CiTIUS at the TREC 2022 Health Misinformation Track

Marcos Fernández-Pichel, Manuel Prada-Corral, David E. Losada and Juan C. Pichel
Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela
15782 Santiago de Compostela, Spain
{marcosfernandez.pichel,manuel.deprada.corral,david.losada,juancarlos.pichel}@usc.es

ABSTRACT

The TREC Health Misinformation Track fosters the development of retrieval methods that promote credible and correct information over misinformation for health-related decision tasks. To make the procedure more realistic, an Answer Prediction challenge was added to this year's track. In these working notes, we describe our endeavours to estimate the correct response to each topic using search engine results and Transformer models. On the other hand, our adhoc retrieval solutions are based on new addons to our pipeline and the weighted fusion of signals.

CCS CONCEPTS

• **Health misinformation, Fusion, SE, GPT-3**;

1 INTRODUCTION

Search engines are frequently used to locate health information online [1]. However, there is a lot of false material on the internet concerning illnesses and remedies [2]. It has been proven that engaging with inaccurate search results causes people to make bad judgments regarding their health [3].

The TREC Health Misinformation Track endeavours to develop retrieval techniques that favour accurate and reliable information over misinformation for needs relating to health-related information. Last year our team, affiliated to CiTIUS at the University of Santiago de Compostela (Spain), presented a complete multistage retrieval system for addressing this task [4, 5]. This software is available to the community[1].

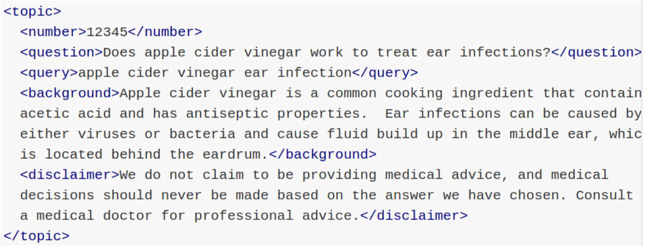
In 2022, we continued exploiting the same architecture as the main basis to approach this problem. However, we tried to go one step further by including new signals to identify misinformation and merging them in a weighted manner. For instance, a credibility estimator and a text readability classifier were added.

The second challenge of this track, topic answer prediction, was addressed in two alternative ways: using a search engine's top results to estimate the correct answer to a query, or prompting a language model (GPT-3) to obtain an estimate of the correct answer.

These working notes are organised as follows: Section 2 briefly presents the data, Section 3 introduces the new answer prediction task, Section 4 presents the traditional search task, Section 5 reports the obtained results, and, finally, Sections 6 and 7 describe how we updated our system architecture and expose some conclusions.

2 DATASET

The organisers of the track opted for the no-clean version of the C4 dataset. This corpus was created by Google to train their sequence-to-sequence T5 model [6]. The collection is formed of text extracts from the April 2019 snapshot of Common Crawl[2], and it contains approximately 1 billion English documents.



```
<topic>
  <number>12345</number>
  <question>Does apple cider vinegar work to treat ear infections?</question>
  <query>apple cider vinegar ear infection</query>
  <background>Apple cider vinegar is a common cooking ingredient that contain
  acetic acid and has antiseptic properties.  Ear infections can be caused by
  either viruses or bacteria and cause fluid build up in the middle ear, whic
  is located behind the eardrum.</background>
  <disclaimer>We do not claim to be providing medical advice, and medical
  decisions should never be made based on the answer we have chosen. Consult
  a medical doctor for professional advice.</disclaimer>
</topic>
```

Figure 1: A TREC 2022 Health Misinformation Track topic.

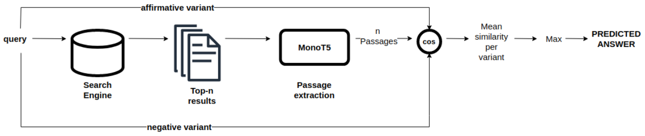


Figure 2: Answer prediction strategy based on search engine's output.

Each topic provided by the organisers consists of a health-related query. The topic represents a user trying to determine whether or not a treatment is useful for a given disease or condition. All topics have a fixed structure (see Figure 1). The main difference with previous years is that the correct answer is not provided.

3 ANSWER PREDICTION

In 2022, the first subtask consisted of predicting the correct answer for a given topic. To that end, we implemented two different strategies, namely: distilling knowledge from the top results of a search engine or prompting a GPT-3 language model.

3.1 Prediction based on Search Engine's output

We wanted to test here the ability of a search engine to retrieve webpages that supply correct answers to the search topics. For that purpose we implemented a Python program that runs queries against Google's API. The queries were obtained from the question field of the TREC topic (no modification was done to this textual question).

For each query, the top n results were obtained and, for each result, the most relevant passage was extracted. Passage relevance was estimated using MonoT5 [7]. Next, a negative and an affirmative variant of the query were generated with a text-oriented parser [8]. The mean similarity between each variant and the top n passages was computed. The variant (affirmative or negative) yielding the

[1] https://github.com/MarcosFP97/Multistage-Retrieval-System

[2] https://commoncrawl.org/

highest average similarity is the one that determines the answer (see Figure 2).

Utilising the TREC 2021 topics as a training set, we found that the most effective approach consists of making the estimation from a single document (the highest ranked webpage). Thus, the correct answer is finally estimated from the similarity between each variant and a single passage (extracted from the top 1 result).

3.2 Prompting GPT-3

The second strategy consisted of prompting GPT-3 with the topic question + "Yes/No?", i.e. "*Does apple cider vinegar work to treat ear infections? Yes/No?*". The language model produces an answer in natural language to the question. We implemented a simple program to process the GPT-3 output and interpret if, according to this transformer, the treatment is *helpful*, *unhelpful* or *inconclusive* (when the model has not enough knowledge to answer this question). To adjust this component, we made some tests with the TREC 2021 topics and the corresponding correct answers [9].

3.3 Runs

Our team submitted three different runs to this task:

- **citius.gpt-3:** utilises GPT-3 prompt to make the prediction. The confidence score was set to 1 for "*yes*" or "*no*" answers. For inconclusive predictions, the answer was set to "*no*" and confidence was set to 0.5. We decided to consider inconclusive cases as negative cases based on the observed answer distributions in previous years.
- **citius.se:** uses the search engine's results to determine the answer. The confidence score is obtained by a min-max scaling of the similarity value. In this case, there are no inconclusive predictions.
- **citius.se&gpt-3:** the answer is set as the conjunction of both predictions (GPT-3 and search engine). The confidence score is the multiplication of both confidence scores.

4 ADHOC RETRIEVAL

The main goal is to identify correct and credible information over misinformation. This year we explored the incorporation of new predictors into our system, and we also made a thorough exploration of estimates of correctness.

4.1 Credibility classifier

In our previous participations, the achieved helpful scores were fairly good but our harmful scores were too high. Our main goal in 2022 was thus to decrease the number of harmful documents retrieved. To this end, the first innovation introduced in our system was a credibility estimator.

Several classifiers, ranging from traditional to new deep-language (Deep-L) algorithms, were trained on the CLEF 2018 Consumer Health Search (CHS) Task dataset [10]. This built "trustworthiness" classifiers, which we adopted as a proxies of credibility estimators. Next, we compared the ability of these classifiers to distinguish credible and non-credible contents from the TREC 2021 Health Misinfo dataset. The best performer was a Random Forest (RF) classifier, which was taken as our reference credibility classifier for the 2022 task.

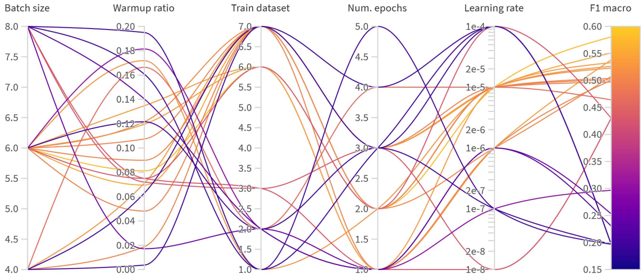


Figure 3: Hyper-parameter sweep for RoBERTa large.

4.2 Readability classifier

As shown in previous studies, the legibility of a text often correlates with its credibility [11]. Thus, a readability estimator was also incorporated as a new signal into our system. Again, several classifiers were tested against the 2021 data (Random Forest, Bert, Naïve-Bayes). The training step was done with the CLEF 2018 CHS Task dataset (and, more specifically, with the readability labels from this collection). The selected readability classifier was a BERT base model.

4.3 Correctness classifier

In previous years, we had focused our efforts on correctness. However, the obtained results and supplementary experiments suggested us that a supervised approach does not achieve a great performance, particularly with scarce training data [4, 5]. This year we decided to further explore some state-of-the-art Deep-L supervised models and tried to enhance their performance. Our work was oriented to correctness classification at passage-level and we explored more-in-depth hyper-parameter selection and included more data.

No information about the correct answer is available (unless it is estimated with some strategy like the one shown in Section 3) and, thus, the correctness classifier is fed with the unmodified question and each passage:

$$[query]\ [SEP]\ [passage] \longrightarrow [correctness\ label]$$

Using this template, the following models were tested: BERT base and large [12], RoBERTa base and large [13], DistilBERT [14], ALBERT [15], XLNet [16], MPNet [17], and MiniLM [18]. For each model, hyper-parameter selection was performed utilising *Sweeps* from the library *Weights & Biases*[3].

Two datasets were utilised for fine-tuning the models: the TREC 2019 Decision Track data [19] and the TREC 2020 Health Misinformation Track [20]. Next, the models were evaluated against a held out partition of the 2021 data. This partition is similar to the 2022 data (the same C4 collection was used for both years). It must also be noticed that no 2021 data was used for training since this dataset was later used for signal fusion (see Section 4.4).

The tool performed a Bayesian approach to optimise these parameters, utilising approximately 20 training executions per model. Finally, we took the model with the best performance over the validation set and calculated its F1 on the test set. Figure 4 shows the optimisation process for RoBERTa large.

[3] https://wandb.com

After setting the hyperparameters, MPNet, BERT large, RoBERTa large, and MiniLM were chosen for the last fusion step.

4.4 Signal fusion

At this point, we had incorporated three new evidences to our pipeline and we wanted to evaluate the performance of the weighted fusion of these signals and pre-existing signals (i.e. document relevante based on Bm25 and passage relevance based on MonoT5).

A 3-fold cross validation on the 2021 data was performed to determine the most promising combinations. Many possible signal combinations were tested (i.e. Bm25 + readability, Bm25 + MiniLM, MonoT5 + credibility, Bm25 + MonoT5 + credibility, etc.) and signals were always combined in a linear way.

For the most promising combinations, a grid search of weights was conducted to select the final parameters[4]. See Table 1 for an example of this optimisation. Note that MonoT5 gets the highest weight, which confirms the conclusions of previous studies [21].

Weights	**Signals**	**Compatibility (help-harm)**
0.99-0.01	MonoT5 question + cred.	0.0586
0.95-0.05	MonoT5 question + cred.	**0.0619**
0.90-0.10	MonoT5 question + cred.	0.0611
0.85-0.15	MonoT5 question + cred.	0.0595
0.80-0.20	MonoT5 question + cred.	0.0553
0.75-0.25	MonoT5 question + cred.	0.0507

Table 1: Grid search to set the weights of two signals.

4.5 Runs

The submitted runs for this task were the following:

- **citius.base**: this run consisted of an initial document level BM25 search followed by a MonoT5 passage re-ranking of the top 100 retrieved documents using the "question" field for each topic.
- **citius.r1**: this run consisted of an initial document level BM25 search followed by a passage re-ranking of the top-100 documents based on the weighted fusion of two scores: **0.95×MonoT5 + 0.05×credibility**.
- **citius.r2**: BM25 initial search followed by a passage re-ranking of the top-100 documents based on the weighted fusion of three scores: **0.95×MonoT5 + 0.025×credibility + 0.025×readability**.
- **citius.r3**: BM25 initial search followed by a MonoT5 passage re-ranking of the top 100 retrieved documents. The MonoT5 model is fed with a derived "correct" sentence from the original question field. The answer estimated as correct is produced from the **GPT-3 prompting** approach, see Section 3. For questions estimated as inconclusive the BM25 original ordering was kept.
- **citius.r4**: the same as citius.r3 but the MonoT5 ordering was maintained for those questions considered as inconclusive.
- **citius.r5**: BM25 initial search followed by a MonoT5 passage re-ranking of the top 100 retrieved documents using the derived correct sentence from the question field. In this case, the predicted correct answer is produced from the **SE** approach, see Section 3.
- **citius.r6**: the same as citius.r5, but to predict the answer **the conjunction of GPT-3 and SE** was used. For questions labeled as inconclusive by GPT-3, the MonoT5 ranking is kept.

Runs	Accuracy	AUC
citius.se	0.62	0.7072
citius.gpt-3	0.76	0.7672
citius.se_gpt	**0.80**	**0.8160**
Median	0.64	0.7072

Table 2: Our results for the Answer Prediction Task.

Runs	Help	Harm	Help - Harm
citius.base	0.2559	0.2148	0.0451
citius.r1	0.1836	0.1533	0.0303
citius.r2	0.1841	0.1457	0.0384
citius.r3	0.2427	0.1463	**0.0964**
citius.r4	0.2607	0.1775	0.0832
citius.r5	0.2579	0.2016	0.0563
citius.r6	0.2601	0.1801	0.0800
Median	0.2455	0.1465	0.0990

Table 3: Our results for the AdHoc Retrieval task.

Despite the effort dedicated to the experimentation with data from previous years and external collections, the correctness feature did not outperform others like credibility or readability. Thus, correctness was not included in the final submitted runs. This suggests that predicting the correctness of an excerpt based only on its content and a given query is still a difficult task, even for state-of-the-art models.

5 RESULTS

5.1 Answer Prediction Task

For this task, our best run is the one that combines the search engine's output and GPT-3 knowledge (see Table 2). This run is well above the median of all participants.

5.2 AdHoc Retrieval Task

For this task, our best run is the one that re-ranks MonoT5 passages based on GPT-3 predicted answers (see Table 3). Our main objective this year was to reduce the number of harmful documents that are retrieved. This was achieved by run citius.r3, which gets a helpful score similar to that of citius.base but a harmful score much lower. The difference of compatibilities (help-harm column) of our best run is close to the median of all participants.

The runs based on the fusion of signals (citius.r1, citius.r2) did not produce good results. This may indicate that the learned weights do

[4] This optimisation was conducted with the entire set of 2021 topics.

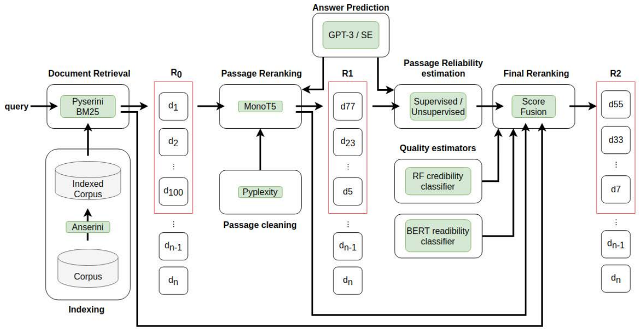


Figure 4: Updated version of the Multistage Retrieval System for Health Misinformation Detection.

not transfer well to the 2022 data. Comparing citius.r1 and citius.r2, it seems that including readability (citius.r2) is slightly beneficial, but this requires further analysis.

6 UPDATED MULTISTAGE RETRIEVAL SYSTEM

Our participation in TREC 2022 also allowed us to further develop our multistage retrieval system for health misinformation detection [5]. We have incorporated the most promising signals into this system. This tool represents the technological foundations for our research on online misinformation detection. The system[5] is available to the scientific community and it can be reused and expanded.

7 CONCLUSIONS AND FUTURE WORK

Our proposal was based on new estimators that might help to identify helpful documents and avoid harmful contents. We have conducted a thorough experimentation to determine the relative importance of each of these signals and to predict the correct answer of a given medical information need.

We can conclude that the most promising results were obtained in the answer prediction task, with estimates obtained from combining GPT-3 and a search engine's output. For the adhoc retrieval task, signal fusion did not work as expected. The best performer was a passage re-ranking strategy where passages are ranked based on their similarity to "correct" answers (and correct answers are produced from GPT-3 estimates and a textual parser).

The GPT-3 strategies are promising, as we managed to improve our base technology by reducing the number of harmful documents without significantly affecting the retrieval of helpful documents.

In the near future, we plan to further study these results and try to find innovative ways to reduce the retrieval of harmful contents. We are also trying to understand what makes a query more likely to retrieve harmful documents. If we succeed, we could define some sort of query-specific technique on the top of our current technology.

[5] https://github.com/MarcosFP97/Multistage-Retrieval-System

ACKNOWLEDGMENTS

The authors thank the support obtained from: i) 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 ii) Consellería de Educación, Universidade e Formación Profesional (accreditation 2019-2022 ED431G-2019/04, ED431C 2022/19) 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] Susannah Fox. *Health topics: 80% of internet users look for health information online*. Pew Internet & American Life Project, 2011.

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

[3] 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.

[4] Marcos Fernández-Pichel, Manuel Prada-Corral, David E Losada, Juan C Pichel, and Pablo Gamallo. CiTIUS at the TREC 2021 Health Misinformation Track. In *The Thirtieth Text REtrieval Conference Proceedings (TREC 2021), NIST Special Publication 500-335*, 2021.

[5] Marcos Fernández-Pichel, David E. Losada, and Juan C. Pichel. A multistage retrieval system for health-related misinformation detection. *Engineering Applications of Artificial Intelligence*, 115:105211, 2022.

[6] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*, 2019.

[7] Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. Document ranking with a pretrained sequence-to-sequence model. *arXiv preprint arXiv:2003.06713*, 2020.

[8] Yu Zhang, Houquan Zhou, and Zhenghua Li. Fast and accurate neural CRF constituency parsing. In *Proceedings of IJCAI*, pages 4046–4053, 2020.

[9] Charles LA Clarke, Saira Rizvi, Mark D Smucker, and Maria Maistro. Overview of the trec 2021 health misinformation track. In *TREC*, 2021.

[10] Jimmy Jimmy, Guido Zuccon, Joao Palotti, Lorraine Goeuriot, and Liadh Kelly. Overview of the clef 2018 consumer health search task. *CLEF 2018 Working Notes*, 2125, 2018.

[11] Alexandra Olteanu, Stanislav Peshterliev, Xin Liu, and Karl Aberer. Web credibility: Features exploration and credibility prediction. In *European conference on information retrieval*, pages 557–568. Springer, 2013.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pretraining of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[13] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[14] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[15] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2019.

[16] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. XLnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

[17] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867, 2020.

[18] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788, 2020.

[19] Mustafa Abualsaud, Christina Lioma, Maria Maistro, Mark D Smucker, and Guido Zuccon. Overview of the trec 2019 decision track. In *Proceedings of the Twenty-Eigth Text REtrieval Conference (TREC 2019)*, 2019.

[20] Charles LA Clarke, Saira Rizvi, Mark D Smucker, Maria Maistro, and Guido Zuccon. Overview of the trec 2020 health misinformation track. In *TREC*, 2020.

[21] Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. Vera: Prediction techniques for reducing harmful misinformation in consumer health search. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2066–2070, 2021.