Evidence-based Fact-Checking of Health-related Claims

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, Dina Demner-Fushman
U.S. National Library of Medicine, National Institutes of Health
sarrouti.mourad@gmail.com
{asma.benabacha, yassine.m'rabet}@nih.gov, ddemner@mail.nih.gov

Abstract

The task of verifying the truthfulness of claims in textual documents, or fact-checking, has received significant attention in recent years. Many existing evidence-based fact-checking datasets contain synthetic claims and the models trained on these data might not be able to verify real-world claims. Particularly few studies addressed evidence-based fact-checking of health-related claims that require medical expertise or evidence from the scientific literature. In this paper, we introduce HEALTHVER, a new dataset for evidence-based fact-checking of health-related claims that allows to study the validity of real-world claims by evaluating their truthfulness against scientific articles. Using a three-step data creation method, we first retrieved real-world claims from snippets returned by a search engine for questions about COVID-19. Then we automatically retrieved and re-ranked relevant scientific papers using a T5 relevance-based model. Finally, the relations between each evidence statement and the associated claim were manually annotated as SUPPORT, REFUTE and NEUTRAL. To validate the created dataset of 14,330 evidence-claim pairs, we developed baseline models based on pretrained language models. Our experiments showed that training deep learning models on real-world medical claims greatly improves performance compared to models trained on synthetic and open-domain claims. Our results and manual analysis suggest that HEALTHVER provides a realistic and challenging dataset for future efforts on evidence-based fact-checking of health-related claims. The dataset, source code, and a leaderboard are available at https://github.com/sarrouti/healthver.

1 Introduction

The exponential growth of textual information in the form of news, forums, and stories on the web has resulted in the explosion of misinformation (Zhang et al., 2019; Da San Martino et al., 2020). While false information could be dangerous in general, medical misinformation, in particular, presents a challenge to human health and could be detrimental when search engines or social media are used to guide health-related decisions (Barua et al., 2020). For instance, the COVID-19 pandemic has caused the spread of false claims about the origin, prevention, diagnosis, and treatment of the disease (Naeem et al., 2020). COVID-19 related misinformation caused people to turn to fake and unproven cures (Pennycook et al., 2020).

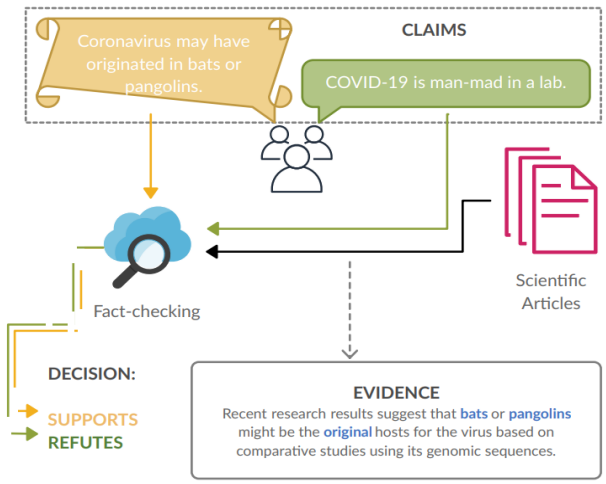

Figure 1: Overview of our evidence-based fact-checking approach with examples of COVID-19 claims, supported and refuted by evidence extracted from a scientific article.

Recently, retrieving and debunking misinformation has received significant attention, especially from fact-checking organizations (e.g. Snopes) that debunk false information. False claims and fake news stories, however, are still spreading on the web (Pennycook et al., 2020). While most fact-checking organizations use human validation of information, the ever-increasing amount of new information on the Internet makes manual verification challenging, time-consuming, and costly (Rashkin et al., 2017; Thorne and Vlachos, 2018; Fan et al.,

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 3499–3512
November 7–11, 2021.

2020). Moreover, in contrast to the open domain, fact-checking medical information requires deep medical knowledge about the topic of the claims (Kotonya and Toni, 2020). As a result, automated fact-checking tools for verifying the veracity of a given health-related claim or news story are needed to assist the information seekers in evaluating the retrieved information from uncontrolled data collections such as the web.

Automated fact-checking systems based on deep learning to identify misinformation distributed on the Internet is a promising approach to counter its spread. Evaluating the veracity of a given claim against textual sources (e.g. scientific articles, Wikipedia) that can support, refute or relate to the claim has been explored to fight the spread of misinformation (Thorne et al., 2018; Wadden et al., 2020). The claims on which the existing fact-checking systems rely for training are synthetic, since they were manually created from the sentences or citations retrieved from a corpus of documents. For instance, claims in the FEVER dataset (Thorne et al., 2018) were manually created by mutating sentences from Wikipedia documents. Whereas the scientific claims in (Wadden et al., 2020) were created from citation sentences by annotators. "Natural" claims, or real-world claims, expressed by Internet users differ from the manually created claims for several reasons. Recent events such as the COVID-19 pandemic showed that real-world claims could include multiple facts (e.g. "Vitamins C and D may help your immune system fight COVID-19") which makes the verification process more complicated as the evidence (i.e. a scientific article) may support the claim about one fact and refute another one that is stated in the same claim. Moreover, most of the claims that spread on the web use speculative and vague language (e.g. "social distancing measures *could be* effective in stopping the spread of the virus"). Therefore, the deep learning models trained on the manually created claims are unlikely to be able to validate claims found on the web. Furthermore, verifying claims in domains such as the medical domain, where medical expertise is needed, makes the task more challenging. The adaptation of fact-checking models trained on open-domain claims to health-related claims might not necessarily work well.

To tackle the aforementioned issues, we introduce a new fact-checking dataset called HEALTHVER, a new dataset for evidence-based fact-checking of health-related claims based on scientific articles. Compared to the existing efforts, we use naturally-occurring claims from the web and scientific articles for verification. As shown in Figure 2, given a claim and a relevant scientific article retrieved from a corpus of scientific articles, our verification system predicts three types of relations between the claim and the evidence extracted from an article: SUPPORTS, REFUTES, and NEUTRAL. In summary, this paper makes the following main contributions:

1. We introduce HEALTHVER, a new manually annotated dataset consisting of 14,330 evidence-claim pairs with their veracity label (i.e. SUPPORTS, REFUTES, and NEUTRAL). To the best of our knowledge, this is the first evidence-based fact-checking study that investigates the veracity of real-world claims against scientific articles.

2. We analyze the complexity of the claims in the HEALTHVER dataset and compare it with the complexity of existing datasets using an ensemble of relation extraction approaches. We also compare the generalization potential of HEALTHVER with existing datasets using pairwise zero-shot accuracy deltas and a new measure that takes into account training set sizes.

3. Our experiments show that training deep learning-based fact-checking models on real-world and in-domain claims substantially improves the performance compared to training on synthetic and open-domain claims. Our results also show that HEALTHVER is a challenging testbed for developing new evidence-based fact-checking systems designed to validate real-world and health-related claims against a corpus of textual documents.

4. We present a detailed error analysis of state-of-the-art models trained and evaluated on HEALTHVER to identify the challenges in real-world claim verification against scientific articles.

2 Related Work

In recent years, there have been growing concerns about the rampant spread of fake news, false claims, and fabricated stories (Derczynski et al., 2017; Poddar et al., 2018; Mishra et al., 2020). Due to the proliferation of misinformation, several efforts have been made to construct fact-checking datasets to advance automated fact-checking systems (Hanselowski et al., 2019; Thorne et al., 2021; Schuster et al., 2021). Vlachos and Riedel (2014)

collected a claim-verification dataset from fact-checking websites (e.g. PolitiFact) that contains a limited number of claims (106 claims).

Wang (2017) constructed a large dataset of 12.8k claims from the PolitiFact website. Nakov et al. (2018) introduced a dataset for the CLEF-2018 CheckThat! shared task on political debates in English and Arabic. A common feature for all the aforecited datasets is that they only contain claims without evidence to support or refute them. For evidence-based fact-checking, Thorne et al. (2018) constructed the FEVER dataset of 185.4k claims generated by mutating sentences extracted from Wikipedia. However, the claims are synthetic since they are created by altering the evidence sentences. Augenstein et al. (2019) introduced MULTIFC, the claim verification dataset of natural claims. It consists of 34,918 claims, collected from 26 fact checking websites in English, the evidence pages to verify the claims, and other metadata information.

In the medical domain, a new dataset was introduced for the TREC 2020 Health Misinformation Track. Documents related to COVID-19 from the CommonCrawl News dataset[1] have been used. In this dataset, the evidence for claim validation was missing. Kotonya and Toni (2020) built a dataset called PUBHEALTH which includes 11.8K claims accompanied by journalists' explanations from fact-checking websites (e.g. Snopes, Politifact). PUBHEALTH is designed to evaluate veracity prediction and explanation generation tasks. The majority of the claims in this dataset are false. Wadden et al. (2020) created SCIFACT, a corpus of 1.4k scientific claims accompanied by abstracts that support or refute each claim. This dataset, however, contains synthetic claims.

The existing datasets for evidence-based fact-checking systems are either based on mutated sentences (e.g. creating claims from Wikipedia and citations sentences), that are not real-world and natural claims, or use journalists' explanations from fact-checking websites. Claims found on the Internet are arguably more complicated and challenging to verify than synthetic claims.

Therefore, to mitigate the above discrepancies between the real-world claims and the training data, we introduce a new dataset of real-world claims related to COVID-19 with associated evidence extracted from scientific articles, manually annotated with three types of relations: SUPPORTS, REFUTES, and NEUTRAL.

[1]From January 1st, 2020 to April 30, 2020.

3 The HEALTHVER Dataset

This section describes our proposed approach to create the HEALTHVER dataset, which consists of three main stages:

1. Claim retrieval: retrieving, extracting, and selecting real-world health-related claims from snippets returned by a search engine for given questions that are asked online.

2. Evidence retrieval: automatically retrieving scientific articles relevant to these claims and then manually extracting evidence from their abstracts.

3. Claim verification: manually verifying whether the real-world claim is supported or refuted by the extracted evidence, or deciding that the information is insufficient to make a decision (i.e. neutral or irrelevant to the claim).

3.1 Claim retrieval

The claim retrieval stage aims to retrieve naturally-occurring claims from the Internet. To do so, we first used a set of most popular questions about COVID-19 asked by information seekers online, e.g., those captured in the TREC-COVID topics[2]. We targeted natural claims related to COVID-19 as the recent COVID-19 pandemic represents a good example of uncontrolled proliferation of false claims and stories which can cause serious consequences for consumer health (Barua et al., 2020). In addition to the questions that we collected, we used questions released by the TREC Health Misinformation Track[3], TREC-COVID and related questions (i.e. "*what people also ask*") generated by the Bing search engine. We have collected 80 questions, listed in Appendix B. We used this set of questions that ask about the origin, spread, prevention, diagnosis, and treatment of COVID-19 since most of the misinformation is related to these topics. For each of these selected questions, we retrieved the associated text snippets from the top-40 Bing search results using the Bing Web Search API[4] (a subscription key is needed to use this service). We did not set any restrictions regarding the source of the snippets, but most claims were found in news articles, blog posts, and social media.

Real-world claims were extracted from the returned snippets and validated manually by the an-

[2]ir.nist.gov/covidSubmit/data.html

[3]trec-health-misinfo.github.io

[4]api.cognitive.microsoft.com/bing/v7.0/search

notators (three authors of this paper who have expertise in biomedical NLP). The annotators were tasked with manually extracting claims related to the original questions. The extracted claims consist of general information about COVID-19 and do not contain any private information (e.g. "*Coronavirus may have originated in bats.*"). We define real-world claims as assertions that express facts without providing evidence. The claims could contain a single or multiple pieces of information from the text snippets. The claims could be either true, false, neutral, or irrelevant to the question. The average length of the collected claims is 19 words.

Figure 2 presents some examples of real-world claims related to the question "What is the origin of COVID-19?".

Question: What is the origin of COVID-19
Snippets:
- Coronavirus may have originated in bats or pangolins. The first known cases of COVID-19 were in Wuhan, China...
- February 18 A group of 27 prominent scientists outside China publishes a statement in The Lancet to condemn conspiracy theories suggesting that COVID-19 does not have a natural origin and ...
Claims:
- Coronavirus may have originated in bats or pangolins.
- The first known cases of COVID-19 were in Wuhan, China.
- COVID-19 does not have a natural origin.

Figure 2: Examples of health-related claims extracted from snippets returned by the Bing search engine.

3.2 Evidence retrieval

The evidence retrieval task aims to retrieve scientific evidence that could support or refute the health-related claims. To this end, we used the COVID-19 Open Research Dataset (CORD-19) as a source of scientific articles on COVID-19 (Wang et al., 2020), where the majority of papers are from PubMed Central. We used the CORD-19 2020-07-16 version to create this dataset. For each question, we first retrieved the relevant scientific articles from the CORD-19 collection using the BM25 model and the Terrier[5] search engine. We then re-ranked the top-1000 documents with the Text-to-Text Transfer Transformer (T5) (Raffel et al., 2020) relevance-based re-ranking model and selected the top-10 relevant articles. T5 was shown to be effective on newswire retrieval and MS MARCO (Nogueira et al., 2020). We fine-tuned the model on MS MARCO passage ranking dataset (Bajaj et al., 2018) by maximizing the log probability of generating the output token "true" when the document is relevant, and the token "false" when the document is not relevant to the query (Nogueira et al., 2020). Once fine-tuned, we first apply a softmax only on the logits of the "true" and "false" generated tokens, and then re-rank the documents using the probabilities of the "true" token. Table 1 presents the article retrieval performance, in terms of precision, recall, and NDCG, of the BM25 and T5 models on the TREC-COVID test set.

Models	P@10	R@10	NDCG@10
BM25	0.674	0.016	0.594
T5	0.796	0.018	0.742

Table 1: Article/evidence retrieval: Performance of the BM25 and T5 models on the TREC-COVID test set.

Re-ranking the search results considerably helped human annotators in finding the best evidence statements to verify the claims. In fact, in our preliminary analysis of the search results, we found that NEUTRAL examples were more frequent than REFUTES or SUPPORTS when using BM25 without re-ranking.

3.3 Claim verification

The annotators were given the collected claims and the associated top-10 abstracts from relevant documents and were asked to extract the evidence statements and label each evidence-claim pair as: SUPPORTS, REFUTES, or NEUTRAL. The label NEUTRAL was used if the evidence was neutral or irrelevant to the claim. The evidence statement could be complete or incomplete. It also could be a sentence, part of the sentence, or a passage. The annotators were asked to extract up to four evidence statements from each abstract for each claim. The title of the document could also be used as evidence. A single claim could be supported and refuted by different evidence statements. Figure 3 shows an example of a claim that is supported and refuted with different evidence statements.

For the claims that include multiple pieces of information, the SUPPORTS label was considered if the evidence supports one of them and the other pieces of information were neutral. Similarly, the REFUTES label was considered if the evidence refutes one of the multiple pieces of information. For NEUTRAL examples, we encouraged the annotators to select sentences that are relevant but do not contain enough information to make a decision. We did

[5] http://terrier.org/

Claim: Vitamin D Lowers Your Risk of COVID-19
Evidence 1: ecological investigation on 51 countries including 408,748 participants, analyses indicated no correlation between vitamin D levels and recovery rate (r= 0.041) as well as mortality rate (r=-0.073) globally (Ghasemian et al., 2021). [REFUTES]
Evidence 2: testing positive for COVID-19 was associated with increasing age (RR(age<50)=1.05, p<0.021; RR(age[]50)=1.02, p<0.064)), non-white race(RR=2.54,p<0.01) and being likely vitamin D deficient (deficient/treatment-not-increased: RR=1.77,p<0.02) as compared to likely vitamin D sufficient(not-deficient/treatment-not-decreased), with predicted COVID-19 rates in the vitamin D deficient group of 21.6]%(95%CI[14.0%-29.2%]) versus 12.2%(95%CI[8.9%-15.4%]) in the vitamin D sufficient group (Meltzer et al., 2020). [SUPPORTS]

Figure 3: Example of a claim that is supported and refuted by different evidence statements.

not set time restrictions for labeling the evidence-claim pairs, but the annotators spent on average less than 1 minute per evidence-claim pair. The average length of evidence statements is 38 words. Compared to the SCIFACT and FEVER datasets which do not include evidence for the NOINFO claims, we provide evidence statements for NEUTRAL/NOINFO examples in HEALTHVER. We provide such annotation since the results of veracity prediction change when using different selection strategies for claims labeled NOINFO (Thorne et al., 2018). NOINFO is used when there is not enough information to make a decision. In (Wadden et al., 2020), the authors observed that the evidence is found in abstracts for the majority of claims. Therefore, in this study, we decided to use the abstracts rather than full articles to verify the truthfulness of real-world claims. Figure 4 shows examples of supported and refuted claims and the associated evidence extracted from the abstracts of CORD-19 articles.

Abstract: https://www.biorxiv.org/content/10.1101/2020.05.12.091397v1
Evidence: Recent research results suggest that bats or pangolins might be the original hosts for the virus based on comparative studies using its genomic sequences.
Claims with labels:
- Coronavirus may have originated in bats or pangolins. [SUPPORTS]
- The first known cases of COVID-19 were in Wuhan, China. [NEUTRAL]
- COVID-19 does not have a natural origin. [REFUTES]

Figure 4: Examples of supported and refuted health-related claims and associated evidence extracted from relevant scientific articles.

4 Dataset Analysis

4.1 Inter-annotator agreement

Due to the complexity of labeling the claim-evidence pairs and following previous efforts (Thorne et al., 2018; Wadden et al., 2020), we only evaluated the agreement between annotators on label assignment. We randomly selected 603 claim-evidence pairs for re-annotation. We obtained a Cohen's Kappa of $k = 0.76$ (Cohen, 1968), which indicates that the inter-annotator reliability is satisfactory, as the obtained k of 0.76 is above the commonly applied criteria of .70; it is also comparable to the 0.75 Cohen's Kappa reported in Wadden et al. (2020).

4.2 Dataset statistics

Table 2 presents the main statistics with: (i) number and distribution of the claims are shown in Table 2a and (ii) number of questions, claims, and evidence statements are shown in Table 2b. We observed that a single evidence statement can support or refute different real-world claims for a given topic. Also, a single claim can be supported or refuted by different evidence statements. As shown in Table 2b, we identified 738 evidence statements for the 1,855 retrieved claims, which yields 14,330 evidence-claim pairs, presented in Table 2a. To split the dataset into training/validation/test sets and to guarantee all claims in the test and validation sets do not appear in the training set, we randomly selected 230 claims and their evidence statements for the test set, 230 claims and their associated evidence statements for the validation set, and the rest for the training set. We split the training/validation/test sets by claims rather than questions to have a balanced dataset class-wise. We selected approximately the same number of SUPPORTS, REFUTES, and NEUTRAL examples in the validation and test sets.

The size of the HEALTHVER dataset is large compared to SCIFACT, and approximately the same as the size of PUBHEALTH.

4.3 Claim complexity analysis

One of the characteristics of the dataset is the complexity of the claims. Complex claims are statements that include multiple pieces of information (facts). For instance, the claim "Dogs and cats can get covid, but cats are more susceptible to infection." contains three facts that need to be checked.

Our basic assumption is that a claim with more relations is a more complex claim To study the

complexity of the claims we rely on automatic relation extraction methods We use an ensemble method that computes the average number of relations extracted using three distinct methods:
1. Dependency-based open relation extraction in the form of (subject, relation, object) triples with OpenIE (Angeli et al., 2015).
2. A supervised BERT model (Soares et al., 2019) trained on abstract relation categories from the SemEval dataset (Hendrickx et al., 2010).
3. A count of all verbal phrases detected by the Stanford CoreNLP parser (Manning et al., 2014).

Each method has different characteristics and led to substantially different ranges in the number of extracted relations. For instance, the BERT model often recognizes multiple relation classes between the same subject and object entities, while OpenIE rarely does so.

To evaluate the relevance and reliability of the averaging ensemble, we annotated manually a subset of the top 370 claims ranked according to the average number of relations by assigning a (1) single fact, or (2) multiple facts label to each claim.

Figure 5 presents the results of this evaluation. We find that the relation averaging ensemble has a high precision in detecting complex claims, ranging between 75.07% for claims with an average of 4 detected relations to 100% for claims with a detected average of more than 10.67.

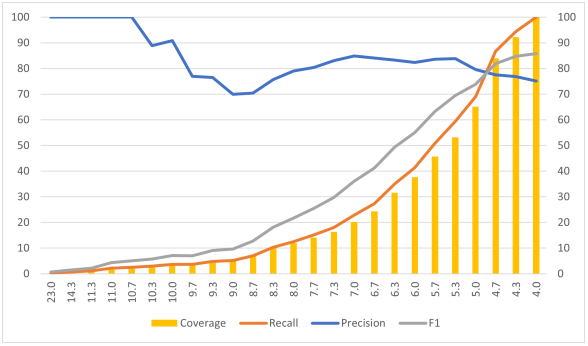

Figure 5: Evaluation of the relation averaging ensemble potential in detecting complex claims. Recall is computed according to the number of reference complex claims. Coverage is the total ratio of all claims that has a relation number greater than x.

Following these observations on the reliability of the averaging ensemble as a complexity indicator (for values exceeding 4.0), we use it to compare HEALTHVER with three existing misinformation datasets in figure 6. We find that HEALTHVER has consistently more complex claims in proportion than FEVER, PUBHEALTH, and SCIFACT, at all complexity levels.

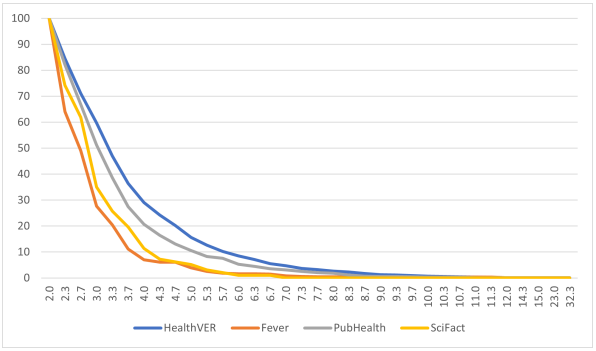

Figure 6: Distribution of Complex Claims in HEALTHVER, FEVER, PUBHEALTH, and SCIFACT according to the average relation number indicator.

N.B. There are three caveats to consider when performing automatic relation extraction to analyze claim complexity. First, some relations are peripheral information and do not actually add complexity to the claim. Second, automatic relation extraction methods are not able to differentiate between important/main relations and peripheral ones as that was not their objective. Third, these methods have both recall and precision errors when it comes to extracting relations from the main facts. Therefore, the aim of our relation extraction ensemble method was not to compute an absolute value of complexity for a given dataset, but rather to provide relative comparisons between dataset pairs.

5 Dataset validation

We validate HEALTHVER and its ability to support the fact-checking task compared to the existing datasets.

5.1 Baseline models

Given a pair (c, e), where c is the health-related claim accompanied by a scientific evidence e, fact-verification models are tasked with predicting a label $\hat{y}(c, e) \in \{\text{SUPPORTS}, \text{REFUTES}, \text{NEUTRAL}\}$.

We examined pretrained BERT (Devlin et al., 2019) as well as two variants trained on scientific and biomedical articles: SciBERT (Beltagy et al., 2019) and BioBERT (Lee et al., 2019). In addition to BERT models, we examined T5 (Raffel et al., 2020) for its state-of-the-art effectiveness. We train the models on claim-evidence pairs. Claims and evidence are concatenated and passed to the models to make the labeling decision. We minimize the

Set	Supports	Refutes	Neutral	Total
Training	3,782	2,411	4,397	10,590
Validation	533	391	993	1,917
Test	671	425	727	1,823
Total	4,986	3,227	6,117	14,330

(a) Claim labels distribution

#Questions	#Claims	#Evidence
80	1,855	738

(b) Evidence and claim counts.

Table 2: Statistics of the new HEALTHVER dataset.

cross-entropy loss between $\hat{y}(c, e)$ and the gold label $y(c, e)$.

For BERT-based models, we concatenate the claim c and its associated evidence e with $[SEP]$, add $[CLS]$ to the sequence $[c, SEP, e]$, and feed the input to BERT. The $[CLS]$ representation is fed into a softmax layer for a three-way classification.

For the T5-based model, the input sequence for the task is "Claim: [c] Evidence: [e] Target:". We fine-tuned T5 to generate the target tokens SUPPORTS, REFUTES, or NEUTRAL which are the ground truth labels. T5 is a sequence-to-sequence model that uses traditional transformer architecture and BERT's masked language modeling.

In our experiments, we used the BERT (base-uncased), SciBERT (scivocab-uncased), BioBERT (v1.0-pubmed-pmc), and T5-base implementations provided in HuggingFace's Transformers package version 2.10 (Wolf et al., 2020). All models were trained with a batch size of 16 and maximum sequence length of 512 tokens for 20 epochs using single P100 GPUs (16 GB VRAM) on a shared cluster. Adam optimiser with a learning rate of 1e-5 was used.

5.2 Evaluation metrics

Claim verification can be seen as a Natural Language Inference (NLI) task. Therefore, we consider the fact-checking task as a multi-class classification problem, as in previous efforts (Thorne et al., 2018; Wadden et al., 2020). For a given claim and its associated evidence, the models assign one of the following labels: SUPPORTS, REFUTES, and NEUTRAL. Macro precision, macro recall, macro F1-score, and accuracy have been used to evaluate the effectiveness of the models.

In addition, to compare the datasets with each other, we evaluate their pairwise zero-shot performance, $Z_{i,j}(m)$, according to an evaluation measure m, when a BERT-base model is fine-tuned on the training set of dataset i, and tested on the test set of dataset j, and the pairwise zero-shot delta as:

$$\Delta_{i,j}(m) = Z_{i,j}(m) - Z_{j,i}(m) \quad (1)$$

To adjust for the discrepancies in training set sizes, noted s_i and s_j, we also compute a size-adjusted delta as:

$$\Delta^a_{i,j}(m) = \Delta_{i,j}(m) \times e^{\delta_{i,j}(m)\frac{s_j - s_i}{s_i + s_j}} \quad (2)$$

with $\delta_{i,j}(m) = sign(\Delta_{i,j}(m))$. The rationale of $\Delta^a_{i,j}$ is that if a dataset i has a better pairwise zero-shot performance than a dataset j, $\Delta_{i,j}$ should be highlighted further if the training set of i is smaller than the training set of j, and highlighted less otherwise.

5.3 Results and discussion

In our experiments, we (1) examine the effect of different training datasets, (2) study the the generality of HEALTHVER, (3) present the results of baseline models, and (4) examine the effect of the model input on the performance of veracity prediction.

Table 3a presents the results of the BERT-base model fine-tuned on FEVER, SCIFACT, PUBHEALTH, and HEALTHVER, and evaluated on the HEALTHVER test set. The NOINFO label in FEVER and SCIFACT is equivalent to the NEUTRAL label in HEALTHVER. We considered the TRUE label as SUPPORTS, the FALSE and MIXTURE labels as REFUTES, and UNPROVEN as NEUTRAL when experimenting with PUBHEALTH. Although the FEVER dataset (145,449 training examples) is much larger than SCIFACT (809 training examples) and HEALTHVER, the F1-score shows that training on SCIFACT and HEALTHVER achieves better results than training on FEVER that is based on Wikipedia sentences. These results support our hypothesis that domain-specific fact verification benefits more from training on in-domain claims. The results also confirm that training the models on synthetic claims does not perform well on the real-world claims. On the other hand, the results

Training data	P	R	F1	Acc.
FEVER	40.63	45.14	36.26	40.59
SCIFACT	37.40	41.50	36.62	39.33
PUBHEALTH	35.33	34.03	25.70	28.63
HEALTHVER	73.45	73.70	73.54	74.82

(a) Results of the BERT-base model fine-tuned on each training set and evaluated on the HEALTHVER test set.

Test set	P	R	F1	Acc.
FEVER	65.71	65.80	60.97	65.77
SCIFACT	54.20	54.55	46.78	54.81
PUBHEALTH	36.85	32.87	29.37	38.57
HEALTHVER	73.45	73.70	73.54	74.82

(b) Results of the BERT-base model fine-tuned on the HEALTHVER training set and tested on each test set.

Table 3: Comparison of zero-shot transfer performance.

Dataset	HEALTHVER		FEVER		SCIFACT		PUBHEALTH	
	$\Delta_{i,j}$	$\Delta^a_{i,j}$	$\Delta_{i,j}$	$\Delta^a_{i,j}$	$\Delta_{i,j}$	$\Delta^a_{i,j}$	$\Delta_{i,j}$	$\Delta^a_{i,j}$
HEALTHVER	0.	0.	-25.18	-58.92	-15.48	-6.56	-9.94	-9.57
FEVER	**25.18**	**58.92**	0.	0.	-18.43	-6.86	0.39	0.92
SCIFACT	15.48	6.56	**18.43**	**6.86**	0.	0.	-0.29	-0.67
PUBHEALTH	**9.94**	**9.57**	-0.39	-0.92	0.29	0.67	0.	0.
Average	**16.86**	**25.01**	-2.38	-17.66	-11.20	-4.25	-3.28	-3.10

Table 4: Pairwise zero-shot accuracy deltas ($\Delta_{i,j}$) and size-adjusted accuracy deltas ($\Delta^a_{i,j}$) for all dataset pairs. Best results are highlighted in bold row-wise.

showed that training on PUBHEALTH leads to poor performance since it is an imbalanced dataset and its main goal is to evaluate the explanations for fact-checking prediction.

To evaluate the generalization potential of HEALTHVER, we tested the BERT-based model trained on HEALTHVER on the existing datasets, as shown in Table 3b. From Table 3a and Table 3b, we can observe that HEALTHVER generalizes better than the existing datasets. For instance, the BERT-based model trained on HEALTHVER and tested on FEVER dev set achieves 60.97% in F1-score, while the BERT-based model trained on FEVER and tested on HEALTHVER test set achieves 36.26% in terms of F1-score.

To investigate further this aspect, we compute the pairwise zero-shot accuracy deltas $\Delta_{i,j}(accuracy)$ and size-adjusted accuracy deltas $\Delta^a_{i,j}(accuracy)$ between all dataset pairs (cf. table 4). The results show that HEALTHVER generalizes substantially better out-of-the-box than all other datasets, with an average accuracy delta of +16.86, and +25.01 when adjusted for training set size, while the average zero-shot deltas for all other datasets was negative and ranged between -17.66 and -2.38. The substantially higher performance from HealthVER in pairwise generalization deltas could be due in part to the higher complexity of the claims. While a very high level of complexity can likely hurt generalization performance, we think that the moderately higher complexity in HealthVER improved the fine-tuning of the BERT transformer by a relevant broadening of the textual context.

Models	P	R	F1	Acc.
BERT-base	73.45	73.70	73.54	74.82
SciBERT	76.62	78.15	77.21	78.11
BioBERT	74.07	75.73	74.59	76.52
T5-base	**80.82**	**79.00**	**79.60**	**80.69**

Table 5: Claim verification: Results of baseline models on the HEALTHVER test set.

Model input	P	R	F1	Acc.
Claim-only	49.92	48.38	48.67	50.00

Table 6: The performance of the "claim-only" model trained and evaluated on HEALTHVER, using T5.

We also explore and compare different baseline models including BERT, SciBERT, BioBERT, and T5 trained and evaluated on HEALTHVER (cf. table 5) and the impact of training the best model on the claims only without the evidence (cf. table 6). The results show (1) that T5 has the best performance, (2) that performance could be improved by using in-domain BERT-based language models such as SciBERT and BioBERT, and (3) that performance drops substantially without the evidence. This indicates that there are no sufficient language cues in the claim text alone for a correct classification, and that the model needs access to the evidence statements to verify the claims.

Error analysis. We performed a manual analysis of the test set where the claim verification model predicted an incorrect label. Table 7 in Appendix A presents some examples. The error analysis has shown that evidence-claim pairs are mostly classi-

fied incorrectly if there is not a significant lexical overlap between the claim and the evidence (example #1). The model also misclassified the evidence-claim pairs due to wrong lexical or semantic relations such as COVID-19 vs. Coronavirus (example #3) and "Diabetes" vs. "Type 2 Diabetic" (example #7). Using different abbreviations in claims and evidence statements is also found to be a cause of error. For instance, the model was not able to interpret abbreviations such as ACEIs and ACE (example #2). We also find that stating multiple facts in a single claim was one of the challenges in verifying real-world claims. For instance, in examples #2, #5, and #6, the evidence verifies one fact of the claim and neutrally states the other facts. To verify such claims, the evidence should be extracted from multiple articles.

6 Conclusion

In this paper, we explored evidence-based fact-checking of real-world and health-related claims found on the Internet. To this end, we introduced HEALTHVER, a new evidence-based fact-checking dataset, to verify the veracity of real-world claims by evaluating their assertions against scientific articles. We analyzed the complexity of the claims in HEALTHVER using a relation extraction ensemble and compared its generalization potential with existing datasets using pairwise zero-shot accuracy deltas and a new measure that takes into account training set sizes. We found that the proportion of complex claims in HEALTHVER is consistently higher than in the existing datasets at all complexity levels. Our experiments showed that training fact-checking models on real-world claims improves the accuracy of these models compared to training on synthetic claims. The results also showed that training models on in-domain data substantially improves health-related claim verification accuracy compared to training on open-domain data. We believe that HEALTHVER will provide a realistic and challenging testbed for new evidence-based fact-checking systems for real-world claims.

Ethics Statement

Data collection process for the HEALTHVER dataset: Questions about COVID-19 were adapted from the TREC-COVID and TREC Health Misinformation Track questions and augmented with the related questions suggested by the Bing search engine. The claims were abstracted from the snippets returned by Bing in response to the above questions submitted as queries. The scientific evidence sentences were extracted from PubMed abstracts.

Biases and limitations: For various reasons, the collection could have some biases. For example, all TREC Health Misinformation Track questions are about natural remedies that might treat COVID-19. We mitigated this bias by using the TREC-COVID questions that were collected at different locations and across several groups of clinicians, patients, and researchers. There is also a bias due to the time in the pandemic during which the questions were collected. As the pandemic evolved, the focus of misinformation shifted from the origins and treatments of COVID-19 towards the effects of vaccination. The snippets that served as the source of the claims may be biased by the Bing search algorithm. The few sentences extracted from the PubMed abstracts were extracted and labeled manually, which is an inherently subjective task. For example, given the claim ID: 673,

- *Claim: COVID-19 Cases Drop in Warm Weather, But Not Much.*
- *Evidence: temperature is the most influential parameter that reduces the growth at the rate of 13-17 cases/day with a 1C rise in temperature.*

one annotator could infer that the rise in the temperature indicates the warm weather and label the evidence as supporting the claim, whereas another annotator might decide to label this evidence as neutral, since it does not state which cases are dropping, or refuting the claim, as the drop in the cases might be considered insignificant.

HEALTHVER shares the above limitations with other datasets that emulate average information seekers who search the online sources to answer their questions and take an additional step to find scientific evidence to verify facts. Using HEALTHVER in conjunction with the other datasets presented in this work should further mitigate the biases.

Acknowledgments

This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health.

References

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697, Hong Kong, China. Association for Computational Linguistics.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset.

Zapan Barua, Sajib Barua, Salma Aktar, Najma Kabir, and Mingze Li. 2020. Effects of misinformation on COVID-19 individual responses and recommendations for resilience of disastrous consequences of misinformation. *Progress in Disaster Science*, 8:100119.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: Pretrained language model for scientific text. In *EMNLP*.

Jacob Cohen. 1968. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Giovanni Da San Martino, Shaden Shaar, Yifan Zhang, Seunghak Yu, Alberto Barrón-Cedeño, and Preslav Nakov. 2020. Prta: A system to support the analysis of propaganda techniques in the news. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 287–293, Online. Association for Computational Linguistics.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, Vancouver, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Angela Fan, Aleksandra Piktus, Fabio Petroni, Guillaume Wenzek, Marzieh Saeidi, Andreas Vlachos, Antoine Bordes, and Sebastian Riedel. 2020. Generating fact checking briefs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7147–7161, Online. Association for Computational Linguistics.

Roya Ghasemian, Amir Shamshirian, Keyvan Heydari, Mohammad Malekan, Reza Alizadeh-Navaei, Mohammad Ali Ebrahimzadeh, Hamed Jafarpour, Arash Rezaei Shahmirzadi, Mehrdad Khodabandeh, Benyamin Seyfari, Alireza Motamedzadeh, Ehsan Dadgostar, Marzieh Aalinezhad, Meghdad Sedaghat, Nazanin Razzaghi, Bahman Zarandi, Anahita Asadi, Vahid Yaghoubi Naei, Reza Beheshti, Amirhossein Hessami, Soheil Azizi, Ali Reza Mohseni, and Danial Shamshirian. 2021. The role of vitamin d in the age of covid-19: A systematic review and meta-analysis. *medRxiv*.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 493–503, Hong Kong, China. Association for Computational Linguistics.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

David O. Meltzer, Thomas J. Best, Hui Zhang, Tamara Vokes, Vineet Arora, and Julian Solway. 2020. Association of vitamin d deficiency and treatment with COVID-19 incidence.

Rahul Mishra, Dhruv Gupta, and Markus Leippold. 2020. Generating fact checking summaries for web claims. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 81–90, Online. Association for Computational Linguistics.

Salman Bin Naeem, Rubina Bhatti, and Aqsa Khan. 2020. An exploration of how fake news is taking over social media and putting public health at risk. *Health Information & Libraries Journal*.

Preslav Nakov, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Wajdi Zaghouani, Pepa Atanasova, Spas Kyuchukov, and Giovanni Da San Martino. 2018. Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims. In *Lecture Notes in Computer Science*, pages 372–387. Springer International Publishing.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pre-trained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718, Online. Association for Computational Linguistics.

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

Lahari Poddar, Wynne Hsu, Mong Li Lee, and Shruti Subramaniyam. 2018. Predicting stances in twitter conversations for detecting veracity of rumors: A neural approach. In *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 65–72.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. *arXiv preprint arXiv:1906.03158*.

James Thorne, Max Glockner, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2021. Evidence-based verification for real world information needs.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2:*

Short Papers), pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Yifan Zhang, Giovanni Da San Martino, Alberto Barrón-Cedeño, Salvatore Romeo, Jisun An, Haewoon Kwak, Todor Staykovski, Israa Jaradat, Georgi Karadzhov, Ramy Baly, Kareem Darwish, James Glass, and Preslav Nakov. 2019. Tanbih: Get to know what you are reading. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 223–228, Hong Kong, China. Association for Computational Linguistics.

Appendix

A Error analysis

Table 7 presents some examples where the claim verification model predicted an incorrect label in the test set.

B Questions

1. Is dexamethasone effective for treating COVID-19?
2. Are Tylenol, Advil and Motrin effective and safe to take for COVID-19 symptoms?
3. Can favipiravir help treat COVID-19?
4. Can animals spread COVID-19?
5. will SARS-CoV2 infected people develop immunity?
6. Do antibiotics work against the coronavirus?
7. what are the mortality rates overall and in specific populations
8. Does a surgical mask help avoid COVID-19?
9. what kinds of complications related to COVID-19 are associated with diabetes
10. does hydroxychloroquine treat COVID-19?
11. Is there a vaccine for the coronavirus disease?
12. Does heat prevent COVID-19?
13. Can I take any vitamins or supplements to prevent COVID-19?
14. what is known about people that have COVID-19 without any symptoms?
15. Which are the first symptoms of the coronavirus disease?
16. which tests indicate severe covid infection?
17. what is the origin of COVID-19
18. are there any clinical trials available for the coronavirus
19. What does SARS-CoV-2 stand for?
20. What vaccine candidates are being tested for Covid-19?
21. Can 5G technology cause COVID-19?
22. How can I reduce the risk of getting COVID-19?
23. are there any drugs that work for SARS-CoV or SARS-CoV-2 in animals?
24. Can acetaminophen (Tylenol) treat the coronavirus disease?
25. what are the best masks for preventing infection by Covid-19?
26. touching a contaminated surface will not make you sick
27. Can drinking alcohol help in preventing COVID-19?
28. Does garlic protect against covid-19
29. has social distancing had an impact on slowing the spread of COVID-19?
30. Can smoking cannabis (weed) help in preventing COVID-19?
31. Where can I buy hand sanitizer and if I can't find it in the store, can I make my own?
32. Does having a weakened immune system increase your risk of illness from COVID-19?
33. Can pets get the coronavirus disease?

	Example
(1)	**Claim:** A report indicates that Acetaminophen (Tylenol) may be preferred over Ibuprofen (Advil) for coronavirus (fever). **Evidence:** Preliminary evidence suggests potential benefit with chloroquine or hydroxychloroquine. **Gold label:** NEUTRAL **Predicted label:** SUPPORTS
(2)	**Claim:** Evidence is currently lacking and it is too early to make robust conclusions on any link between use of angiotensin-converting enzyme (ACE) inhibitors and angiotensin II type-I receptor blockers with risk or severity of novel coronavirus disease 2019 (COVID-19) infection. **Evidence:** Patients who take ACEIs and ARBS may be at increased risk of severe disease outcomes due to SARS-CoV-2 infections. **Gold label:** REFUTES **Predicted label:** SUPPORTS
(3)	**Claim:** coronavirus is man-made. **Evidence:** This provides evidences strongly supporting scientific hypotheses that bats and pangolins are probable hosts for the COVID-19 virus. At the whole genome analysis level, our findings also indicate that bats are more likely the hosts for the COVID-19 virus than pangolins. **Gold label:** REFUTES **Predicted label:** NEUTRAL
(4)	**Claim:** Surgical Masks Stop Transmission Of COVID-19 From Symptomatic People. **Evidence:** Surgical mask partition for challenged index or nave hamsters significantly reduced transmission to 25% (6/24, P=0.018). Surgical mask partition for challenged index hamsters significantly reduced transmission to only 16.7% (2/12, P=0.019) of exposed nave hamsters. **Gold label:** REFUTES **Predicted label:** SUPPORTS
(5)	**Claim:** Ferrets can catch the coronavirus and might give it to other ferrets. But poultry and pigs don't appear to be at risk. **Evidence:** Experimental data showed ferrets and cats are highly susceptible to SARS-CoV-2 as infected by virus inoculation and can transmit the virus directly or indirectly by droplets or airborne route. **Gold label:** SUPPORTS **Predicted label:** REFUTES
(6)	**Claim:** No experts are remotely advocating for people to take up smoking to prevent COVID-19, but some researchers have theorized nicotine may be playing some role in keeping the virus at bay. **Evidence:** Cannabis smoking is linked with poor respiratory health, immunosuppression and multiple contaminants. Potential synergism between the two epidemics would represent a major public health convergence. Cigarettes were implicated with disease severity in Wuhan, China. **Gold label:** REFUTES **Predicted label:** SUPPORTS
(7)	**Claim:** People with Diabetes May Have Higher Risk for COVID-19. **Evidence:** Type 2 diabetic patients were more susceptible to COVID-19 than overall population, which might be associated with hyperglycemia and dyslipidemia. **Gold label:** SUPPORTS **Predicted label:** NEUTRAL

Table 7: Examples of HEALTHVER claims that were incorrectly classified by the BERT-based system.

34. Which organs are most affected by COVID-19?
35. Do COVID-19 and SARS-CoV-2 mean the same thing?
36. What are the possible symptoms of COVID-19 in children?
37. are patients taking Angiotensin-converting enzyme inhibitors (ACE) inhibitors at increased risk for COVID-19?
38. what hand sanitizers kill COVID-19?
39. are heart complications likely in patients with COVID-19?
40. Do antibodies make you immune to COVID-19?
41. Does UV light help in preventing covid-19?
42. How does the coronavirus differ from seasonal flu?
43. Can vitamin C treat COVID-19?
44. Can taking medication to lower fever, such as paracetamol (tylenol) and ibuprofen (advil) worsen COVID-19?
45. Can wearing masks help in preventing the spread of the coronavirus disease?
46. Are there natural remedies that will prevent me from getting infected with COVID-19?
47. What are some of the more severe symptoms of COVID-19?
48. what evidence is there for dexamethasone as a treatment for COVID-19?
49. what is a cytokine storm and how is it related to COVID-19?
50. How dangerous is COVID-19?
51. how do people die from the coronavirus?

52. how does the coronavirus respond to changes in the weather
53. Can drinking hot green tea help in preventing COVID-19?
54. Can coronaviruses spread from people to animals?
55. Can animals spread COVID-19 to people?
56. Can smoking help in preventing COVID-19?
57. What psychological effects could the COVID-19 pandemic cause?
58. How has the COVID-19 pandemic impacted violence in society, including violent crimes?
59. Are you immune to COVID-19 after recovering from it?
60. is remdesivir an effective treatment for COVID-19?
61. Is it safe to go outside during COVID-19 pandemic?
62. Are there any antiviral drugs to treat the coronavirus disease?
63. what are the early symptoms of COVID-19?
64. Can children get COVID-19?
65. can bcg vaccine cure covid-19
66. Can COVID-19 spread through food?
67. Is a headache sign of the coronavirus disease?
68. How has the COVID-19 pandemic impacted mental health?
69. How to stay mentally healthy during COVID-19 crisis?
70. what types of rapid testing for Covid-19 have been developed?
71. Does drinking lots of water help flush out COVID-19?
72. Does Vitamin D impact COVID-19 prevention and treatment?
73. How much impact do masks have on preventing the spread of the COVID-19?
74. When was the COVID-19 pandemic declared?
75. Can people recover from COVID-19?
76. what kinds of complications related to COVID-19 are associated with hypertension?
77. what are the health outcomes for children who contract COVID-19?
78. Can face masks protect me from the coronavirus disease?
79. Does Vitamin C impact COVID-19 prevention and treatment?
80. Can vinegar help in preventing COVID-19?