Vera: Prediction Techniques for Reducing Harmful Misinformation in Consumer Health Search

Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin

David R. Cheriton School of Computer Science
University of Waterloo, Ontario, Canada

ABSTRACT
The COVID-19 pandemic has brought about a proliferation of harmful news articles online, with sources lacking credibility and misrepresenting scientific facts. Misinformation has real consequences for consumer health search, i.e., users searching for health information. In the context of multi-stage ranking architectures, there has been little work exploring whether they prioritize correct and credible information over misinformation. We find that, indeed, training models on standard relevance ranking datasets like MS MARCO passage—which have been curated to contain mostly credible information—yields models that might also promote harmful misinformation. To rectify this, we propose a label prediction technique that can separate helpful from harmful content. Our design leverages pretrained sequence-to-sequence transformer models for both relevance ranking and label prediction. Evaluated at the TREC 2020 Health Misinformation Track, our techniques represent the top-ranked system: Our best submitted run was 19.2 points higher than the second-best run based on the primary metric, a 68% relative improvement. Additional post-hoc experiments show that we can boost effectiveness by another 3.5 points.

CCS CONCEPTS
• Information systems → Users and interactive retrieval.

KEYWORDS
Multi-Stage Ranking; Sequence-to-Sequence Models

1 INTRODUCTION
The Internet has rapidly grown into an influential medium for producing and disseminating content to broad audiences. With uncontrolled growth come opportunists who use this advantage to distribute misinformation for personal gain. Search engines, which form the gateway to much of this information, can significantly influence user behavior. Thus, systems today face the monumental task of discerning authoritative and correct from dubious and incorrect information. In the current environment amidst the COVID-19 pandemic, there has been an increase in the general public’s interest in consumer health issues. System responses to such user queries should attempt to promote helpful results while flagging (or even suppressing) harmful results to assist users in making informed decisions based on scientific consensus.

The related task of fact verification has been widely studied by the NLP community on corpora such as Wikipedia and discussion blogs [8, 19]. Fact verification systems must predict a claim’s veracity and in some cases must provide relevant support from a corpus. More recently, researchers have built various fact verification datasets [9, 20] grounded on scientific corpora, such as the literature on coronaviruses and COVID-19 [22]. However, there are added complexities in consumer health search: Systems need to navigate a larger space of content which contains bad actors spreading misinformation and return documents that capture some output text. Since this architecture focuses exclusively on relevance ranking, we propose an additional label prediction technique called Vera, which is inspired by the success of VerT5erini [15], a state-of-the-art fact verification system on the SciFACT task [20] that also uses T5. In our implementation, Vera takes advantage of effectiveness judgments from the TREC 2019 Decision Track [1] to promote helpful content and suppress harmful content. A linear combination of prediction scores from Vera and relevance scores from our two-stage reranking design produced the best system at the TREC 2020 Health Misinformation Track [4] in terms of the primary metric [5, 6].

2 TASK DEFINITION
The context of this work is the ad hoc retrieval task in the TREC 2020 Health Misinformation Track [4], where systems are provided with a corpus of new articles C and are tasked to return a ranked list of 1000 documents for a set of topics. Recognizing limitations in
the standard notion of topical relevance, the evaluation explicitly assesses whether correct and credible information are prioritized over incorrect information. The document collection comprises news articles from the CommonCrawl news crawl spanning the first four months of 2020 (January 1st, 2020 to April 30th, 2020), covering the onset of the COVID-19 pandemic. This corpus has over 65M articles and is about 1.7 TB in size.

Each of the 46 topics in the evaluation has a description field comprising a question of the form: "Can A B COVID-19?", where A is a treatment and B is one of five effect terms (cause, cure, help, prevent, and worsen). Each topic has an answer field, which is either a "yes" or "no" that corresponds to the medical consensus at the time of topic creation. NIST assessors first judged documents based on three aspects:

- **Usefulness**, which considers if the document includes content that a user might find useful in answering the topic.
- **Correctness**, which verifies if the document’s answer aligns with the topic answer (i.e., medical consensus). Note that there is a difference between a document not answering the question and providing an incorrect answer.
- **Credibility**, which assesses the document’s credibility.

These aspects are mapped into the graded relevance scale in Table 1. From this, the organizers created "helpful" and "harmful" qrels by taking only the documents with positive and negative grades, respectively. For evaluation, the harmful qrels took the absolute value of the relevance grade (since tools like 𝐻𝑅 explicitly separate helpful from harmful content).

<table>
<thead>
<tr>
<th>Description</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Useful; correct; credible</td>
<td>4</td>
</tr>
<tr>
<td>Useful; correct; not credible or not judged</td>
<td>3</td>
</tr>
<tr>
<td>Useful; no answer or not judged; credible</td>
<td>2</td>
</tr>
<tr>
<td>Useful; no answer or not judged; not credible or not judged</td>
<td>1</td>
</tr>
<tr>
<td>Not useful; ignored; ignored;</td>
<td>0</td>
</tr>
<tr>
<td>Useful; incorrect; not credible or not judged</td>
<td>-1</td>
</tr>
<tr>
<td>Useful; incorrect; credible</td>
<td>-2</td>
</tr>
</tbody>
</table>

Table 1: Relevance grades for the TREC 2020 Health Misinformation Track.

The output of first-stage retrieval (𝐻₀) is passed to a reranking pipeline comprised of a pointwise reranker, monoT5 (𝐻₁), and then a pairwise reranker, duoT5 (𝐻₂). This basic design was outlined in Pradeep et al. [16]. Note, critically, as demonstrated by our experiments, this design is not sufficient for our task, as it has a tendency to retrieve topically relevant but harmful information. In what follows, we describe not only the basic design but necessary modifications for our task.

3.1 𝐻₀: Keyword Retrieval

The candidate generation stage 𝐻₀ (also called first-stage retrieval) receives as input the user query 𝑞 and produces top 𝑘₀ candidates 𝑅₀. In our implementation, the query is treated as a bag of words for ranking documents from the corpus using a standard inverted index based on BM25 [18]. All our experiments used the Pyserini IR toolkit [2, 10], which provides a Python interface to Anserini [24, 25], itself built on the popular open-source Lucene search engine. At search time, we retrieve the top-100 documents per query.

3.2 𝐻₁: Pointwise Reranking with monoT5

In stage 𝐻₁, documents retrieved in 𝐻₀ are reranked by a pointwise reranker called monoT5. The model estimates a score 𝑞mono quantifying how relevant a candidate 𝑑𝑖 ∈ 𝑅𝑛−1 is to a query 𝑞, that is, 𝑃(Relevant = 1|𝑑𝑖, 𝑞). Details of monoT5 are described in Nogueira et al. [14]; here, we only provide a short overview.

The monoT5 model uses T5-3B [17] and formulates the problem as a sequence-to-sequence task. Specifically, ranking is performed using the following input sequence template, as suggested by Nogueira et al. [14]:

\[
\text{Query: } q \quad \text{Document: } d \quad \text{Relevant: } \quad q \rightarrow q \text{ monoT5 \rightarrow } q
\]

where 𝑞 and 𝑑 are the query and document texts, respectively. The model is fine-tuned to produce the token “true” or “false” depending on whether the document is relevant or not to the query.

For the TREC 2020 Health Misinformation Track, the default question text is the topic description. We call this standard template the monoT5base variant. Alternatively, we also consider a variant, monoT5NL, where we rephrase the question "Can A B COVID-19?" and the answer field in a natural language sentence form, i.e., as “A can B COVID-19” if the answer field is “yes” and as “A can not B COVID-19” if the answer field is “no”. The goal of this template is to see if there are any improvements to be gained by aligning the query text with the answer field.

We train the monoT5 model by first fine-tuning on the MS MARCO passage dataset and then fine-tuning it again on MedMARCO, which is a subset of the MS MARCO passage dataset where only queries containing medical terms are kept [12]. Zhang et al. [26] called this training strategy "pre-fine-tuning"; see Pradeep et al. [16] for additional details.

At inference time, to compute probabilities for each query–document pair, we apply a softmax only to the logits of the “true” and “false” tokens and rerank the top-100 documents according to the probabilities assigned to the “true” token.

As discussed in Lin et al. [11], one reoccurring theme in the application of transformers to text ranking is the handling of texts that are longer than the input sequences that the models were
designed to handle (typically, 512 tokens). Following Pradeep et al. [16], we first segment each document into passages by applying a sliding window of six sentences with a stride of three. Each passage was then prepended with the title of the document. We obtain a probability of relevance for each passage by performing inference on it independently, and then select the highest probability among the passages as the relevance score of the document; this technique has been called MaxP [7, 26].

3.3 $H_2$: Pairwise Reranking with duoT5

The output $R_1$ from the previous stage serves as input to the pairwise reranker we call duoT5. In this pairwise approach, the reranker considers a pair of documents $(d_i, d_j)$ and estimates the probability $p_{i,j}$ that candidate $d_i$ is more relevant than $d_j$ to query $q$, that is, $P(d_i > d_j | d_i, d_j, q)$, where $d_i > d_j$ denotes that $d_i$ is more relevant than $d_j$ (with respect to the query $q$).

Details of duoT5, including the default hyperparameters used in our work, are described in Pradeep et al. [16]; here, we only provide a short overview. As the name suggests, duoT5 is also based on the pairwise approach. To address this issue, we devised an alternative model pairs of representative passages from the documents under consideration to compute the pairwise scores, which are then aggregated to yield the relevance score of each document. We increase the maximum input tokens from the default of 512 to 1024 to account for pairs of passages being twice as long.

3.4 Label Prediction

We cast the problem of separating helpful from harmful content as a label prediction task. Our Vera model, also based on T5-3B, was inspired by Pradeep et al. [15] and T5’s pretraining on MNLI [23].

Given the topic $q$ and the highest monoT5 scoring segment $s_i$ from a document $d_i$, the model is tasked to predict a label $\hat{y}(q, s_i) \in \{true, false\}$. Here, we use the following input sequence:

Query: $q$  Document: $s_i$  Relevant:

For the TREC 2020 Health Misinformation Track, we have the two variants, duoT5base and duoT5NL, taking the same query templates as the pointwise reranker.

At inference time, we aggregate the pairwise scores $p_{i,j}$ so that each document receives a single score $s_i$ using the Sym-Sum method proposed by Pradeep et al. [16], where $J_i = \{0 \leq j < k_1, j \neq i\}$:

$$
\text{SYM-SUM} : s_i = \sum_{j \in J_i} (p_{i,j} + (1 - p_{j,i}))
$$

In previous work, the top-50 candidates in $R_1$ are reranked according to their scores $s_i$ to obtain a ranked list of candidates $R_2$ designed for final consumption (thus requiring $50 \times 49$ individual inferences). However, in the TREC 2020 Health Misinformation Track, the ad hoc retrieval task required a ranked list of 1000 documents per topic. Since Vera performs a linear combination using the scores from multi-stage ranking and label prediction, to keep the combination meaningful we need top-1000 candidates, but this would be very computationally expensive due to the quadratic nature of the pairwise approach. To address this issue, we devised an alternative solution: We still only rerank the top-50 $R_1$ candidates, but form an intermediate ranked list, denoted by $R'_2$, the scores of which we post-process to combine with the scores of $R_1$ to obtain a final ranked list of 1000 candidates.

Let the score of a document $d_j$ in $R_1$ be $s_{i}^{\text{mono}}$, and the highest and lowest monoT5 scores of candidates in $R'_2$ be $s_{\text{max}}$ and $s_{\text{min}}$, respectively. Similarly, let the highest and lowest scores after aggregating over candidates in $R'_2$ be $s_{\text{max}}$ and $s_{\text{min}}$, respectively. Then we calculate the final duoT5 scores using one of two methods:

$$
\begin{align*}
\text{duo}_1 & : s_i = \begin{cases} 
    s_{\text{mono}}^{\text{max}} + & \frac{\left(s_i - s_{\text{min}}\right)\left(s_{\text{max}} - s_{\text{min}}\right)}{s_{\text{max}} - s_{\text{min}}}, \\
    s_{i}^{\text{mono}}, & d_i \not\in R'_2 \\
\end{cases}, \\
\text{duo}_2 & : s_i = \begin{cases} 
    s_{i}^{\text{mono}} + s_i, & d_i \not\in R'_2 \\
    s_i, & d_i \not\in R'_2 \\
\end{cases}
\end{align*}
$$

These scores determine the final top-1000 ranked list, $R_2$.

At inference time, we use the highest scoring monoT5 passage as the representative passage for each document. We feed the duoT5 model pairs of representative passages from the documents under consideration to compute the pairwise scores, which are then aggregated to yield the relevance score of each document. We increase the maximum input tokens from the default of 512 to 1024 to account for pairs of passages being twice as long.

4 RESULTS

Table 2 reports results from the TREC 2020 Health Misinformation Track. For reference, row (a) provides the median score across 51 submissions from eight groups for the evaluation. Rows (b)–(d) present the three top-scoring submitted runs (per group); note that row (c) represents a manual submission. As we can see, the Vera technique described here was the top-scoring run submitted to the evaluation by a large margin. Rows (e)–(j) denote additional runs that were part of our official submission; rows (e)–(i) show the results of different configurations that used only relevance ranking (i.e., no label prediction). Rows (b) and (i)–(m) represent variants that combine both relevance and label prediction scores, as described in Section 3.4. Rows (k)–(m) show results of the highest scoring configuration on top of each of the three relevance ranking methods, discovered by an evenly spaced sweep of the linear
This motivates our linear combination approach described in Table 2: Compatibility scores on the TREC 2020 Health Misinformation Track. Results of significance tests ($t$-tests, $p < 0.05$) are denoted by superscripts.

<table>
<thead>
<tr>
<th>Model</th>
<th>COMPHelp</th>
<th>COMPHarm</th>
<th>COMPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Median</td>
<td>0.334</td>
<td>0.075</td>
<td>0.259</td>
</tr>
<tr>
<td>(b) Vera ($\lambda = 0.5, z = \text{mono})$ = h2oloo.m8</td>
<td>0.490$^e$</td>
<td>0.016$^f$</td>
<td>0.474$^g$</td>
</tr>
<tr>
<td>(c) cr-kq-td (Webis)</td>
<td>0.334</td>
<td>0.052</td>
<td>0.282</td>
</tr>
<tr>
<td>(d) adhoc_run3 (KU)</td>
<td>0.401</td>
<td>0.121</td>
<td>0.280</td>
</tr>
<tr>
<td>(e) BM25 = h2oloo.m1</td>
<td>0.368</td>
<td>0.120</td>
<td>0.248</td>
</tr>
<tr>
<td>(f) + monoT5harm = h2oloo.m2</td>
<td>0.440</td>
<td>0.113</td>
<td>0.327</td>
</tr>
<tr>
<td>(g) + duoT5harm = h2oloo.m4</td>
<td>0.466$^e$</td>
<td>0.129</td>
<td>0.346$^f$</td>
</tr>
<tr>
<td>(h) + monoT5harm = h2oloo.m3</td>
<td>0.511$^e$</td>
<td>0.075$^f$</td>
<td>0.436$^g$</td>
</tr>
<tr>
<td>(i) + duoT5harm = h2oloo.m5</td>
<td>0.549$^f$</td>
<td>0.080$^f$</td>
<td>0.469$^f$</td>
</tr>
<tr>
<td>(j) Vera ($\lambda = 0.0, z = \text{mono})$ = h2oloo.m7</td>
<td>0.449</td>
<td>0.015$^f$</td>
<td>0.434$^f$</td>
</tr>
<tr>
<td>(k) Vera ($\lambda = 0.95, z = \text{mono})$</td>
<td>0.567$^e$</td>
<td>0.019$^f$</td>
<td>0.484$^g$</td>
</tr>
<tr>
<td>(l) Vera ($\lambda = 0.95, z = \text{duo})$</td>
<td>0.520$^{efgh}$</td>
<td>0.018$^{fgh}$</td>
<td>0.502$^{efg}$</td>
</tr>
<tr>
<td>(m) Vera ($\lambda = 0.75, z = \text{duo})$</td>
<td>0.546$^{efghijk}$</td>
<td>0.037$^{fghj}$</td>
<td>0.509$^{efghj}$</td>
</tr>
</tbody>
</table>

Table 2: Compatibility scores on the TREC 2020 Health Misinformation Track. Results of significance tests ($t$-tests, $p < 0.05$) are denoted by superscripts.

The combination parameter $\lambda$. These configurations were not official submissions to the evaluation and come with the added benefit of hindsight. We applied $t$-tests ($p < 0.05$) to determine the statistical significance of metric differences, except for (a), (c), and (d); these results are denoted by the standard superscript notation.

Let us begin by focusing only on relevance ranking. We see that pointwise reranking helps on top of the BM25 baseline, row (e) vs. (f), and pairwise reranking helps on top of pointwise reranking, row (f) vs. (g), as expected. These three settings all have similar COMPHarm scores that are higher than the median, row (a). That is, our runs are surfacing not only more helpful content, but more harmful results as well. This comes as no surprise since our models are only trained on relevance ranking, and indeed, topically relevant information can be harmful. Note here also that these settings used only the topic description.

Both COMPHelp and COMPHarm scores improve when the query is rephrased to align with the topic answer, comparing the "base" and "NL" input template variants, row (h) vs. (i) and row (i) vs. (g). This suggests that these models are capturing notions of "answer correctness" despite being trained on relatively clean relevance ranking datasets like MS MARCO passage. Since we notice improved effectiveness across the board, all further experiments use this input template. Note that here, we still have not added label prediction, and our runs are already substantially better than other submissions to the evaluation, rows (c) and (d). However, these runs still score above the median in COMPHarm, which is concerning.

The label prediction model in isolation (i.e., $\lambda = 0$) results in a large reduction in COMPHarm compared to the pointwise ranker, row (j) vs. (h). However, this comes with a drop in COMPHelp as well. This motivates our linear combination approach described in Section 3.4 that incorporates relevance and label prediction signals. Our top submission, Vera ($\lambda = 0.5, z = \text{mono})$, row (b), greatly improves upon the label prediction model, row (j), in terms of COMPHarm with only a negligible increase in COMPHelp.

In Figure 1, we plot COMPHelp and COMPA scores as a function of $\lambda$ for three different settings. First, we note that for both mono and duo1, COMPHelp increases as we increase the weight on relevance ranking; COMPA increases all the way until 0.95, after which the measure drops because we’re reverting to pointwise and pairwise reranking, respectively. Second, we see that mono and duo1 follow similar trajectories across both measures until $\lambda = 0.8$, after which duo1’s higher COMPHelp, which we can see from row (j) vs. (h), “kicks in.” We note an improvement of around 1.5 points in both COMPHelp and COMPA at $\lambda = 0.95$, row (i) vs. (k).

The duo2 relevance setting behaves differently from mono and duo1: COMPHelp appears convex, with a maximum at $\lambda = 0.9$. As a result, COMPA also follows a different trajectory in that the curve is flatter for intermediate $\lambda$ values compared to other metrics, but performs consistently better than the other configurations. COMPA is maximized at $\lambda = 0.75$, shown in row (m), and this represents our most effective system configuration. Note that this constitutes an 80% relative improvement compared the cr-kq-td manual run by the Webis team, shown in row (c). Overall, our experimental results show that adapting a multi-stage ranking pipeline to incorporate a harmful information classifier like Vera is an easy and effective solution to reduce misinformation in consumer health search.

5 CONCLUSIONS

In this paper, we analyze how multi-stage neural reranking designs perform at prioritizing correct and credible information over misinformation. We find that since these models focus on relevance ranking, they have a tendency to return both helpful information as well as topically relevant but harmful misinformation.

To combat this, we introduced Vera, a label prediction model that exploits a generation-based approach to rerank candidates from pure relevance ranking models to suppress harmful content. Experiments show that our system outperforms other systems submitted to the ad hoc retrieval task in the TREC 2020 Health Misinformation Track by a large margin. Our design can potentially improve consumer health search to combat misinformation, a challenge recently amplified by the COVID-19 pandemic.

ACKNOWLEDGEMENTS

This research was supported in part by the Canada First Research Excellence Fund, the Natural Sciences and Engineering Research Council (NSERC) of Canada, and the Waterloo–Huawei Joint Innovation Laboratory. Additionally, we would like to thank Google for computational resources in the form of Google Cloud credits.
REFERENCES


