Scientific Fact-Checking: A Survey of Resources and Approaches

Juraj Vladika and **Florian Matthes**
Department of Computer Science
Technical University of Munich
Garching, Germany
{juraj.vladika, matthes}@tum.de

Abstract

The task of fact-checking deals with assessing the veracity of factual claims based on credible evidence and background knowledge. In particular, scientific fact-checking is the variation of the task concerned with verifying claims rooted in scientific knowledge. This task has received significant attention due to the growing importance of scientific and health discussions on online platforms. Automated scientific fact-checking methods based on NLP can help combat the spread of misinformation, assist researchers in knowledge discovery, and help individuals understand new scientific breakthroughs. In this paper, we present a comprehensive survey of existing research in this emerging field and its related tasks. We provide a task description, discuss the construction process of existing datasets, and analyze proposed models and approaches. Based on our findings, we identify intriguing challenges and outline potential future directions to advance the field.

1 Introduction

In today's digital age, vast amounts of data are generated and new scientific breakthroughs achieved at a rapid pace. With millions of scientific articles being published annually, it has become increasingly challenging for researchers and the general public to stay informed about the latest developments and discoveries across various fields. On top of that, an especially challenging task for researchers is finding appropriate evidence for scientific claims and research hypotheses they are currently investigating. Exploring large academic databases and thoroughly examining scientific publications in them in order to verify specific facts is a time-consuming process. Automating the process of fact-checking scientific claims using methods based on Natural Language Processing (NLP) for knowledge exploration and evidence mining can greatly aid researchers in these efforts.

One way how the Internet has benefited society is by making scientific knowledge easily accessible, transferable, and searchable in a matter of seconds. Inevitably, this has introduced new risks and challenges – it has become difficult to discern reliable sources from dubious content. Many scientific claims found in online articles, social media posts, or news reports are not always trustworthy and backed by reliable evidence. Furthermore, not only are humans prone to creating inaccurate information – modern generative language models can also produce misleading text that sounds convincing. All of these factors, combined with the quick pace at which content is proliferated online, contribute to the spread of misinformation, which has negative societal consequences (West and Bergstrom, 2021).

Fact-checking is the task of assessing the veracity of factual claims appearing in written or spoken sources. It is traditionally performed manually by experts in journalism and dedicated applied fields. Automated fact-checking appeared as an approach where methods of Natural Language Processing (NLP) and Machine Learning (ML) are used to assist experts in making these decisions or completely automating the whole process (Nakov et al., 2021). Fact-checking becomes especially relevant during major political events like elections or referendums because of a sharp increase in deceptive and propagandist content. Most recently, the COVID-19 pandemic has brought the scientific discourse and the misinformation that comes with it into the spotlight. Medical misinformation is especially dangerous because it has influenced people to try unproven cures and treatments and make harmful health-related decisions. (Roozenbeek et al., 2020; Pennycook et al., 2020).

We define *scientific fact-checking* as a subset of the fact-checking task concerned with verifying the veracity of claims related to scientific knowledge. While the primary role of general fact-checking is

Findings of the Association for Computational Linguistics: ACL 2023, pages 6215–6230
July 9-14, 2023

to help detect misinformation and curb its spread, scientific fact-checking additionally aids scientists in testing their hypotheses and helps wider audiences contextualize new scientific findings. The most popular scientific domain in scientific NLP research is the biomedical domain (Rajpurkar et al., 2022), but insights learned from it can be generalized to other scientific domains. Scientific fact-checking can be performed both over the highly structured and complex language of science found in research publications and over the more easily understandable language found in news articles and online postings meant for lay audiences. Many scientists have decried the misinterpretation of their work when presented in the news press (Yavchitz et al., 2012), which makes scientific fact-checking even more relevant in bridging the gap between these two registers by performing an evidence-based assessment of scientific discoveries.

Considering the constantly increasing amount of misinformation in the digital era and the expanding number of scientific publications, the interest in developing automated fact-checking solutions and efficient resources for it is on the rise. We present this survey to systematize the existing work in this area. To the best of our knowledge, this is the first survey on fact-checking with a specific focus on the scientific domain. Our three main contributions are:

1. We describe existing datasets for scientific fact-checking, including their construction process and main characteristics.
2. We analyze the developed approaches and models for solving the task of scientific fact-checking, focusing on their components and design choices.
3. We outline general findings, identify challenges, and highlight promising future directions for this emergent task.

2 Task Definition

2.1 General Fact-checking

In general, *fact-checking* can be defined as the task of assessing whether a factual claim is valid based on evidence. It is a time-consuming task that is still usually performed manually by journalists. Automated approaches based on NLP have emerged to help assist humans in parts of the fact-checking process. Popular datasets used for benchmarking this task in NLP contain rewritten Wikipedia sentences as claims and annotated articles as evidence (Thorne et al., 2018; Jiang et al., 2020). For real-world settings, datasets were constructed by collecting claims and expert-written verdicts from dedicated fact-checking websites, such as PolitiFact (Vlachos and Riedel, 2014), Snopes (Hanselowski et al., 2019), or MultiFC (Augenstein et al., 2019) which draws from 26 fact-checking portals. This type of datasets usually contains claims currently trending in society, related to topics from world news, politics, media, or online rumors and hoaxes.

2.2 Scientific Fact-checking

We define *scientific fact-checking* as a variation of the fact-checking task that deals with assessing claims rooted in scientific knowledge. The dominant purpose of general fact-checking is to combat the spread of misinformation, while scientific fact-checking has the additional motive of helping scientists verify their research hypotheses, discover evidence, and facilitate scientific work. Scientific fact-checking comes with specific challenges not always present in general fact-checking, such as:

- **Claims:** Facts to be checked can be research hypotheses that scientists want to verify, claims made by everyday social media users, or queries posed to search engines dealing with scientific concepts (e.g., health-related concerns).
- **Evidence:** Scientific knowledge is constantly evolving when new research is conducted, which can make previous evidence obsolete and invalid. Moreover, different studies can come to diverging conclusions which complicates the final assessment of a claim. In clinical settings, this obstacle is facilitated by systematic reviews, which provide levels of evidence and strength of recommendations for any decision.
- **Domain:** The scientific language used in research publications is highly complex and contains domain-specific terminology, which presents a challenge for a general-purpose language model. This requires adapting the NLP systems to the scientific domain. On top of that, scientific text often contains relations between concepts spanning multiple sentences, which makes representation of the full context and long-text modeling an essential aspect.

- **Structure:** The highly structured nature of scientific knowledge makes it convenient to model it with structured representations like knowledge graphs, which can aid the fact-checking process. On the other hand, scientific publications commonly include different visualization techniques like tables, charts, and figures, all of which introduce additional multimodal challenges to verification.

These characteristics and other challenges with scientific fact-checking will be discussed in more detail in the following sections, especially in the Discussion section.

3 Related Tasks

In this section, we present tasks related to scientific fact-checking. We group them into three categories: (1) tasks related to misinformation detection; (2) retrieval of claims, arguments, and evidence from text; and (3) NLP tasks in the scientific domain.

3.1 Misinformation Detection

Since the principal function of fact-checking is to curb the spread of misinformation, it naturally belongs to a group of NLP tasks concerned with misinformation detection. Related tasks in this domain include fake news detection (Zhou and Zafarani, 2020), propaganda detection (Da San Martino et al., 2021), rumor detection (Bian et al., 2020), or stance detection (Hardalov et al., 2022). While most of these tasks deal with misinformation related to politics and society, recently, there has been an increase in scientific and health-related misinformation detection, especially pertaining to content related to the COVID-19 pandemic (Shahi and Nandini, 2020; Hossain et al., 2020; Antypas et al., 2021).

3.2 Claim Detection and Evidence Mining

A crucial prerequisite for automated fact-checking is devising methods that detect claims in the open domain. To achieve this, Yuan and Yu (2019) used a rule-based system to identify health claims in news headlines, while Wührl and Klinger (2021) develop a BERT-based model to detect biomedical claims in social media posts. After the claims are detected, an important next step is determining whether a claim is check-worthy since all claims are deemed relevant or interesting enough to be fact-checked. Check-worthiness for scientific claims was studied in the shared task CLEF-CheckThat! (Nakov et al., 2022) and by Zuo et al. (2022), where annotators helped construct a dataset of health-related claims from news articles.

Automatic gathering of evidence for scientific claims constitutes another line of research. There is work in this area focusing on humanities and social sciences (Stahlhut, 2021), although the majority of work we found is once again in life sciences. Numerous tools have been developed for searching PubMed, the largest database of biomedical publications (Lu, 2011), such as PubTator (Wei et al., 2019), Textpresso (Müller et al., 2018), LitSense (Allot et al., 2019), and EvidenceMiner (Wang et al., 2020). These methods usually look at the posed query (claim) and detect named entities, keywords, or metadata patterns to retrieve relevant results from the database. The end goal of this process is to help scientists gather evidence for their research, while in fact-checking, evidence retrieval is just one component of the whole process.

3.3 Scientific NLP Tasks

Scientific fact-checking belongs to a group of NLP tasks dealing with scientific text understanding. These tasks share a common challenge: working with highly complex scientific language and specific terminology. This has become even more apparent with the underwhelming performance of large language models, pre-trained on vast amounts of news data and web content, on NLP tasks in the scientific domain. Domain adaption is an essential cornerstone of modern NLP models working with specialized domains.

The task of Natural Language Inference (NLI), commonly equated with Recognizing Textual Entailment (RTE), is the task of inferring whether a premise entails or contradicts a given hypothesis. This task is a crucial component of automated fact-checking since predicting the final veracity of the claim is modeled entailment recognition between a claim and found evidence. For the scientific domain, datasets like MedNLI, which features medical claims rooted in the medical history of patients (Romanov and Shivade, 2018); SciNLI, which has claims from the domain of computational linguistics (Sadat and Caragea, 2022); and NLI4CT, with claims and evidence that originate from clinical trials reports of breast cancer patients (Vladika and Matthes, 2023).

Another knowledge-intensive NLP task related to fact-checking is question answering. In par-

Dataset	# Claims	Claim Origin	Evidence Source	Domain
SciFact (Wadden et al., 2020)	1,409	Researchers	Research papers	Biomedical
PubHealth (Kotonya and Toni, 2020b)	11,832	Fact-checkers	Fact-checking sites	Public health
Climate-FEVER (Diggelmann et al., 2020)	1,535	News articles	Wikipedia articles	Climate change
HealthVer (Sarrouti et al., 2021)	1,855	Search queries	Research papers	Health
COVID-Fact (Saakyan et al., 2021)	4,086	Reddit posts	Research, news	COVID-19
CoVERT (Mohr et al., 2022)	300	Twitter posts	Research, news	Biomedical

Table 1: Datasets for the task of scientific fact-checking and claim verification

ticular, open-domain question answering aims to find answers to given questions in unstructured textual corpora (Karpukhin et al., 2020), reminiscent of the process of finding relevant evidence for given claims in fact-checking. Popular datasets for biomedical QA are BioASQ (Tsatsaronis et al., 2015) and PubMedQA (Jin et al., 2019). Another important benchmark is BLURB (Biomedical Language Understanding and Reasoning Benchmark), introduced by Gu et al. (2022) to measure the performance of models in six different natural language understanding tasks over biomedical text. Finally, automated evidence synthesis is a task that aims to automate the process of creating systematic reviews for clinical trials (Brassey et al., 2021).

3.4 Related Surveys

There are already existing surveys that cover general automated fact-checking (Thorne and Vlachos, 2018; Zeng et al., 2021; Guo et al., 2022) by formalizing the task, outlining the most important datasets and proposed solutions, and discussing challenges. The survey by Kotonya and Toni (2020a) focuses on explainability methods in existing fact-checking approaches and present the most important explainability aspects these systems should satisfy. The survey by Bekoulis et al. (2021) focuses on approaches for tackling FEVER, the most popular dataset for fact verification (Thorne et al., 2018).

4 Datasets

In this section, we outline the existing datasets for scientific fact-checking that we found in the literature. The discovery process started with querying the well-known databases ACL Anthology,[1] IEEE Explore,[2] and ACM Digital Library[3] with the search string *("scientific" OR "biomedical") AND ("fact checking" OR "fact verification" OR "claim verification").* Retrieved articles were collected and the list was further expanded with any cited or citing paper from the initial batch of articles, according to Semantic Scholar.[4] In order for a dataset to be considered a fact-checking dataset, we stipulate it needs to provide claims, evidence (either documents or sentences), and final veracity labels. Such a dataset enables both the task of evidence retrieval and verdict prediction. This is important because the end goal of many automated fact-checking systems is to emulate the work of experts, where both seeking the evidence and making conclusions based on them constitute the process. This requirement narrowed the final list to the datasets summarized in Table 1. In the remainder of the section, we will describe the process and challenges related to constructing datasets.

4.1 Claim Creation

The starting point in the dataset construction process is collecting the claims that will later be factcheked. Claims in fact-checking are usually divided into synthetic, referring to claims written by annotators (e.g., by modifying sentences from Wikipedia), and natural, which are those claims crawled from real-world sources like fact-checking sites or social media posts. The first type of claims end up being fluent, atomic, and decontextualized, which is very appropriate for processing by NLP models (Wright et al., 2022b). Other authors focus on more organic and noisy claims found in online

[1] https://aclanthology.org/

[2] https://ieeexplore.ieee.org/

[3] https://dl.acm.org/

[4] https://www.semanticscholar.org//

posts since such claims are usually relevant and interesting to be fact-checked automatically (Mohr et al., 2022).

One common approach is to take original sentences from an appropriate source and have annotators reformulate them to a cleaner form. The dataset SCIFACT features biomedical claims that originate from human-written citation sentences in research articles but with the final form rewritten by annotators to make them more atomic and easily processed. Similarly, claims in HEALTHVER originate from Bing snippets of the most-searched user queries related to health and COVID-19, eventually reformulated by annotators. In the same vein, CLIMATE-FEVER contains sentences related to climate change extracted from online blogs and news websites, rewritten by annotators.

The remaining datasets from Table 1 relied on completely automatically retrieving claims. PUBHEALTH used news titles from fact-checking articles related to public health as its claims. This assumption works in many cases where titles are indeed factual claims, but some examples in the final dataset are generic titles with no relevance for fact-checking. COVID-FACT scraped claims from posts of a highly moderated subreddit *r/COVID19*, where users were already required to make atomic claims in their post titles. They also automatically constructed all of their negative (refuted) claims with word in-filling from masked language models, which ended with some unusable examples. Finally, COVERT is the only dataset in the list that features completely organic claims found in Twitter posts. They used a biomedical claim detection model Wührl and Klinger (2021) to extract claims that feature a causative relation and also included mentions of any biomedical entities.

4.2 Evidence Set Construction

Once the claims are collected, the next step is pairing them with appropriate evidence that addresses their veracity. The evidence source are often scientific publications, featuring highly complex and structured scientific language, or more easily understandable sources like news articles and Wikipedia articles. While working with text from scientific publications is more challenging both for humans and NLP models alike, they provide more rigorous scientific evidence. On the other hand, the general-purpose text provides evidence in a more explainable and intuitive form to a wider audience.

The SCIFACT dataset pairs the claims with abstracts of those scientific publications where they originated from, adding distractor abstracts to make detecting appropriate evidence more challenging. Likewise, claims in HEALTHVER are also mapped to appropriate scientific publications found by the annotators. Datasets COVID-FACT and COVERT feature a combination of both scientific publications and news articles as their evidence source, while PUBHEALTH uses solely the web articles from fact-checking websites where their claims originated from. In the same way as the original FEVER dataset, CLIMATE-FEVER uses Wikipedia articles as its evidence source.

4.3 Class Labels

Another integral component of dataset construction is labeling the claims with appropriate veracity labels. Following the tradition set by the FEVER dataset (Thorne et al., 2018), most of the datasets include three labels: SUPPORTED, REFUTED, and NOT ENOUGH INFORMATION (NEI). The definition of the NEI label has a different meaning in different datasets. In SCIFACT , this label refers to those claims for which none of the candidate abstracts contain suitable evidence to make a decision. In other datasets, it refers to the case where relevant evidence itself implies or states that there is currently not enough information to make a reliable and informed conclusion about the claim's veracity. Additionally, the dataset PUBHEALTH is the only one to feature a MIXED label, a label denoting a claim that consists of multiple factual statements with opposite veracity labels.

5 Approaches

In this section, we describe different modeling approaches devised for the task of scientific fact-checking. The standard framework usually consists of three major components that can all be modeled as well-established NLP tasks: document retrieval, evidence (rationale) selection, and verdict prediction (Zeng et al., 2021). This framework is visualized in Figure 1.

Table 2 summarizes the models we found in the literature, developed for the scientific fact-checking datasets from the previous chapter, with three framework components in each of them highlighted. While the most common approach is building separate models for each element and applying them in a pipeline, the best-performing systems

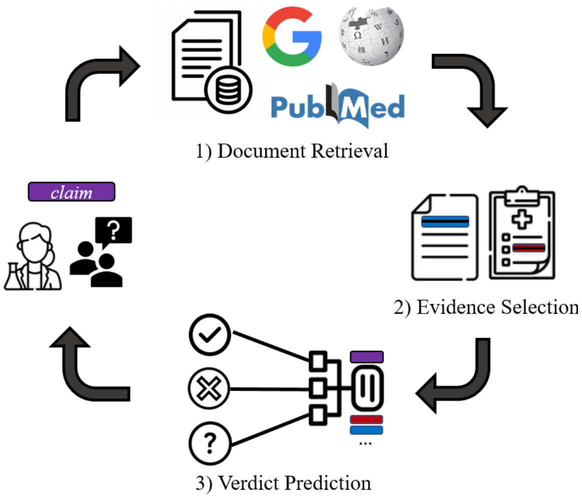

Figure 1: The standard three components of the framework for automated scientific fact-checking

jointly learn the rationale selection and verdict prediction with a shared representation. The dataset SciFact has the most models developed for it, partly owing to the shared task SCIVER (Wadden and Lo, 2021). For some of the datasets, we did not find dedicated models other than baselines provided in their originating papers. We analyze each part of the framework in more detail.

5.1 Document Retrieval

Given a corpus of documents that serve as the knowledge source, document retrieval is concerned with retrieving the relevant documents that might contain evidence related to the claim. It is usually solved with approaches typical for Information Retrieval. These can be separated into sparse retrieval and dense retrieval approaches. Sparse retrieval uses ranking functions such as TF-IDF and BM25, which match exact keywords in a query with an inverted index of all words from the corpus. Conversely, dense retrieval deploys dense vector representations of queries, which consider the semantic meaning of the query and can catch synonyms and related concepts (Karpukhin et al., 2020).

For SciFact, the document retrieval task focuses on retrieving relevant abstracts from a corpus of around 5 thousand given scientific abstracts. The baseline model VeriSci uses the simple TF-IDF metric to retrieve top *k* relevant abstracts. The models VerT5erini and later MultiVerS use the approach of first retrieving top *k* relevant abstracts using the BM25 metric and then adjusting the rankings using a T5 (Raffel et al., 2020) neural pointwise re-ranker based on (Nogueira et al., 2020), which is trained on the MS MARCO passage dataset used for machine reading comprehension (Nguyen et al., 2016). On the other hand, ParagraphJoint and ARSJoint used the dense vector representation BioSentVec (Chen et al., 2019), which was trained from 30 million biomedical publications and clinical notes.

Searching for evidence in a small corpus of documents (5k in SciFact) is useful for experimental settings but not realistic for real-world settings where large databases with millions of scientific publications have to potentially be queried to find appropriate evidence. When expanding document retrieval for SciFact to 500k documents in (Wadden et al., 2022a) and using the same BM25 + T5 re-ranking approach, the authors noticed performance drops of at least 15 points in the final F1 score of veracity prediction. This shows the need for a more precise semantic search of evidence documents. The authors of COVID-Fact tackle this by using snippets of the top 10 results returned by Google Search API for a given claim. This mimics how humans would approach fact-checking, but usually, additional verification of source quality and trustworthiness is needed in such an approach.

5.2 Evidence Selection

Evidence selection is the task of selecting relevant rationale sentences from the previously retrieved documents to be used as evidence for claim veracity prediction in the next step. Even though this step can be modeled as a span detection task, evidence is usually modeled at a sentence level. It can then be taken as a binary classification task of predicting whether a sentence is relevant or irrelevant. Most commonly, top *k* sentences are selected, similarly to the document retrieval step.

A common approach to evidence selection is to deploy models for sentence similarity and take those sentences that are the most similar to the claim being checked. The baselines for PubHealth and COVID-Fact both use the Sentence-BERT model (Reimers and Gurevych, 2019) to retrieve the top 5 most similar sentences. Sentence-BERT is a model based on siamese networks and provides semantically rich sentence embeddings that can easily be compared using cosine-similarity. VerT5erini uses a T5 model fine-tuned on MS MARCO (same as in the previous step) for this task.

While using sentence similarity for evidence selection is a straightforward and intuitive approach,

Dataset	Model	Document Retrieval	Rationale Selection	Verdict Prediction	Result (F1)
SciFact	VeriSci (Wadden et al., 2020)	TF-IDF	BERT	BERT	0.395
	ParagraphJoint (Li et al., 2021)	BioSentVec	BERT + MLP / BERT + KGAT	BERT + MLP	0.609
	VerT5erini (Pradeep et al., 2021)	BM25 + T5 re-ranker (tuned on MS MARCO)	T5 (tuned on MS MARCO)	T5 (no fine-tuning)	0.634
	ARSJoint (Zhang et al., 2021)	BioSentVec	BioBERT, MLP	BioBERT, MLP	0.655
	MultiVerS (Wadden et al., 2022b)	BM25 + T5 re-ranker	Longformer (binary head)	Longformer (ternary head)	0.672
CoVERT	Zero-shot MultiVerS (Wührl and Klinger, 2022)	BM25 + T5 re-ranker	Longformer (binary head)	Longformer (ternary head)	0.620
PubHealth	Baseline (Kotonya and Toni, 2020b)	provided	Sentence-BERT	SciBERT	0.705
Climate-FEVER	ClimateBERT (Webersinke et al., 2021)	provided	provided	ClimateBERT	0.757
HealthVer	Baseline (Sarrouti et al., 2021)	provided	provided	T5-base	0.796
COVID-Fact	Baseline (Saakyan et al., 2021)	Google Search	Sentence-BERT	RoBERTa (fine-tuned on GLUE)	0.820

Table 2: Models developed for scientific fact-checking with three pipeline components and verdict prediction performance on their respective dataset

it can fall short because evidence sentences could be paraphrased or use rather different wording from the original claim. Consequently, Wright et al. (2022a) improve the performance of evidence selection on CoVERT and COVID-Fact datasets by fine-tuning sentence similarity models on pairs of sentences about scientific findings from scientific articles matched with paraphrased sentences from news and social media reporting on these findings.

In all mentioned approaches, evidence selection and verdict prediction are made with two separate models, which means that the final claim veracity predictor might not have knowledge of the full context of evidence. ParagraphJoint, ARSJoint, and MultiVerS are so-called joint models because they all use multi-task learning to jointly learn the tasks of rationale selection and verdict prediction. For this purpose, they use a shared representation of the claim and the abstract obtained by concatenating the claim with the full abstract of a candidate document and converting it to a dense representation. This alleviates the problem of missing context during final label prediction. ParagraphJoint uses BERT (Devlin et al., 2019) as the encoder model, while ARSJoint uses the domain-specific BioBERT model (Lee et al., 2020), pre-trained on the text of biomedical research publications. Evidence selection is performed by passing the representation of each candidate sentence (extracted from the full abstract representation) to a multi-layer perceptron (MLP) classifier. Likewise, MultiVerS obtains the joint claim-abstract representations and perform rationale selection with the Longformer model (Beltagy et al., 2020), a transformer model for long documents that takes up to 4096 tokens.

5.3 Verdict Prediction

The final step of the fact-checking pipeline is for a model to produce the verdict on a given claim's veracity. As mentioned in the datasets section, the most common setting is to have three labels (Supported, Refuted, Not Enough Information), although models developed for one set of labels can be adapted to a dataset with a different set of labels. This component can easily be modeled as a classification task where the classifier learns to predict one of the three classes. All the baselines from Table 2 perform this task by fine-tuning large language models for label prediction on their respective datasets. The base models used include the general-purpose BERT or T5 and the domain-specific BioBERT and SciBERT (Beltagy et al., 2019) models. These models receive as their input pairs of claims and accompanying rationale sentences selected in the previous step and then give the final output as output.

As described previously, the joint models developed for solving SCIFACT use multi-task learning to learn both the evidence selection and verdict prediction steps with a shared claim+abstract representation. Both ParagraphJoint and ARSJoint again use a dedicated MLP that takes the previous step's representation. At the same time, ParagraphJoint also experimented with Kernel Graph Attention Network (KGAT), which performed well for general fact-checking datasets by learning relations between evidence sentences using a graph structure (Liu et al., 2020). MultiVerS once again uses the Longformer model, this time with a three-way classification head over the encoding of the entire claim and rationale sentences.

The MultiVerS model was also used in a zero-shot setting by Wührl and Klinger (2022) to fact-check the COVERT dataset. Since this dataset consists of tweets and is pretty noisy when compared to expert-written claims found in SCIFACT , the authors transformed the tweets into atomic claims consisting of triples (entity, cause, entity). Such a representation significantly improved the performance on this dataset and showed that models developed for one scientific fact-checking dataset can provide promising results for other datasets when the claims are represented in an appropriate form.

6 Discussion

In this chapter, we discuss the current challenges in scientific fact-checking and provide directions for future work and trends.

Evidence quality. A common challenge in fact-checking is ensuring that the evidence used for making veracity decisions is appropriate and of high quality. Especially in scientific fact-checking, the nature of scientific knowledge is such that it is updated and readjusted as new discoveries appear, so a claim that was once refuted by evidence could become supported with more substantial, more recent evidence. Time-aware scoring for evidence ranking was explored for general fact-checking (Allein et al., 2021). Additionally, scientific sources can contradict one another and give differing results for the same research hypotheses, which is related to the ML concept of learning with label disagreement (Uma et al., 2021). In the medical field, systematic reviews provide evidence-based clinical recommendations with the level of evidence (how much testing was performed) and the strength of recommendation (is it just a hint or a strict medical recommendation) (Cro et al., 2020). So far, none of the datasets have taken into account the evidence that is changing with time, disagreeing evidence, or differing levels and strength of evidence. A promising research direction is constructing resources and benchmarks that would consider these intricacies of scientific fact-checking.

Reasoning and Explainability. Fact-checking is one of the NLP tasks where making the models and their decision process transparent and explainable to humans is of high importance for their wide-scale adoption (Augenstein, 2021). Modern deep neural models for NLP tasks are generally described as black-box models, and their inner workings are still hard to grasp completely. While there have been explainable approaches for general fact-checking, the only explainable method in this survey was proposed by Kotonya and Toni (2020b). It uses a combination of extractive and abstractive text summarization of evidence source documents to provide end users with a concise explanation of why a certain verdict was produced. Considering that scientists often present their thoughts with argumentative structures (Lauscher et al., 2018), a promising research approach is learning the conceptual relations between multiple pieces of evidence to come up with a conclusion. This was used by Krishna et al. (2022) to develop a neuro-symbolic model that learns logical relations between evidence sentences for FEVER. Another promising research avenue is using counterfactual explanations, which have proven useful in many NLP tasks (Keane et al., 2021).

Dataset size. A common obstacle in fact-checking for all domains and related misinformation detection tasks is the small size of existing datasets. One way to overcome this performance hindrance is combining multiple scientific fact-checking datasets or datasets for related NLP tasks that deal with seeking rationale in text. The model MultiVerS described in the previous chapter utilized this approach by combining datasets HEALTHVER, COVID-FACT, and SCIFACT together with FEVER, PubMedQA, and EvidenceInference datasets to improve the final performance on the fact-checking task. Other than combining datasets for training purposes, another emerging approach to mitigate the lack of training data is generating new scientific claims to augment the existing data. Wright et al. (2022b) apply this approach by using the generative model BART and external biomedical knowledge sources to construct claims while showing promising zero-shot performance.

External knowledge. Scientific knowledge is complex and contains lots of interconnected concepts. This makes it suitable for representation with structures like Knowledge Graphs (KGs) that model world knowledge in the form of entities and relations between them. KGs have been constructed for various scientific disciplines, while the most well-known one for biomedical knowledge is Unified Medical Language System (UMLS) (Bodenreider, 2004), which models various interactions between proteins, drugs, diseases, genes, and other concepts. KGs have proven useful in enhancing a wide array of NLP tasks (Schneider et al., 2022). Enhancing BERT with infused disease knowledge from MeSH (He et al., 2020b) and structured medical knowledge from UMLS (He et al., 2020a) showed improved performance over knowledge-intensive biomedical NLP tasks, as well as for the open-domain question answering (Yu et al., 2021). Recent work has shown that reasoning over knowledge graphs can improve encyclopedic fact verification (Kim et al., 2023).

Multimodality and multilinguality. Misinformation is increasingly being spread in forms other than text, including misleading images, artificially constructed videos, or incorrect figures (Nielsen and McConville, 2022). Visuals were an especially popular tool for spreading misinformation about the COVID-19 pandemic (Brennen et al., 2020). Particularly in scientific publications, authors present their data in the forms of figures, tables, and other visualizations. The FEVEROUS shared task (Aly et al., 2021) made progress in this direction by requiring participants to develop systems that verify claims over evidence in the structured format (tables and lists). Other than multiple modalities, online claims are made in a multitude of world languages, which calls for the development of efficient multilingual models for scientific fact-checking.

Human-centered fact-checking. Most of the developed fact-checking systems are still limited in practical use because their system design often does not take into account how fact-checking is done in the real world (Glockner et al., 2022) and ignores the insights and needs of various stakeholder groups core to the fact-checking process (Juneja and Mitra, 2022). Several works started to investigate human evaluation in fact-checking systems. Examples include effectively delivering the misinformation detection results to users (Seo et al., 2019) or guiding the user toward fact-checked news (Lo et al., 2022). Making the process of NLP-based fact-checking more human-centered is a promising future direction that will make it more reliable, trustworthy, and easier for wide-scale adoption.

7 Conclusion

In this survey, we reviewed and systematized existing datasets and solutions for the task of scientific fact-checking. We introduced the task and compared it to its related NLP endeavors, described the existing datasets and their construction process, and explained the models used for scientific fact-checking with their pipeline components. Finally, we provided a critical discussion of current challenges and highlighted promising future directions for the task of scientific fact-checking.

8 Limitations

Even though we performed a rigorous literature search to try to cover all existing work on scientific fact-checking, there is possibly work that was left uncovered due to different keywords, naming conventions (e.g., fact-checking vs. claim verification). Whenever possible, we tried covering all related work and all relevant cited papers.

All approaches for automated scientific fact-checking described in this work are still not safe for widespread adoption in practice due to constraints to their performance. Having deployed automated fact-checking systems that would produce incorrect verdicts could lead to mistrust in their usefulness and the process of fact-checking itself, including the work of dedicated manual fact-checkers.

Acknowledgements

This research has been supported by the German Federal Ministry of Education and Research (BMBF) grant 01IS17049 Software Campus 2.0 (TU München). We would like to thank the anonymous reviewers for their helpful feedback.

References

Liesbeth Allein, Isabelle Augenstein, and Marie-Francine Moens. 2021. Time-aware evidence ranking for fact-checking. *Journal of Web Semantics*, 71:100663.

Alexis Allot, Qingyu Chen, Sun Kim, Roberto Vera Alvarez, Donald C Comeau, W John Wilbur, and Zhiyong Lu. 2019. LitSense: making sense of biomedical literature at sentence level. *Nucleic Acids Research*, 47(W1):W594–W599.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13, Dominican Republic. Association for Computational Linguistics.

Dimosthenis Antypas, Jose Camacho-Collados, Alun Preece, and David Rogers. 2021. COVID-19 and misinformation: A large-scale lexical analysis on Twitter. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 119–126, Online. Association for Computational Linguistics.

Isabelle Augenstein. 2021. *Towards Explainable Fact Checking*. Ph.D. thesis.

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 *Conference on Empirical Methods in Natural Language Processing*.

Giannis Bekoulis, Christina Papagiannopoulou, and Nikos Deligiannis. 2021. A review on fact extraction and verification. *ACM Comput. Surv.*, 55(1).

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wen bing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor detection on social media with bi-directional graph convolutional networks. In *AAAI Conference on Artificial Intelligence*.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl1):D267–D270.

Jon Brassey, Christopher Price, Jonny Edwards, Markus Zlabinger, Alexandros Bampoulidis, and Allan Hanbury. 2021. Developing a fully automated evidence synthesis tool for identifying, assessing and collating the evidence. *BMJ Evidence-Based Medicine*, 26(1):24–27.

J Brennen, Felix Simon, and Rasmus Nielsen. 2020. Beyond (mis)representation: Visuals in covid-19 misinformation. *The International Journal of Press/Politics*, 26.

Qingyu Chen, Yifan Peng, and Zhiyong Lu. 2019. BioSentVec: creating sentence embeddings for biomedical texts. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE.

Suzie Cro, Tim P Morris, Michael G Kenward, and James R Carpenter. 2020. Sensitivity analysis for clinical trials with missing continuous outcome data using controlled multiple imputation: a practical guide. *Statistics in medicine*, 39(21):2815–2842.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2021. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20.

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.

Thomas Diggelmann, Jordan L. Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. *ArXiv*, abs/2012.00614.

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.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

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

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.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. A survey on stance detection for mis- and disinformation identification. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1259–1277, Seattle, United States. Association for Computational Linguistics.

Bin He, Di Zhou, Jinghui Xiao, Xin Jiang, Qun Liu, Nicholas Jing Yuan, and Tong Xu. 2020a. Bert-mk: Integrating graph contextualized knowledge into pretrained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2281–2290.

Yun He, Ziwei Zhu, Yin Zhang, Qin Chen, and James Caverlee. 2020b. Infusing Disease Knowledge into BERT for Health Question Answering, Medical Inference and Disease Name Recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4604–4614, Online. Association for Computational Linguistics.

Tamanna Hossain, Robert L. Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. COVIDLies: Detecting COVID-19 misinformation on social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. 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.

Prerna Juneja and Tanushree Mitra. 2022. Human and technological infrastructures of fact-checking. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).

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.

Mark T. Keane, Eoin M. Kenny, Eoin Delaney, and Barry Smyth. 2021. If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4466–4474. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Jiho Kim, Sungjin Park, Yeonsu Kwon, Yohan Jo, James Thorne, and Edward Choi. 2023. Factkg: Fact verification via reasoning on knowledge graphs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, Toronto, Canada. Association for Computational Linguistics.

Neema Kotonya and Francesca Toni. 2020a. Explainable automated fact-checking: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5430–5443, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Neema Kotonya and Francesca Toni. 2020b. 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.

Amrith Krishna, Sebastian Riedel, and Andreas Vlachos. 2022. Proofver: Natural logic theorem proving for fact verification. *Transactions of the Association for Computational Linguistics*, 10:1013–1030.

Anne Lauscher, Goran Glavaš, and Simone Paolo Ponzetto. 2018. An argument-annotated corpus of scientific publications. In *Proceedings of the 5th Workshop on Argument Mining*, pages 40–46, Brussels, Belgium. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Xiangci Li, Gully A Burns, and Nanyun Peng. 2021. A paragraph-level multi-task learning model for scientific fact-verification. In *SDU@ AAAI*.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained fact verification with kernel graph attention network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7342–7351, Online. Association for Computational Linguistics.

Kuan-Chieh Lo, Shih-Chieh Dai, Aiping Xiong, Jing Jiang, and Lun-Wei Ku. 2022. Victor: An implicit approach to mitigate misinformation via continuous verification reading. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 3511–3519, New York, NY, USA. Association for Computing Machinery.

Zhiyong Lu. 2011. PubMed and beyond: a survey of web tools for searching biomedical literature. *Database*, 2011. Baq036.

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.

H.-M Müller, Kimberly Auken, Y. Li, and P. Sternberg. 2018. Textpresso central: A customizable platform for searching, text mining, viewing, and curating biomedical literature. *BMC Bioinformatics*, 19.

Preslav Nakov, Alberto Barrón-Cedeño, Giovanni da San Martino, Firoj Alam, Julia Maria Struß, Thomas Mandl, Rubén Míguez, Tommaso Caselli, Mucahid Kutlu, Wajdi Zaghouani, Chengkai Li, Shaden Shaar, Gautam Kishore Shahi, Hamdy Mubarak, Alex Nikolov, Nikolay Babulkov, Yavuz Selim Kartal, Michael Wiegand, Melanie Siegel, and Juliane Köhler. 2022. Overview of the clef–2022 checkthat! lab on fighting the covid-19 infodemic and fake news detection. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5–8, 2022, Proceedings*, page 495–520, Berlin, Heidelberg. Springer-Verlag.

Preslav Nakov, David 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 *IJCAI*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. In *CoCo@ NIPs*.

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*, SIGIR '22, page 3141–3153, New York, NY, USA. Association for Computing Machinery.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 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. PMID: 32603243.

Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. 2021. Scientific claim verification with VerT5erini. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, pages 94–103, online. Association for Computational Linguistics.

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. *J. Mach. Learn. Res.*, 21(1).

Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. 2022. Ai in health and medicine. *Nature medicine*, 28(1):31–38.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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 3982–3992.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596.

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.

Mobashir Sadat and Cornelia Caragea. 2022. Scinli: A corpus for natural language inference on scientific text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7399–7409.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, 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, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Phillip Schneider, Tim Schopf, Juraj Vladika, Mikhail Galkin, Elena Simperl, and Florian Matthes. 2022. A decade of knowledge graphs in natural language processing: A survey. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 601–614, Online only. Association for Computational Linguistics.

Haeseung Seo, Aiping Xiong, and Dongwon Lee. 2019. Trust it or not: Effects of machine-learning warnings in helping individuals mitigate misinformation. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, page 265–274, New York, NY, USA. Association for Computing Machinery.

Gautam Kishore Shahi and Durgesh Nandini. 2020. *FakeCovid- A Multilingual Cross-domain Fact Check News Dataset for COVID-19*. ICWSM.

Chris Stahlhut. 2021. *Interactive Evidence Detection*. Ph.D. thesis, Technische Universität, Darmstadt.

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, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):1–28.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

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.

Juraj Vladika and Florian Matthes. 2023. Sebis at semeval-2023 task 7: A joint system for natural language inference and evidence retrieval from clinical trial reports. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Toronto, Canada. 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.

David Wadden and Kyle Lo. 2021. Overview and insights from the SCIVER shared task on scientific claim verification. In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 124–129, Online. Association for Computational Linguistics.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022a. 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.

David Wadden, Kyle Lo, Lucy Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022b. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 61–76, Seattle, United States. Association for Computational Linguistics.

Xuan Wang, Yingjun Guan, Weili Liu, Aabhas Chauhan, Enyi Jiang, Qi Li, David Liem, Dibakar Sigdel, John Caufield, Peipei Ping, and Jiawei Han. 2020. EVIDENCEMINER: Textual evidence discovery for life sciences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 56–62, Online. Association for Computational Linguistics.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2021. Climatebert: A pretrained language model for climate-related text. *arXiv preprint arXiv:2110.12010*.

Chih-Hsuan Wei, Alexis Allot, Robert Leaman, and Zhiyong Lu. 2019. PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Research*, 47(W1):W587–W593.

Jevin D. West and Carl T. Bergstrom. 2021. Misinformation in and about science. *Proceedings of the National Academy of Sciences*, 118.

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 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.

Amelie Wührl and Roman Klinger. 2021. Claim detection in biomedical Twitter posts. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 131–142, Online. Association for Computational Linguistics.

Amelie Wührl and Roman Klinger. 2022. Entity-based claim representation improves fact-checking of medical content in tweets. In *Proceedings of the 9th Workshop on Argument Mining*, pages 187–198, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Amélie Yavchitz, Isabelle Boutron, Aida Bafeta, Ibrahim Marroun, Pierre Charles, Jean Mantz, and Philippe Ravaud. 2012. Misrepresentation of randomized controlled trials in press releases and news coverage: A cohort study. *PLOS Medicine*, 9(9):1–11.

Donghan Yu, Chenguang Zhu, Yuwei Fang, Wenhao Yu, Shuohang Wang, Yichong Xu, Xiang Ren, Yiming Yang, and Michael Zeng. 2021. Kg-fid: Infusing knowledge graph in fusion-in-decoder for open-domain question answering. *arXiv preprint arXiv:2110.04330*.

Shi Yuan and Bei Yu. 2019. Hclaime: A tool for identifying health claims in health news headlines. *Inf. Process. Manage.*, 56(4):1220–1233.

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.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.*, 53(5).

Chaoyuan Zuo, Kritik Mathur, Dhruv Kela, Noushin Faramarzi, and Ritwik Banerjee. 2022. Beyond belief: a cross-genre study on perception and validation of health information online. *International Journal of Data Science and Analytics*, 13:1–16.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
8

☑ A2. Did you discuss any potential risks of your work?
8

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
No response.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☒ Did you run computational experiments?

Left blank.

□ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

□ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

□ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

□ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.