Fogg, Prominence-interpretation theory: Explaining how people assess credibility online, in: Extended abstracts on Human Factors in Computing Systems, 2003, pp. 722–723.
[18] D. H. McKnight, C. J. Kacmar, Factors and effects of information credibility, in: Proceedings of the ninth international conference on Electronic commerce, 2007, pp. 423–432.
[19] Y. Yamamoto, K. Tanaka, Enhancing credibility judgment of web search results, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2011, pp. 1235–1244.
[20] C. Hahnel, F. Goldhammer, U. Kröhne, J. Naumann, The role of reading skills in the evaluation of online information gathered from search engine environments, Computers in Human Behavior 78 (2018) 223–234.
[21] A. L. Ginsca, A. Popescu, M. Lupu, Credibility in information retrieval, Found. Trends Inf. Retr. 9 (2015) 355–475.
[22] R. Urena, F. Chiclana, E. Herrera-Viedma, DecitrustNET: A graph based trust and reputation framework for social networks, Information Fusion 61 (2020) 101–112.
[23] M. Kattenbeck, D. Elsweiler, Understanding credibility judgements for web search snippets, Aslib Journal of Information Management 71 (2019) 368–391.
[24] K. M. Griffiths, T. T. Tang, D. Hawking, H. Christensen, Automated assessment of the quality of depression websites, Journal of Medical Internet Research 7 (2005) e59.
[25] S. C. Matthews, A. Camacho, P. J. Mills, J. E. Dimsdale, The internet for medical information about cancer: help or hindrance?, Psychosomatics 44 (2003) 100–103.
[26] Q. V. Liao, W.-T. Fu, Age differences in credibility judgments of online health information, ACM Transactions on Computer-Human Interaction (TOCHI) 21 (2014) 1–23.
[27] J. Schwarz, M. Morris, Augmenting web pages and search results to support credibility assessment, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2011, pp. 1245–1254.
[28] M. Fernández-Pichel, D. E. Losada, J. C. Pichel, D. Elsweiler, Reliability prediction for health-related content: a replicability study, in: European Conference on Information Retrieval, Springer, 2021, pp. 47–61.
[29] M. Fernández-Pichel, D. E. Losada, J. C. Pichel, D. Elsweiler, Comparing traditional and neural approaches for detecting health-related misinformation, in: Int. Conf. of the Cross-Language Evaluation Forum for European Languages, Springer, 2021, pp. 78–90.
[30] Y. Zhao, J. Da, J. Yan, Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning based approaches, Information Processing & Management 58 (2021) 102390.
[31] A. Olteanu, S. Peshterliev, X. Liu, K. Aberer, Web credibility: Features exploration and credibility prediction, in: P. Serdyukov, P. Braslavski, S. O. Kuznetsov, J. Kamps, S. Rüger, E. Agichtein, I. Segalovich, E. Yilmaz (Eds.), Advances in Information Retrieval, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 557–568.
[32] C. Edwards, P. Spence, C. Gentile, A. Edwards, A. Edwards, How much klout do you have...a test of system generated cues on source credibility, Computers in Human Behavior 29 (2013) A12–A16.
[33] J. ODonovan, B. Kang, G. Meyer, T. Höllerer, S. Adalii, Credibility in context: An analysis of feature distributions in twitter, in: 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing, 2012, pp. 293–301. doi:10.1109/SocialCom-PASSAT.2012.128.
[34] S. Sikdar, B. Kang, J. ODonovan, T. Höllerer, S. Adah, Understanding information credibility on twitter, in: 2013 International Conference on Social Computing, 2013, pp. 19–24. doi:10.1109/SocialCom.2013.9.
[35] J. Lin, R. Nogueira, A. Yates, Pretrained transformers for text ranking: Bert and beyond, Synthesis Lectures on Human Language Technologies 14 (2021) 1–325.
[36] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, arXiv:1910.10683 (2019).
[37] R. Nogueira, Z. Jiang, J. Lin, Document ranking with a pretrained sequence-to-sequence model, arXiv:2003.06713 (2020).
[38] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, C. Potts, Learning word vectors for sentiment analysis, in: Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human language technologies, 2011, pp. 142–150.
[39] S. Tahvili, L. Hatvani, E. Ramentol, R. Pimentel, W. Afzal, F. Herrera, A novel methodology to classify test cases using natural language processing and imbalanced learning, Engineering Applications of Artificial Intelligence 95 (2020) 103878.
[40] D. Valcarce, A. Landin, J. Parapar, Á. Barreiro, Collaborative filtering embeddings for memory-based recommender systems, Engineering Applications of Artificial Intelligence 85 (2019) 347–356.

[41] C. Porcel, A. Ching-López, G. Lefranc, V. Loia, E. Herrera-Viedma, Sharing notes: An academic social network based on a personalized fuzzy linguistic recommender system, Engineering Applications of Artificial Intelligence 75 (2018) 1–10.
[42] J. M. Chenlo, J. Parapar, D. E. Losada, J. Santos, Finding a needle in the blogosphere: An information fusion approach for blog distillation search, Information Fusion 23 (2015) 58–68.
[43] H. Wang, L. Jiang, Q. Zhao, H. Li, K. Yan, Y. Yang, S. Li, Y. Zhang, L. Qiao, C. Fu, et al., Progressive structure network-based multiscale feature fusion for object detection in real-time application, Engineering Applications of Artificial Intelligence 106 (2021) 104486.
[44] E. B. Varghese, S. M. Thampi, A multimodal deep fusion graph framework to detect social distancing violations and fcgs in pandemic surveillance, Engineering Applications of Artificial Intelligence 103 (2021) 104305.
[45] E. A. Fox, J. A. Shaw, Combination of multiple searches, NIST special publication SP 243 (1994) 243–252.
[46] J. De Borda, Mémoire sur les élections au scrutin, Histoire de l'Academie Royale des Sciences pour 1781 (Paris, 1784) (1784).
[47] J. A. Aslam, M. Montague, Models for metasearch, in: Proceedings of the 24th Int. ACM SIGIR Conference on Research and Development in Information Retrieval, 2001, pp. 276–284.
[48] T.-Y. Liu, Learning to rank for information retrieval, Springer Science & Business Media, 2011.
[49] R. Benham, J. S. Culpepper, Risk-reward trade-offs in rank fusion, in: Proceedings of the 22nd Australasian Document Computing Symposium, 2017, pp. 1–8.
[50] R. Pradeep, X. Ma, X. Zhang, H. Cui, R. Xu, R. Nogueira, J. Lin, H2oloo at TREC 2020: When all you got is a hammer... deep learning, health misinformation, and precision medicine, in: Proceedings of the 29th Text REtrieval Conference (TREC), 2020.
[51] J. Bevendorff, M. Völske, B. Stein, A. Bondarenko, M. Fröbe, S. Günther, M. Hagen, Webis at TREC 2020: Health Misinformation Track, in: Proceedings of the 29th Text REtrieval Conference (TREC), 2020.
[52] J. Bevendorff, B. Stein, M. Hagen, M. Potthast, Elastic chatnoir: Search engine for the clueweb and the common crawl, in: European Conference on Information Retrieval, Springer, 2018, pp. 820–824.
[53] L. C. Lima, D. B. Wright, I. Augenstein, M. Maistro, University of Copenhagen participation in TREC Health Misinformation Track 2020, arXiv:2103.02462 (2021).
[54] N. Asadi, J. Lin, Document vector representations for feature extraction in multi-stage document ranking, Information Retrieval 16 (2013) 747–768.
[55] N. Asadi, J. Lin, Effectiveness/efficiency tradeoffs for candidate generation in multi-stage retrieval architectures, in: Proceedings of the 36th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, 2013, pp. 997–1000.
[56] J. S. Culpepper, C. L. Clarke, J. Lin, Dynamic cutoff prediction in multi-stage retrieval systems, in: Proceedings of the 21st Australasian Document Computing Symposium, 2016, pp. 17–24.
[57] P. Yang, H. Fang, J. Lin, Anserini: Enabling the use of lucene for information retrieval research, in: Proceedings of the 40th Int. ACM SIGIR Conference on Research and Development in Information Retrieval, 2017, pp. 1253–1256.
[58] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, M. Gatford, et al., Okapi at TREC-3, NIST Special Publication Sp 109 (1995) 109.
[59] C. Kamphuis, A. P. de Vries, L. Boytsov, J. Lin, Which BM25 do you mean? a large-scale reproducibility study of scoring variants, in: European Conference on Information Retrieval, Springer, 2020, pp. 28–34.
[60] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, arXiv:1706.03762 (2017).
[61] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, arXiv:1810.04805 (2018).
[62] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, L. Deng, MS MARCO: A human generated machine reading comprehension dataset, in: CoCo@ NIPS, 2016.
[63] R. Pradeep, R. Nogueira, J. Lin, The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models, arXiv:2101.05667 (2021).
[64] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese bert-networks, arXiv:1908.10084 (2019).
[65] P. Gamallo., M. Corral., M. Garcia., Comparing dependency-based compositional models with contextualized word embeddings, in: Proceedings of the 13th Int. Conf. on Agents and Artificial Intelligence (ICAART) - Volume 2, SciTePress, 2021, pp. 1258–1265.
[66] W. Chu, Z. Ghahramani, C. K. Williams, Gaussian processes for ordinal regression., Journal of Machine Learning Research 6 (2005) 1019–1041.
[67] W. Chu, Z. Ghahramani, Preference learning with gaussian processes, in: Proceedings of the 22nd Int. Conf. on Machine Learning, 2005, pp. 137–144.
[68] W. Chu, S. S. Keerthi, New approaches to support vector ordinal regression, in: Proceedings of the 22nd Int. Conf. on Machine Learning, 2005, pp. 145–152.
[69] B. Bartell, G. W. Cottrell, R. Belew, Learning to retrieve information, in: Proceedings of the Swedish Conference on Connectionism, 1995, p. 27.
[70] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, G. Hullender, Learning to rank using gradient descent, in: Proceedings of the 22nd Int. Conf. on Machine learning, 2005, pp. 89–96.
[71] Y. Cao, J. Xu, T.-Y. Liu, H. Li, Y. Huang, H.-W. Hon, Adapting ranking svm to document retrieval, in: Proceedings of the 29th Int. ACM SIGIR Conference on Research and Development in Information Retrieval, 2006, pp. 186–193.
[72] E. Agichtein, E. Brill, S. Dumais, R. Ragno, Learning user interaction models for predicting web search result preferences, in: Proceedings of the 29th Int. ACM SIGIR Conference on Research and Development in Information Retrieval, 2006, pp. 3–10.
[73] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, H. Li, Learning to rank: from pairwise approach to listwise approach, in: Proceedings of the 24th Int. Conf. on Machine Learning, 2007, pp. 129–136.
[74] T. Qin, T.-Y. Liu, M.-F. Tsai, X.-D. Zhang, H. Li, Learning to search web pages with query-level loss functions, Technical Report 156 (2006) 28.
[75] W. B. Croft, D. Metzler, T. Strohman, Search engines: Information retrieval in practice, volume 520, Addison-Wesley Reading, 2010.

[76] C. Lioma, J. G. Simonsen, B. Larsen, 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.
[77] C. L. Clarke, M. D. Smucker, A. Vtyurina, Offline evaluation by maximum similarity to an ideal ranking, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 225–234.
[78] W. Webber, A. Moffat, J. Zobel, A similarity measure for indefinite rankings, ACM Transactions on Information Systems (TOIS) 28 (2010) 1–38.
[79] J. Parapar, D. E. Losada, Á. Barreiro, Testing the tests: simulation of rankings to compare statistical significance tests in information retrieval evaluation, in: Proceedings of the 36th Annual ACM Symposium on Applied Computing, 2021, pp. 655–664.
[80] J. Parapar, D. E. Losada, M. A. Presedo-Quindimil, A. Barreiro, Using score distributions to compare statistical significance tests for information retrieval evaluation, Journal of the Association for Information Science and Technology 71 (2020) 98–113.
[81] Y. Zhang, H. Zhou, Z. Li, Fast and accurate neural crf constituency parsing, arXiv:2008.03736 (2020).
[82] S. R. Bowman, G. Angeli, C. Potts, C. D. Manning, A large annotated corpus for learning natural language inference, arXiv:1508.05326 (2015).
[83] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, L. Specia, Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation, arXiv:1708.00055 (2017).
[84] O. Chapelle, Y. Chang, Yahoo! learning to rank challenge overview, in: Proceedings of the Learning to Rank Challenge, PMLR, 2011, pp. 1–24.
[85] A. Arampatzis, J. Kamps, S. Robertson, Where to stop reading a ranked list? Threshold optimization using truncated score distributions, in: Proceedings of the 32nd Int. ACM SIGIR Conference on Research and Development in Information Retrieval, 2009, pp. 524–531.
[86] G. Wenzek, M.-A. Lachaux, A. Conneau, V. Chaudhary, F. Guzmán, A. Joulin, E. Grave, CCNet: Extracting high quality monolingual datasets from web crawl data, arXiv:1911.00359 (2019).
[87] A. E. Ezugwu, A. M. Ikotun, O. O. Oyelade, L. Abualigah, J. O. Agushaka, C. I. Eke, A. A. Akinyelu, A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects, Engineering Applications of Artificial Intelligence 110 (2022) 104743.
[88] R. Sicilia, M. Merone, R. Valenti, P. Soda, Rule-based space characterization for rumour detection in health, Engineering Applications of Artificial Intelligence 105 (2021) 104389.
[89] A. Hussain, E. Cambria, S. Poria, A. Hawalah, F. Herrera, Information fusion for affective computing and sentiment analysis, Information Fusion 71 (2021) 97–98.
[90] N. Burkart, M. F. Huber, A survey on the explainability of supervised machine learning, Journal of Artificial Intelligence Research 70 (2021) 245–317.