Learning Trustworthy Web Sources to Derive Correct Answers and Reduce Health Misinformation in Search

Dake Zhang[1], Amir Vakili Tahami[2], Mustafa Abualsaud[1], Mark D. Smucker[2]
[1] David R. Cheriton School of Computer Science, University of Waterloo
[2] Department of Management Sciences, University of Waterloo

ABSTRACT

When searching the web for answers to health questions, people can make incorrect decisions that have a negative effect on their lives if the search results contain misinformation. To reduce health misinformation in search results, we need to be able to detect documents with correct answers and promote them over documents containing misinformation. Efforts to use signals of document quality as a proxy for correctness have had limited success, as evidenced by the large gap between automatic and manual methods in the TREC Health Misinformation Track. In the 2021 track, automatic runs were not allowed to use the known answer to a topic's health question, and as a result, the top automatic run had a compatibility-difference score of 0.043 while the top manual run, which used the known answer, had a score of 0.259. The compatibility-difference measures the ability of methods to rank correct and credible documents before incorrect and non-credible documents. By using an existing set of health questions and their known answers, we show it is possible to learn which web hosts are trustworthy, from which we can predict the correct answer to the 2021 health questions with an accuracy of 76%. Using our predicted answers, we can promote documents that we predict contain this answer and achieve a compatibility-difference score of 0.129, which is a three-fold increase in performance over the best previous automatic method.

KEYWORDS

Health Misinformation; Stance Detection; Web Search

1 INTRODUCTION

Search engine results can either help or hinder people's ability to correctly answer health-related questions [15]. When a search engine's results are biased toward correct information, people are more likely to make a correct decision, but when biased toward incorrect information, people are more likely to make an incorrect decision than if they had not searched in the first place. As shown by White and Hassan [21], biased search results can come about from bias in the document collection, to how people formulate their queries, to how the retrieval algorithm functions. Likewise, how people respond to search results can be affected by the prevalence of certain answers in the results [7, 9, 20] as well as their own personal biases [2, 11, 22].

If people make incorrect decisions with regard to their health queries, these decisions may have a serious negative impact on their lives. Approaches to reducing the rate at which people make incorrect decisions include changes to the search process [12], alerting users to bias in results [8], providing answers directly [5, 10] and the ranking of search results [18]. In this paper, we focus on the latter approach, i.e., ranking correct information before incorrect information.

The TREC Health Misinformation Track (2019-2021) provides a framework for evaluating ranking approaches to reducing health misinformation in search results [3]. The track's search topics are formulated as questions regarding the effectiveness of treatments for health issues. Each search topic is supplied with a *stance* that is to be taken as the correct answer. For example, the topic, "Should I apply ice to a burn?", has a given stance of *unhelpful*.

When a ranker is given the correct answer to a TREC Health Misinformation search topic, Pradeep et al. [18] have shown that using T5 sequence-to-sequence models to determine documents' stances and to rerank documents can produce superior results. Unfortunately, Pradeep et al.'s method lacks a way to automatically determine the correct answer, which limits the approach.

In this paper, we show a simple method to predict the correct answer from a web collection, and using the predicted answer, we can then rank documents in a fashion similar to that of Pradeep et al. [18], and we can do this in a fully automatic manner.

Automatic fact verification has been extensively studied in recent years. Stance detection is an important component in these fact and rumour verification pipelines, whose aim is to find the alignment of a text with respect to a claim. A common practice is to detect stances from various controlled sources and aggregate them to reach a final prediction. Using the web as a source complicates this approach, for web documents are prone to include misinformation.

Popat et al. [16, 17] jointly assessed linguistic features of documents to gauge credibility and stance, before aggregating them to get a final prediction for claims on the web. In addition to stance detection, they also weighted the credibility of articles by the trustworthiness of sources. This trustworthiness was derived from the number of times a web source supported a true claim or denied a false claim. In a similar work, Dong et al. [6] first automatically extracted numerous facts from websites and then estimated the trustworthiness and accuracy of sources using an iterative process, based on the assumption that trustworthy sources contain accurate facts and accurate facts come from trustworthy sources.

In this paper, we take the idea of learning trustworthy sources via exogenous signals and utilize it here in a simpler form, where we train a *trust model* to learn trustworthy hostnames by seeing if they contain information consistent with a set of health search topics with known answers.

Using this trust model, we can predict a *helpful probability* for a new topic and rerank documents based on the extent to which the documents align with our prediction. Our method is automatic and does not use the correct stance of topics to produce a run. We are able to find correct (helpful) information on par with the best automatic runs submitted to the TREC 2021 Health Misinformation Track while reducing the amount of incorrect (harmful) results to a level similar to that of manual runs that utilized knowledge of the correct answer. This performance places our method far

above the existing automatic methods and is comparable to some strong manual runs. To our knowledge, our method represents a new state-of-the-art for an automatic method on the TREC 2021 Health Misinformation task.

2 DATA

We use health topics from the TREC 2019 and 2021 Health Misinformation Track [1, 3], and White and Hassan [21]. Each topic is comprised of a single health issue and a treatment (e.g., "antibiotics common cold" [21]). Each topic also comes with a label that describes the true efficacy of the treatment according to trusted medical sources (e.g., cochrane.org). These labels are categorized into helpful, unhelpful, and inconclusive. We only use topics that are labeled as helpful or unhelpful. Some of the topics from White and Hassan [21] overlap with topics in the Health Misinformation Track. We remove those overlapping topics and sample a balanced subset of helpful/unhelpful topics from White and Hassan [21]. In total, for each helpful and unhelpful category, we have 17 topics from [1], 25 topics from [3], and 45 topics from [21]. For document judgments, we use qrels from the TREC 2019 and 2021 Health Misinformation Track [1, 3], in particular the supportiveness aspect of the judgments. Supportiveness (denoted as effectiveness in [1]) refers to a document's stance on the efficacy of the treatment for the health issue. We use 2019 qrels for training and validation, and the 2021 qrels for evaluation. Both 2019 track and 2021 track used web-based document collections: ClueWeb12-B13 in 2019 and CommonCrawl C4.en.noclean [19] in 2021. We describe how this data is used in more detail in the next section.

3 METHODS

Our pipeline depends mainly on two models, a stance detection model (SDM) that is based on the T5 [19] language model, and a trust model (TM) that we build using the logistic regression model. The stance detection model is used to detect a document's stance on the efficacy of the topic, whereas the trust model aggregates stance scores from different hostnames to learn which hostnames to trust and predict an answer of whether or not the treatment is helpful to the health issue.

3.1 Stance Detection Model (SDM)

Inspired by the Vera system from Pradeep et al. [18], we fine-tune the pre-trained T5 language model [19] to detect stances. We formulate this task as a binary classification task: given the health topic and a relevant document, the model aims to detect the document's stance towards the treatment of the health issue, i.e., whether or not the document supports the use of the treatment.

For fine-tuning, we use the 2019 qrels. The qrels for those 34 topics (17 helpful topics and 17 unhelpful topics), are heavily imbalanced, with a total of 2,078 supportive documents and only 144 dissuasive documents. To prevent the model from biasing towards the majority and improve its generalizability, we sample an equal number of supportive and dissuasive documents for each topic and remove document judgments of 5 topics that only have supportive or dissuasive documents.

To overcome the 512 tokens limit of T5 [19], we design a heuristic approach for selecting relevant sentences from documents. In our approach, we define a list of indicator words that include the topic's query tokens and a set of pre-selected stance-related words [1]. We split the document into sentences and score each sentence by counting the total number of occurrences of indicator words, where we do word matching in their stemmed form. Finally, we concatenate those top-scoring sentences in their original order in the document. If the total number of tokens is still below 512, we repeatedly add the sentences following the first selected sentence.

The T5 [19] language model reframes multiple natural language processing (NLP) tasks into a unified text-to-text framework where different NLP tasks can be prompted. We adopt a method similar to Nogueira et al. [14] in constructing our input to prompt T5 for stance detection. Our input to the model is:

```
stance topic: {query} document: {selected sentences}
```

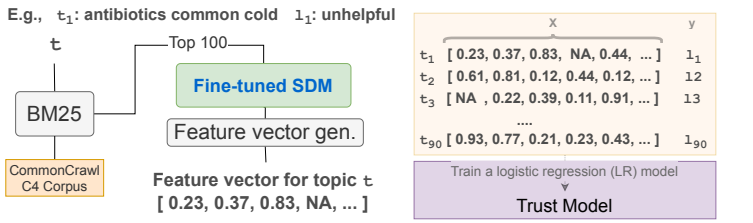
Where `{query}` is the topic's query and `{selected sentences}` is the concatenation of high-scoring sentences described earlier. To obtain binary classification scores, we use an approach similar to Pradeep et al. [18]. Specifically, we apply a softmax function on the logits of the "favor" and "against" found in T5's first generated token. This allows our model to output a supportive score and a dissuasive score that sum to 1. These scores indicate the extent to which the document supports/dissuades the use of the treatment for the health issue.

3.2 Trust Model

Our trust model is designed to learn from trustworthy sources that often give correct health information. For example, WebMD is known to offer credible and in-depth medical information written by trusted medical professionals, and therefore documents from this hostname are likely to be more accurate than some other sources (e.g., blogs or treatment marketing websites). The goal of the trust model is to automatically learn which hostnames to trust and predict the true efficacy of the health treatment based on those sources. Using our stance model, we can detect whether a document supports or dissuades the use of a treatment. Given the true efficacy of the training topics, we can further infer whether this document is correct or not. With labeled training topics and their relevant documents with predicted stances, we are able to train a logistic regression (LR) model to act as the trust model.

Figure 1 shows an overview of the procedure to build the training set and to train the trust model. For each of the total 90 sampled helpful and unhelpful topics from White and Hassan [21], we retrieve the top 100 documents using BM25 as implemented in Pyserini [13]. We then obtain stances of these 100 documents using our SDM. These stances are used to construct a feature vector for the topic, where each feature value corresponds to a hostname and its value is a re-scaled supportive score from our SDM ($2 \times$ `doc_supportive_score` $- 1$). If the top 100 includes multiple documents from the same hostname, we choose the topmost document for that hostname (i.e., the most relevant document according to BM25 scores). The size of the feature vector is the number of distinct hostnames in the top 100 results across all 90 topics. In our

[1] Stance words: `help, treat, benefit, effective, safe, improve, useful, reliable, evidence, prove, experience, find, conclude, ineffective, harm, hurt, useless, limit, insufficient, dangerous, bad.`



We build a feature vector for each topic t from [21], where each feature value represents a distinct hostname (e.g., webmd.com), and its value is a re-scaled supportive score from the stance detection model (SDM) for topic t on the most relevant document from the hostname. The size of the vector is the # of distinct hostnames in the top 100 results across all topics. We use the feature vector with the topic label as training samples for the LR Trust Model.

Figure 1: Procedure of building the LR-based trust model, which is trained using predicted stances of documents from hostnames.

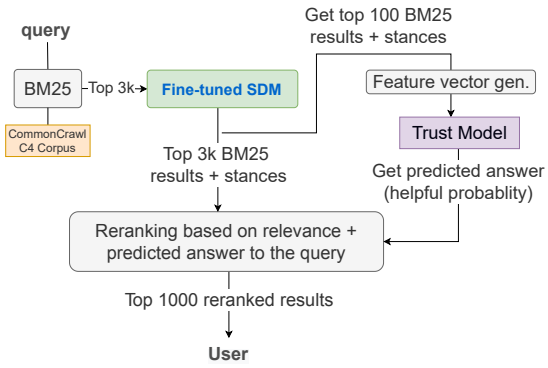


Figure 2: Pipeline for predicting answers to new queries and reranking search results based on correctness and relevance.

case, the size is 3847. The default feature value is 0 if that value's hostname does not appear in the top 100 BM25 results for a topic. In total, we have 90 feature vectors, one for each training topic, and 90 corresponding efficacy labels. We train the LR model using the feature vector as the independent variable and the binary topic label (1 for helpful and 0 for unhelpful) as the dependent variable. We denote the positive probability output from the LR as the `helpful_probability` of the topic.

Our training set is built such that the trust model learns which hostnames provide more correct health information and updates its weights accordingly. For predicting answers to new health questions, the trust model uses the learned weights to aggregate stances from hostnames to predict its final answer. In essence, the model utilizes the wisdom of the crowd, i.e., hostnames, while emphasizing trustworthy sources when making its prediction. While building the trust model, we have experimented with different machine learning models (SVMs, RandomForest, etc), but choose LR for its best performance and good interpretability. We have also experimented with different methods to re-scale SDM scores, and choose the described one as it yields the best performance.

3.3 Pipeline

Figure 2 gives an overview of the pipeline for predicting answers and reranking search results. Given a query, we first retrieve the top 3k BM25 results and use our SDM to get stances for each document. We use stances of the top 100 results to create the feature vector for the query and ignore hostnames that are not part of the 3847 distinct hostnames in the training set. The trust model takes the feature vector as input and outputs the predicted answer (i.e., the helpful probability) to the treatment in the query. Finally, we rerank search results using stance scores from the SDM, the predicted answer from the Trust Model, and the BM25 score. We assign each document with a `correct_score` that reflects our confidence that this document provides the correct information.

$$\texttt{correct\_score} = \texttt{supportive\_score} \times \texttt{helpful\_probability} + \texttt{dissuasive\_score} \times (1 - \texttt{helpful\_probability})$$

Where `supportive_score` and `dissuasive_score` are from our SDM, and `helpful_probability` is the answer probability from our trust model. Finally, we combine `correct_score` with the BM25 score to rerank search results using the formula below:

$$\texttt{final\_score} = \texttt{BM25\_score} \times e^{\texttt{correct\_score} - 0.5}$$

4 EXPERIMENT

We evaluate the effectiveness of our pipeline on the TREC 2021 Health Misinformation Track. We use the primary evaluation measure set by the track's organizers, i.e., the compatibility-difference measure [4]. Compatibility measures the similarity of a given ranking to the ideal ranking using a rank-biased overlap metric. The ideal ranking is constructed by sorting the qrels based on NIST's judgments of usefulness, credibility, and correctness of documents [3]. Helpful and harmful documents are respectively defined as documents with stances matching or opposing the topic's stance field. The helpful compatibility is the compatibility of a run to an ideal ranking of only helpful documents, while the harmful compatibility compares the run to the worst ranking of only harmful documents. The compatibility difference is the difference between these two compatibility scores. The greater this difference is, the better the run is in promoting correct and credible information over misinformation. Of the track's 50 topics, NIST judged 35 topics, among which there are 3 topics that do not have any documents considered harmful. As the track's organizers did [3], we only consider the remaining 32 topics for evaluation purposes.

4.1 Experiment Settings

Overall, we train and validate our models on 2019's data and report the test result on 2021's data. Specifically, we have the following two experiment settings:

(a) **2019 cross-validation**: We randomly split 2019 topics into five folds and performed cross-validation to find the best set of hyperparameters for our SDM and the trust model.
(b) **2021 test**: Using the hyperparameters obtained above, we evaluate our pipeline on 2021's data.

4.2 Model Hyperparameters

For our SDM, we use the 2019 sampled qrels for fine-tuning. We train and test this model in a zero-shot setting, meaning the dataset splitting is by topics. We fine-tune `T5-Large` using AdamW optimizer with a learning rate of 2e-5 and a batch size of 16, with Early Stopping based on the F1-macro on the validation set (random 10% of the training set) with a patience of 5. For the trust model, we use the default configuration of LR in the `scikit-learn` python library except that we set the norm of the penalty to be none.

5 RESULTS

Table 1 shows that our stance detection model (SDM) is able to detect the document's stance (either supportive or dissuasive) with good performance, and that our trust model (TM) can derive the correct answer to health questions in the majority of cases.

Model	Data	TPR	FPR	Acc	AUC
SDM	2019 sampled qrels	0.800	0.449	0.665	0.807
	2021 qrels	0.897	0.182	0.882	0.930
TM	2019 topics	0.471	0.176	0.647	0.682
	2021 topics	0.640	0.120	0.760	0.822

Table 1: Classification Performance. SDM: Stance Detection Model, TM: Trust Model, TPR: True Postive Rate, FPR: False Positive Rate, Acc: Accuracy, AUC: Area Under the Curve.

To illustrate what the trust model has learned, we list the top 5 hostnames ranked by weights in Table 2. We observe that those high-ranked hostnames are indeed credible sources. Space limitations prevent us from showing more of the learned weights, but in our inspection of them, we see a mix of very sensible results as well as mistakes. While some host weights appear wrong, in the aggregate the trust model is able to predict correct answers.

	Hostname	Weight
1	www.cochrane.org	2.7860
2	emedicine.medscape.com	2.3674
3	patient.info	1.9708
4	experts.mcmaster.ca	1.8269
5	www.everydayhealth.com	1.6509

Table 2: Top 5 hostnames ranked by weights learned by the trust model trained on White and Hassan [21] topics.

Table 3 lists the best automatic run, `best_auto`, and the best manual run, `best_manual`, submitted in the 2021 Health Misinformation Track as well as the following:

- `BM25-Baseline` [Automatic]: Top 1000 BM25 results.
- `Trust-Pipeline` [Automatic]: Output from our pipeline (this paper's main results).
- `Correct-Stance` [Manual]: In place of the trust model's predicted stance, we use the correct stance in our pipeline to show the upper bound of our method.

Run Identifier	Fields Used	C(help)	C(harm)	C(Δ)
`BM25-Baseline`	query	0.122	0.144	-0.022
`best_auto`	desc	0.195*	0.153*	0.043*
`Trust-Pipeline`	query	**0.198**†	**0.069**†	**0.129**†
`Correct-Stance`	query, stance	0.223†	0.047†	0.176†
`best_manual`	desc, stance	**0.297***	**0.038***	**0.259***

Table 3: Overall performance using the Compatibility metric. Top three runs are automatic. Bottom two runs are manual. († indicates significant difference from `BM25-Baseline`, $p < 0.01$; * indicates the value is from [3]; bold font indicates the best automatic/manual performance).

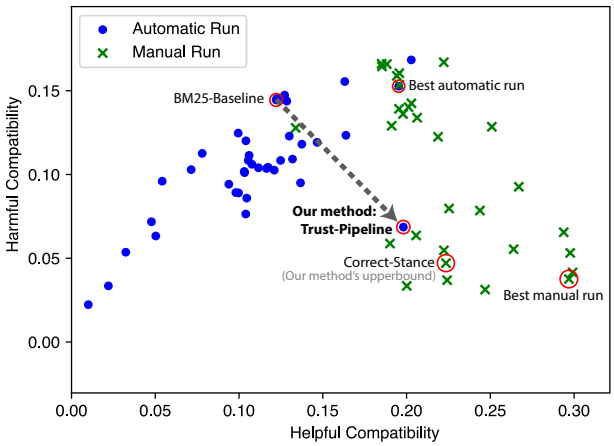


Figure 3: Comparison with automatic and manual runs submitted in the TREC 2021 Health Misinformation Track.

Table 3 and Figure 3 show that our method (`Trust-Pipeline`) achieves a new high for automatic runs on this task. Our method has a helpful-compatibility comparable to the previous best automatic run, and our method reduces the harmful-compatibility to levels on par with strong manual runs. Our fully automatic method is able to rerank the `BM25-Baseline` and move it from the cluster of automatic runs into the cluster of strong manual runs with low harmful-compatibility.

6 CONCLUSION

Our work demonstrates that in the limited domain of the TREC Health Misinformation Track, we can predict answers to unseen questions from the misinformation laden web by learning trustworthy web hosts, and then use these predicted answers to reduce misinformation in search results. Using the top 100 BM25 ranked documents with their predicted stances towards health questions, our trust model can predict correct answers with an accuracy of 76% to 50 health questions from the TREC 2021 Health Misinformation Track. With the predicted answers and predicted document stances, we are able to rerank a BM25 baseline and obtain an automatic run that achieves a significant increase in performance over the best previous automatic run.

REFERENCES

[1] Mustafa Abualsaud, Christina Lioma, Maria Maistro, Mark D. Smucker, Guido, and Zuccon. 2020. Overview of the TREC 2019 Decision Track. In *TREC*.

[2] Leif Azzopardi. 2021. *Cognitive Biases in Search: A Review and Reflection of Cognitive Biases in Information Retrieval.* ACM, 27–37. https://doi.org/10.1145/3406522.3446023

[3] Charles LA Clarke, Mark D Smucker, and Maria Maistro. 2021. Overview of the TREC 2021 Health Misinformation Track. In *TREC*.

[4] Charles LA Clarke, Alexandra Vtyurina, and Mark D Smucker. 2021. Assessing Top-Preferences. *ACM Transactions on Information Systems (TOIS)* 39, 3 (2021), 1–21.

[5] Dina Demner-Fushman, Yassine Mrabet, and Asma Ben Abacha. 2020. Consumer health information and question answering: helping consumers find answers to their health-related information needs. *Journal of the American Medical Informatics Association* 27, 2 (2020), 194–201.

[6] Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, and Wei Zhang. 2015. Knowledge-based trust: estimating the trustworthiness of web sources. *Proceedings of the VLDB Endowment* 8, 9 (2015), 938–949.

[7] Tim Draws, Nava Tintarev, Ujwal Gadiraju, Alessandro Bozzon, and Benjamin Timmermans. 2021. This Is Not What We Ordered: Exploring Why Biased Search Result Rankings Affect User Attitudes on Debated Topics. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 295–305.

[8] Robert Epstein, Ronald E Robertson, David Lazer, and Christo Wilson. 2017. Suppressing the search engine manipulation effect (SEME). *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–22.

[9] Amira Ghenai, Mark D. Smucker, and Charles L.A. Clarke. 2020. *A Think-Aloud Study to Understand Factors Affecting Online Health Search.* ACM, 273–282. https://doi.org/10.1145/3343413.3377961

[10] Anat Hashavit, Hongning Wang, Raz Lin, Tamar Stern, and Sarit Kraus. 2021. Understanding and Mitigating Bias in Online Health Search. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 265–274.

[11] Annie Y.S. Lau and Enrico W. Coiera. 2007. Do People Experience Cognitive Biases while Searching for Information? *Journal of the American Medical Informatics Association* 14, 5 (09 2007), 599–608. https://doi.org/10.1197/jamia.M2411 arXiv:https://academic.oup.com/jamia/article-pdf/14/5/599/2139239/14-5-599.pdf

[12] Annie Y.S. Lau and Enrico W. Coiera. 2009. Can Cognitive Biases during Consumer Health Information Searches Be Reduced to Improve Decision Making? *Journal of the American Medical Informatics Association* 16, 1 (1 2009), 54–65. https://doi.org/10.1197/jamia.M2557 arXiv:https://academic.oup.com/jamia/article-pdf/16/1/54/2572282/16-1-54.pdf

[13] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. *Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations.* Association for Computing Machinery, New York, NY, USA, 2356–2362. https://doi.org/10.1145/3404835.3463238

[14] Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document Ranking with a Pretrained Sequence-to-Sequence Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020.* Association for Computational Linguistics, Online, 708–718. https://doi.org/10.18653/v1/2020.findings-emnlp.63

[15] Frances A Pogacar, Amira Ghenai, Mark D Smucker, and Charles LA Clarke. 2017. 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.* 209–216.

[16] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion.* 1003–1012.

[17] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2018. CredEye: A credibility lens for analyzing and explaining misinformation. In *Companion Proceedings of the The Web Conference 2018.* 155–158.

[18] Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. 2021. 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.* 2066–2070.

[19] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv e-prints* (2019). arXiv:1910.10683

[20] Qiurong Song and Jiepu Jiang. 2022. How Misinformation Density Affects Health Information Search. In *The World Wide Web Conference* (Virtual Event, Lyon, France) *(WWW '22).* Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3485447.3512141

[21] Ryen W. White and Ahmed Hassan. 2014. Content Bias in Online Health Search. *ACM Trans. Web* 8, 4, Article 25 (nov 2014), 33 pages. https://doi.org/10.1145/2663355

[22] Ryen W White and Eric Horvitz. 2015. Belief dynamics and biases in web search. *ACM Transactions on Information Systems (TOIS)* 33, 4 (2015), 1–46.