CiTIUS at the TREC 2020 Health Misinformation Track

Marcos Fernández-Pichel
marcosfernandez.pichel@usc.es
Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela
15782 Santiago de Compostela, Spain

David E. Losada
david.losada@usc.es
Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela
15782 Santiago de Compostela, Spain

Juan C. Pichel
juancarlos.pichel@usc.es
Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela
15782 Santiago de Compostela, Spain

David Elsweiler
david@elsweiler.co.uk
Chair for Information Science, University of Regensburg
Regensburg, Germany

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 document-level 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 [3]. The results provided can, however, often be unreliable [4], inaccurate [5], or of poor quality [6]. Results with any of these attributes can be referred to as examples of **misinformation**.

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. [7] proved in their user study. Medical hoaxes, miracle diets, or advice provided by unqualified people abound in all digital media [8]. These contents can be highly dangerous if taken as true and applied without professional medical supervision [9]. 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 [10].

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] and to test it with TREC Health Misinformation data.

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 attempt 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 [11] or Lucene [12]. However, we decided to use **Anserini** [13], which is Lucene-based, but offers practical advantages 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** [14] approach, setting normalisation parameter (b) to 0.75 and TF weight upper-bounded limit (k_1) to 1.2. The first retrieval baseline was generated using Pyserini[1], 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.

[1] https://github.com/castorini/pyserini

```
<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: 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. [15], and CLEF eHealth consumer health task 2018 [16]. Each of these collections contains web pages related to the field of health, but the second dataset additionally features pages related to politics, finance, environment, and news about famous people.

Although the Schwarz et al. [15] and CLEF eHealth [16] 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 [17, 18]. 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 [19]. 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 [20, 21]. However, nowadays most unreliable sites replicate these characteristics with great success.
  Based on these criteria five features were defined to be taken into account: normalised value of internal links, normalised value of external links, normalised 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 normalisation, the original authors analysed a random sample of documents and they experimentally chose a large normalisation denominator (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 [17, 19]. Therefore, two characteristics have been defined to be taken into account: the normalised value of commercial links and the normalised 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 normalisation, the normalised 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 length.
- **Word-based features**: textual content and style are often good indicators of a website's reliability or reputation [22, 23]. Therefore, each word in a document was considered as a different dimension, taking its normalised frequency score. Since the original 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 length.

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

When performing the experiments, 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).

[2] https://github.com/MarcosFP97/Health-Rel/blob/master/lexicon/privacy.txt
[3] https://github.com/MarcosFP97/Health-Rel/blob/master/lexicon/contact.txt
[4] https://github.com/MarcosFP97/Health-Rel/blob/master/lexicon/comm_list.txt
[5] https://www.nltk.org/nltk_data/
[6] https://github.com/MarcosFP97/Health-Rel/blob/master/lexicon/stopwords.txt
[7] 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 *pred_CLEF*, *pred_Sondhi* and *pred_Schwarz* are each model's prediction for the test document and the weights were set to the relative size of these three training collecions:

$$Rel(doc) = 0,97 \times pred\_CLEF + 0,027 \times pred\_Sondhi + 0,006 \times pred\_Schwarz \quad (1)$$

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 one (**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 [24], 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 *Vitamin D cures COVID-19*, since we are looking to promulgate misinformation. 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 (where sentences were represented using BERT). To this aim, we used Sentence Transformers[8] Python library, which offers several pre-trained models for embeddings generation, and then we applied cosine similarity between sentences.

5.2 AdHoc Retrieval Task

Unlike the previous task, here the main goal was to recover **correct information**. To that end, sites supporting the topic's answers were assumed to be relevant. We submitted four different runs or solutions to this problem.

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 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 [24], 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 *Vitamin D does not cure COVID-19*, since we are looking to promulgate correct 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

Runs	Rprec
CiTIUSCrdTot	0.0105
CiTIUSCrdRelTot	**0.0354**
CiTIUSSimTot	0.0332
Median	0.0976

Table 1: Our results for the 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 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*.

[8] https://github.com/UKPLab/sentence-transformers

Runs	CAM_MAP_three	NDCG (us, co, cr)	Compatibility (harmful-only)	Compatibility (helpful-only)
CiTIUSCrdAdh	0.0037	0.0412	0.0082	0.0586
CiTIUSCrdRelAdh	0.0355	0.1393	0.0475	0.1721
CiTIUSSimAdh	0.0252	0.1212	0.0351	0.1207
CiTIUSSimRelAdh	**0.0793**	**0.2353**	**0.0600**	**0.2376**
Median	0.1389	0.3308	0.0747	0.337

Table 2: Our results for the AdHoc Retrieval Task.

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 **CiTIUSSimRelAdh** run, which combined BERT-based similarity with the relevance ranking, produced our best results. The classifier-based variant was our worst performer.

7 CONCLUSIONS

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

We developed a document-level reliability classifier using previously annotated Web Search datasets. However, this strategy generalized poorly when applied to TREC data. We additionally proposed a naive sentence similarity solution based on BERT. This solution seems to perform better, but it is nevertheless still too simple.

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

8 FUTURE WORK

As the first next step, we intend to try some **passage retrieval** techniques to extract on-topic information from larger documents. This might help to improve performance by removing noise. Afterward, several **sources of evidence** can be combined to better detect misinformation (objectivity classifiers, fact-checkers, readability estimators, etc.)

Another 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.

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-2019/04, 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

[1] Parikshit Sondhi, V. G. Vinod Vydiswaran, and ChengXiang Zhai. Reliability Prediction of Webpages in the Medical Domain. In Ricardo Baeza-Yates, Arjen P. de Vries, Hugo Zaragoza, B. Barla Cambazoglu, Vanessa Murdock, Ronny Lempel, and Fabrizio Silvestri, editors, *Advances in Information Retrieval*, pages 219–231, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[2] Marcos Fernández-Pichel, David Losada, Juan C. Pichel, and David Elsweiler. Reliability Prediction for Health-related Content: A Replicability Study. In *European Conference on Information Retrieval*, Lucca, Tuscany, Italy, 2021. Springer.

[3] Susannah Fox. *Health topics: 80% of internet users look for health information online.* Pew Internet & American Life Project, 2011.

[4] Mustafa Abualsaud and Mark D Smucker. Exposure and order effects of misinformation on health search decisions. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Rome, 2019.

[5] Gunther Eysenbach. Infodemiology: The epidemiology of (mis) information. *The American journal of medicine*, 113(9):763–765, 2002.

[6] Soo Young Rieh. Judgment of information quality and cognitive authority in the web. *Journal of the American society for information science and technology*, 53(2):145–161, 2002.

[7] Frances A Pogacar, Amira Ghenai, Mark D Smucker, and Charles LA 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 International Conference on Theory of Information Retrieval*, pages 209–216, 2017.

[8] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(3):1–42, 2019.

[9] BBC. Alcohol-related deaths increasing in the United States. https://www.nih.gov/news-events/news-releases/alcohol-related-deaths-increasing-united-states, 2020. [Online; accessed 05-October-2020].

[10] Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G Lu, and David G Rand. Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science*, 31(7):770–780, 2020.

[11] Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Douglas Johnson. Terrier information retrieval platform. In *European Conference on Information Retrieval*, pages 517–519. Springer, 2005.

[12] Michael McCandless, Erik Hatcher, Otis Gospodnetić, and O Gospodnetić. *Lucene in action*, volume 2. Manning Greenwich, 2010.

[13] Peilin Yang, Hui Fang, and Jimmy Lin. Anserini: Enabling the use of lucene for information retrieval research. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 1253–1256, New York, NY, USA, 2017. Association for Computing Machinery.

[14] Stephen Robertson and Hugo Zaragoza. *The probabilistic relevance framework: BM25 and beyond.* Now Publishers Inc, 2009.

[15] Julia Schwarz and Meredith Morris. Augmenting Web Pages and Search Results to Support Credibility Assessment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, page 1245–1254, New York, NY, USA, 2011. Association for Computing Machinery.

[16] Hanna Suominen, Liadh Kelly, Lorraine Goeuriot, Aurélie Névéol, Lionel Ramadier, Aude Robert, Evangelos Kanoulas, Rene Spijker, Leif Azzopardi, Dan Li, Jimmy, João Palotti, and Guido Zuccon. Overview of the CLEF ehealth Evaluation Lab 2018. In Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian Yun Nie, Laure Soulier, Eric SanJuan, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 286–301, Cham, 2018. Springer International Publishing.

[17] Luca Becchetti, Carlos Castillo, Debora Donato, Ricardo Baeza-Yates, and Stefano Leonardi. Link analysis for web spam detection. *ACM Transactions on the Web (TWEB)*, 2(1):1–42, 2008.

[18] Allan Borodin, Gareth O Roberts, Jeffrey S Rosenthal, and Panayiotis Tsaparas. Link analysis ranking: algorithms, theory, and experiments. *ACM Transactions on Internet Technology (TOIT)*, 5(1):231–297, 2005.

[19] Wei Zha and H Denis Wu. The impact of online disruptive ads on users' comprehension, evaluation of site credibility, and sentiment of intrusiveness. *American*

Communication Journal, 16(2), 2014.

[20] Alexandru L. Ginsca, Adrian Popescu, and Mihai Lupu. Credibility in information retrieval. *Found. Trends Inf. Retr.*, 9(5):355–475, December 2015.

[21] Michal Kakol, Radoslaw Nielek, and Adam Wierzbicki. Understanding and predicting web content credibility using the content credibility corpus. *Information Processing & Management*, 53(5):1043–1061, 2017.

[22] David Matsumoto, Hyisung C Hwang, and Vincent A Sandoval. Cross-language applicability of linguistic features associated with veracity and deception. *Journal of Police and Criminal Psychology*, 30(4):229–241, 2015.

[23] Subhabrata Mukherjee and Gerhard Weikum. Leveraging joint interactions for credibility analysis in news communities. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 353–362, 2015.

[24] David E Losada, Javier Parapar, and Alvaro Barreiro. A rank fusion approach based on score distributions for prioritizing relevance assessments in information retrieval evaluation. *Information Fusion*, 39:56–71, 2018.