Scientific Web Claims: A survey of definitions, tasks, datasets and methods

Salim Hafid, Sandra Bringay, Konstantin Todorov

Scientific Web Claims: A survey of definitions, tasks, datasets and methods

Salim Hafid[1*], Sandra Bringay[2,3] and Konstantin Todorov[2]

[1*]médialab, Sciences Po, Paris, France.
[2]LIRMM, University of Montpellier, CNRS, Montpellier, France.
[3]University of Paul-Valery, Montpellier, France.

*Corresponding author(s). E-mail(s): salim.hafid@sciencespo.fr;
Contributing authors: sandra.bringay@lirmm.fr; todorov@lirmm.fr;

Abstract

Scientific web claims are seen as scientific claims as observed on the Web, across social media, online news, and other platforms. The growing prevalence of scientific discussions on the Web has intensified the need to process and assess this specific type of claims. Unlike claims from scientific publications, scientific web claims are expressed in lay terms, are often decontextualized, and typically lack proper citations, which poses unique challenges for their identification, verification, and communication. Nevertheless, the correct processing of scientific web claims is crucial to keeping online science discussions accurate and informed, for instance through fact-checking. This survey provides the first systematic overview dedicated specifically to scientific web claims. We review and compare existing definitions, task formulations, datasets, and methodological approaches across three major perspectives: (1) Scientific fact-checking on the Web, (2) Scientific citations on the Web, and (3) Science communication on the Web. Our interdisciplinary analysis integrates insights from natural language processing, information retrieval, artificial intelligence, social sciences, and science communication. We identify major methodological challenges, including the lack of unified definitions, domain-agnostic corpora, and foundational models tailored to science-related online discourse. We also discuss challenges related to the existing interplay between emotions and distortions of science online. By mapping current research efforts and highlighting open problems, this survey lays the groundwork for developing robust datasets, methods, and evaluation frameworks to advance the automated processing of scientific web claims, a necessary capability for strengthening the reliability of science-related online discourse at scale.

Keywords: Scientific Web claims, Fact-checking, Science communication, Scientific journalism, Online discourse

1 Introduction

The advent of social media has led to an increased participation of the general public in discussions around scientific topics, claims, and resources [1, 2]. As a consequence, online discussions beyond the evidence-based boundaries of "traditional science"[1] are characterized by misleading and polarized scientific online discourse [3, 4], where public perceptions of science depend on social and political dynamics [5]. Moreover, findings have shown the existence of tendencies to favor conflict in online discussions around science, as well as a compromise of accuracy by lack of details which might be relevant to scientists [6].

Scientific web claims, seen as scientific claims as observed on social media and in online press, are an important part of scientific online discourse. They include online posts featuring verifiable scientific claims like *"A study shows that COVID vaccines cause cancer, we're in big trouble!"* but exclude posts like *"My father got COVID"* which merely contain a scientific topic without conveying actual scientific knowledge. Processing and fact-checking scientific web claims is crucial to keeping online discussions around scientific topics accurate and informed. However, processing such claims presents several methodological challenges. First, they can take various forms [1]: from conveying actual scientific knowledge (e.g., by stating a scientific claim) to citing scientific knowledge (e.g., through a scientific reference). Second, these claims are typically uttered informally, where examples include claims such as *"covid vaccines just don't work on children"*, and tend to contain fuzzy or incomplete citations such as *"Stanford study shows that vaccines don't work"*, where the actual study is never cited (more examples of such differences are shown in Figure 1). Such challenges are exacerbated by the scale and heterogeneity of scientific web claims across disciplines.

In this context, the interdisciplinary and cross-disciplinary relevance of the analysis of scientific claims on the Web has been studied in depth, for instance in Hafid et al. [7]. Recent efforts in the fact-checking community have produced datasets targeting scientific web claims [8–11], acknowledging their distinct nature and the need for specialized analytical methods. Complementary perspectives from social sciences, psychology, and computational linguistics further reinforce this view. Studies show that scientific topics on news and social media are often sensationalized, politicized, and contribute to post-normal science communication [2, 3, 12, 13], where boundaries between fact-based knowledge and opinion become blurred [4]. From a psychological standpoint, research

[1] "Traditional science" is defined by Brüggemann et al. [3] in opposition to "Post-normal science" (seen as science in the social media era). In "traditional" science, scientists pursue research "disconnected from society and the world of values, interests and political conflicts", whereas actors in "post-normal" science can be seen as "scientist citizens". More generally, authors differentiate between "traditional" and "post-normal" science in terms of norms, roles, and communicative practices.

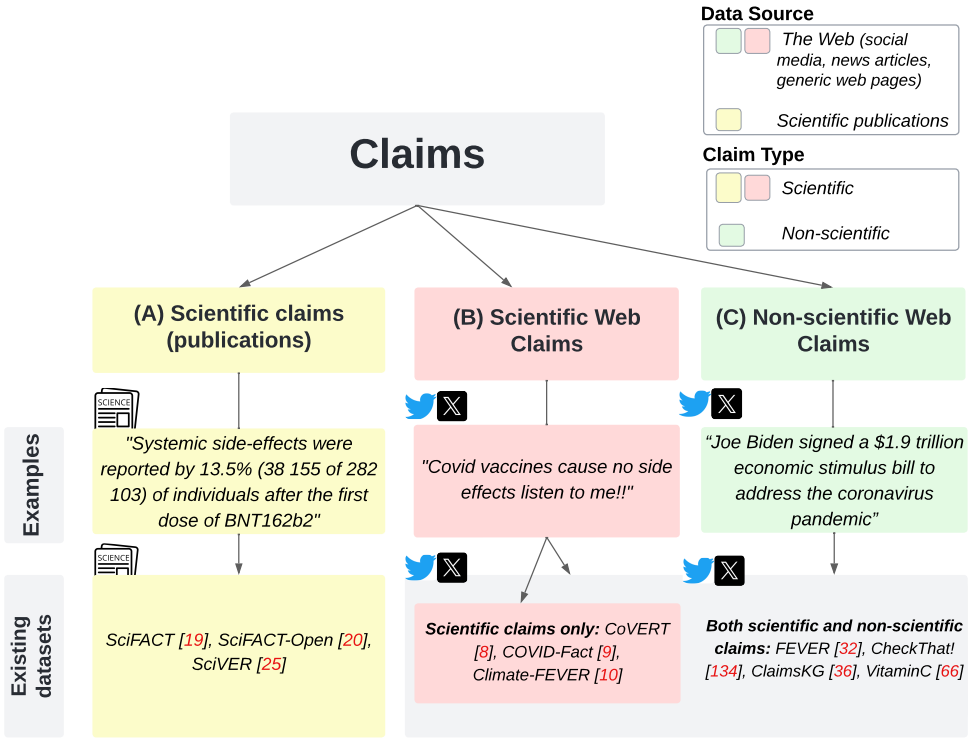


Fig. 1 Types of claims by data source and science-relatedness (taken from Hafid et al. [7]).

highlights the evolving trust and accessibility of scientific discourse online [14, 15]. Linguistic analyses likewise confirm that scientific online language differs from general web or scholarly discourse [11, 16]. Together, these works indicate the emergence of a distinct form of scientific online discourse, motivating a focused computational study.

Furthermore, claims, as an important part of online discourse, have been studied extensively by a wide variety of AI-related disciplines including automatic fact-checking, question answering, argument mining, and stance detection [17]. More specifically, existing research into claim-related tasks can be divided in three distinct categories in terms of data source and science-relatedness, as presented in Hafid et al. [7] (Figure 1): **(1) Scientific claims (publications):** (Figure 1.A) defined as claims which are extracted from scientific publications. Existing works include dedicated datasets such as SciTAB[18], SciFACT [19], SciFACT-Open [20], SciVer [21], and several dedicated models [21–26]. This claim category is **outside the scope of this work**. Our work aligns with most recent human fact-checking efforts which are focused on web data in societal contexts rather than in dedicated scientific communities [27]. This survey is positioned in this context, where we consider scientific claims from the Web and not from scientific publications. **(2) Scientific web claims:** (Figure 1.B) defined as claims including scientific knowledge that could be scientifically verified [28]. Thus,

scientific web claims are *claims from the Web which can only be verified with the help of scientific publications or research data used in the scientific process.* An example is shown in Figure 1, which underlines the difference between scientific claims from the Web (1.B) and scientific claims from research publications (1.A): while in both examples the claims are scientifically verifiable, the scientific publication claim is written in a way that reflects the language specificity found in scientific publications, while the scientific web claim is phrased in layman's terms. **(3) Non-scientific web claims:** (Figure 1.C) complementary to the scientific web claim definition, we define non-scientific web claims as assertions made on the Web and for which the verification does not involve a science-knowledge collection or scientific domain-knowledge. An example of a non-scientific web claim is shown in Figure 1.C, where verifying the claim does not involve a science-knowledge collection.

Web claims (the union of groups B and C in Figure 1) span various fact-checking-related downstream tasks with datasets such as CheckThat! [29–31], FEVER [32], MultiFC [33], FEVEROUS [34, 35], ClaimsKG [36], as well as several dedicated models [37–41]. In this survey, we focus exclusively on the scientific subset of web claims (see Figure 1.B). Existing work on scientific web claims has used various definitions and problem settings, and spans various tasks, datasets and methods, which we will be reviewing in this survey.

In this context, robust methods are required to correctly process scientific web claims. Training and evaluation of new methodological approaches for scientific web claims requires reliable large-scale ground-truth corpora that are based on sound definitions of scientific online discourse. While existing works from communities such as fact-checking and science communication have started investigating scientific web claims as a research object [8–11], definitions and corpora for scientific web claims remain domain-specific (e.g., COVID-19 [9], Climate change [10], or Medicine [42]), where generalisability is limited. More generally, the lack of a macro-perspective on the existing definitions and methodological approaches towards robust processing of scientific web claims is a crucial obstacle for advancing research in this area, as it impedes both constructing large-scale corpora that are based on unified definitions and fairly evaluating and benchmarking existing methods in this context.

In this work, we aim at bridging this research gap by systematically surveying and comparing existing definitions, tasks, datasets and methods that aim at the processing of scientific web claims. The organization of the paper is as follows: In Section 2, we define the scope and methodology of this survey. We namely differentiate between scientific claims from publications, scientific claims from the Web, and non-scientific claims from the Web. We position this survey as one focusing on scientific web claims. In Section 3, we present results of the survey. We systematically screen publications from nine venues corresponding to five distinct scientific communities. We start by reviewing existing definitions, where we aim at understanding how the concept of scientific web claims can be defined in relation to existing works from various disciplines. Then, we discuss existing tasks, datasets and methods related to scientific web claims

with a focus on three distinct perspectives: Scientific Fact-checking on the Web, Scientific Citations on the Web, and Scientific Communication on the Web. Finally, in Section 4, we highlight problems that remain open and challenges that need to be addressed towards robust and automated processing of scientific web claims.

2 Methodology of the Survey and Related Surveys

2.1 Methodology

This systematic review was conducted following the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines [43].

In this section, we describe this methodology used to extract publications from existing relevant literature, and the process by which the selected publications were reviewed. Our methodology also extends that of an existing survey on online claims [17].[2] An illustration of our methodology's pipeline is given in Figure 2.

Research fields selection: First, we identified research fields which are involved with scientific web claims in particular, and with scientific online discourse in general. To do so, we investigated which research fields have been collecting and annotating data relevant to scientific web claims. Our results identified the following three distinct research areas: (1) Fact-checking on the Web, where data is led mostly by fact-checking portals and shared tasks,[3] (2) Scientific citations on the Web, and (3) Science communication on the Web.

Retrieval and reviewing of relevant publications: After identifying the relevant research fields, we determined the scientific communities where works addressing the aforementioned fields can be found. In particular, we focused on the following scientific communities: Natural Language Processing (NLP), Web Mining, Information Retrieval (IR), Artificial Intelligence (AI), and Science Communication. An initial set of publications dealing with scientific web claims was extracted using a keyword-based search on a scholarly search engine (Google Scholar). The set of keywords used when querying the engine is the following: *fact, check, discourse, claim, argument, evidence, journalism, public trust, cite, citation, science, scientific, entailment, fake news, stance, opinion, rumour / rumor, viewpoint, question answering*. Based on the retrieved publications from the keyword-based queries, we selected venues from the most relevant papers for systematic screening. We show the selected venues for each of the identified scientific communities in Table 1. For each selected venue, we screened the proceedings of the years 2021-2025 and selected papers relevant to scientific web claims. The relevance criterium was the following: a publication is relevant if it contains either a definition, a dataset, or a task which includes scientific web claims. While reviewing publications from proceedings, relevant cited papers were also taken

[2] We discuss how our survey is positioned with regard to existing related surveys in Section 2.2.

[3] Established fact-checking portals include Snopes (www.snopes.com), FactCheck (https://www.factcheck.org), and AFP (https://www.afp.com/fr/produits-services/afp-fact-check). Shared tasks/labs include FEVER (https://fever.ai/task.html) and CheckThat! (https://checkthat.gitlab.io/clef2026/)

into account, regardless of publication year and venue. To review a publication, we extracted the definition used for scientific web claim and scientific web discourse, as well as the tasks, datasets and methods used by the publication. The data collection process resulted in several hundred publications, of which a total of 165 publications were deemed relevant to scientific web claims (see Figure 2). We show how the relevant publications are distributed per venue in Figure 3.

Table 1 Venues analyzed for the survey of tasks, datasets and methods related to scientific web claims. Note that the scientific communities as presented by this table are not disjoint. For instance, works published at a venue from the AI community can contain contributions relevant to the NLP community.

Scientific community	Venues
Natural Language Processing (NLP)	EMNLP, ACL
Web Mining	WebConf, TWEB
Information Retrieval (IR)	SIGIR, CIKM
Artificial Intelligence (AI)	AAAI, IJCAI
Science communication	JCOM

2.2 Related surveys

Existing surveys on claim-related tasks can be divided in two categories: the **first category** contains surveys on automated fact-checking of web claims [44–46]. Dmonte et al. [47] reviewed claim verification pipelines focusing on LLM-based systems, while Augenstein et al. [48] discussed both opportunities and challenges for fact-checking in the era of LLMs. Both covered LLM-specific issues such as hallucination, evidence sourcing, and retrieval challenges. Akhtar et al. [49] surveyed existing works on automated fact-checking focusing on multimodality and on how benchmark datasets and models for multimodality are both important yet challenging to construct given the multiple challenges of accounting for multiple modalities across all subtasks related to fact-checking. Nakov et al. [50] focused on how current automated systems can best assist human fact-checkers. Boland et al. [17] went beyond fact-checking and contributed a conceptual model of online claims that relates claims to other downstream tasks such as opinion mining, stance detection, and argument mining.

The **second category** of surveys focuses on scientific claims from scientific publications. Vladika and Matthes [51] reviewed tasks, datasets and methods related to fact-checking claims originating from scientific publications. However, to our best knowledge, no existing surveys focus specifically on **scientific web claims**. In this paper, we review existing definitions, tasks, datasets, methods and approaches developed specifically to deal with scientific claims uttered in online web contexts (e.g., on social media and in online news articles).

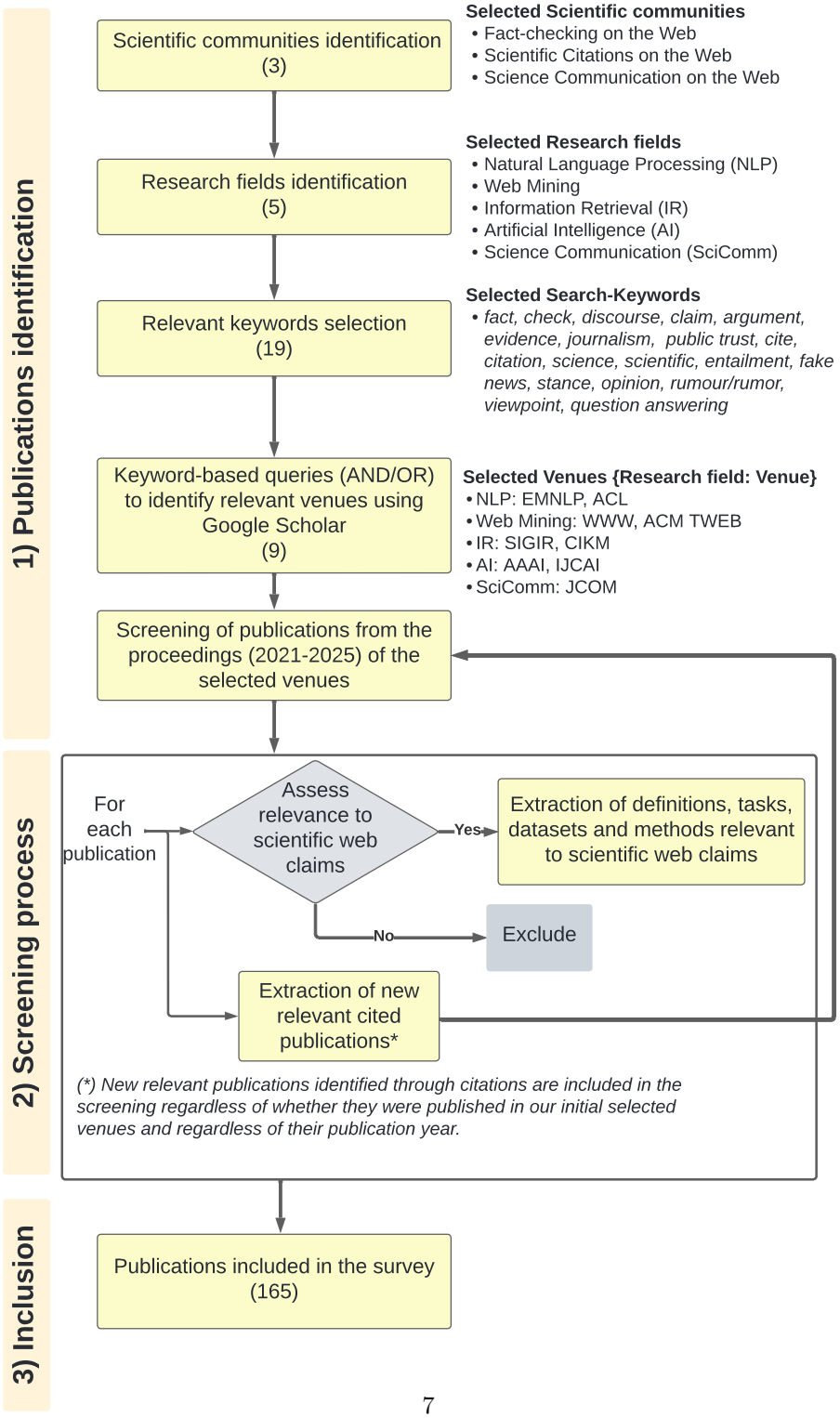


Fig. 2 Methodology used by the survey to collect and screen publications, following the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines [43]. The criterium used to assess relevance of a publication (gray in the figure) was the following: a publication is relevant if it contains a definition, a dataset, or a task which includes (not exclusively) scientific web claims.

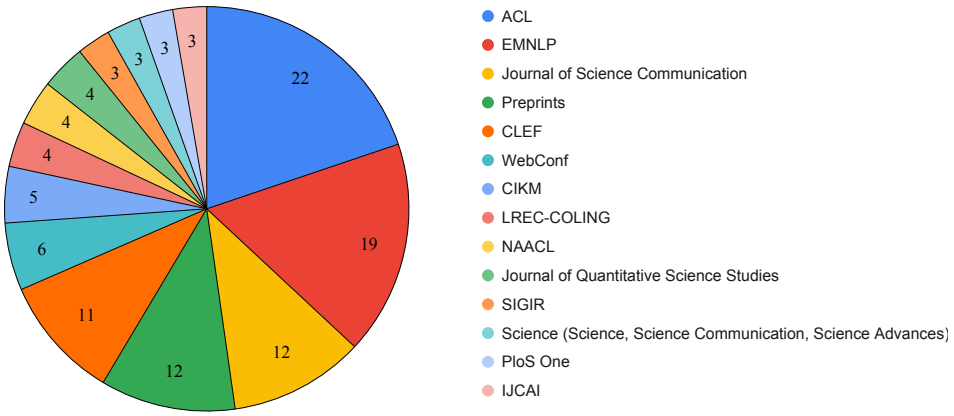


Fig. 3 Distribution of analyzed publications over venues for all venues with at least 3 publications (111 publications are shown out of the total 165 included in the survey).

3 Definitions, Tasks, Datasets and Methods

We select all relevant tasks from the surveyed literature and divide them in three categories:

- **Category 1** - Scientific Fact-checking on the Web (Figure 4)[4]
- **Category 2** - Scientific Citations on the Web (Figure 7)
- **Category 3** - Science Communication on the Web (Figure 8)

We chose the aforementioned three categories in order to align our survey's granularity with that of existing research fields (e.g., the fact-checking field and the science communication field). Moreover, our analysis showed that the existing differences in definitions, tasks, datasets and methods are well-represented by our proposed three categories insofar as our screened papers can be exhaustively reviewed and discussed using these categories.

The following subsections are structured as follows: For each category, we summarize existing relevant tasks illustrated by a dedicated figure. We then provide an overview of the definitions, problem settings, datasets, methods and approaches used by existing literature to address each of these tasks in their relation to scientific web claims.

3.1 Scientific Fact-checking on the Web

(a) Definitions: From the fact-checking and journalistic perspectives, a claim is seen as a statement that must be *eligible for fact-checking* [17]. The focus is less on the role of the claim in the context of the discourse and more on its content. Thus, a claim is usually defined as a *"piece of information provided by a source towards an entity"*

[4]Note that this category does not include fact-checking of scientific claims from scientific publications, see the definitions given in Section 1 for a detailed explanation of the scope of this work.

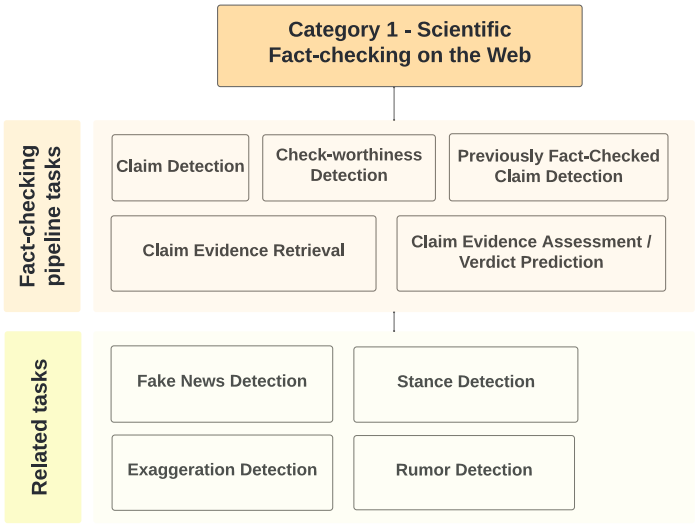


Fig. 4 Illustration of the first category (Scientific Fact-checking on the Web) of existing tasks related to scientific web claims. We distinguish between tasks that are part of the established fact-checking pipeline (see Figure 5), and tasks which can overlap with tasks from the fact-checking pipeline, and which we refer to as "related tasks". For instance, the task of Stance Detection can overlap with the Claim Detection task when candidate texts include scientifically verifiable claims containing a stance towards a science-related topic (e.g., stance detection for climate change). We discuss in more detail in Section 3.1 similarities and differences between the categories of tasks illustrated in this figure.

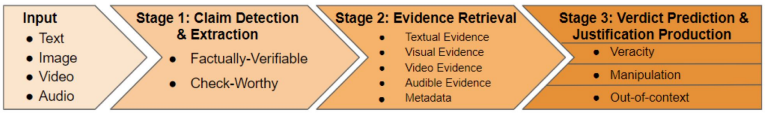
[17]. From the perspective of Scientific Fact-checking on the Web (Figure 4), some works define scientific web claims as claims that contain scientific entities [11], whereas others define scientific web claims by restricting the topic of the claims to the one they consider as scientific (e.g., COVID-19 [9], Climate Change [10], Biology/Medicine [8, 52]), or by focusing exclusively on numerical claims [53, 54]. Works have also combined topic filters with heuristics such as: the presence of causal relations [8], the presence of *"a proposition whose truthfulness can only be determined by additional evidence"* [9], or the presence of text for which *"evidence could be retrieved from a knowledge document collection that decreases the investigators uncertainty about the truthfulness (or falsehood) of the statement."* [10]. Beyond entity- and topic-specific formulations, Hafid et al. [28] have defined scientific web claims as claims from the Web which contain scientific knowledge that can only be verified with the help of scientific publications or research data used in the scientific process. We summarize existing definitions in Figure 6.

(b) Tasks & Datasets: Most tasks in Scientific Fact-checking on the Web (Figure 4) can be linked to the well-studied automated fact-checking pipeline (Figure 5). First, claims have to be detected. Related tasks involve:

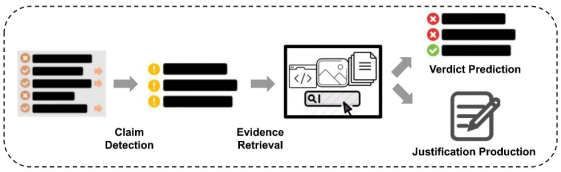
- **Scientific claim detection:** given an input statement, the system has to be predict whether it contains a scientific claim. The task of scientific claim detection overlaps with existing related tasks from the literature such as Rumor detection, where candidates for scientific web claims include scientifically verifiable rumors on

science-related topics [55]; Stance detection, where overlapping candidates include scientifically verifiable statements showing an attitude towards a science-related topic (e.g., Climate Change [56–58], COVID-19 [59]); and Exaggeration Detection, where overlapping candidates include exaggerations of scientific findings in online contexts such as science press releases [60, 61].

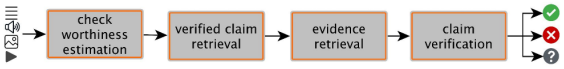
- **Check-worthiness detection:** given an input text, the system has to be predict whether it contains check-worthy statements, i.e., statements that are worth being verified by a professional fact-checker. This implies asking whether the statement contains a verifiable factual claim, is of public interest, and/or appears to be harmful [30]. This task has been traditionally applied to generic claims on the Web, i.e., with no distinction between scientific web claims (Figure 1.B) and non-scientific web claims (Figure 1.C), where applications include established shared tasks and competitions such as the CLEF CheckThat! tasks [31, 50, 62–65]. More recently, check-worthiness detection has been extended to scientific web claims, where use-cases include detecting claims from the Web related to prevention, diagnoses, risks, treatments, and cures of medical conditions [66].



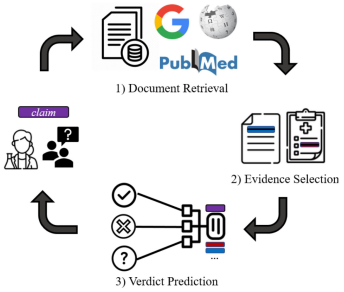
Pipeline (A): Multimodal fact-checking pipeline, from Akhtar et al. [48]



Pipeline (B): NLP framework for automated fact-checking, from et Guo, Schlichtkrull, and Vlachos [44]



Pipeline (C): Fact-checking pipeline, from Nakov et al. [49]



Pipeline (D): The standard three components of the framework for automated scientific fact-checking, from Vladika and Matthes [50]

Fig. 5 Web Claims Fact-checking pipelines as illustrated by four distinct existing surveys. The individual figures were extracted from [49] (A), [45] (B), [50] (C), and [51] (D). The standard pipeline consists of four distinct tasks: 1) Claim detection (based on criteria such as factual-verifiability or check-worthiness), 2) Evidence retrieval (either from a pool of candidate evidence documents, or in an open setting), 3) Evidence Selection (usually in the form of rationales, i.e., sentence-snippets from retrieved documents that support or refute the claim), 4) Verdict prediction (usually indicating whether the claim is supported by, refuted by, or unrelated to the retrieved evidence snippet). Task 1 can be skipped in datasets which already contain check-worthy/factual claims, and tasks 3 and 4 can be merged in a single task coined as Claim Verification.

After detecting scientific web claims, the next step in the automated fact-checking pipeline is to retrieve relevant evidence which supports or refutes the claim (see Figure 5).

- **Claim Evidence Retrieval:** Given a scientific claim from the Web, the system has to first retrieve candidate evidence statements, then label each (claim, evidence) pair to determine whether a given evidence supports the claim, refutes it, or does not contain enough information about it. This task has been traditionally applied to generic claims on the Web, where applications include datasets such as ClaimsKG [36] and VitaminC [67], and established shared tasks and competitions such as the FEVER shared tasks [35, 68, 69], the CheckThat! tasks [30, 70, 71], and the Fake News Challenge (FNC-1).[5] More recently, existing works have been curating datasets for claim evidence retrieval for scientific web claims, specifically around health-related claims (e.g., CoVERT [8], HEALTHVER [72], CHECK-Covid [73]) and climate-related claims (e.g., ClimateFEVER [10]). After retrieving candidate evidence documents, depending on the dataset, an additional step can be necessary: selecting relevant rationale sentences from previously retrieved documents. This step is called Evidence selection or Rationale selection. Datasets which require Rationale selection include COVID-FACT [9] and CoVERT [8], whereas in datasets like Climate-FEVER [10] and HEALTHVER [72], evidence documents are already in the form of short snippets and thus do not require Rationale selection.

Lastly, once evidence snippets are retrieved, the final step is to verify the veracity of the claim.

- **Verdict Prediction:** Given a scientific claim from the Web, the system has to predict a verdict with regards to its veracity. The veracity values vary between existing datasets [36, 74]. For instance, the ClaimsKG dataset normalized veracity values across 15 fact-checking websites into the following four normalized veracity ratings: True, False, Mixture, Other. Due to the increasing popularity of fact-checking portals such as FactCheck, Snopes, and FullFact [75, 76],[6] a new task coined Previously Fact-Checked Claim Detection emerged, where systems have to pair new claims with existing fact-checked claims to determine their veracity. In the case of Previously Fact-Checked Claim Detection, the verdict on the new claim is based on the verdict the previously fact-checked claim which most closely matches the new claim [77–79], whereas in standard Verdict Prediction, the verdict is based on assessment of the retrieved evidence statements (see Claim Evidence Retrieval task). In contrast to non-scientific claims, where the veracity or falsehood of a claim, once established, remains unchanged, the veracity value of a scientific claim depends entirely on variables related to the experimental setting inside which the findings behind the claim were produced [80, 81]: e.g., representativeness of the sampled data, statistical significance, rigor and extensiveness of the evaluation protocol. Moreover, even when assuming that a scientific claim is to be proven true or false at a given moment in time, its veracity value can nonetheless change at any given moment in the future

[5] http://www.fakenewschallenge.org
[6] https://factcheck.org, https://snopes.com, https://fullfact.org

<table>
<tr><th colspan="3" rowspan="2">Scientific Web Claims</th><th colspan="3">Tasks</th></tr>
<tr><th>Claim Detection</th><th>Evidence Retrieval</th><th>Evidence Assessment / Verdict Prediction</th></tr>
<tr><td colspan="3" rowspan="6">Definitions</td><td>"Sentence-level segments that involve one or more scientific entities and are eligible for fact-checking" [11]</td><td colspan="2">Climate-related claims which are "well-formed and subjectively investigable". [10]
A claim is well-formed if it's a "single English sentence, consistent, unambiguous, and complete". A claim is subjectively investigable if "evidence could be retrieved from a knowledge document collection (kdc) that decreases the investigators uncertainty about the truthfulness (or falsehood) of the statement."</td></tr>
<tr><td rowspan="2">COVID-19-related claims which are factually verifiable and/or check-worthy (i.e., are likely to be false, are of public interest, are potentially harmful and are not easy to fact-check by a layperson) [63]</td><td colspan="2">COVID-19-related "propositions whose truthfulness can only be determined by additional evidence" [9]</td></tr>
<tr><td colspan="2">Numeric claims, seen as "statements needing verification of any explicit or implicit quantitative or temporal content" [52], or as statements containing numerical expressions which can be verified against time-series evidence (procedural definition) [53]</td></tr>
<tr><td rowspan="3">Check-worthy claims (following [63]) which have to do with "prevention, diagnoses,
risks, treatments, and cures of medical conditions" [65]</td><td colspan="2">Biomedical COVID-19-related claims which contain causal relations [8]</td></tr>
<tr><td colspan="2">COVID-19-related "assertions that express facts without providing evidence" [71]</td></tr>
<tr><td colspan="2">"Atomic factual statements describing one aspect of a scientific entity or process related to COVID-19 (such that they can be fact-checked against primary research)" [72]</td></tr>
<tr><td colspan="3">Datasets</td><td>SciClops [11], CW-CURE [65], CheckThat! 2022 Task 1 [63]</td><td colspan="2">BEAR-FACT[51], CovidFACT [9], ClimateFEVER [10], CoVERT [8], HEALTHVER [71], Check-COVID [72], QuanTemp [52], TSVer [53]</td></tr>
<tr><td rowspan="4">Methods / Models</td><td rowspan="3">Transformer based</td><td>Encoder-only models (e.g., BERT-based)</td><td>BERT, SciBERT, NewsBERT, SciNewsBERT [11], XLM-RoBERTa [190], BERTweet, ConvBERT, Electra [191]</td><td>Sentence-BERT [9], Longformer [8], MultiVerS [8], RoBERTa-Large [72]</td><td>SciBERT [140], ClimateBERT [96], RoBERTa [9, 72], Longformer [8], MultiVerS [8]</td></tr>
<tr><td>Encoder-decoder models (e.g., T5, BART)</td><td>T5-large, BART-large [65]</td><td>T5 ranker [8]</td><td>T5-base [Sar+21]</td></tr>
<tr><td>Decoder-only models (e.g., GPT-3, GPT-4)</td><td>GPT-3 [85], XLNet [192]</td><td>GPT-3.5 [72], GPT4 [52]</td><td>GPT-3.5 [72], LLaMa3 [94]</td></tr>
<tr><td colspan="2">Other methods/models</td><td>Random Forest [11], SVM [192], Syntax-based heuristics [11]</td><td colspan="2">Google Search [9], BM25 [8]</td></tr>
</table>

Fig. 6 Detailed overview of existing definitions, tasks, datasets, methods and models for scientific web claims from a fact-checking perspective (Section 3.1). We structure the table around three tasks (Claim Detection, Evidence Retrieval, and Verdict Prediction) which correspond to the three main steps of the fact-checking pipeline previously presented in Figure 5. We **highlight** in the definitions words which correspond to three heuristics used by existing work to define and annotate scientific web claims: domain, entities, and factuality. We explain in the Challenges Section of this survey (Section 4) why such heuristics have limited reliability and do not account for the complexity and diversity of scientific web claims. We also summarize in the same section how future research could address this challenge.

upon presence of more recent and more robust contradicting evidence [80, 82]. For this reason, recent research has argued for an alignment of automated fact-checking methods with human fact-checking efforts, by retrieving "source-guarantee evidence" [83], i.e., the evidence used by the claimant to make the claim. This way, discussions and assessments of the veracity value of the claim/finding can be had within the context of the research publication from which the findings stem. The necessity of retrieving source-guarantee evidences is what motivates the task of Citation-source retrieval, which we discuss further in Section 3.2.

(c) Methods: For each of the tasks above, we summarize the methods used by existing works in Figure 6.

To detect scientific web claims, existing work has relied largely on language models to classify textual statements into verifiable scientific claims. More specifically, works have used large pre-trained language models (i.e., models trained on large general-domain corpora to capture general features of the language) which are then fine-tuned to the desired target task. Gollapalli et al. [66] fine-tuned a T5 model [84] to detect check-worthy health-related claims from X (ex-Twitter), while Smeros et al. [11] fine-tuned a BERT model and a SciBERT model to detect claims containing scientific entities in online news articles and social media postings. For multimodal content, to our best knowledge, no existing dataset is specifically dedicated to scientific web claims. However, many existing multimodal fact-checking-related datasets contain generic web claims (i.e., contain both scientific and non-scientific web claims, see Figure 1) and have implemented various methods to detect check-worthy multimodal claims. Works from the CheckThat! 2023 shared task on multimodal check-worthiness detection [31] have used different models spanning various architectures to extract both textual and visual features from multimodal web claims. Textual features were extracted mainly using models based on the Transformer architecture [85], an architecture which is based on a multi-head attention mechanism, allowing for key tokens (words, sentences) to be perceived by models as more important than other tokens. Existing works have used GPT-3 [86], BERT [87], RoBERTa [86] and BERTweet [88] to extract textual features from check-worthy claims. On the other hand, visual features were extracted mainly using models based on Convolutional Neural Networks (CNN) [89]. Existing works have used ResNet [90] and ConvNext [88] to extract visual features from check-worthy claims.

To automatically attribute veracity values to the detected scientific web claims, three steps are necessary: Evidence retrieval, Rationale selection, and Verdict prediction (see Figure 5). To retrieve evidence documents, existing work has relied on tools such as Google Search [9] or ranking algorithms such as BM25 [8]. Additional intermediate steps present in some works involve re-ranking evidence documents with slower but more complex models, including encoder-decoder models like T5 [8, 91] and decoder models like GPT4 [53], or performing style transfer techniques, e.g., rewriting a scientific web claim in more formal language using LLaMa3 to enhance retrieval of candidate scientific publications [92]. Once the initial pool of candidate

documents is retrieved and optionally re-ranked, rationale snippets must be selected. To do so, existing work has mainly relied on two types of language models. Some works have used language models whose architectures enable processing of long documents, thus enabling retrieval and assessment of full documents (e.g., news articles, scientific papers) as candidate evidence documents from which rationale sentences can be extracted. Examples include works which have used the Longformer model [8], a model with an attention mechanism that scales linearly with sequence length (as opposed to quadratic scaling with sequence length in standard transformer-based models) [93], thus enabling the processing of documents of thousands of tokens or longer. Other works have used language models whose architectures can derive semantically meaningful sentence embeddings which can be compared using similarity metrics such as cosine similarity. Using cosine similarity at the sentence-level enables the process of comparing an input claim to a pool of candidate evidence sentences to be much more computationally efficient [94], thus optimizing the time necessary to perform Rationale selection. Existing works have mainly used Sentence-BERT [9], a model which uses siamese and triplet network structures to generate embeddings at the sentence-level much more efficiently than standard transformer-based models [94].

Finally, to predict verdicts for scientific web claims, existing work has relied on a variety of pre-trained language models, which are either fine-tuned on the target dataset or used in a zero-shot setting. Similar to the evidence retrieval step, some works include an intermediate paraphrasing step to match the claim's structure with the retrieved evidence before verdict prediction. Such works typically use LLMs either in a zero-shot setting or fine-tuned with efficient techniques like Low Rank Adaptation (LORA) and Direct Preference Optimization (DPO) loss [95]. For the final verdict prediction, existing works have used encoder-only models such as SciBERT [96], ClimateBERT [97], or RoBERTa [9]; encoder-decoder models such as T5-base [72]; and decoder-only models such as GPT [53].

Besides developing methods and models for specific subtasks of the aforementioned fact-checking pipeline separately, a related line of research has tried to automate all of the pipeline into a single system. In such a unified system, the input is a scientific claim the user wants to fact-check, and the output is typically a verdict along with relevant sources. The system connects to a scholarly database through an API (e.g., SemanticScholar or GoogleScholar), and performs search, evaluation, and summarization of evidence based on the academic databases. Existing systems divide in two: Open- and closed-source systems. Examples of closed-source systems include Scite,[7], Elicit,[8] and Consensus.[9] Such systems exhibit some efforts towards transparency, e.g., Scite shows the user how the natural language query is turned into a keyword-based query with logic operators. However, they remain black-box systems where the weighting of retrieved sources is typically unbeknowst to the user. Additionally, some systems impose constraints on the number of papers that can be processed and retrieved.

[7] https://www.scite.ai
[8] https://www.elicit.com/
[9] https://www.consensus.app

On the open-source side, examples of existing systems include VALSCI [98], a self-hostable system based on LLMs which includes bibliometric scoring, Veracity [99], which encompasses a reliability score based on sources, and traditional Retrieval-Augmented Generation (RAG) systems such as LitLLM [100] or LLAssist [101]. In both closed- and open-source variants, systems suffer from typical issues of LLM-based systems like hallucination [48], but also suffer from scalability issues, namely with high-throughput verification, and from bibliometric bias.

3.2 Scientific Citations on the Web

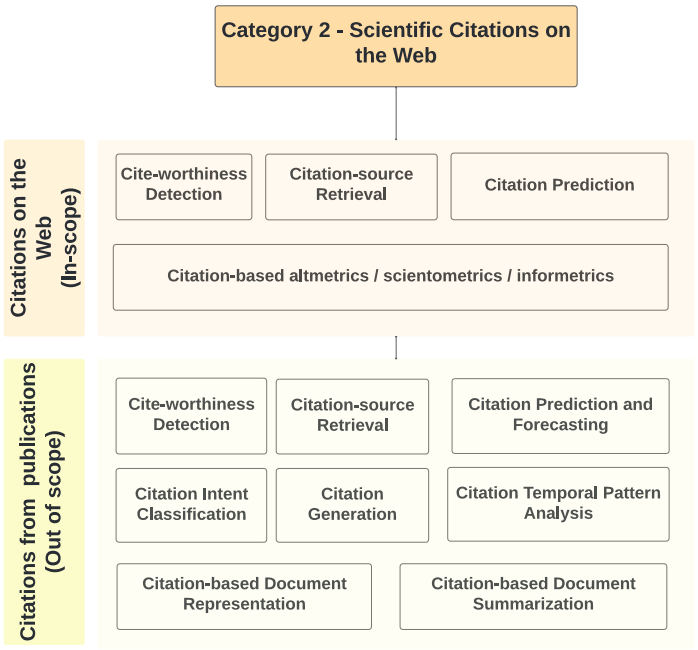


Fig. 7 Illustration of the second category (Scientific Citations on the Web) of existing tasks related to scientific web claims. We differentiate between tasks related to citations in online web contexts, which are within the scope of this survey, and tasks related to citations from scientific publications, which are out of scope. For Citations on the Web, existing works include specific NLP/IR downstream tasks (e.g., cite-worthiness detection, citation prediction), but also include more general scientific subcommunities with works on citation-based altmetrics, scientometrics and informetrics.

A substantive amount of literature exists on the topic of scientific citations in scientific publications, with works in cite-worthiness prediction [102], citation prediction and forecasting [103–105], citation-based impact prediction [106], citation intent classification [107–109], analysis of temporal patterns of citations [110, 111], citation generation [112–114], citation-based scientific document representation [115], citation-based paper summarization [116–119], and citation falsification/manipulation [120–123].

On the other hand, research on citations in online web contexts (e.g., social media, online press articles) also spans various research topics related to altmetrics, where citations are used as a proxy to measure the dissemination of scientific results on social media [124–126], the impact of a scientific work [127, 128], or to study the user communities which interact with scientific findings online [129, 130]. We summarize existing citation-related tasks in Figure 7.

In this section, we will focus on a subset of this research, namely the research done on citations in online web contexts applied to scientific dis- and misinformation detection in general, and to scientific web claims in particular. Using the publications retrieved and analyzed in our survey, we will first review existing definitions. Then, we will list existing relevant tasks, where for each task we define the problem setting and describe existing datasets from the literature. Finally, we will review existing methods and approaches used by related work to solve existing tasks.

(a) Definitions: From the perspective of Scientific Citations on the Web (Figure 7), the focus is less on the content of the claim and more on whether it cites -or ought to cite- a scientific reference that justifies the claim formulated in the text. This gave rise to the notion of cite-worthiness. The notion of cite-worthiness relates to the notion of check-worthiness, which has been extensively researched by fact-checking-related studies over the years.[10] A sentence is defined as "check-worthy" if it is worth fact-checking (e.g., contains a verifiable factual claim, is potentially harmful, and is of general interest) [30, 31], whereas a sentence is "cite-worthy" if it contains (or ought to contain) a reference to an external source [102, 131]. While check-worthiness detection can help professional fact-checkers detect which claims to focus on, cite-worthiness detection can be used to flag scientific results which are presented without references.

(b) Tasks & Datasets: In the context of scientific dis- and misinformation detection, similar to automated fact-checking, scientific citations on the Web can be tracked using a pipeline which comprises the following two tasks:

- **Cite-worthiness detection:** given an input statement, the system has to predict whether it contains a reference to an external scientific source (e.g., a scientific publication, a dataset, statistics). This task was first coined by Wright and Augenstein [102], who curated CITE-WORTH, a multidomain dataset specifically dedicated to cite-worthiness detection of sentences from scientific publications, and was extended to an online social media context (with texts from X) by Hafid et al. [131].
- **Citation-source retrieval:** given an input document (e.g., an online news article) which is cite-worthy, the system has to retrieve the original URL of the cited scientific reference. This task is motivated by fuzzy citation habits in online news and on social media [132], e.g., *"I read a Harvard study showing that masks can't stop the virus from spreading"*, where the actual publication is not properly cited. Existing work [133] found that, in a pool of 8,600 Reuters articles, over 25% of news

[10] See the CheckThat! Lab editions hosted by the CLEF conference https://checkthat.gitlab.io/clef2024/task1/

articles which cite a scientific publication do not cite the correct link to the actual publication. In this context, existing works at the intersection of Natural Language Processing (NLP) and Information Retrieval (IR) have developed datasets, models, and retrieval algorithms to disambiguate fuzzy citations in online news articles [132–135].

(c) Methods: In this subsection, we summarize the methods used to detect cite-worthy statements and to retrieve citation sources.

Determining whether a scientific text lacks, and hence requires a citation, has been a challenge in the NLP community. Existing approaches have tackled cite-worthiness detection in the context of scientific publications, using corpora constructed from academic articles in specific fields. Existing works have used statistical learning, traditional machine learning models as well as recent large language models. For instance, Sugiyama et al. [136] used both a Support Vector Machine (SVM) model and a Maximum Entropy (ME) model which they combined with simple textual features such as unigrams, bigrams, and proper nouns to extract cite-worthy statements from scientific publications. They created a dataset of cite-worthy and non-cite-worthy sentences extracted from the ACL Anthology Reference corpus [137]. More recently, more advanced approaches [138] have measured the performance of a Convolutional Recurrent Neural Network on the ACL Arc dataset as well as the arXivCS dataset [139] and the Scholarly Dataset 2.[11] The limitations of these works are mainly related to domain-specificity, class imbalances, and little to no presence of data quality analysis. These issues were addressed in Wright and Augenstein [102], where the authors build and share a curated multi-domain dataset specifically dedicated to the task of cite-worthiness detection, that is used to evaluate a number of language models such as BERT [140], SciBERT [141] and Longformer [93] against a logistic regression baseline. However, existing literature has not yet focused on cite-worthiness detection in online web contexts, where to our best knowledge, the only work has been by Hafid et al. [131], where findings showed that models trained on scientific publications performed less well on text from X.

To retrieve the original source URLs of fuzzy citations in online web contexts, existing research has worked on both online news articles and on social media. For online news, existing methods have focused on either extracting DOIs/links directly from the news articles' body [142], or, when no links can be found in the articles' body, on extracting metadata entities (e.g., author names) from the news articles' body, then using the extracted metadata to query scholarly search engines (e.g., Microsoft Academic, Scopus) and thus infer the cited scientific publication [132]. To our best knowledge, the latest work on news citation prediction was done by Wang and Yu [133]. It involves indexing the PubMed corpus in ElasticSearch, then developing NER-based queries (using author names, affiliation, journal) based on a dataset of science health news articles, and using the queries with an enhanced BM25 retrieval algorithm where the

[11] The mentioned datasets (ACL Arc, Scholarly Dataset 2, arXivCS) are available online at http://citation-recommendation.org/publications

score is weighted by features such as the scientific studies' publication date. For social media, existing work has focused on X and has developed baselines based on several models (BM25, SBERT, e5) [143] in both zero-shot and fine-tuned settings to retrieve the original scientific publications mentioned by tweets. In this case, the retrieval is performed in an experimental closed setting, where authors construct TweetCite, a dataset based on a query set (tweets with claims that lack proper citations) and a collection set (candidate scientific publications). The task is to retrieve, for each query tweet, the missing original publication based on the collection set. More recently, the task of citation-source retrieval has been included in the CheckThat! tasks [70, 135],[12] leading to the development of multiple new models and approaches mixing traditional BM25-based retrieval with large language models and neural re-rankers [134].

3.3 Science communication on the Web

We previously discussed how research from the social sciences has taken interest in scientific online discourse and in scientific web claims (Section 1). Specifically, research on scientific web claims from a science communication perspective presents multiple theoretical and methodological similarities with existing research from NLP/ML, thus making it a relevant perspective for this survey.

In science communication, scientific web claims are studied from perspectives such as scientific journalism (Figure 8.A), public perception of science (Figure 8.B), and public engagement with science (Figure 8.C). In this section, we give an overview of definitions, tasks, datasets and methods used by recent research to address related research problems.

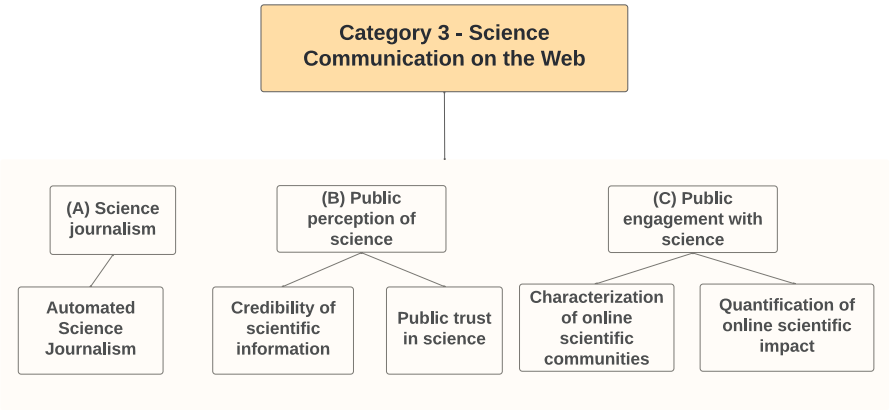


Fig. 8 Illustration of the third category (Science Communication on the Web) of existing tasks related to scientific web claims. In comparison with former categories of this survey where most works are organized around specific NLP/ML downstream tasks (e.g., in Figure 4), tasks in science communication refer to scientific subcommunities and not necessarily to a unique task formulation of a research problem. In this context, we further divide this category in three subcategories: Science journalism (A), Public perception of science (B), and Public engagement with science (C).

[12] https://checkthat.gitlab.io/clef2026/

(a) Definitions: From a Science Journalism perspective (Figure 8.A), research aims at processing and summarizing research findings to generate press releases that can briefly yet faithfully report recent scientific findings in layman's terms. Generated press releases contain scientific web claims, where the focus is on notions such as abstractiveness, seen as how much a headline or a short summary captures the salient ideas of a scientific paper [144, 145], and faithfulness, seen as the accuracy of the press release with regard to the original scientific study. Faithfulness has been approached from various angles such as sensationalism, misrepresentation, and subjectivity [3, 146, 147].

From the perspective of Public perception of science (Figure 8.B), the focus is not on the content of the claims as much as it is on how users perceive them. In this context, works have focused on notions such as credibility and trust [148–151]. LeBel et al. [152] proposed a unified framework for scientific credibility, where a credible finding is defined as one that has *"repeatedly survived risky falsification attempts"*. They proposed a framework based on four falsifiable dimensions: transparency of methods and data, reproducibility of results, robustness of results to different data-processing and analytic decisions, and replicability of the effects. Augenstein [153] proposed citeworthiness (see the definition given in Section 3.2) and exaggeration (seen as the exaggeration of statistical relationships such as correlation, conditional causality, and causality) as two key notions towards better credibility assessment of online reporting of scientific findings.

From the perspective of Public engagement with science (Figure 8.C), the focus is both on the outreach and impact of scientific web claims [154–156] and on the online communities which interact with them [157]. The impact of science is typically tracked through engagement types [154] such as behavioral engagement (e.g., actions or interactions resulting in concrete behavior change, for instance pro-environmental behavior), attitudinal engagement (e.g., trust in science/scientists, valuing science, expressing interest or exhibiting emotional reactions), and cognitive engagement (e.g., knowledge- or awareness-gain about a scientific topic or field).

(b) Tasks, Datasets & Methods:

Science journalism: (Figure 8.A) From a science journalism perspective, with the recent advancements in deep learning and large language models, research has investigated the question of automating the reporting process, with works on summarizing research findings for non-expert readers, e.g., by using layman terms [144, 158], or by generating entire science news reports [159, 160]. Such works have relied on both extractive methods (directly retrieving relevant sentences from scientific publications and including them in the summary report) and abstractive methods (first comprehending the scientific publication, then generating a report shortening and paraphrasing its content). Extractive methods involve algorithms such as Latent

Semantic Analysis (LSA) [161], a frequency-based algorithm which first models relations between words and sentences in a given corpus, then generates vectors that represent salient concepts in the corpus, then outputs the subset of sentences which are related to the most important concepts. Another extractive method is PacSum [162], an algorithm which combines a BERT language model with a modelisation of the scientific publication's text as a directed graph. Abstractive methods involve Sequence-to-Sequence language models such as Longformer [93], and Generative pre-trained transformers such as GPT-4 [163] or LLaMa3 [164]. Other works in science journalism have investigated the patterns used by journalists to selectively report on scientific findings [165], as well as patterns of misrepresentation, sensationalism and harmful recommendation [3, 146, 147, 166].

Public perception of science: (Figure 8.B) To study the public perception of science in online contexts, research has worked on both the perceived credibility of scientific information and the public trust in science online. Research on the credibility of scientific information has worked on establishing useful frameworks and downstream tasks towards credibility assessment. We previously discussed how LeBel et al. [152] proposed a unified framework for scientific credibility which is based on four falsifiable dimensions: transparency of methods and data, reproducibility of results, robustness of results to different data-processing and analytic decisions, and replicability of the effects. More recently, Augenstein [153] proposed Cite-worthiness detection and Exaggeration detection as two candidate tasks whose addressing has direct impacts towards a more robust credibility assessment of scientific statements in online web contexts (e.g., news articles, social media). Another dimension of public perception of science in online contexts is public trust in science. Research has investigated how factors such as narratives, emotionalization, visual representations, professional background, and message style influence source trustworthiness, message credibility, and behavioral intentions [167, 168]. Other works have investigated how uncertainty, as communicated in research findings, impacts laypeople's trust in online science. Investigated factors include framing of findings, emphasis on the limited reliability of the finding, people's domain-related epistemological beliefs, criticism from political elites and perceived personal benefits [15, 169–173].

Public engagement with science: (Figure 8.C) Lastly, to study the public engagement with science online, research has used various methods. For instance, research has used methods such as community networks analysis, topic networks analysis and audience segmentation to quantify and contextualize the impact of scientific publications within a specific research topic [155, 156, 174, 175], or within the global context of researcher communities online [176]. Works have also used a variety of metrics towards characterizing online communities of attention around science, e.g., click metrics [177], citation indicators [178], and social media metrics (e.g., number of tweets/retweets, number of followers) [179]. Beyond online communities, works have investigated who exactly engages with scientific claims made by scientists online. Suleski and Ibaraki [180] found that an increasing amount of research fails to gain attention from researchers outside the specialized fields. However, in more recent

research, Côté and Darling [181] found that, on X (ex-Twitter), beyond a range of circa 1,000 followers, follower-types included educational organizations, media, members of the public, and small numbers of decision-makers. Tweeting, therefore, *"has the potential to disseminate scientific information widely after initial efforts to gain followers"* [181]. Authors argue that such results should encourage scientists to invest in building a strong social media presence for an enhanced scientific outreach, where scientific findings communicated online have the potential of reaching laypeople and decision makers. However, a strong scientific outreach might not guarantee the correct perception of scientific claims and findings from the public. Roedema et al. [157] held semi-structured interviews with 26 European scientists who actively engage with citizens in online communities. Participants reported that online communities *"provide hollow interactions, devaluate scientific expertise or even represent a hostile environment"*. Authors argue for an alignment between the scientists' intended contribution to online interactions and the scientists' perspective and deployed online repertoires.

4 Challenges

Based on our literature review, we identify a set of methodological challenges in dealing with scientific online discourse in general, and with scientific web claims in particular. We proceed to present these as open issues.

(1) Existing definitions and ground-truth corpora for scientific web claims are mostly domain-specific. Very few domain-agnostic definitions and corpora exist for scientific web claims [28]. This is particularly true for fact-checking-related tasks (see Section 3.1). Instead, scientific web claims are usually defined using simple filtering heuristics such as the presence of scientific entities (e.g., *"vaccine"*, *"COVID-19"*) [11], however with no clear definition of what a scientific entity is, or by restricting claims to specific science-related topics (e.g., COVID-19 [9], Climate Change [10], Biology/Medicine [8]). More generally, existing definitions do not account for the various forms that scientific online discourse can take (e.g., differentiating between texts which actually contain scientific knowledge and text which merely reference scientific knowledge). As a consequence, existing ground-truth corpora do not account for the diversity and the complexity of scientific web claims. Moreover, the heterogeneity of definitions on which existing corpora are based makes it difficult to compare performance of methods and models across datasets, and is thus a crucial obstacle for robust benchmarking and evaluation of models in this context. Future work could address this challenge by focusing on building domain-agnostic definitions, datasets and benchmarks that account for the complexity of scientific web claims.

(2) Fact-checking-related tasks do not differentiate between scientific and non-scientific web claims. Most recent fact-checking efforts have been focused on web data in societal contexts rather than in dedicated scholar communities [27]. However, most existing fact-checking-related tasks focus on web claims (i.e., claims at the union of groups B and C in Figure 1), and do not differentiate between scientific web claims (Figure 1.B) and non-scientific web claims (Figure 1.C). Fact-checking

a web claim can be very different depending on whether the claim is scientific or not: existing literature has shown that scientific discourse from the Web has unique linguistic characteristics which differ from traditional scientific publications' text and also differ from generic social-media text [7, 16, 175, 182]. For instance, recent work by Hafid et al. [7] found that scientific web claims used more analytical speech but also more sentiment-related speech than non-scientific web claims. More intuitively, when compared to a non-scientific web claim, fact-checking a scientific web claim relies on distinct sources of evidences, tools for verification, and credibility indicators. As a consequence, current methods and models trained only on generic web claims (the union of groups B and C in Figure 1) might not be efficient in fact-checking scientific web claims (Figure 1.B). In a study comparing the empirical performance of models in fact-checking-related tasks in both scientific and non-scientific web claims, results showed that BERT-based language models performed worse by up to 17 F1 points on the scientific subset of web claims compared to the non-scientific one [7]. Addressing this challenge requires more empirical work towards a better understanding of scientific web claims. It may also ultimately require the development of separate task formalizations and benchmarks for accurate fact-checking of scientific web claims.

(3) Citation-related tasks are not extended enough to social media. We previously discussed how tasks such as cite-worthiness detection and citation-source retrieval are essential to a better credibility assessment of scientific discourse (see Sections 3.3 and 3.2). However, existing datasets and methods have focused almost exclusively on claims and citations from scientific publications [102] and news articles [132, 133], and have only recently started to get extended to social media [70, 135]. We previously stated how on social media, research has shown the existence of tendencies to favor conflict and compromise accuracy by lack of details which might be relevant to scientific findings [6]. In this context, it becomes crucial to keep extending research efforts in credibility assessment of scientific discourse to social media. For instance, recent studies have extended the tasks of cite-worthiness detection [131] and citation-source retrieval [134, 135, 143] to X (ex-Twitter), where scientific web claims are often uttered without references, thus making the tasks of flagging non-referenced scientific claims and retrieving their original scientific publications especially relevant. Future work could extend such studies to more social media contexts, more modalities, and more languages. A separate yet equally important line of research is to build a better understanding of citation habits of social media users through large-scale data analyses across platforms and languages. Existing work has focused only on the English language [1, 143] and on specific platforms like X and Facebook [183].

(4) Lack of large-scale ground-truth corpora for scientific online discourse. We previously discussed how existing datasets for scientific online discourse (a) are domain-specific, (b) do not account for the various forms scientific online discourse can take, (c) are limited in size. While challenges (a) and (b) could be addressed by contributing ground-truth annotated corpora that are based on robust domain-agnostic definitions of scientific online discourse, scaling-up such corpora remains a

challenge. Future work could use existing definitions and annotation protocols to scale-up the data construction process by exploring weak-labeling or LLM-based annotation to contribute large-scale ground-truth corpora for scientific online discourse.

(5) Lack of foundational general purpose models for scientific online discourse. This challenge comes as a direct consequence of the former. Scientific online discourse lacks general-purpose language models. Such models exist for scientific discourse from publications (e.g., SciBERT [141], SPECTER [184], BioGPT [185]), and for social media discourse (e.g., BERTweet [186], TimeLMs [187]), but are still lacking for scientific online discourse, i.e., the intersection of scientific discourse and social media discourse. General purpose models can serve as a strong task-agnostic basis on top of which task-specific fine-tuning can be performed. This might be especially relevant for scientific online discourse, where research has shown that the alternative instruction-tuned LLMs struggle with context-dependent classification tasks on social media [188, 189], especially in multi-lingual settings [190]. Existing works in scientific online discourse have contributed multi-label classifiers to differentiate various forms of scientific online discourse [28], and classifiers for specific downstream tasks such as cite-worthiness detection [131, 143], but have yet to contribute general purpose foundation models.

(6) The interplay between emotions and distortions of scientific online discourse is understudied. The analysis of the role of emotions in online scientific discussions is an understudied research field [191]. In such environments, emotions can be understood as a functional and instrumental means of disseminating potentially distorted scientific knowledge [191]. For instance, a recent analysis by Hafid et al. [7] on fact-checking data found that scientific web claims used more sentiment-related speech than non-scientific web claims. In a distinct study on health crisis policymaking, research shows the existence of an interplay between emotional and cognitive dynamics when perceiving scientific discourse [192]. Together, these results emphasize the need for a nuanced understanding of how science may be distorted in online discussions, especially in contexts of uncertainty such as global pandemics. To bridge this gap, future research should focus on the link between emotions and informal online scientific discourse by integrating emotion detection within the analysis of scientific web claims. Novel annotated data, models and tools are needed for a fine-grained analysis of the type of emotions involved in scientific distortion. Beyond tools, user-studies are necessary to investigate the effects of sentiment-based distortions on users' perceived trustworthiness of scientific online discourse. Ultimately, such work should lead to comprehensive audits of online platform mechanisms and new recommendations for platform design and user literacy to minimize the distortion of scientific findings online.

5 Conclusion

Unlike prior surveys on scientific fact-checking or citation analysis, this work unifies these perspectives under the concept of scientific web claims. In this survey, we provided the first systematic overview dedicated to scientific web claims, a specific yet underexplored subset of online discourse. By distinguishing them from both scientific claims originating in publications and from generic non-scientific web claims, we highlighted their unique linguistic, methodological, and societal challenges. Our review mapped existing definitions, task formulations, datasets, and methods across three major perspectives: Scientific fact-checking on the Web, Scientific citations on the Web, and Science communication on the Web. We showed that while each of these communities has begun to address scientific web claims, their efforts remain fragmented, domain-specific, and often limited in scale.

Future work should therefore focus on building large-scale, multilingual, and multimodal datasets (text, image, video, social metadata) in order to capture the heterogeneity of scientific web claims. This would make it possible to develop common benchmarks and to robustly compare existing models. Moreover, while pretrained models exist for many sub-domains, there is still no pretrained language model specifically dedicated to the intersection of science and social media. A promising perspective is therefore the creation of foundational models for online scientific discourse, capable of generalizing across a variety of tasks such as fact-checking, citation-source retrieval, and credibility assessment. Such models could be inspired from existing open-source/open-weights LLMs such as BioMedLM, BLOOM, and LLaMa, and from existing architectures for multimodal learning such as CLIP or LLaVA.

In the long term, methods and models for analyzing scientific web claims will need to be integrated into open science infrastructures (e.g., arXiv, PubMed, HAL) in order to facilitate verification and transparency. In this perspective, participatory verification may also be envisaged: by mobilizing collective intelligence, it can improve information reliability by detecting, contextualizing, and correcting false assertions more rapidly and at scale, while making uncertainties and evidence trails more visible to the public.

Beyond the technical dimension, advancing this field also requires an interdisciplinary dialogue with social sciences and communication studies to better understand public perceptions of science and the societal impacts of scientific misinformation.

By bridging computational, social, and communicative perspectives, we argue that the study of scientific web claims is a critical step towards ensuring accurate, reliable, and socially responsible science communication in the digital age, and that achieving this goal will require sustained interdisciplinary collaboration, particularly between AI (NLP/ML/IR) researchers, social scientists, and communication scholars to jointly define the theoretical foundations, develop robust methods, and align evaluation with real-world needs.

References

[1] Hafid, S.: Detection, linking and interpretation of science-related claims and their contexts from online discourse. PhD thesis, University of Montpellier (2024)

[2] Iyengar, S., Massey, D.S.: Scientific communication in a post-truth society. Proceedings of the National Academy of Sciences **116**(16), 7656–7661 (2019)

[3] Brüggemann, M., Lörcher, I., Walter, S.: Post-normal science communication: exploring the blurring boundaries of science and journalism. Journal of Science Communication **19**(3), 02 (2020)

[4] Neuberger, C., Weingart, P., Fähnrich, B., Fecher, B., Schäfer, M.S., Schmid-Petri, H., Wagner, G.G.: Der Digitale Wandel der Wissenschaftskommunikation. Berlin-Brandenburgische Akademie der Wissenschaften, ??? (2021)

[5] Lerner, B., Hubner, A.Y., Shulman, H.C.: Science populism impacts perceptions of credibility across scientific professions. Scientific Reports **15**(1), 28465 (2025) https://doi.org/10.1038/s41598-025-14115-8

[6] Dunwoody, S.: Science journalism: Prospects in the digital age. In: Routledge Handbook of Public Communication of Science and Technology, pp. 14–32. Routledge, ??? (2021)

[7] Hafid, S., Schellhammer, S., Kartal, Y.S., Papastergiou, T., Dietze, S., Bringay, S., Todorov, K.: An in-depth analysis of the linguistic characteristics of science claims on the web and their impact on fact-checking. ACM Trans. Web **19**(3) (2025) https://doi.org/10.1145/3746170

[8] Mohr, I., Wührl, A., Klinger, R.: Covert: A corpus of fact-checked biomedical covid-19 tweets. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 244–257 (2022)

[9] Saakyan, A., Chakrabarty, T., Muresan, S.: 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), pp. 2116–2129 (2021)

[10] Diggelmann, T., Boyd-Graber, J., Bulian, J., Ciaramita, M., Leippold, M.: Climate-fever: A dataset for verification of real-world climate claims. arXiv preprint arXiv:2012.00614 (2020)

[11] Smeros, P., Castillo, C., Aberer, K.: Sciclops: Detecting and contextualizing scientific claims for assisting manual fact-checking. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pp. 1692–1702 (2021)

[12] West, J.D., Bergstrom, C.T.: Misinformation in and about science. Proceedings of the National Academy of Sciences **118**(15), 1912444117 (2021)

[13] De Semir, V.: Scientific journalism: Problems and perspectives (2000)

[14] Banerjee, R., Kelkar, A.H., Logan, A.C., Majhail, N.S., Pemmaraju, N.: The democratization of scientific conferences: Twitter in the era of covid-19 and beyond. Current hematologic malignancy reports **16**, 132–139 (2021)

[15] Kreps, S.E., Kriner, D.L.: Model uncertainty, political contestation, and public trust in science: Evidence from the covid-19 pandemic. Science advances **6**(43), 4563 (2020)

[16] August, T., Card, D., Hsieh, G., Smith, N.A., Reinecke, K.: Explain like i am a scientist: The linguistic barriers of entry to r/science. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–12 (2020)

[17] Boland, K., Fafalios, P., Tchechmedjiev, A., Dietze, S., Todorov, K.: Beyond facts–a survey and conceptualisation of claims in online discourse analysis. Semantic Web **13**(5), 793–827 (2022)

[18] Lu, X., Pan, L., Liu, Q., Nakov, P., Kan, M.-Y.: Scitab: A challenging benchmark for compositional reasoning and claim verification on scientific tables. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 7787–7813 (2023)

[19] Wadden, D., Lin, S., Lo, K., Wang, L.L., Zuylen, M., Cohan, A., Hajishirzi, H.: Fact or fiction: Verifying scientific claims. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 7534–7550 (2020)

[20] Wadden, D., Lo, K., Kuehl, B., Cohan, A., Beltagy, I., Wang, L.L., Hajishirzi, H.: Scifact-open: Towards open-domain scientific claim verification. In: Findings of the Association for Computational Linguistics: EMNLP 2022, pp. 4719–4734 (2022)

[21] Wadden, D., Lo, K.: Overview and insights from the sciver shared task on scientific claim verification. In: Proceedings of the Second Workshop on Scholarly Document Processing, pp. 124–129 (2021)

[22] Pradeep, R., Ma, X., Nogueira, R., Lin, J.: Scientific claim verification with vert5erini. arXiv preprint arXiv:2010.11930 (2020)

[23] Zhang, Z., Li, J., Fukumoto, F., Ye, Y.: Abstract, rationale, stance: A joint model for scientific claim verification. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 3580–3586 (2021)

[24] Li, X., Burns, G., Peng, N.: A paragraph-level multi-task learning model for scientific fact-verification. arXiv preprint arXiv:2012.14500 (2020)

[25] Wadden, D., Lo, K., Wang, L., Cohan, A., Beltagy, I., Hajishirzi, H.: Multivers: Improving scientific claim verification with weak supervision and full-document context. In: Findings of the Association for Computational Linguistics: NAACL 2022, pp. 61–76 (2022)

[26] Ortega, R., Gómez-Pérez, J.M.: Sciclaims: An end-to-end generative system for biomedical claim analysis. arXiv preprint arXiv:2503.18526 (2025)

[27] Juneja, P., Mitra, T.: Human and technological infrastructures of fact-checking. Proceedings of the ACM on Human-Computer Interaction **6**(CSCW2), 1–36 (2022)

[28] Hafid, S., Schellhammer, S., Bringay, S., Todorov, K., Dietze, S.: Scitweets - a dataset and annotation framework for detecting scientific online discourse. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management. CIKM '22, pp. 3988–3992. Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3511808.3557693 . https://doi.org/10.1145/3511808.3557693

[29] Köhler, J., Shahi, G.K., Struß, J.M., Wiegand, M., Siegel, M., Mandl, T., Schütz, M.: Overview of the clef-2022 checkthat! lab task 3 on fake news detection. Working Notes of CLEF (2022)

[30] Nakov, P., Barrón-Cedeño, A., San Martino, G., Alam, F., Struß, J.M., Mandl, T., Míguez, R., Caselli, T., Kutlu, M., Zaghouani, W., *et al.*: Overview of the clef–2022 checkthat! lab on fighting the covid-19 infodemic and fake news detection. In: International Conference of the Cross-Language Evaluation Forum for European Languages, pp. 495–520 (2022). Springer

[31] Alam, F., Barrón-Cedeño, A., Cheema, G.S., Hakimov, S., Hasanain, M., Li, C., Míguez, R., Mubarak, H., Shahi, G.K., Zaghouani, W., et al.: Overview of the clef-2023 checkthat! lab task 1 on check-worthiness in multimodal and multigenre content. Working Notes of CLEF (2023)

[32] Thorne, J., Vlachos, A., Christodoulopoulos, C., Mittal, A.: 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), pp. 809–819 (2018)

[33] Augenstein, I., Lioma, C., Wang, D., Lima, L.C., Hansen, C., Hansen, C., Simonsen, J.G.: 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), pp. 4685–4697 (2019)

[34] Aly, R., Guo, Z., Schlichtkrull, M.S., Thorne, J., Vlachos, A., Christodoulopoulos, C., Cocarascu, O., Mittal, A.: Feverous: Fact extraction and verification over unstructured and structured information. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1) (2021)

[35] Aly, R., Guo, Z., Schlichtkrull, M.S., Thorne, J., Vlachos, A., Christodoulopoulos, C., Cocarascu, O., Mittal, A.: The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In: Aly, R., Christodoulopoulos, C., Cocarascu, O., Guo, Z., Mittal, A., Schlichtkrull, M., Thorne, J., Vlachos, A. (eds.) Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER), pp. 1–13. Association for Computational Linguistics, Dominican Republic (2021). https://doi.org/10.18653/v1/2021.fever-1.1 . https://aclanthology.org/2021.fever-1.1

[36] Tchechmedjiev, A., Fafalios, P., Boland, K., Gasquet, M., Zloch, M., Zapilko, B., Dietze, S., Todorov, K.: Claimskg: A knowledge graph of fact-checked claims. In: The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part II 18, pp. 309–324 (2019). Springer

[37] Wang, L., Shi, L., Kou, F., Zhu, L., Ma, C., Zhang, P., Xu, M., Li, Z.: Evicheck: Evidence-driven independent reasoning and combined verification method for fact-checking. In: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, pp. 3380–3388 (2025)

[38] Ge, Z., Wu, Y., Chin, D.W.K., Lee, R.K.-W., Cao, R.: Resolving conflicting evidence in automated fact-checking: A study on retrieval-augmented llms. In: Kwok, J. (ed.) Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25, pp. 9656–9664. International Joint Conferences on Artificial Intelligence Organization, ??? (2025). https://doi.org/10.24963/ijcai.2025/1073 . AI and Social Good. https://doi.org/10.24963/ijcai.2025/1073

[39] Wang, X., Cabrio, E., Villata, S.: Safe: Structured argumentation for fact-checking with explanations. In: 34th International Joint Conference on Artificial Intelligence (IJCAI-25), pp. 11114–11118 (2025). International Joint Conferences on Artificial Intelligence

[40] Altoe, F., Pinto, S.M.G., Pinto, H.S.: Explainable automatic fact-checking for journalists augmentation in the wild. In: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, pp. 10262–10270 (2025)

[41] Gong, H., Xu, W., Wu, S., Liu, Q., Wang, L.: Heterogeneous graph reasoning for fact checking over texts and tables. Proceedings of the AAAI Conference on

Artificial Intelligence **38**(1), 100–108 (2024) https://doi.org/10.1609/aaai.v38i1.27760

[42] Srba, I., Pecher, B., Tomlein, M., Moro, R., Stefancova, E., Simko, J., Bielikova, M.: Monant medical misinformation dataset: Mapping articles to fact-checked claims. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2949–2959 (2022)

[43] Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., et al.: The prisma 2020 statement: an updated guideline for reporting systematic reviews. bmj **372** (2021)

[44] Eldifrawi, I., Wang, S., Trabelsi, A.: Automated justification production for claim veracity in fact checking: A survey on architectures and approaches. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 6679–6692 (2024)

[45] Guo, Z., Schlichtkrull, M., Vlachos, A.: A survey on automated fact-checking. Transactions of the Association for Computational Linguistics **10**, 178–206 (2022)

[46] Thorne, J., Vlachos, A.: Automated fact checking: Task formulations, methods and future directions. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 3346–3359 (2018)

[47] Dmonte, A., Oruche, R., Zampieri, M., Calyam, P., Augenstein, I.: Claim verification in the age of large language models: A survey. arXiv preprint arXiv:2408.14317 (2024)

[48] Augenstein, I., Baldwin, T., Cha, M., Chakraborty, T., Ciampaglia, G.L., Corney, D., DiResta, R., Ferrara, E., Hale, S., Halevy, A., *et al.*: Factuality challenges in the era of large language models and opportunities for fact-checking. Nature Machine Intelligence **6**(8), 852–863 (2024)

[49] Akhtar, M., Schlichtkrull, M., Guo, Z., Cocarascu, O., Simperl, E., Vlachos, A.: Multimodal automated fact-checking: A survey. In: Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 5430–5448 (2023)

[50] Nakov, P., Corney, D., Hasanain, M., Alam, F., Elsayed, T., Barron-Cedeno, A., Papotti, P., Shaar, S., Da San Martino, G., *et al.*: Automated fact-checking for assisting human fact-checkers. In: IJCAI, pp. 4551–4558 (2021). International Joint Conferences on Artificial Intelligence

[51] Vladika, J., Matthes, F.: Scientific fact-checking: A survey of resources and approaches. In: Findings of the Association for Computational Linguistics: ACL 2023, pp. 6215–6230 (2023)

[52] Wührl, A., Resendiz, Y.M., Grimminger, L., Klinger, R.: What makes medical claims (un) verifiable? analyzing entity and relation properties for fact verification. In: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2046–2058 (2024)

[53] Anand, A., Anand, A., Setty, V., et al.: Quantemp: A real-world open-domain benchmark for fact-checking numerical claims. arXiv preprint arXiv:2403.17169 (2024)

[54] Strong, M., Vlachos, A.: Tsver: A benchmark for fact verification against time-series evidence. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pp. 29894–29914 (2025)

[55] Solovev, K., Pröllochs, N.: Moral emotions shape the virality of covid-19 misinformation on social media. In: Proceedings of the ACM Web Conference 2022, pp. 3706–3717 (2022)

[56] Bai, N., Silva Torres, R., Fensel, A., Metze, T., Dewulf, A.: Inferring climate change stances from multimodal tweets. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '24, pp. 2467–2471. Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3626772.3657950 . https://doi.org/10.1145/3626772.3657950

[57] Upadhyaya, A., Fisichella, M., Nejdl, W.: Intensity-valued emotions help stance detection of climate change twitter data. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI, pp. 6246–6254 (2023)

[58] Upadhyaya, A., Fisichella, M., Nejdl, W.: A multi-task model for emotion and offensive aided stance detection of climate change tweets. In: Proceedings of the ACM Web Conference 2023, pp. 3948–3958 (2023)

[59] Weinzierl, M., Harabagiu, S.: Identifying the adoption or rejection of misinformation targeting covid-19 vaccines in twitter discourse. In: Proceedings of the ACM Web Conference 2022, pp. 3196–3205 (2022)

[60] Wright, D., Augenstein, I.: Semi-supervised exaggeration detection of health science press releases. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 10824–10836 (2021)

[61] Li, Y., Zhang, J., Yu, B.: An nlp analysis of exaggerated claims in science news. In: Proceedings of the 2017 EMNLP Workshop: Natural Language Processing Meets Journalism, pp. 106–111 (2017)

[62] Barrón-Cedeño, A., Alam, F., Caselli, T., Da San Martino, G., Elsayed, T.,

Galassi, A., Haouari, F., Ruggeri, F., Struß, J.M., Nandi, R.N., *et al.*: The clef-2023 checkthat! lab: Checkworthiness, subjectivity, political bias, factuality, and authority. In: European Conference on Information Retrieval, pp. 506–517 (2023). Springer

[63] Arslan, F., Hassan, N., Li, C., Tremayne, M.: A benchmark dataset of check-worthy factual claims. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 14, pp. 821–829 (2020)

[64] Nakov, P., Barrón-Cedeño, A., Da San Martino, G., Alam, F., Struß, J.M., Mandl, T., Míguez, R., Caselli, T., Kutlu, M., Zaghouani, W., Li, C., Shaar, S., Shahi, G.K., Mubarak, H., Nikolov, A., Babulkov, N., Kartal, Y.S., Beltrán, J.: The clef-2022 checkthat! lab on fighting the covid-19 infodemic and fake news detection. In: Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part II, pp. 416–428. Springer, Berlin, Heidelberg (2022). https://doi.org/10.1007/978-3-030-99739-7_52 . https://doi.org/10.1007/978-3-030-99739-7_52

[65] Sundriyal, M., Akhtar, M.S., Chakraborty, T.: Leveraging social discourse to measure check-worthiness of claims for fact-checking. arXiv preprint arXiv:2309.09274 (2023)

[66] Gollapalli, S.D., Du, M., Ng, S.-K.: Identifying checkworthy cure claims on twitter. In: Proceedings of the ACM Web Conference 2023, pp. 4015–4019 (2023)

[67] Schuster, T., Fisch, A., Barzilay, R.: 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, pp. 624–643 (2021)

[68] Akhtar, M., Aly, R., Chen, Y., Deng, Z., Schlichtkrull, M., Whitehouse, C., Vlachos, A.: The 2nd automated verification of textual claims (averitec) shared task: Open-weights, reproducible and efficient systems. In: Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER), pp. 201–223 (2025)

[69] Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., Mittal, A.: The fact extraction and verification (fever) shared task. In: Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), pp. 1–9 (2018)

[70] Alam, F., Struß, J.M., Chakraborty, T., Dietze, S., Hafid, S., Korre, K., Muti, A., Nakov, P., Ruggeri, F., Schellhammer, S., *et al.*: Overview of the clef-2025 checkthat! lab: Subjectivity, fact-checking, claim normalization, and retrieval. In: International Conference of the Cross-Language Evaluation Forum for European Languages, pp. 199–223 (2025). Springer

[71] Nakov, P., Da San Martino, G., Elsayed, T., Barrón-Cedeño, A., Míguez, R., Shaar, S., Alam, F., Haouari, F., Hasanain, M., Mansour, W., *et al.*: Overview

of the clef–2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings 12, pp. 264–291 (2021). Springer

[72] Sarrouti, M., Abacha, A.B., M'rabet, Y., Demner-Fushman, D.: Evidence-based fact-checking of health-related claims. In: Findings of the Association for Computational Linguistics: EMNLP 2021, pp. 3499–3512 (2021)

[73] Wang, G., Harwood, K., Chillrud, L., Ananthram, A., Subbiah, M., Mckeown, K.: Check-covid: Fact-checking covid-19 news claims with scientific evidence. In: Findings of the Association for Computational Linguistics: ACL 2023, pp. 14114–14127 (2023)

[74] Wang, W.Y.: "liar, liar pants on fire": A new benchmark dataset for fake news detection. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 422–426 (2017)

[75] Lowrey, W.: The emergence and development of news fact-checking sites: Institutional logics and population ecology. Journalism studies **18**(3), 376–394 (2017)

[76] Graves, L., Cherubini, F.: The rise of fact-checking sites in europe. Digital News Project Report (2016)

[77] Shaar, S., Haouari, F., Mansour, W., Hasanain, M., Babulkov, N., Alam, F., Da San Martino, G., Elsayed, T., Nakov, P.: Overview of the clef-2021 checkthat! lab task 2 on detecting previously fact-checked claims in tweets and political debates. In: CLEF (working Notes), pp. 393–405 (2021)

[78] Nakov, P., Da San Martino, G., Alam, F., Shaar, S., Mubarak, H., Babulkov, N.: Overview of the clef-2022 checkthat! lab task 2 on detecting previously fact-checked claims (2022)

[79] Shaar, S., Babulkov, N., Da San Martino, G., Nakov, P.: That is a known lie: Detecting previously fact-checked claims. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 3607–3618 (2020)

[80] Anderson, S.T., Lin, K.K.: Chapter 3 - scientific method. In: Eltorai, A.E.M., Bakal, J.A., Haglin, J.M., Abboud, J.A., Crisco, J.J. (eds.) Translational Orthopedics. Handbook for Designing and Conducting Clinical and Translational Research, pp. 13–15. Academic Press, ??? (2024). https://doi.org/10.1016/B978-0-323-85663-8.00014-3 . https://www.sciencedirect.com/science/article/pii/B9780323856638000143

[81] Pryce Davis, R.R.: Evaluating Claims in Popular Science Media: Nature of Science Versus Dynamic Epistemological Knowledge. https://repository.isls.org/bitstream/1/2334/1/477-478.pdf. [Online; accessed 23-Feb-2024] (2012)

[82] Sober, E.: Evidence and Evolution: The Logic Behind the Science. Cambridge University Press, ??? (2008)

[83] Glockner, M., Hou, Y., Gurevych, I.: Missing counter-evidence renders nlp fact-checking unrealistic for misinformation. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 5916–5936 (2022)

[84] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research **21**(140), 1–67 (2020)

[85] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)

[86] Däniken, P., Deriu, J.M., Cieliebak, M.: Zhaw-cai at checkthat! 2023: Ensembling using kernel averaging. In: 14th Conference and Labs of the Evaluation Forum (CLEF), Thessaloniki, Greece, 18-21 September 2023, pp. 534–545 (2023). CEUR Workshop Proceedings

[87] Frick, R.A., Vogel, I., Choi, J.-E.: Fraunhofer sit at checkthat! 2023: enhancing the detection of multimodal and multigenre check-worthiness using optical character recognition and model souping. Working Notes of CLEF (2023)

[88] Aziz, A., Hossain, M., Chy, A.: Csecu-dsg at checkthat! 2023: transformer-based fusion approach for multimodal and multigenre check-worthiness. Working Notes of CLEF (2023)

[89] Li, Z., Liu, F., Yang, W., Peng, S., Zhou, J.: A survey of convolutional neural networks: analysis, applications, and prospects. IEEE transactions on neural networks and learning systems **33**(12), 6999–7019 (2021)

[90] Sadouk, H.T., Sebbak, F., Zekiri, H.E.: Es-vrai at checkthat! 2023: Analyzing checkworthiness in multimodal and multigenre (2023)

[91] Schofield, J., Tian, S., Truong, H.T.T., Heil, M.: Ds@ gt at checkthat! 2025: exploring retrieval and reranking pipelines for scientific claim source retrieval on social media discourse. arXiv preprint arXiv:2507.06563 (2025)

[92] Schreieder, T., Färber, M.: Claim2source at checkthat! 2025: zero-shot style transfer for scientific claim-source retrieval (2025)

[93] Beltagy, I., Peters, M.E., Cohan, A.: Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150 (2020)

[94] Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084 (2019)

[95] Wührl, A., Klinger, R.: Self-adaptive paraphrasing and preference learning for improved claim verifiability. arXiv preprint arXiv:2412.11653 (2024)

[96] Kotonya, N., Toni, F.: Explainable automated fact-checking for public health claims. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 7740–7754 (2020)

[97] Webersinke, N., Kraus, M., Bingler, J.A., Leippold, M.: Climatebert: A pretrained language model for climate-related text. arXiv preprint arXiv:2110.12010 (2021)

[98] Edelman, B., Skolnick, J.: Valsci: an open-source, self-hostable literature review utility for automated large-batch scientific claim verification using large language models. BMC bioinformatics **26**(1), 140 (2025)

[99] Curtis, T.L., Touzel, M.P., Garneau, W., Gruaz, M., Pinder, M., Wang, L.W., Krishna, S., Cohen, L., Godbout, J.-F., Rabbany, R., Pelrine, K.: Veracity: an open-source ai fact-checking system. In: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence. IJCAI '25 (2025). https://doi.org/10.24963/ijcai.2025/1254 . https://doi.org/10.24963/ijcai.2025/1254

[100] Agarwal, S., Sahu, G., Puri, A., Laradji, I.H., Dvijotham, K.D., Stanley, J., Charlin, L., Pal, C.: Litllm: A toolkit for scientific literature review. arXiv preprint arXiv:2402.01788 (2024)

[101] Haryanto, C.Y.: Llassist: simple tools for automating literature review using large language models. arXiv preprint arXiv:2407.13993 (2024)

[102] Wright, D., Augenstein, I.: Citeworth: Cite-worthiness detection for improved scientific document understanding. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 1796–1807 (2021)

[103] Soni, S., Bamman, D., Eisenstein, J.: Predicting long-term citations from short-term linguistic influence. In: Findings of the Association for Computational Linguistics: EMNLP 2022, pp. 5700–5716 (2022)

[104] Geng, H., Wang, D., Zhuang, F., Ming, X., Du, C., Jiang, T., Guo, H., Liu, R.: Modeling dynamic heterogeneous graph and node importance for future citation prediction. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, pp. 572–581 (2022)

[105] He, F., Lee, W.-C., Fu, T.-Y., Lei, Z.: Cines: Explore citation network and event sequences for citation forecasting. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 798–807 (2021)

[106] Yan, P., Kang, Y., Jiang, Z., Song, K., Lin, T., Sun, C., Liu, X.: Modeling scholarly collaboration and temporal dynamics in citation networks for impact prediction. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '24, pp. 2522–2526. Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3626772.3657926 . https://doi.org/10.1145/3626772.3657926

[107] Kunnath, S.N., Pride, D., Knoth, P.: Prompting strategies for citation classification. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, pp. 1127–1137 (2023)

[108] Tsai, H.-J., Yen, A.-Z., Huang, H.-H., Chen, H.-H.: Citation intent classification and its supporting evidence extraction for citation graph construction. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, pp. 2472–2481 (2023)

[109] Ji, T., Self, N., Fu, K., Chen, Z., Ramakrishnan, N., Lu, C.-T.: Dynamic multi-context attention networks for citation forecasting of scientific publications. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 7953–7960 (2021)

[110] Singh, J., Rungta, M., Yang, D., Mohammad, S.: Forgotten knowledge: Examining the citational amnesia in nlp. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 6192–6208 (2023)

[111] Jiang, S., Koch, B., Sun, Y.: Hints: Citation time series prediction for new publications via dynamic heterogeneous information network embedding. In: Proceedings of the Web Conference 2021, pp. 3158–3167 (2021)

[112] Li, X., Ouyang, J.: Related work and citation text generation: A survey. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 13846–13864 (2024)

[113] Aly, R., Tang, Z., Tan, S., Karypis, G.: Learning to generate answers with citations via factual consistency models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 11876–11896 (2024)

[114] Gu, N., Hahnloser, R.H.: Scilit: A platform for joint scientific literature discovery, summarization and citation generation. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System

Demonstrations), pp. 235–246 (2023)

[115] Ostendorff, M., Rethmeier, N., Augenstein, I., Gipp, B., Rehm, G.: Neighborhood contrastive learning for scientific document representations with citation embeddings. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 11670–11688 (2022)

[116] Luo, Z., Xie, Q., Ananiadou, S.: Citationsum: Citation-aware graph contrastive learning for scientific paper summarization. In: Proceedings of the ACM Web Conference 2023, pp. 1843–1852 (2023)

[117] Chen, X., Li, M., Gao, S., Yan, R., Gao, X., Zhang, X.: Scientific paper extractive summarization enhanced by citation graphs. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 4053–4062 (2022)

[118] Mao, Y., Zhong, M., Han, J.: Citesum: Citation text-guided scientific extreme summarization and domain adaptation with limited supervision. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 10922–10935 (2022)

[119] An, C., Zhong, M., Chen, Y., Wang, D., Qiu, X., Huang, X.: Enhancing scientific papers summarization with citation graph. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 12498–12506 (2021)

[120] Besançon, L., Cabanac, G., Labbé, C., Magazinov, A.: Sneaked references: Fabricated reference metadata distort citation counts. Journal of the Association for Information Science and Technology (2024)

[121] Cabanac, G., Labbé, C., Magazinov, A.: Tortured phrases: A dubious writing style emerging in science. evidence of critical issues affecting established journals. arXiv preprint arXiv:2107.06751 (2021)

[122] Wren, J.D., Georgescu, C.: Detecting anomalous referencing patterns in pubmed papers suggestive of author-centric reference list manipulation. Scientometrics **127**(10), 5753–5771 (2022)

[123] Ibrahim, H., Liu, F., Zaki, Y., Rahwan, T.: Google scholar is manipulatable. arXiv preprint arXiv:2402.04607 (2024)

[124] Mobarak, S., Stott, M.C., Lee, W.-J., Davé, M.S., Tarazi, M., Macutkiewicz, C.: The importance of social media to the academic surgical literature: Relationship between twitter activity and readership metrics. Surgery **170**(3), 650–656 (2021)

[125] Bornmann, L., Haunschild, R.: How to normalize twitter counts? a first attempt based on journals in the twitter index. Scientometrics **107**, 1405–1422 (2016)

[126] Eysenbach, G., *et al.*: Can tweets predict citations? metrics of social impact based on twitter and correlation with traditional metrics of scientific impact. Journal of medical Internet research **13**(4), 2012 (2011)

[127] Kousha, K., Thelwall, M.: Covid-19 publications: Database coverage, citations, readers, tweets, news, facebook walls, reddit posts. Quantitative Science Studies **1**(3), 1068–1091 (2020)

[128] Hassan, S.-U., Imran, M., Gillani, U., Aljohani, N.R., Bowman, T.D., Didegah, F.: Measuring social media activity of scientific literature: An exhaustive comparison of scopus and novel altmetrics big data. Scientometrics **113**, 1037–1057 (2017)

[129] Alperin, J.P., Fleerackers, A., Riedlinger, M., Haustein, S.: Second-order citations in altmetrics: A case study analyzing the audiences of covid-19 research in the news and on social media. bioRxiv, 2023–04 (2023)

[130] Didegah, F., Mejlgaard, N., Sørensen, M.P.: Investigating the quality of interactions and public engagement around scientific papers on twitter. Journal of informetrics **12**(3), 960–971 (2018)

[131] Hafid, S., Ammar, W., Bringay, S., Todorov, K.: Cite-worthiness detection on social media: A preliminary study. In: International Workshop on Natural Scientific Language Processing and Research Knowledge Graphs, pp. 19–30 (2024). Springer

[132] Ravenscroft, J., Clare, A., Liakata, M.: Harrigt: Linking news articles to scientific literature. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics-System Demonstrations (2018), pp. 19–24 (2018)

[133] Wang, J., Yu, B.: Linking health news to research literature. arXiv preprint arXiv:2107.06472 (2021)

[134] Hafid, S., Kartal, Y.S., Schellhammer, S., Boland, K., Dimitrov, D., Bringay, S., Todorov, K., Dietze, S.: Overview of the clef-2025 checkthat! lab task 4 on scientific web discourse. linguistics **12**, 13 (2025)

[135] Struß, J.M., Schellhammer, S., Dietze, S., Setty, V., Chakraborty, T., Nakov, P., Anand, A., Chungkham, P., Hafid, S., Sahnan, D., et al.: The clef-2026 checkthat! lab: Advancing multilingual fact-checking. arXiv preprint arXiv:2602.09516 (2026)

[136] Sugiyama, K., Kumar, T., Kan, M.-Y., Tripathi, R.C.: Identifying citing sentences in research papers using supervised learning. In: 2010 International Conference on Information Retrieval & Knowledge Management (CAMP), pp. 67–72 (2010). IEEE

[137] Bird, S., Dale, R., Dorr, B.J., Gibson, B.R., Joseph, M.T., Kan, M.-Y., Lee, D., Powley, B., Radev, D.R., Tan, Y.F., *et al.*: The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In: LREC (2008)

[138] Färber, M., Thiemann, A., Jatowt, A.: To cite, or not to cite? detecting citation contexts in text. In: Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings 40, pp. 598–603 (2018). Springer

[139] Färber, M., Thiemann, A., Jatowt, A.: A high-quality gold standard for citation-based tasks. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (2018)

[140] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

[141] Beltagy, I., Lo, K., Cohan, A.: 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), pp. 3615–3620. Association for Computational Linguistics, Hong Kong, China (2019). https://doi.org/10.18653/v1/D19-1371 . https://aclanthology.org/D19-1371

[142] Hendricks, G., Tkaczyk, D., Lin, J., Feeney, P.: Crossref: The sustainable source of community-owned scholarly metadata. Quantitative Science Studies **1**(1), 414–427 (2020)

[143] Hafid, S., Kartal, Y.S., Schellhammer, S., Jacot, V., Bringay, S., Dietze, S., Todorov, K.: Disambiguation of implicit scientific references on x. In: Proceedings of the 36th ACM Conference on Hypertext and Social Media. HT '25, pp. 165–170. Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3720553.3746676 . https://doi.org/10.1145/3720553.3746676

[144] Dangovski, R., Shen, M., Byrd, D., Jing, L., Tsvetkova, D., Nakov, P., Soljačić, M.: We can explain your research in layman's terms: Towards automating science journalism at scale. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 12728–12737 (2021)

[145] Nallapati, R., Zhou, B., Santos, C., Gulçehre, Ç., Xiang, B.: Abstractive text summarization using sequence-to-sequence rnns and beyond. In: Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, pp. 280–290 (2016)

[146] Dempster, G., Sutherland, G., Keogh, L.: Scientific research in news media: a

case study of misrepresentation, sensationalism and harmful recommendations. Journal of Science Communication **21**(1), 06 (2022)

[147] Wijnker, W., Smeets, I., Burger, P., Willems, S.: Debunking strategies for misleading bar charts. Journal of Science Communication **21**(7), 07 (2022)

[148] Reif, A., Guenther, L., Yokoyama, H.M.: Public (dis) trust in science in digital media environments. SISSA Medialab srl (2024)

[149] Schäfer, M.S., Kremer, B., Mede, N.G., Fischer, L.: Trust in science, trust in chatgpt? how germans think about generative ai as a source in science communication. Journal of Science Communication **23**(9), 04 (2024)

[150] Zimmermann, F., Petersen, C., Kohring, M.: Who, if not science, can you trust to guide you through a crisis? the relationship between public trust in science and exposure to established and alternative online sources in times of crisis. Journal of Science Communication: JCOM **23**(09: Public (dis) trust in science in digital media environments, A05), 1–19 (2024)

[151] Guenther, L., Schröder, J.T., Reif, A., Brück, J., Taddicken, M., Weingart, P., Jonas, E.: Intermediaries in the limelight: how exposure to trust cues in content about science affects public trust in science. Journal of Science Communication **23**(9), 06 (2024)

[152] LeBel, E.P., McCarthy, R.J., Earp, B.D., Elson, M., Vanpaemel, W.: A unified framework to quantify the credibility of scientific findings. Advances in Methods and Practices in Psychological Science **1**(3), 389–402 (2018)

[153] Augenstein, I.: Determining the credibility of science communication. In: Proceedings of the Second Workshop on Scholarly Document Processing, pp. 1–6 (2021)

[154] Portman, J., Miara Ms, V.Y., Baram-Tsabari, A.: How does social-media-based science communication affect young audiences? a scoping review of impact making. Journal of Science Communication **24**(5), 02 (2025)

[155] Haunschild, R., Bornmann, L., Potnis, D., Tahamtan, I.: Investigating dissemination of scientific information on twitter: A study of topic networks in opioid publications. Quantitative Science Studies, 1–56 (2021)

[156] Carlson, J., Harris, K.: Quantifying and contextualizing the impact of biorxiv preprints through automated social media audience segmentation. PLoS Biology **18**(9), 3000860 (2020)

[157] Roedema, T., Broerse, J.E., Kupper, F.: "who is going to believe me, if i say 'i'm a researcher?'"—scientists' role repertoires in online public engagement. Journal of Science Communication **20**(3), 1–19 (2021)

[158] Fatima, M., Strube, M.: Cross-lingual science journalism: Select, simplify and rewrite summaries for non-expert readers. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1843–1861 (2023)

[159] Pu, D., Wang, Y., Loy, J.E., Demberg, V.: Scinews: From scholarly complexities to public narratives–a dataset for scientific news report generation. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 14429–14444 (2024)

[160] Cardenas, R., Yao, B., Wang, D., Hou, Y.: 'don't get too technical with me': A discourse structure-based framework for automatic science journalism. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 1186–1202 (2023)

[161] Steinberger, J., Jezek, K., *et al.*: Using latent semantic analysis in text summarization and summary evaluation. Proc. ISIM **4**(93-100), 8 (2004)

[162] Zheng, H., Lapata, M.: Sentence centrality revisited for unsupervised summarization. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (2019). Association for Computational Linguistics

[163] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)

[164] AI, M. https://ai.meta.com/blog/meta-llama-3/. [Online; accessed 27-May-2024] (2024)

[165] Kohler, S., Promies, N., Lehmkuhl, M.: Patterns in the journalistic selection of neuroscientific research results

[166] Cao, Y., Nair, A.M., Eyimife, E., Soofi, N.J., Subbalakshmi, K., Wullert II, J.R., Basu, C., Shallcross, D.: Can large language models detect misinformation in scientific news reporting? arXiv preprint arXiv:2402.14268 (2024)

[167] Flemming, D., Cress, U., Kimmig, S., Brandt, M., Kimmerle, J.: Emotionalization in science communication: The impact of narratives and visual representations on knowledge gain and risk perception. Frontiers in Communication **3**, 3 (2018)

[168] König, L., Breves, P.: Providing health information via twitter: professional background and message style influence source trustworthiness, message credibility and behavioral intentions. Journal of science communication **20**(4), 04 (2021)

[169] Kimmerle, J., Flemming, D., Feinkohl, I., Cress, U.: How laypeople understand the tentativeness of medical research news in the media: An experimental study on the perception of information about deep brain stimulation. Science Communication **37**(2), 173–189 (2015)

[170] Bromme, R., Mede, N.G., Thomm, E., Kremer, B., Ziegler, R.: An anchor in troubled times: Trust in science before and within the covid-19 pandemic. PloS one **17**(2), 0262823 (2022)

[171] Robbins, M., Calabrese, C., Featherstone, J.D., Barnett, G.A.: Understanding knowledge and perceptions of genome editing technologies: a textual analysis of major agricultural stakeholder groups. Journal of Science Communication **20**(5), 07 (2021)

[172] Klaus, G., Oswald, L., Ernst, A., Merk, C.: Effects of opinion statements on laypeople's acceptance of a climate engineering technology. comparing the source credibility of researchers, politicians and a citizens' jury. Journal of Science Communication **20**(1), 03 (2021)

[173] Frewer, L., Hunt, S., Brennan, M., Kuznesof, S., Ness, M., Ritson, C.: The views of scientific experts on how the public conceptualize uncertainty. Journal of risk research **6**(1), 75–85 (2003)

[174] Moukarzel, S., Rehm, M., Del Fresno, M., Daly, A.J.: Diffusing science through social networks: The case of breastfeeding communication on twitter. PloS one **15**(8), 0237471 (2020)

[175] Walter, S., Lörcher, I., Brüggemann, M.: Scientific networks on twitter: Analyzing scientists' interactions in the climate change debate. Public Understanding of Science **28**(6), 696–712 (2019)

[176] Horta, V., Ströele, V., Braga, R., David, J.M.N., Campos, F.: Analyzing scientific context of researchers and communities by using complex network and semantic technologies. Future Generation Computer Systems **89**, 584–605 (2018)

[177] Fang, Z., Costas, R., Tian, W., Wang, X., Wouters, P.: How is science clicked on twitter? click metrics for bitly short links to scientific publications. Journal of the Association for Information Science and Technology **72**(7), 918–932 (2021)

[178] Aksnes, D.W., Langfeldt, L., Wouters, P.: Citations, citation indicators, and research quality: An overview of basic concepts and theories. Sage Open **9**(1), 2158244019829575 (2019)

[179] Díaz-Faes, A.A., Bowman, T.D., Costas, R.: Towards a second generation of 'social media metrics': Characterizing twitter communities of attention around science. PloS one **14**(5), 0216408 (2019)

[180] Suleski, J., Ibaraki, M.: Scientists are talking, but mostly to each other: a quantitative analysis of research represented in mass media. Public Understanding of Science **19**(1), 115–125 (2010)

[181] Côté, I.M., Darling, E.S.: Scientists on twitter: Preaching to the choir or singing from the rooftops? Facets **3**(1), 682–694 (2018)

[182] Chandrasekharan, E., Samory, M., Jhaver, S., Charvat, H., Bruckman, A., Lampe, C., Eisenstein, J., Gilbert, E.: The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. Proceedings of the ACM on Human-Computer Interaction **2**(CSCW), 1–25 (2018)

[183] Alperin, J.P., Fleerackers, A., Riedlinger, M., Haustein, S.: Second-order citations in altmetrics: A case study analyzing the audiences of covid-19 research in the news and on social media. Quantitative Science Studies, 1–17 (2024)

[184] Cohan, A., Feldman, S., Beltagy, I., Downey, D., Weld, D.S.: Specter: Document-level representation learning using citation-informed transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 2270–2282 (2020)

[185] Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., Liu, T.-Y.: Biogpt: generative pre-trained transformer for biomedical text generation and mining. Briefings in bioinformatics **23**(6), 409 (2022)

[186] Nguyen, D.Q., Vu, T., Tuan Nguyen, A.: BERTweet: A pre-trained language model for English tweets. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 9–14. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.emnlp-demos.2 . https://aclanthology.org/2020.emnlp-demos.2

[187] Loureiro, D., Barbieri, F., Neves, L., Anke, L.E., Camacho-Collados, J.: Timelms: Diachronic language models from twitter. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 251–260 (2022)

[188] Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., Yang, D.: Can large language models transform computational social science? Computational Linguistics **50**(1), 237–291 (2024)

[189] Mu, Y., Wu, B.P., Thorne, W., Robinson, A., Aletras, N., Scarton, C., Bontcheva, K., Song, X.: Navigating prompt complexity for zero-shot classification: A study of large language models in computational social science. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 12074–12086 (2024)

[190] Mohta, J., Ak, K., Xu, Y., Shen, M.: Are large language models good annotators? In: Proceedings On, pp. 38–48 (2023). PMLR

[191] Taddicken, M., Reif, A.: Between evidence and emotions: emotional appeals in science communication. Media and Communication **8**(1), 101–106 (2020)

[192] Lemor, A., Montpetit, É.: Exploring the role of uncertainty, emotions, and scientific discourse during the covid-19 pandemic. Policy and Society **43**(3), 289–303 (2024)

[193] Savchev, A.: Ai rational at checkthat!-2022: Using transformer models for tweet classification. In: CLEF (Working Notes), pp. 656–659 (2022)

[194] Clark, K., Luong, M.-T., Le, Q.V., Manning, C.D.: Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555 (2020)

[195] Kavatagi, S., Rachh, R., Mulimani, M.: Vtu_bgm at checkthat! 2022: An autoregressive encoding model for detecting check-worthy claims (2022)