ABSTRACT
The TREC Health Misinformation track focuses on discerning reliable from unreliable information and correct from incorrect information. This problem is very common in Web Search results and it is especially critical when it is related to health content [1]. This year’s task focuses on COVID-19 and SARS-CoV-2 misinformation. In our experiments, we applied a BM25 retrieval baseline as a first step. Afterwards, we used a reliability classifier recently developed by our team [2]. Finally, we also experimented with BERT-based variants that attempt to estimate similarity between sentences.

CCS CONCEPTS
• misinformation, COVID-19, reliability;

1 INTRODUCTION
Search engines represent a powerful tool for end-users to find information related to different topics easily and quickly. However, the results offered can be unreliable, inaccurate, or highly technical [3]. This phenomenon is called misinformation and it is very common within huge data collections, such as Web Search results [4].

Misinformation can have a greater or lesser impact depending on the topic, but it is especially sensitive when it comes to health-related content, as Pogacar et al. [5] proved in their user study. Medical hoaxes, miracle diets, or advice provided by unqualified people abound in all digital media [6]. These contents can be highly dangerous if taken as true and applied without professional medical supervision [7]. This has become particularly evident in the context of the pandemic we are facing in 2020, with substantial information about COVID-19 being either dubious or of poor quality [8].

The TREC 2020 Health Misinformation Track focuses on misinformation related to COVID-19 and SARS-CoV-2. Our understanding of this disease is constantly evolving, so tracking objective information should be based on developing a retrieval system able to return scientific accurate documents.

In this report, we explain the characteristics of the runs submitted by our team, CiTIUS, for the TREC 2020 Health Misinformation Track, and discuss our results. Our runs represent an exploratory approach to leverage existing labelled data to build a reliability classifier [2] adapted to the TREC Health Misinformation tasks.

2 DOCUMENTS AND TOPICS
In the TREC 2020 Health Misinformation Track, a news corpus from January 2020 to April 2020 was provided. The documents were obtained from CommonCrawl News, which contains news articles from all over the world.

Topics try to model how people search for health advice online. Fifty topics with a fixed structure were provided. All include number, title, description, answer, evidence, and narrative, as it can be seen in Figure 1. The title field has the form of a pair of treatment and disease, where the disease is always COVID-19. The description is formulated as a question, which contains treatment, effect, and disease. The answer corresponds to the medical consensus at the time of topic creation. Finally, the remaining fields were not intended to be used by the systems, but only by human assessors to produce qrels.

3 RETRIEVAL BASELINE
For indexing and processing the collection, we considered different state-of-the-art tools, such as Terrier [9] or Lucene [10]. However, we decided to use Anserini [11], which uses Lucene underneath, and it was highly practical to support the needs of this track.

For all the runs, the title field was used to produce the search query. We decided to use a bag-of-words approach, where at least one term or clause must match for a document to appear in the results. We selected a classical BM25 [12] approach, setting normalization parameter \(b\) to 0.75 and TF weight upper-bounded limit \(k_1\) to 1.2. The first retrieval baseline was generated using Pyserini1, Anserini’s Python implementation. This facilitated the integration with the rest of the elements in our technology (our reliability classifier is also developed in Python).

The baseline was combined with other techniques, such as BERT sentence-similarity or our reliability classifier in order to produce a final estimation of the presence of misinformation.

---

1Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

1https://github.com/castorini/pyserini
Figure 1: A TREC 2020 Health Misinformation Track topic (Topic 13).

4 RELIABILITY CLASSIFIER

In previous research [2], we developed a predictive technology able to distinguish among reliable and unreliable web documents, based on Natural Language Processing (NLP) and Machine Learning techniques. To that end, three different Web Search datasets were used: Sondhi et al. [1], Schwarz et al. [13], and CLEF eHealth consumer health task 2018 [14]. All of them contain web pages related to the field of health, but the second dataset also has pages belonging to the topics of politics, finance, environment, and news about famous people.

Although the Schwarz et al. [13] and CLEF eHealth [14] collections were labelled in terms of credibility and trustworthiness, respectively, we considered these concepts as proxies of reliability for our experiments.

Our main goal was to build a document-level classifier using a standard supervised learning approach. More specifically, we followed the methodology designed in [1]. In this previous work, webpages were represented following a number of features, namely:

- **Link-based features**: the number and type of links are usually a good indicator of the type of website we are dealing with. For example, as Sondhi and his colleagues exposed, a more reputable or reliable site tends to have more internal links, while a less reliable site tends to have more external links and advertisements. On the other hand, the presence or not of privacy or contact links can be an indicator of reliability. This is because the presence of these types of elements gives a sense of confidence to the user who consults the resource. However, nowadays most unreliable sites replicate these characteristics with great success.

Based on these criteria five features were defined to be taken into account: normalized value of internal links, normalized value of external links, normalized value of total links, the presence or not of contact link (boolean), and the presence or not of privacy link (boolean). For the latter two, the original paper did not explain how they had been computed. Therefore, we manually defined two list of privacy\(^2\) and contact\(^3\) expressions, such as Privacy Policy or Contact Us, after performing a first exploratory analysis over the documents. For normalization, a random sample of documents was analysed and a large normalization denominator was chosen (the link count was divided by \(Z_1\), which was set to 200).

- **Commercial features**: the presence of commercial interest often indicates a low reputation. Therefore, two characteristics have been defined to be taken into account: the normalized value of commercial links and the normalized frequency of commercial words on the website.

For the latter, an initial list of indicative words of commercial interest was proposed in the article. Our contribution was to manually extend it by adding more words\(^4\). Since the original article was not explicit about word preprocessing, we followed a naive approach, in which a word must match exactly with some word in the list to be taken into account for the final metric. This strategy can be improved in future versions by applying lemmatization techniques, for example. Regarding normalization, the normalized value of commercial links was obtained dividing by the same \(Z_1\) used above. The second feature consisted of dividing the number of commercial words found by the document’s length.

- **Word-based features**: textual content and style are often good indicators of a website’s reliability or reputation. Therefore, each word in a document was considered as a different dimension, taking its normalized frequency score. Since the authors did not declare the use of any preprocessing stage, we applied no stemming or lemmatization.

Similarly, we considered two alternative pre-processing strategies, with and without stopword removal. To this aim, the NLTK\(^5\) English stoplist was manually extended\(^6\) after some preliminary exploration over the documents. Finally, for each word we divided the number of occurrences of the word by the document’s length.

Besides testing the feature sets in isolation, we also tested a final combination that merged all features together. Moreover, we tested two variants of it: one with word features extracted with stopword removal and another one with word features extracted with no stopword removal.

When carrying out the experimentation, a vector support machine was used as a learning method. More specifically, we employed Python’s implementation of SVMlight\(^7\). To compare the different feature sets, we used a weighted accuracy metric and, in case of a tie, the F1-metric of the minority (non-reliable) class was given priority.

To determine the best reliability detection features, a stratified 5-fold cross validation strategy was used with each dataset (except for the Schwarz et al. collection, which is very small and, thus, we used a 2-fold cross validation).

\(\footnote{https://github.com/MarcosFP97/Health-Rel/blob/master/lexicon/privacy.txt} \footnote{https://github.com/MarcosFP97/Health-Rel/blob/master/lexicon/contact.txt} \footnote{https://github.com/MarcosFP97/Health-Rel/blob/master/lexicon/comm_list.txt} \footnote{https://www.nltk.org/nltk_data/} \footnote{https://github.com/MarcosFP97/Health-Rel/blob/master/lexicon/stopwords.txt} \footnote{https://bitbucket.org/wcauchois/pysvmlight}
Experiments with all three datasets suggested that the best reliability detection models were those based on word features or based on combining all features together. Keeping or not stopwords had a slight impact on performance. However, this impact varied among each dataset. More details about the metrics used and the experiments can be found in [2].

Finally, we built a model for each collection and its best feature combination. Given a test document, Equation 1 determined its final class, where \( \text{pred\_CLEF} \), \( \text{pred\_Sondhi} \) and \( \text{pred\_Schwarz} \) are each model’s prediction for the test document and the weights were set to the relative size of these three training collections:

\[
\text{Rel}(\text{doc}) = 0.97 \times \text{pred\_CLEF} + 0.027 \times \text{pred\_Sondhi} + 0.006 \times \text{pred\_Schwarz}
\]  

\( (1) \)

5 SUBMITTED RUNS

5.1 Total Recall Task

In this task, the main goal was to retrieve documents that promulgated misinformation. To that end, documents contradicting the topic’s answers were assumed to be misinformation. We submitted three different runs or solutions to this problem.

The first run (CiTIUSCrdTot) applied the BM25 retrieval baseline described before. After that, we ranked the \( n \) retrieved documents based on our reliability classifier’s output (ranked by increasing reliability) and we kept the top ten thousand non-reliable documents in the ranking. We are aware of this being a naive method (it ignores the matching between the description field and the retrieved pages, and just estimates misinformation based on the reliability of the entire page). In any case, we thought it was a natural baseline against which more sophisticated baselines could be tested.

The second run (CiTIUSCrdRelTot) applied the same strategy, but it also used a voting method, Borda Count [15], to combine both rankings, relevance and reliability, and kept the top ranking documents.

The last run (CiTIUSSimTot) was the most sophisticated variant. A hand-crafted expression was created for each topic by combining description and answer fields. An example could be \( \text{Vitamin D does not cure COVID-19} \), but it also used a voting method, Borda Count [15], to combine both rankings, relevance and reliability, and kept the top ranking documents.

The first one (CiTIUSCrdAdh) applied the BM25 retrieval baseline described before. After that, we ranked the \( n \) retrieved documents based on our reliability classifier’s output but, in this case, we promoted highly reliable sites (the top ten thousand documents were kept from a ranking of documents organized by decreasing reliability).

The second run (CiTIUSCrdRelAdh) applied the same strategy, but it also used a voting method, Borda Count [15], to combine both rankings, relevance and reliability, and kept the top ranking documents.

The third run (CiTIUSSimAdh) consisted of producing a hand-crafted expression for each topic by combining description and answer fields. An example could be \( \text{Vitamin D does not cure COVID-19} \), since we are looking to promulgate misinformation correctly and relevant information. After obtaining the title-based BM25 baseline, we ranked the \( n \) retrieved documents based on maximum sentence similarity between the new hand-crafted expression and all sentences in each document. As in the previous task, we used the Sentence Transformers library and cosine similarity.

Finally, the last solution (CiTIUSSimRelAdh) applied a sentence-similarity strategy again. However, it also used Borda Count to combine both rankings, relevance and similarity.

6 RESULTS

6.1 Total Recall Task

The R-Precision results for the total recall task are shown in Table 1. All our methods performed worse than the median performance of the participants in the task. The classifier-based strategy (CiTIUSCrdTot) was the worst performer. It appears that this word-based document-level classification is too rough (and perhaps biased towards the topical words used in the training data). It must also be noted that the estimation of relevance combined with the reliability classifier (CiTIUSCrdRelTot) yields to better performance than the reliability classifier alone. This suggests that the relevance estimation should be kept as an integral part of the system.

The embedding-based approach (CiTIUSSimTot) worked better than the classifier-based strategy but we did not combine it with any relevance information (because we could only submit three official runs). We expect that the combination of CiTIUSSimTot with relevance information leads to further benefits in terms of performance.

6.2 AdHoc Retrieval Task

This task was focused on obtaining credible and correct information. To that end, the assessments were created based on the concepts of usefulness, correctness, and credibility.

---

\( ^6 \)https://github.com/UKPLab/sentence-transformers
Organizers designed specific measures to account for these aspects (e.g., CAM_MAP_three), but they also evaluated runs in terms of traditional relevance measures (e.g., NDCG). Our results are shown in Table 2.

Again, our basic strategies fared worse than the median participant. The CIITUSSimRelAdh run, which combined BERT-based similarity with the relevance ranking, produced our best results. The classifier-based variant was our worst performer.

### 7 FUTURE WORK

As the first next step, we intend to improve our classifier’s performance and its capacity for generalization. To that end, we want to train the datasets with BERT models [16], which have been proved to be successful for Natural Language Processing (NLP) tasks. Moreover, we also intend to perform transfer learning experiments for the sake of achieving generalisation [17, 18].

On the other hand, an interesting approach could be to determine the impact of this news in social media, and see if it exists a correlation between reliable information and its presence on this kind of media.

Finally, we also want to try some sentence retrieval techniques to extract on-topic information from larger documents. This might help to improve performance by removing noise.

### 8 CONCLUSIONS

The TREC 2020 Health Misinformation Track focused on COVID-19 misinformation. To solve this problem, we presented different simple strategies.

We developed a reliability classifier using a previously annotated Web Search dataset. However, this strategy generalized poorly when applied to TREC data.

On the other hand, we proposed a naive sentence similarity solution based on BERT. This solution seems to perform better, but it is still too simple.

Finally, it must be noticed that combining relevance output to any of the previous strategies improves the final performance.

### ACKNOWLEDGMENTS

This work was funded by FEDER/Ministerio de Ciencia, Innovación y Universidades – Agencia Estatal de Investigación/ Project (RTI2018-093336-B-C21). This work has received financial support from the Consellería de Educación, Universidade e Formación Profesional (accreditation 2019-2022 ED431G-2019/84, ED431C 2018/29, ED431C 2018/19) and the European Regional Development Fund (ERDF), 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


