SciTweets - A Dataset and Annotation Framework for Detecting Scientific Online Discourse

Salim Hafid, Sebastian Schellhammer, Sandra Bringay, Konstantin Todorov, Stefan Dietze

SciTweets - A Dataset and Annotation Framework for Detecting Scientific Online Discourse

Salim Hafid*
salim.hafid@lirmm.fr
LIRMM, CNRS, University of Montpellier
Montpellier, France

Sebastian Schellhammer*
sebastian.schellhammer@gesis.org
GESIS - Leibniz Institute for the Social Sciences
Cologne, Germany

Sandra Bringay
LIRMM, CNRS, University of Montpellier
Montpellier, France

Konstantin Todorov
LIRMM, CNRS, University of Montpellier
Montpellier, France

Stefan Dietze
GESIS - Leibniz Institute for the Social Sciences
Cologne & Heinrich-Heine-University Düsseldorf, Germany

ABSTRACT

Scientific topics, claims and resources are increasingly debated as part of online discourse, where prominent examples include discourse related to COVID-19 or climate change. This has led to both significant societal impact and increased interest in scientific online discourse from various disciplines. For instance, communication studies aim at a deeper understanding of biases, quality or spreading pattern of scientific information whereas computational methods have been proposed to extract, classify or verify scientific claims using NLP and IR techniques. However, research across disciplines currently suffers from both a lack of robust definitions of the various forms of science-relatedness as well as appropriate ground truth data for distinguishing them. In this work, we contribute (a) an annotation framework and corresponding definitions for different forms of scientific relatedness of online discourse in Tweets, (b) an expert-annotated dataset of 1261 tweets obtained through our labeling framework reaching an average Fleiss Kappa κ of 0.63, (c) a multi-label classifier trained on our data able to detect science-relatedness with 89% F1 and also able to detect distinct forms of scientific knowledge (claims, references). With this work we aim to lay the foundation for developing and evaluating robust methods for analysing science as part of large-scale online discourse.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; **Machine learning**; **Discourse, dialogue and pragmatics**; • **Networks** → *Online social networks.*

KEYWORDS

Science Discourse, Scientific Claims, Dataset Annotation

ACM Reference Format:
Salim Hafid, Sebastian Schellhammer, Sandra Bringay, Konstantin Todorov, and Stefan Dietze. 2018. SciTweets - A Dataset and Annotation Framework for Detecting Scientific Online Discourse. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Scientific topics, claims and resources are increasingly debated as part of societal discourse in online news and social media. Examples include the increased participation of journalists, policy makers, scientists, celebrities and the general public in scientific online discourse [13], where Twitter in particular is used widely for discussing scientific insights (see examples in Table 1). Specifically for emerging topics such as COVID-19, an elevated role of preliminary scientific results beyond the traditional peer review system can be observed, for instance, as part of preprints, opinion pieces and informal utterances in scientific online debates [27].

Table 1: Examples (tweets 1 to 4) and Counterexamples (tweet 5) of scientific online discourse tweets

(1) Donating blood not only helps others, but reduces the rate of cancer and heart disease in the donor.
(2) via @medical_xpress A new in vitro (test tube) study, ""Dietary functional benefits of Bartlet http://t.co/Qv1C1GjQin #UFO4UBlogHealth
(3) How is @UChicagoIME shaping the future of science? Find out on April 6!
(4) Study: Shifts in electricity generation spur net job growth, but coal jobs decline - via @DukeU http://t.co/AXGmKUPata
(5) My father got COVID-19.

While it has been recognised that online discourse as observed in news and social Web platforms produces phenomena such as misinformation spread [28] or reinforcement of biases [10] with

*Both authors contributed equally to this research.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

widely assumed harmful effects for democratic societies [2], misinformation on scientific topics such as COVID-19 or climate change has particularly detrimental effects on society and public health.

This has led to research into scientific online discourse across various disciplines. From a social sciences perspective, works measure the engagement with scientific publications on social media [5, 6, 12] or investigate the role of social media in facilitating the flow of scientific information [3]. In science communication, research discusses implications of risk communication [16] or the spreading pattern associated with preliminary scientific results and the diffusion of science through social networks [17, 30], while research in cognitive and social psychology investigates the perceived trustworthiness of scientific online discourse [15].

Methodological research at the intersection of NLP, information retrieval and machine learning is aimed at detecting, classifying or verifying (scientific) claims and discourse [11, 18, 19, 24, 25], and is a key facilitator for large-scale interdisciplinary analysis of science discourse. Prior works often focus on actual scholarly publications [14, 21], where the formality of language differs substantially from science claims in online news and social media, e.g., Twitter.

Datasets are crucial to facilitate such research and were proposed with various definitions of science-relatedness that each are based on specific assumptions. Some works define scientific claims as claims expressing an aspect of one or more scientific entities [24, 29], however with no robust definition of what a scientific entity is. Other works selected scientific claims by restricting the domain to one they deemed scientific (e.g COVID-19 [23], climate change [7], or medicine [25]), where generalisability is limited.

Moreover, claims may be synthetically generated [29, 31] or ground truth data is constructed using simple heuristics exploiting keywords or referenced pay-level-domains (PLDs) based on narrow predefined dictionaries [22], and predicates [21]. Generally, robust definitions of science-relatedness are lacking that distinguish between items that actually convey scientific knowledge, e.g., a science claim or a reference to a scientific resource, and other forms of science-relatedness, for instance, items stating a fact about a particular scientist without actually conveying scientific knowledge. Hence, the lack of datasets that are based on sound definitions of science-relatedness is a crucial obstacle for advancing research into scientific online discourse and for fairly evaluating and benchmarking existing NLP and IR methods in this context.

In order to address these challenges, we propose *SciTweets*, a publicly available dataset and annotation framework for science discourse on Twitter.

In particular, we make the following contributions:

(1) **A hierarchical definition of science-relatedness.** Through an iterative process of literature review, data exploration, expert labeling and deliberation, we devise a set of definitions of science relatedness distinguishing between tweets that convey scientific knowledge in the form of claims or scientific references and tweets with a broader relatedness to research contexts and processes (Section 2).

(2) **Annotation framework.** Building on our set of reusable definitions, we provide an annotation framework consisting of iteratively improved and evaluated labeling instructions and a data sampling strategy informed by heuristics and a weakly supervised classifier for ensuring a balanced set of labels.

(3) **Ground truth dataset.** We provide a dataset of 1261 tweets, labeled using our annotation framework by four expert annotators who each labeled the whole set, reaching Fleiss κ inter-annotator agreements between 0.61 and 0.66. All data is made publicly available under a CC-BY Creative Commons license.

(4) **Baseline classification models.** Demonstrating the utility of our dataset and definitions, we train a baseline classifier achieving approx. 89% F1 in distinguishing science from non-science-related tweets and 78 % F1 in detecting science claims, references and otherwise related tweets (macro average in all cases).

2 CONSTRUCTING THE SCITWEETS CORPUS

This section describes the annotation framework, sampling strategy, the annotation process and the resulting *SciTweets* dataset.

2.1 Category Definitions & Annotation Framework

Given the lack of robust definitions of science-relatedness, we followed an iterative process of data exploration, literature review and preliminary labeling rounds. We started by selecting and observing samples of science-related texts coming from science-related datasets [7, 14, 21, 23–25, 29] and reviewing related definitions together with researchers from various disciplines. We then manually classified them into categories, and held intermediate annotation rounds with new samples to test the agreement across categories. We then identified difficult examples that had high interannotator disagreement and updated categories and annotation guidelines accordingly. In total, we held two intermediate annotation rounds with three to four annotators to improve the robustness of the categories. Definitions and labeling guidelines were considered final and ready for annotating actual ground truth data (see Section 2.2) after we obtained a satisfactory inter-annotator agreement and they facilitated an exhaustive labeling of all tweets. Categories and their definitions are described in full here[1], depicted in Figure 1 and summarised below.

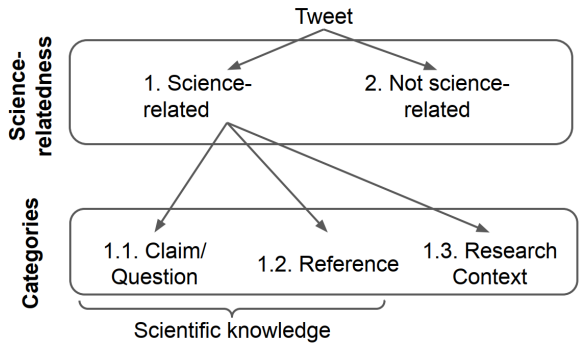

Figure 1: Categories of science-relatedness

Category 1 - Science-related: Texts that fall under at least one of the following categories:

> **Category 1.1 - Scientific knowledge (scientifically verifiable claims):** Does the text include a claim or a question that could be scientifically verified? (see Tweet 1 in Table 1)

[1] Code and data available at: https://github.com/AI-4-Sci/SciTweets

Category 1.2 - Reference to scientific knowledge: Does the text include at least one reference to scientific knowledge? References can either be direct, e.g., DOI, title of a paper or indirect, e.g., a link to an article that includes a direct reference (see Tweet 2 in Table 1).

Category 1.3 - Related to scientific research in general: Does the text mention a scientific research context (e.g., mention of a scientist, scientific research efforts, research findings)? (see Tweet 3 in Table 1)

Category 2 - Not science-related: Texts that don't fall under either of the 3 previous categories. (see Tweet 5 in Table 1)

One of the main findings from our intermediate annotation rounds was that science-relatedness cannot be defined through the presence of specific entities (e.g., "COVID-19", "vaccine") or specific domains (e.g., medicine, biology, climate) as done by several related works [24, 29], given that the notion of scientific entity itself is ambiguous and therefore ill-defined, and hence, may lead to both false-positives and false-negatives. The reason for that is that science-relatedness is not an inherent attribute that an entity or a domain can have or not have, but rather a volatile attribute that strongly depends on the context. For instance, the word "blood" is a scientific entity in the sentence "More money is put into research efforts trying to create artificial blood", but is not a scientific entity in the sentence "He's so good at playing the guitar, it's like it's in his blood!". Therefore, for our Category 1.1 which is about texts containing scientific claims or questions, we applied the criterium of scientific verifiability, which we define as the "possibility of being verified in a document created by scientists (e.g., a scientific paper or statistics from a research institution), or verified in a document that could in theory be created by scientists, regardless of how hard that verification might be".

Category 1.1 is crucial to distinguishing discourse that carries scientific knowledge (see Tweet 1 in Table 1) from discourse that is just related to science in general, thereby enabling important tasks such as scientific claim retrieval, claim verification and claim linking. Category 1.2 (see Tweet 2 in Table 1) is crucial for research that aims at understanding the role different sources play in online science discourse and the impact of science in various sources (scientific journals, online news articles, preprints, blogs). Together, Categories 1.1 and 1.2 are crucial to identifying online discourse that carries scientific knowledge in order to facilitate research into the evolution of scientific discourse in online environments. Category 1.3 is important for distinguishing discourse that does not carry scientific knowledge or a reference to scientific knowledge but nonetheless clearly mentions a scientific research context (see Tweet 3 in Table 1). Tweets in that category are able to facilitate research into public perception of discourse about science and the scientific process rather than actual scientific insights. These three subcategories are not mutually exclusive, e.g., Tweet 4 in Table 1 belongs to Categories 1.1, 1.2 and 1.3. We also introduce additional *Confidence Score*, *Compound Claim* and *Irony* labels.

2.2 Data and Sampling Strategy

To create our expert-annotated *SciTweets* dataset, we sample tweets from the full text archive underlying TweetsKB [8], a public knowledge graph containing metadata of more than 2 billion English tweets created from archiving 10 billion raw tweets through the 1% Twitter API stream between February 2013 to December 2020. We extract URLs from the tweets text and resolve shortened URLs for all tweets prior to April 2018, since later ones are already extracted and resolved in the corpus.

Preliminary data exploration has shown that the percentage of science-related tweets is very small, where random sampling would surface tweets dominated by negative cases. Hence, we do not sample tweets randomly but aim to ensure a more balanced ratio of science-related and unrelated tweets. Further, we aim at including hard negative examples (e.g., "My second shot of COVID-19 vaccine gave me headache.") instead of tweets that are obviously unrelated (e.g., "I like pop music."). Using this approach, labeling efforts are steered towards potentially relevant cases rather than towards obviously unrelated tweets that can be obtained with high precision through random sampling combined with minimal labeling or simple heuristics. We deploy a two-stage annotation process.

Sampling & Annotation Stage 1. First, we apply basic heuristics (see details here[2]) on the Twitter corpus, to identify potentially science-related tweets. These identify tweets for Category 1.1 looking for patterns like nouns that are connected with argumentative predicates like 'cause' or 'lead to', filtered by a predefined list of scientific terms. Category 1.2 tweets are selected by filtering tweets that contain a URL with a subdomain that is included in a predefined list of 17,500 scientific subdomains from open access repositories, science newspaper sections and science magazines (e.g., "link.springer.com", "sciencedaily.com"). For Category 1.3, we retrieve tweets that mention terms and phrases related to scientists, the scientific research process and publications. Applying this approach on a randomly sampled set of 5 million tweets obtains 18,000 tweets that are likely to fall into either subcategory of Category 1. Given that these heuristics employ a strict pattern-matching, likely leading to high precision and low recall results, we expect the identified 18,000 tweets to lack diversity. To obtain more diverse candidate tweets, we finetune a BERTweet [20] multi-label classifier on the 18,000 tweets as positive examples and a random sample of 18,000 tweets that were not identified by the heuristics as negative examples. Our intuition is that classification results from this classifier will lead to less precise but more diverse tweets in the final set.

After training, we use both the heuristics and the classifier to label the tweets in a randomly selected set of 100 K tweets, where each tweet is assigned two labels per category, i.e., one through the heuristics and one from the classifier. Assuming that classifier predictions will result in different tweets than the heuristics and to ensure diversity in the dataset to be annotated, we obtain all tweets where the labels of the heuristics and classifier differ for at least one of the three categories 1.1, 1.2 and 1.3, resulting in a set of 1046 tweets that were annotated in a first labeling step (Section 2.3).

Sampling & Annotation Stage 2. These 1046 expert annotations obtained in the first annotation stage are used to train a new multi-label classifier for the three science-relatedness subcategories (see details in Section 3). After applying the classifier to a new set of 100K randomly selected tweets, we obtain the 100 tweets with

[2]https://github.com/AI-4-Sci/SciTweets

the highest confidence from the resulting predictions for each category, i.e., 300 tweets in total. After filtering out duplicates, this results in 215 additional tweets, that were again annotated by expert annotators 2.3), resulting in a total set of 1261 expert-annotated tweets.

2.3 Annotation & Quality Assurance

All tweets are labeled by the same four annotators, including the two main co-authors of this paper as well as a PhD student and a bachelor's student, both from the field of Computer Science. Before annotating, we held individual training sessions with the annotators in which we examined examples and counterexamples of each category of the labeling framework to ensure the labeling task was understood correctly. Cases of weak disagreement, i.e., where the label of one annotator differs to the labels of the three remaining annotators, were consolidated using a majority vote. For tweets from the second annotation stage, cases of high disagreement, i.e., where both labels *Yes* and *No* were each selected by two annotators, were resolved in a discussion between all four annotators, whereas we did not resolve high disagreements for tweets from the first annotation stage. We measured the inter-annotator agreement by computing Fleiss Kappa κ [9] with the labels of all four annotators to evaluate the annotation quality. Agreement scores of 0.61, 0.63 and 0.66 for categories 1.1, 1.2 and 1.3 averaging a score of 0.63 are comparable to results on similar tasks[3]. Since we estimate our task to be more difficult than those tasks and because the fleiss score has been shown to be different based on the number of annotators [9], we estimate our results to be comparable to the mentioned papers and thus to be encouraging.

2.4 Statistics

The *SciTweets* dataset consists of 1261 human-annotated tweets including the labels of the individual annotators, as well as the consolidated ground-truth label for each category. Table 2 shows the distributions of the ground-truth labels with a high ratio (31.88%) of science-related tweets (Category 1), where the consolidated label to at least one science-relatedness subcategory is *Yes*, and a balanced distribution of subcategories ranging from 15.65% for Category 1.2 to 23.82% for Category 1.1.

Table 2: Distribution of the ground-truth labels

Labels	Category 1	Category 1.1	Category 1.2	Category 1.3
Yes	402 (31.88%)	283 (23.82%)	190 (15.65%)	259 (21.32%)
No	859 (68.12%)	905 (76.18%)	1024 (84.35%)	956 (78.68%)

3 CLASSIFICATION OF SCIENCE-RELATEDNESS

We evaluate a single multi-label classifier for both the binary task of classifying if a tweet is science-related as well as the multi-label task of assigning one or more subcategories of science-relatedness to a tweet. Experimenting with different base models showed that SciBERT [4] provides superior performance on the tasks. To evaluate the multi-label classifier on the binary task we map the classifier's multi-label predictions to a binary prediction, i.e., the classifier predicts a tweet to be science-related if it assigns at least one of the three subcategories 1.1, 1.2 or 1.3. Table 3 shows the performance of the classifier for both tasks applying 10-fold cross validation using all *SciTweets* tweets without high disagreements. As expected, the classifier performs better on the binary task, because a false positive prediction for categories 1.1, 1.2, and 1.3 could still be a true positive prediction for the binary task. Given the high ratio of science-related tweets in *SciTweets* compared to TweetsKB, the precision for both tasks is expected to be lower when performed on a random sample of TweetsKB, because of the increase of false positives. Hence, to get a more representative performance estimate, we train a new multilabel classifier on the 1046 tweets from annotation stage 1 and set a prediction threshold for each subcategory, so that the classifier makes only 100 positive predictions per subcategory out of 100K tweets. Table 4 shows the precision for these positive predictions (215 tweets, labeled in the second annotation stage).

Table 3: Classifier performance (binary and multilabel tasks)

Task	Category	Precision	Recall	F1
binary	1 - Science-related	84.70	83.99	84.34
	2 - Not Science-related	92.67	93.03	92.85
multi	1.1 - Scientific Claim	75.00	81.18	77.97
	1.2 - Reference	76.19	77.01	76.60
	1.3 - Research Context	81.06	79.65	80.35

Table 4: Classifier Performance for multilabel task

Metric	Category 1.1	Category 1.2	Category 1.3
Precision@100	85.00%	74.00%	86.00%

4 CONCLUSION

Resources and claims related to science contribute significantly to online discourse, in particular with respect to emerging topics of high societal importance, such as climate change or COVID-19. The understanding of science-related online discourse can help prevent the spread of science-related misinformation through the help of computational methods that facilitate research across various disciplines. Foundations of such research and methods are sound definitions of science relatedness and reliable and publicly available datasets that enable the advancement and evaluation of methods dealing with downstream tasks such as (scientific) claims detection, retrieval, classification or verification. In this paper, we propose a hierarchical definition of science-relatedness underlying an annotation framework for science-related tweets that forms the basis of *SciTweets*, an unprecedented annotated ground-truth dataset for science discourse on Twitter. Based on this data, we train a baseline classifier for detection of science-relatedness, showing promising initial results. Whereas *SciTweets* is a comparably small corpus, it provides a high-quality ground truth for testing models, where

[3]Thorne et al. [26] achieved a fleiss score of 0.68 on a fact verification task with five annotators and Alam et al. [1] achieved a score of 0.75 on a task of determining whether a tweet contains a verifiable factual claim.

the heuristics used as part of our sampling methodology open directions for future work by obtaining large-scale weakly labeled training data for training models.

ACKNOWLEDGMENTS

This work is supported by the AI4Sci grant, co-funded by MESRI (France, grant UM-211745), BMBF (Germany, grant 01IS21086), and the French National Research Agency (ANR).

REFERENCES

[1] Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, et al. 2020. Fighting the COVID-19 infodemic: modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. *arXiv preprint arXiv:2005.00033* (2020).

[2] Hunt Allcott and Matthew Gentzkow. 2017. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives* 31, 2 (May 2017), 211–36. https://doi.org/10.1257/jep.31.2.211

[3] Rahul Banerjee, Amar H Kelkar, Aaron C Logan, Navneet S Majhail, and Naveen Pemmaraju. 2021. The democratization of scientific conferences: Twitter in the era of COVID-19 and beyond. *Current hematologic malignancy reports* 16, 2 (2021), 132–139.

[4] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3615–3620. https://doi.org/10.18653/v1/D19-1371

[5] Jedidiah Carlson and Kelley Harris. 2020. Quantifying and contextualizing the impact of bioRxiv preprints through automated social media audience segmentation. *PLoS Biology* 18, 9 (2020), e3000860.

[6] Adrián A Díaz-Faes, Timothy D Bowman, and Rodrigo Costas. 2019. Towards a second generation of 'social media metrics': Characterizing Twitter communities of attention around science. *PloS one* 14, 5 (2019), e0216408.

[7] Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2021. CLIMATE-FEVER: A Dataset for Verification of Real-World Climate Claims. *arXiv:2012.00614 [cs]* (Jan. 2021). http://arxiv.org/abs/2012.00614 arXiv: 2012.00614.

[8] Dimitar Dimitrov, Erdal Baran, Pavlos Fafalios, Ran Yu, Xiaofei Zhu, Matthäus Zloch, and Stefan Dietze. 2020. TweetsCOV19 - A Knowledge Base of Semantically Annotated Tweets about the COVID-19 Pandemic. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Virtual Event, Ireland) *(CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 2991–2998. https://doi.org/10.1145/3340531.3412765

[9] Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76 (1971), 378–382.

[10] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Quantifying Controversy on Social Media. *Trans. Soc. Comput.* 1, 1, Article 3 (jan 2018), 27 pages. https://doi.org/10.1145/3140565

[11] Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, Vikas Sable, Chengkai Li, and Mark Tremayne. 2017. ClaimBuster: the first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment* 10, 12 (Aug. 2017), 1945–1948. https://doi.org/10.14778/3137765.3137815

[12] Robin Haunschild, Lutz Bornmann, Devendra Potnis, and Iman Tahamtan. 2021. Investigating dissemination of scientific information on Twitter: A study of topic networks in opioid publications. *Quantitative Science Studies* (2021), 1–56.

[13] Shanto Iyengar and Douglas S. Massey. 2019. Scientific communication in a post-truth society. *Proceedings of the National Academy of Sciences* 116, 16 (2019), 7656–7661. https://doi.org/10.1073/pnas.1805868115 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1805868115

[14] Tom Jansen and Tobias Kuhn. 2016. Extracting core claims from scientific articles. In *Benelux Conference on Artificial Intelligence*. Springer, 32–46.

[15] S. E. Kreps and D. L. Kriner. 2020. Model uncertainty, political contestation, and public trust in science: Evidence from the COVID-19 pandemic. *Science Advances* 6, 43 (2020), eabd4563. https://doi.org/10.1126/sciadv.abd4563 arXiv:https://www.science.org/doi/pdf/10.1126/sciadv.abd4563

[16] Nan Li, Heather Akin, Leona Yi-Fan Su, Dominique Brossard, Michael A Xenos, and Dietram Scheufele. 2016. Tweeting disaster: An analysis of online discourse about nuclear power in the wake of the Fukushima Daiichi nuclear accident. *Journal of Science Communication* 15, 5 (2016), A02.

[17] Sara Moukarzel, Martin Rehm, Miguel Del Fresno, and Alan J Daly. 2020. Diffusing science through social networks: The case of breastfeeding communication on Twitter. *PloS one* 15, 8 (2020), e0237471.

[18] Preslav Nakov, Alberto Barrón-Cedeño, Giovanni Da San Martino, Firoj Alam, Julia Maria Struß, Thomas Mandl, Rubén Míguez, Tommaso Caselli, Mucahid Kutlu, Wajdi Zaghouani, Chengkai Li, Shaden Shaar, Gautam Kishore Shahi, Hamdy Mubarak, Alex Nikolov, Nikolay Babulkov, Yavuz Selim Kartal, and Javier Beltrán. 2022. 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* (Stavanger, Norway). Springer-Verlag, Berlin, Heidelberg, 416–428. https://doi.org/10.1007/978-3-030-99739-7_52

[19] Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated Fact-Checking for Assisting Human Fact-Checkers. *arXiv:2103.07769 [cs]* (May 2021). http://arxiv.org/abs/2103.07769 arXiv: 2103.07769.

[20] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 9–14. https://doi.org/10.18653/v1/2020.emnlp-demos.2

[21] José María González Pinto, Janus Wawrzinek, and Wolf-Tilo Balke. 2019. What Drives Research Efforts? Find Scientific Claims that Count! *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)* (2019), 217–226.

[22] Angelika Romanou, Panayiotis Smeros, Carlos Castillo, and Karl Aberer. 2020. SciLens News Platform: A System for Real-Time Evaluation of News Articles. *Proc. VLDB Endow.* 13, 12 (2020), 2969–2972. http://www.vldb.org/pvldb/vol13/p2969-romanou.pdf

[23] Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-Fact: Fact Extraction and Verification of Real-World Claims on COVID-19 Pandemic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 2116–2129. https://doi.org/10.18653/v1/2021.acl-long.165

[24] Panayiotis Smeros, Carlos Castillo, and Karl Aberer. 2021. SciClops: Detecting and Contextualizing Scientific Claims for Assisting Manual Fact-Checking. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (Oct. 2021), 1692–1702. https://doi.org/10.1145/3459637.3482475 arXiv: 2110.13090.

[25] Ivan Srba, Branislav Pecher, Matus Tomlein, Robert Moro, Elena Stefancova, Jakub Simko, and Maria Bielikova. 2022. Monant Medical Misinformation Dataset: Mapping Articles to Fact-Checked Claims. *arXiv:2204.12294 [cs]* (April 2022). https://doi.org/10.1145/3477495.3531726 arXiv: 2204.12294.

[26] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355* (2018).

[27] Francois van Schalkwyk and Jonathan Dudek. 2022. Reporting preprints in the media during the COVID-19 pandemic. *Public Understanding of Science* 31, 5 (2022), 608–616. https://doi.org/10.1177/09636625221077392 arXiv:https://doi.org/10.1177/09636625221077392 PMID: 35196912.

[28] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151. https://doi.org/10.1126/science.aap9559 arXiv:https://www.science.org/doi/pdf/10.1126/science.aap9559

[29] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or Fiction: Verifying Scientific Claims. *arXiv:2004.14974 [cs]* (Oct. 2020). http://arxiv.org/abs/2004.14974 arXiv: 2004.14974.

[30] Stefanie Walter, Ines Lörcher, and Michael Brüggemann. 2019. Scientific networks on Twitter: Analyzing scientists' interactions in the climate change debate. *Public Understanding of Science* 28, 6 (2019), 696–712.

[31] Dustin Wright, David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Isabelle Augenstein, and Lucy Wang. 2022. Generating Scientific Claims for Zero-Shot Scientific Fact Checking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 2448–2460. https://aclanthology.org/2022.acl-long.175