Exploring Open-Domain Fact Verification of Scientific Claims: A Comparative Analysis of Knowledge Sources

Anonymous ACL submission

Abstract

The increasing rate at which medical information and health claims are produced and shared online has highlighted the importance of efficient fact verification systems. The usual setting for this task in the literature assumes the documents containing the evidence for claims are already provided and annotated, or they operate over a limited corpus. While this helps improve the reading comprehension abilities of developed systems, it renders them unrealistic for real-world settings where knowledge sources with potentially millions of documents need to be queried to find relevant evidence. In this paper, we perform an array of experiments to test the performance of open-domain fact verification systems. We test the final verdict prediction of systems on four established datasets of biomedical and health-related claims in different settings. While keeping the evidence sentence selection and label prediction parts of the pipeline constant, document retrieval is performed over three common knowledge sources (PubMed, Wikipedia, Google) and using two different information retrieval techniques. We discuss the results, detect important challenges, outline common retrieval patterns, and provide promising future directions.

1 Introduction

The fast promulgation of knowledge in the digital world has made keeping track of information trustworthiness a challenging endeavor. In particular, health has become a popular talking point and brought with it an abundance of medical advice that permeate online resources (Swire-Thompson et al., 2020). A report by the Pew Research Center (Fox and Duggan, 2013) found that over one-third of American adults have searched the Internet for medical conditions and asked it medical questions before going to a medical professional. The sought information ranged from self-diagnosis to finding medications. Medical misinformation in the pandemic of COVID-19 has led people to turn to unproved and unsafe treatments or make harmful health-related decisions (Roozenbeek et al., 2020).

Automated solutions for fact verification based on Natural Language Processing (NLP) have emerged as a potential aid to help with bringing light into the information overload (Nakov et al., 2021). While most work in the automated fact-checking domain is concerned with claims related to politics, society, rumors, and general misinformation, there has been an increasing interest in fact-checking of scientific and biomedical claims (Kotonya and Toni, 2020; Wright et al., 2022b). The task of automated fact verification consists of retrieving evidence for a claim being checked and then predicting a veracity label based on the discovered evidence. The most common setting for this task either already provides the source document that will contain evidence for the claim or works over a limited, manually constructed collection of documents (Saakyan et al., 2021). While this is an important step in developing models capable of reading comprehension and detecting which spans provide evidence in a given context, this is not a realistic setting for automated fact verification systems deployed in the real world. In such a scenario, the documents containing evidence are not known and knowledge bases containing them can possibly contain millions of documents. Moreover, with the rise of medical assistants and conversational agents in healthcare, many users are turning to these systems as a source of health-related information and medical support (Valizadeh and Parde, 2022).

To address these research gaps, we perform an array of experiments that test the performance of NLP systems for fact verification in the open domain. In the experiments, we keep the parts of the fact verification pipeline concerned with evidence sentence selection and verdict prediction fixed, and vary the knowledge source being used and information retrieval techniques being deployed to

query the databases. Since the final goal of fact-verification systems is to provide a verdict on the correctness of a claim, we measure the usefulness of knowledge sources and retrieval techniques by looking at verdict prediction scores. For this purpose, we leverage four English datasets of biomedical and health claims that contain gold annotations stemming from domain experts. We use the veracity labels of claims in datasets as ground truth.

We opt for three large-scale knowledge sources: PubMed, the cardinal collection of biomedical research publications; Wikipedia, as the largest publicly curated encyclopedia of human knowledge; and Google search results (representing "the whole web"), which is a straightforward and intuitive way how users seek information. Finally, we perform a qualitative analysis of retrieved evidence for some interesting example claims, present the insights from results, and provide future directions.

Our contributions are as follows:

1. We test the claim verdict prediction performance of a fixed fact-verification system on four biomedical fact-checking datasets by using three different knowledge sources (PubMed, Wikipedia, Google Search).

2. We compare the final label prediction performance by retrieving evidence using different techniques (sparse retrieval with BM25 and semantic search with dense vectors).

3. We provide a qualitative error analysis of retrieved evidence for different types of claims and provide insights and future directions for open-domain fact verification.

2 Foundations

2.1 Pipeline for Automated Fact-Checking

The systems for automated fact-checking are usually modeled as a framework with three components, where each component is a well-established NLP task (Zeng et al., 2021). This framework is a three-component pipeline consisting of (1) document retrieval; (2) evidence selection; (3) verdict prediction. We are mostly concerned with how the document retrieval part affects the further entailment process. That is why we fix the evidence-sentence selection model and the entailment prediction model. This way, the quality of the data source and the retrieval technique are the most important variables being tested. Two of these subtasks, or even all three (Zhang et al., 2021), can be learned together in a joint system with a shared representation. For the sake of simplicity of testing, we choose a pipeline system that performs each task sequentially.

In document retrieval, given a corpus of n documents $D = \{d_1, d_2, ..., d_n\}$, the task is to select top k most relevant documents $g_1, ..., g_k$ with a function $w(c, d)$. After the documents are retrieved, the next step is to select evidence sentences that serve as rationale in making a decision regarding the claim's veracity. From m candidate sentences $s_1, s_2, ..., s_m$ comprising the selected documents, top j sentences are selected as evidence sentences $\vec{e} = e_1, e_2, ..., e_n$ with a function $z(c, s)$. Finally, a verdict prediction function is trained to predict $y(c, \vec{e}) \in \{\text{SUPPORTED}, \text{REFUTED}\}$.

Since we are focusing on testing the influence of the knowledge source on the final claim verdict prediction, we experiment with different knowledge sources D and retrieval functions $w(c, d)$. Other components of the pipeline are fixed to make a fair comparison. After testing the values of k and j with different values in the set of $\{1, 3, 5, 10, 20\}$, we set both to be 10 since it provided the best F1 performance and the best trade-off between covering enough content while not cluttering with too much noise. This means we retrieve the top 10 documents and then select the top 10 sentences from them. For $z(c, s)$, we select the model SPICED (Wright et al., 2022a), which is a sentence similarity model that catches paraphrases of scientific claims well and recently set state-of-the-art performance in evidence selection on a couple of scientific fact-verification datasets. For the verdict predictor $y(c, \vec{e})$, we choose the DeBERTa-v3 model (He et al., 2021), since it was shown to be an exceptionally powerful model for textual entailment recognition on the GLUE benchmark – we use a version additionally fine-tuned on various NLI datasets.[1] It should be noted that we use these two models out-of-the-box and do not fine-tune them on any of our datasets in any experiment. This is an intentional zero-shot setting that aims to verify the real-world situation of using a system on yet unknown claims.

2.2 Datasets

We choose four English datasets of biomedical and health claims, built for different purposes.

[1] https://huggingface.co/MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli

SciFact (Wadden et al., 2020) is a dataset of 1,109 claims (in test and dev set) which were expert-written from citation sentences found in biomedical research publication abstracts. These publications originate from PubMed, which is one of the databases queried in our paper.

PubMedQA (Jin et al., 2019) is a dataset of 1,000 labeled claims that were generated from abstract of biomedical papers originating from PubMed. Even though more of a question-answering dataset in nature, it also provides yes/no/maybe labels which make it usable as a fact-checking dataset.

HealthFC (Vladika et al., 2023) is a dataset of 750 claims concerning everyday health and spanning various topics like nutrition, immune system, mental health, and physical activity. The claims originate from user inquiries and they were checked by a team of medical experts using clinical trial reports and systematic reviews as the main evidence source. All the claim verdict explanations are described in a user-friendly language.

CoVert (Mohr et al., 2022) is a dataset of 300 claims related to health and medicine, which are all causative in nature (such as "*vaccines cause autism*"). All the claims originate from Twitter, which means some claims are written informally and thus make an additional challenge by providing a real-world scenario of misinformation checking.

For all of the datasets, we leave out any claims labeled with Not Enough Information (NEI) label. This is because some datasets do not include this label, and those that do include it define it differently. For SciFact, NEI means no evidence documents are present in their internal corpus. For HealthFC, NEI means no conclusive evidence for the claim was found in any clinical trials.

	Gold Evidence		
Dataset	**Precision**	**Recall**	**F1 Score**
SciFact	77.9	88.4	**82.8**
HealthFC	80.5	83.4	**81.9**
CoVert	80.7	86.4	**83.4**
PubMedQA	74.4	80.4	**77.3**

Table 1: Results of final verdict prediction over four datasets using the gold evidence sentences provided with the datasets.

3 Experiment Setup

3.1 Knowledge Sources

For testing on Wikipedia, we used the latest available dump of English Wikipedia that we found, from May 20th, 2023, containing 6.6 million articles.[2] For PubMed, the US National Library of Medicine provides MEDLINE, a snapshot of currently available abstracts in PubMed that is updated once a year. We used the 2022 version found at their website.[3] While this yields 33.4M abstracts, we pre-processed the data following González-Márquez et al. (2023) and removed non-English papers, papers with no abstracts, and papers with unfinished abstracts, which yields 20.6M abstracts. For Google results, we used Google's publicly available Custom Search JSON API.[4]

3.2 Document Retrieval Techniques

We test the performance of two different document retrieval techniques, namely a sparse one and a dense one. Since both types of approaches are deployed in modern search systems, we want to see how much of a difference they make in finding appropriate documents that can verify a claim. As a representative sparse technique, we opt for BM25, an improvement over TF-IDF that takes into account term frequency, document length, and inverse document frequency. Despite its simplicity, it has proven to be a cornerstone of information retrieval approaches due to comparative performance to more sophisticated neural approaches (Kamphuis et al., 2020).

Recently, with the advance of large language models, encoding both the claim and documents with dense vector embeddings and then searching most similar vectors with cosine similarity has proven to be a powerful retrieval method (Karpukhin et al., 2020). A particularly successful recent approach is SimCSE (Gao et al., 2021), which uses contrastive learning and entailment-based training to enhance similarity scoring. We chose a biomedical variation BioSimCSE (Kanakarajan et al., 2022) which fits our use case. For dense retrieval, we encode the entirety of our PubMed corpus and Wikipedia corpus with BioSimCSE and store the embeddings. For sparse

[2] https://dumps.wikimedia.org/enwiki/20230520/

[3] https://www.nlm.nih.gov/databases/download/pubmed_medline.html

[4] https://developers.google.com/custom-search/v1/overview

	BM25			Semantic		
Dataset	**Precision**	**Recall**	**F1 Macro**	**Precision**	**Recall**	**F1 Macro**
SciFact	79.9	72.6	76.1	73.7	80.0	**76.8**
PubMedQA	70.0	70.6	70.3	66.7	84.4	**74.5**
HealthFC	62.7	78.7	69.7	62.6	84.6	72.0
CoVert	76.0	83.3	79.5	75.6	76.8	76.2

Table 2: Results of final verdict prediction over four datasets using evidence retrieved from **PubMed**.

encoding, we construct an inverted index out of Wikipedia and PubMed corpora and later query it using BM25 metrics. After selecting the top 10 documents in each method, the top 10 most similar sentences were taken and jointly with claim the verdict was predicted based on the entailment relation. For Google Search, we took the top 10 returned Google snippets as "evidence sentences" that we then concatenate and use as the evidence block for label prediction. All the experiments were run on a single Nvidia V100 GPU card, in a single run.

3.3 Baseline

To establish a baseline, we first run the system with the gold evidence provided with each of the four datasets. These are the sentences or snippets that were given by the annotators or creators of the respective datasets. The performance is shown in Table 1. It should be noted that this performance is different from those reported in papers introducing these datasets because we remove the claims labeled with NEI and we also did not fine-tune the model on the datasets. This is an intentional choice because the idea is to test the systems in a zero-shot / off-the-shelf setting. We expect the results in our experiments to be lower than this because having annotated evidence is an easier setting than the open-domain fact verification where the evidence needs to be discovered.

4 Results

Tables 2 and 3 show the performance of the fact verification system using evidence retrieved with two different techniques from two different knowledge sources, PubMed and Wikipedia. As expected, the F1 scores are lower than the oracle setting of using gold evidence from Table 1. Still, they come remarkably close to it, taking into account the complexity of finding relevant documents in a sea of 6 and 20 million articles, and further selection of relevant sentences from them, to produce a final verdict. This indicates the open-domain setting is a promising endeavor in scientific fact verification.

In both tables the evidence from documents retrieved using semantic search outperformed the standard sparse metric BM25. Still, BM25 fares well compared to the relatively recent semantic approaches. It is also notable to observe in Table 2 that BM25 excels in precision more so than recall, always beating semantic retrieval in this metric. This is not too surprising considering BM25 relies on exact keyword matching and is better suited for this use case. While BM25 slightly beats the dense BioSimCSE in precision, it is significantly outperformed in recall in the first three datasets. In deeper analysis, as we will show in the next section, we saw the dense technique would more often retrieve articles talking about the claim content using alternate naming for diseases, which led to picking up more supporting arguments for positive claims. CoVert is the only dataset for which BM25 performed better in the PubMed setting, in both precision and recall. Considering the noisy nature of this dataset (tweets and informal language), the inverse document frequency (idf) feature of BM25 was better at finding exact matches for important but rare keywords mentioned in the tweets and ignoring the more common but unimportant words. On the other hand, the poorer performance of the dense technique could be because the embedding was swayed in vector space due to noisy irrelevant topics from tweets.

When looking at the performance of fact verification systems over Wikipedia in Table 3, it is once again apparent that dense retrieval found more relevant documents with better evidence and outperformed the sparse retrieval. Nevertheless, in this case, the precision of BM25 was worse than BioSimCSE. In general, recall in all settings was higher than the ones from PubMed and precision

	BM25			Semantic		
Dataset	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
SciFact	67.9	83.3	74.8	68.8	83.6	75.4
PubMedQA	65.3	83.0	73.1	68.3	78.5	73.2
HealthFC	62.9	87.4	73.1	65.2	92.6	**76.5**
CoVert	72.4	78.3	75.2	78.5	86.8	**82.5**

Table 3: Results of final verdict prediction over four datasets using evidence retrieved from **Wikipedia**.

lower, which shows better prediction of the positive (supported) class but also its over-prediction.

	Google Snippets		
Dataset	**Precision**	**Recall**	**F1 Score**
SciFact	75.5	91.5	82.7
PubMedQA	66.7	95.6	78.5
HealthFC	62.3	92.6	74.5
CoVert	76.4	68.7	72.3

Table 4: Results of final verdict prediction over four datasets using evidence retrieved from "the whole web" (using Google).

Another experiment consisted of querying "the whole web" in order to find relevant evidence for a verdict. This is a common setting explored as one of the straightforward baselines in some fact verification papers (Gupta and Srikumar, 2021; Hu et al., 2022) and it mimics how humans would begin the process of a claim checking. Table 4 reports on this performance. Considering the short nature of Google snippets, they usually do not actually provide "evidence" but commonly the verdict on the claim itself as reported on the source website containing the snippet. At first glance, the performance on this dataset seems impressive, especially considering that for the two toughest datasets, SciFact and PubMedQA, the performance is improved when compared to the first two tables. A more careful look reveals this to be an artefact of data leakage and the way these two datasets were constructed (a similar phenomenon already observed in fact-checking datasets, Glockner et al. (2022)). Considering that in both of them the claims originate from sentences actually contained in PubMed abstracts, Google Search is powerful enough to be able to find the exact sentence that was the origin of these claims. The two other datasets, HealthFC and CoVert, give a more realistic picture of the performance of Google snippets considering they contain organic claims that originated from online users. It is interesting to see that for these two datasets Google beats both settings of PubMed but succumbs to Wikipedia as the knowledge source. This can be attributed to the fact that the simple language of claims in these two datasets can be easier to verify with Google results like blogs and news portals, as opposed to the complex language found in PubMed research publications.

5 Discussion

In this section, we provide further insights and a deeper look into the performance of our pipeline for open-domain fact verification of scientific claims in large knowledge sources. We do this with a qualitative analysis where we looked at what kind of documents and sentences are retrieved from different knowledge sources with different retrieval techniques and outline some common patterns with representative examples.

5.1 Popular and Specialized Claims

The performance of Wikipedia and PubMed as a knowledge source is considerably close to each other when looking at Tables 2 and 3. Nonetheless, there are differences with respect to the claim's domain and popularity. It is evident from the tables that Wikipedia slightly outperformed PubMed for HealthFC, the dataset dealing with everyday consumer health questions, and CoVert, with social media claims related to the COVID-19 pandemic. The simple language in which these claims are posed (e.g., *Does regular consumption of coffee increase the risk of heart disease such as heart attack or stroke?* as opposed to *Omnivores produce less trimethylamine N-oxide from dietary l-carnitine than vegetarians*) corresponds to the more user-friendly language of Wikipedia, when compared to the often highly technical language of medical

Claim	Wikipedia (semantic)	PubMed (semantic)
Can regular intake of vitamin C prevent colds? (Refuted)	Nevertheless, given the consistent effect of vitamin C on the duration and severity of colds in the regular supplementation studies, and the low cost and safety, it may be worthwhile for common cold patients to test on an individual basis whether therapeutic vitamin C is beneficial for them. (Hemilä and Chalker, 2013) (Supported)	According to the most recently published **Cochrane review** on vitamin C and the common cold, one gram per day or more of vitamin C does not influence common cold incidence in the general community, i.e., it does not prevent colds. (en wiki: Vitamin C and the common cold) (Refuted)
Can lung cancer screening by computed tomography (CT) also do harm? (Supported)	Lung cancer screening with low dose computed tomography (ct) is the only method ever proven to reduce lung cancer-specific mortality in high-risk current and former cigarette smokers. We aim to explain why the risks associated with radiation exposure from lung cancer screening are very low and should not be used to avoid screening or dissuade... (Frank et al., 2013) (Reftued)	Low-dose CT screening has been associated with falsely positive test results which may result in unneeded treatment. In a **series of studies** assessing the frequence of false positive rates, results reported that rates ranged from 8-49%.(en wiki: Lung cancer screening) (Supported)
Can ginkgo extract relieve the symptoms of tinnitus? (Refuted)	Ginkgo biloba is a plant extract used to alleviate symptoms associated with cognitive deficits, e.g., decreased memory performance, lack of concentration, decreased alertness, tinnitus, and dizziness. Pharmacologic studies have shown that the therapeutic effect of ginkgo... (Søholm, 1998) (Supported)	Ginkgo leaf extract is commonly used as a dietary supplement, but there is no scientific evidence that it supports human health or is effective against any disease. **Systematic reviews** have shown there is no evidence for effectiveness of ginkgo in treating high blood pressure, menopause-related cognitive decline, tinnitus, post-stroke recovery, or altitude sickness. (en wiki: Gingko Bilboa) (Refuted)
The most prevalent adverse events to Semaglutide are gastrointestinal. (Supported)	We evaluated gastrointestinal (GI) adverse events (AEs) with once-weekly Semaglutide 2.4 mg in adults with overweight or obesity and their contribution to weight loss (WL). GI AEs were more common with semaglutide 2.4 mg than placebo, but typically mild-to-moderate and transient. (Wharton et al., 2022) (Supported)	Possible side effects include nausea, diarrhea, vomiting, constipation, abdominal pain, headache, fatigue, indigestion/heartburn, dizziness, abdominal distension, belching, hypoglycemia (low blood glucose) in patients with type 2 diabetes, flatulence, gastroenteritis, and gastroesophageal reflux disease (GERD) (en wiki: Semaglutide) (Refuted)
Macrolides protect against myocardial infarction. (Refuted)	Our findings indicate that macrolide antibiotics as a group are associated with a significant risk for MI but not for arrhythmia and cardiovascular mortality. (Gorelik et al., 2018) (Refuted)	Macrolides are a class of natural products that consist of a large macrocyclic lactone ring to which one or more deoxy sugars, usually cladinose and desosamine, may be attached. (en wiki: Macrolide) (Supported)

Table 5: Example claims and retrieved evidence from the two different knowledge bases, where only one of them provided a correct final verdict.

research found at PubMed.

Other than the simpler language of claims, another factor for using Wikipedia as a knowledge source is the claim's popularity and established research on it. Most claims in HealthFC are common health concerns people search for on the Internet. This means there is often systematic reviews done on these claims and Wikipedia encourages citing systematic reviews in its articles when available. We noticed our system often retrieved sentences mentioning reviews. Table 5 shows in first three rows how this led to the correct verdict prediction for Wikipedia, but incorrect for PubMed, since the PubMed retriever found standalone studies that might disagree from the research consensus. For 327 claims in HealthFC, combined evidence retrieved from Wikipedia mentions "systematic review" 89 times, while "Cochrane review"[5] is mentioned 60 times (for 1000 claims in PubMedQA, the number is 29 and 11). On the other hand, row 4 of Table 5 shows an example of evidence from Wikipedia being too broad and generalized, while row 5 shows a claim for which there was simply no relevant evidence in the Wikipedia article. For specialized claims concerning deeper medical knowledge or specific research hypotheses, PubMed is a superior knowledge base.

[5]Cochrane is an international organization formed to synthesize medical research findings.

5.2 Precision and Coverage

The comparison between the two retrieval techniques in Tables 2 and 3 shows that semantic search outperforms BM25 in all cases, except for CoVERT on Wikipedia. Considering that most systems from existing work on automated fact-checking use only BM25 in their pipelines, these results can motivate future research towards deploying semantic search with different sentence embedding models. Dense retrieval's ability to deal with synonyms and paraphrases is especially important in the medical field where numerous diseases, drugs, chemical compounds can have multiple names and symbols.

While semantic search provides higher coverage, BM25 offers better precision. Table 2 shows that for PubMed, using BM25 as a retrieval technique achieves higher precision for all datasets, with an especially high improvement for SciFact. The exact match of words posed in the query helps retrieving studies and documents that deal with concepts mentioned in the claim. When looking at Table 6, the first three rows show examples of claims for which dense retrieval got swayed into similar but irrelevant documents, while BM25 managed to un-

Claim	BM25 (PubMed)	Semantic (PubMed)
Do heat patches containing capsaicin help with neck pain? (Refuted)	The objective of this study was to evaluate the efficacy of a hydrogel patch containing capsaicin 0.1% compared with a placebo hydrogel patch without **capsaicin** to treat chronic myofascial **neck pain** (...) There was no significant difference between the two groups in any of the outcome measures. (**Refuted**)	In two randomized trials, a single 60-min application of the **capsaicin** 8% patch reduced pain scores significantly more than a low-concentration (0.04%) capsaicin control patch in patients **with PHN**. (**Supported**)
Does a herbal combination preparation of rosemary, lovage, and centaury effectively relieve symptoms of uncomplicated cystitis? (Supported)	The herbal medicinal product Canephron® N contains BNO 2103, a defined mixture of pulverized **rosemary** leaves, **centaury** herb, and **lovage** root(...) When given orally, BNO 2103 reduced inflammation and hyperalgesia in experimental cystitis in rats. (**Supported**)	Rosmarinus officinalis l., rosemary, is traditionally used to treat headache and improve cardiovascular disease partly due to its vasorelaxant activity, while the vasorelaxant ingredients remain unclear. (**Refuted**)
The extracellular domain of TMEM27 is cleaved in human beta cells. (Supported)	Here, we report the identification and characterization of transmembrane protein 27 (**TMEM27**, collectrin) in pancreatic beta cells. (**Supported**)	We also show that **TMEM2** is strongly expressed in endothelial cells in the subcapsular sinus of lymph nodes and in the liver sinusoid, two primary sites implicated in systemic HA turnover. (**Refuted**)
Normal expression of RUNX1 causes tumorsupressing effects. (Supported)	RUNX1 is a well characterized transcription factor essential for hematopoietic differentiation and RUNX1 mutations are the cause of leukemias. runx1 is highly expressed in normal epithelium of most glands and recently has been associated with solid tumors. (**Refuted**)	RUNX gene over-expression inhibits growth of primary cells but transforms cells with **tumor suppressor defects**, consistent with reported associations with tumor progression. (**Supported**)

Table 6: Example claims and retrieved evidence from PubMed, using the two different retrieval techniques, where only one of them provided a correct final verdict.

cover the correct ones. In the first row, capsaicin is mentioned in both, but only the one from BM25 is about neck pain. In the second row, the exact drug with specified ingredients is uncovered by BM25, while semantic search did not retrieve it. The third row shows an example of when an exact match can be important (TMEM27 vs. TMEM2). On the other hand, the fourth row shows an example of a common use case where semantic matching is beneficial – for this claim to be matched with BM25, "tumorsuppressing" and "effects" would have to be mentioned, but dense retrieval can catch paraphrases like "tumor suppressor defects".

5.3 Future Directions

Based on our findings and discussion, we see the future work could focus on these direction:

- **Modeling disagreement.** We observed how different studies and sources can come to differing conclusions regarding a claim. In this paper, we chose the majority vote among the top 10 documents as the final decision, but this diminishes the information about the prediction uncertainty. This is part of the broader ML problem of learning with disagreements (Leonardelli et al., 2023). The end users of fact-checking systems could appreciate the added interpretability of seeing the level of disagreement among different sources.

- **Assessing evidence quality.** When it comes to medical research articles from PubMed, not all of them hold the same weight, considering the research relevance. While it is hard to assess their validity of results automatically, modeling metadata aspects could give a hint on how to differently weight certain publications. Parameters such as the number of citations, the impact score and reputation of the source journal, the institutions of the authors could lead to more trustworthy results. Similarly, the sources of Wikipedia articles and Google search results contain website of differing reputation and credibility – filtering to trusted domains such as university websites or academic publishers could enhance the level of trust and performance (Kotonya and Toni, 2020). Lastly, temporal aspect (date of publication) is very important since research on certain topics advances and changes with time.

- **Retrieval-augmented generation for FV.** Modern generative large language models (LLMs) have shown the power to both exhibit reasoning capabilities and generate coherent text for users. They already possess learned medical knowledge in their internal weights, but are prone to hallucinations. Therefore, a promising research avenue is to amplify LLMs by passing the retrieved evidence passages from sources like Wikipedia and PubMed to them (Pan et al., 2023). How to effectively combine this, while balancing the trade-off of readability and factuality, is

an open challenge.

6 Related work

6.1 Fact-checking

The task of automated fact-checking refers to verifying the truthfulness of a given claim using background knowledge and relevant evidence (Guo et al., 2022). It is still mostly done manually by dedicated experts, but ongoing research efforts try to automate parts of it with NLP methods. For this purpose, many datasets have been constructed. They contain either synthetic claims generated from Wikipedia (Thorne et al., 2018; Schuster et al., 2021) or real-world claims found on portals dedicated to fact-checking of trending political and societal claims (Augenstein et al., 2019) or appearing in social media (Nielsen and McConville, 2022). This task is also increasingly concerned with assessing scientific claims, where most prominent domains are health (Sarrouti et al., 2021) and climate science (Diggelmann et al., 2021).

6.2 Open-domain fact verification

Fact verification is similar to the NLP task of question answering, where the goal is to either retrieve or generate an answer to a question based on discovered evidence (Rogers et al., 2023), and it can also be analyzed in a closed domain or open domain. In the closed domain, the evidence comes from an already provided source document. This setting is also called Machine Reading Comprehension (MRC) since the goal is to build models that are efficient in recognizing which parts of text correspond to a given query (Baradaran et al., 2022). In the open domain, only the final answer is known and it is the goal of a system to find appropriate evidence in a large corpus of documents or other type of resources (Chen and Yih, 2020).

There have also been efforts in open-domain scientific fact verification. Wadden et al. (2022) expand the corpus of evidence research documents for the dataset SciFact of biomedical claims, from the original 5k to about 500k. In such a setting, they discovered significant performance drops in F1 scores of final verdict predictions. This work analyzed only one knowledge source (biomedical abstracts) and focused on data annotation in such a setting, while our paper expands the research paper corpus even further to 20 million abstracts, and analyzes other knowledge sources and retrieval methods. In Pugachev et al. (2023), the authors take consumer-health question datasets and test the predictive performance of a system using PubMed and Wikipedia. A bigger focus was put on fine-tuning the models on different datasets and testing the efficiency of built-in searche engines of the respective databases. Furthermore, Stammbach et al. (2023) test the performance of six fact-checking datasets from different domains (including encyclopedic, political) using evidence retrieved from three different knowledge bases, while looking solely at one biomedical dataset and one retrieval technique (BM25). Also related is the work by Sauchuk et al. (2022), which shows the clear importance of the document-retrieval component of the fact-checking pipeline on the performance of the whole system.

To the best of our knowledge, our paper features the biggest corpora (using the entirety of available PubMed and Wikipedia dumps, with 20.6M and 6.6M articles), searches "the whole web", analyzes different retrieval techniques (BM25 and semantic), and analyzes datasets of different type and purpose: expert-geared research claims (SciFact and PubMedQA), and organic user-posed health claims (HealthFC and CoVert).

7 Conclusion

In this paper, we conducted a number of experiments assessing the performance of a fact-verification system in an open-domain setting. Moving away from the standard setup of a small evidence corpus, we expand the knowledge sources to two large document bases (PubMed and Wikipedia), searching the whole web via Google Search API, and experiment with two retrieval techniques (sparse and dense). We measured the verdict prediction performance over four established fact-checking datasets. Our results show that searching for evidence in the open domain provides satisfyingly high F1 performance, not far from the closed-domain setting, with a room for further improvement. We conclude that the knowledge source perform comparably, with Wikipedia being better for popular and trending claims and PubMed for technical inquiries. We demonstrate the general superiority of dense retrieval techniques, with examples of where it falls short and BM25 retrieval would be beneficial. We hope our research will encourage more exploration of the open-domain setting in the NLP fact-checking community and addressing real-world misinformation scenarios.

Limitations

In this study, we performed automatic assessment of claims related to medicine and health. These are two sensitive fields where misinformation, model hallucination, and incorrect evidence retrieval can lead to harmful consequences, misinformation spread, and societal effects. The automated scientific fact-verification system described in this work is still far from being safe and consistent for adoption in the real world, due to imperfect performance and drawbacks that arise. In case such an automated fact-verification system would be deployed and produce misleading verdicts, this could decrease the trust in the potential use and development of such solutions.

In our work, for easier comparison we disregard claims annotated with NOT ENOUGH INFORMATION due to different definitions of this label across different datasets and also absence of it in some datasets. This is an important label in fact verification, since not all claims can be conclusively assessed for their veracity. Future work should find a way to effectively include this label into model predictions. This is especially important in the scientific domain considering the constantly evolving and changing nature of scientific knowledge, and sometimes conflicting evidence from different research studies.

Lastly, the fact-checking pipeline used in this paper is a complex system with multiple factors – the choice of the retrieval method, of the sentence selection model, the *top k* value, the NLI model, and the prediction threshold. Some incorrect predictions could have come from, e.g., faulty entailment prediction of the NLI model or other factors that do not necessarily stem from the choice of the knowledge base. Still, we put strict attention to keeping all the factors constant and frozen, to ensure a comparable setup. We focused on reporting only those phenomena and patterns that we observed were commonly occurring, after a thorough analysis of retrieved evidence for each claim.

References

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.

Razieh Baradaran, Razieh Ghiasi, and Hossein Amirkhani. 2022. A survey on machine reading comprehension systems. *Natural Language Engineering*, 28(6):683–732.

Danqi Chen and Wen-tau Yih. 2020. Open-domain question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 34–37, Online. Association for Computational Linguistics.

Thomas Diggelmann, Jordan L. Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2021. Climate-fever: A dataset for verification of real-world climate claims. *ArXiv*, abs/2012.00614.

Susannah Fox and Maeve Duggan. 2013. Health online 2013. *Health*, 2013:1–55.

Luba Frank, Emmanuel Christodoulou, and Ella A Kazerooni. 2013. Radiation risk of lung cancer screening. *Semin. Respir. Crit. Care Med.*, 34(6):738–747.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Max Glockner, Yufang Hou, and Iryna Gurevych. 2022. Missing counter-evidence renders NLP fact-checking unrealistic for misinformation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5916–5936, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rita González-Márquez, Luca Schmidt, Benjamin M. Schmidt, Philipp Berens, and Dmitry Kobak. 2023. The landscape of biomedical research. *bioRxiv*.

Einat Gorelik, Reem Masarwa, Amichai Perlman, Victoria Rotshild, Mordechai Muszkat, and Ilan Matok. 2018. Systematic review, meta-analysis, and network meta-analysis of the cardiovascular safety of macrolides. *Antimicrob. Agents Chemother.*, 62(6).

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Ashim Gupta and Vivek Srikumar. 2021. X-fact: A new benchmark dataset for multilingual fact checking. In *Annual Meeting of the Association for Computational Linguistics*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Harri Hemilä and Elizabeth Chalker. 2013. Vitamin C for preventing and treating the common cold. *Cochrane Database Syst. Rev.*, 2013(1):CD000980.

Xuming Hu, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen, and Philip Yu. 2022. CHEF: A pilot Chinese dataset for evidence-based fact-checking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3362–3376, Seattle, United States. Association for Computational Linguistics.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. 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 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Chris Kamphuis, Arjen P de Vries, Leonid Boytsov, and Jimmy Lin. 2020. Which bm25 do you mean? a large-scale reproducibility study of scoring variants. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*, pages 28–34. Springer.

Kamal raj Kanakarajan, Bhuvana Kundumani, Abhijith Abraham, and Malaikannan Sankarasubbu. 2022. BioSimCSE: BioMedical sentence embeddings using contrastive learning. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 81–86, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

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.

Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. SemEval-2023 task 11: Learning with disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318, Toronto, Canada. Association for Computational Linguistics.

Isabelle Mohr, Amelie Wührl, and Roman Klinger. 2022. Covert: A corpus of fact-checked biomedical covid-19 tweets. In *Proceedings of the Language Resources and Evaluation Conference*, pages 244–257, Marseille, France. European Language Resources Association.

Preslav Nakov, David P. A. Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barr'on-Cedeno, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. In *International Joint Conference on Artificial Intelligence*.

Dan S Nielsen and Ryan McConville. 2022. Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3141–3153.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004, Toronto, Canada. Association for Computational Linguistics.

Alexander Pugachev, Ekaterina Artemova, Alexander Bondarenko, and Pavel Braslavski. 2023. Consumer health question answering using off-the-shelf components. In *European Conference on Information Retrieval*, pages 571–579. Springer.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *ACM Computing Surveys*, 55(10):1–45.

Jon Roozenbeek, Claudia R. Schneider, Sarah Dryhurst, John Kerr, Alexandra L. J. Freeman, Gabriel Recchia, Anne Marthe van der Bles, and Sander van der Linden. 2020. Susceptibility to misinformation about covid-19 around the world. *Royal Society Open Science*, 7(10):201199.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2116–2129, Online. Association for Computational Linguistics.

Mourad Sarrouti, Asma Ben Abacha, Yassine M'rabet, and Dina Demner-Fushman. 2021. Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512.

Artsiom Sauchuk, James Thorne, Alon Halevy, Nicola Tonellotto, and Fabrizio Silvestri. 2022. On the role

of relevance in natural language processing tasks. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1785–1789.

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.

B Søholm. 1998. Clinical improvement of memory and other cognitive functions by ginkgo biloba: review of relevant literature. *Adv. Ther.*, 15(1):54–65.

Dominik Stammbach, Boya Zhang, and Elliott Ash. 2023. The choice of textual knowledge base in automated claim checking. *ACM Journal of Data and Information Quality*, 15(1):1–22.

Briony Swire-Thompson, David Lazer, et al. 2020. Public health and online misinformation: challenges and recommendations. *Annu Rev Public Health*, 41(1):433–451.

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.

Mina Valizadeh and Natalie Parde. 2022. The ai doctor is in: A survey of task-oriented dialogue systems for healthcare applications. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6638–6660.

Juraj Vladika, Phillip Schneider, and Florian Matthes. 2023. Healthfc: A dataset of health claims for evidence-based medical fact-checking.

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.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. SciFact-open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sean Wharton, Salvatore Calanna, Melanie Davies, Dror Dicker, Bryan Goldman, Ildiko Lingvay, Ofri Mosenzon, Domenica M Rubino, Mette Thomsen, Thomas A Wadden, and Sue D Pedersen. 2022. Gastrointestinal tolerability of once-weekly semaglutide 2.4 mg in adults with overweight or obesity, and the relationship between gastrointestinal adverse events and weight loss. *Diabetes Obes. Metab.*, 24(1):94–105.

Dustin Wright, Jiaxin Pei, David Jurgens, and Isabelle Augenstein. 2022a. Modeling information change in science communication with semantically matched paraphrases. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1783–1807, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dustin Wright, David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Isabelle Augenstein, and Lucy Lu Wang. 2022b. Generating scientific claims for zero-shot scientific fact checking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2448–2460, Dublin, Ireland. Association for Computational Linguistics.

Xia Zeng, Amani S Abumansour, and Arkaitz Zubiaga. 2021. Automated fact-checking: A survey. *Language and Linguistics Compass*, 15(10):e12438.

Zhiwei Zhang, Jiyi Li, Fumiyo Fukumoto, and Yanming Ye. 2021. Abstract, rationale, stance: A joint model for scientific claim verification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3580–3586.

A Example Appendix

This is a section in the appendix.