Overview of the CLEF-2025 CheckThat! Lab Task 4 on Scientific Web Discourse

Notebook for the CheckThat! Lab at CLEF 2025

Salim Hafid[1,4,*], Yavuz Selim Kartal[2], Sebastian Schellhammer[2,*], Katarina Boland[3], Dimitar Dimitrov[2], Sandra Bringay[1], Konstantin Todorov[1] and Stefan Dietze[2,3]

[1]LIRMM, CNRS, University of Montpellier, Montpellier, France
[2]GESIS - Leibniz Institute for the Social Sciences, Cologne, Germany
[3]Heinrich-Heine-University, Düsseldorf, Germany
[4]médialab Sciences Po, Paris, France

Abstract

We present an overview of Task 4 of the CheckThat! lab at the 2025 edition of the Conference and Labs of the Evaluation Forum (CLEF). Task 4 focuses on scientific web discourse and consists of two subtasks: detecting and differentiating between different forms of scientific web discourse (task **4a**), and retrieving the scientific publication given a social media post with an implicit reference (task **4b**). Within the context of automated fact-checking, these tasks contribute to the detection of scientific claims as well as the retrieval of scientific evidence for their verification. In total, 10 teams participated in task **4a** and 30 in **4b**, with 6 and 7 teams, respectively, submitting system description papers. The participants in task **4a** primarily used transformer-based approaches, with some teams also experimenting with LLMs for data augmentation or classification. The best-performing team achieved a macro-average F1-score close to 0.8. In task **4b**, most teams employed two-stage retrieval and re-ranking pipelines, including the use of various LLMs, with the best team reaching an MRR@5 score of 0.68. This paper presents a detailed overview of the two tasks, including the datasets and evaluation settings, along with a description of the participants' approaches.

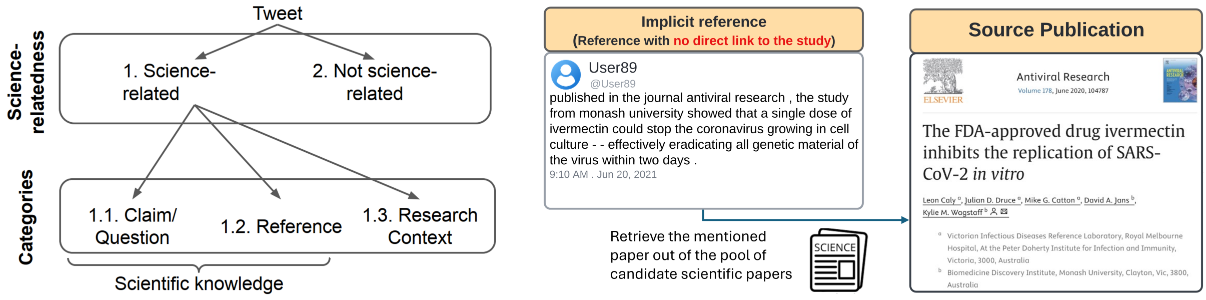
Keywords

scientific web discourse, scientific claims, evidence retrieval

1. Introduction

Scientific web discourse, i.e., discourse on the social web about scientific knowledge, resources, or other research-related information, has increased substantially throughout the past years [1, 2]. Understanding the topics, claims, and studies that are being discussed, the citation habits of users, and the evolution of these habits and the discourse more generally is critical for various tasks and disciplines. For instance, identifying text that conveys scientific knowledge is essential for claim detection [3] and claim verification [4, 5] in this domain. Given that phenomena such as fake news propagation [6] and bias reinforcement [7] may have harmful effects for society [8], especially when coupled with potentially sensitive and controversial topics such as COVID-19 or climate change, tackling such fact-checking-related tasks for scientific web discourse is crucial. However, current state-of-the-art language models have been shown to perform worse for downstream tasks involving scientific claims compared to other domains [9]. Detecting references to scientific studies on social media is crucial for computing altmetrics and assessing the credibility of information. Furthermore, it facilitates research into the evolution of scientific discourse in online environments as studied by various disciplines such

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
salim.hafid@sciencespo.fr (S. Hafid); yavuzselim.kartal@gesis.org (Y. S. Kartal); sebastian.schellhammer@gesis.org (S. Schellhammer); katarina.boland@hhu.de (K. Boland); dimitar.dimitrov@gesis.org (D. Dimitrov); bringay@lirmm.fr (S. Bringay); todorov@lirmm.fr (K. Todorov); stefan.dietze@hhu.de (S. Dietze)
0000-0002-1775-8542 (S. Hafid); 0000-0002-2146-2680 (Y. S. Kartal); 0009-0001-6413-5823 (S. Schellhammer); 0000-0003-2958-9712 (K. Boland); 0000-0002-4504-5144 (D. Dimitrov); 0000-0002-2830-3666 (S. Bringay); 0000-0002-9116-6692 (K. Todorov); 0009-0001-4364-9243 (S. Dietze)



(a) (b)

Figure 1: Fig. (a) shows the different categories of scientific web discourse that are relevant to task **4a** (from [15]). Fig. (b) shows a claim with an implicit reference (box to the left) for which the source publication (box to the right) needs to be retrieved in task **4b** (adapted from [16]).

as communication studies [2, 1], social sciences, and social psychology [10, 11], as well as computational linguistics [12, 13].

Scientific web discourse is often informal, e.g., *"covid vaccines just don't work on children"*, and displays fuzzy/incomplete citation habits, such as *"Stanford study shows that vaccines don't work"*, where the actual study is never cited through explicit references. These characteristics pose challenges both from a societal perspective, leading to poorly informed online debates [8], and from a computational perspective, requiring robust datasets and methods to detect and analyse such discourse.

Therefore, the CLEF-2025 CheckThat! lab contains two tasks that are relevant to the CheckThat! verification pipeline [14], tailored specifically to the analysis of scientific web discourse:

- **Task 4a - Scientific Web Discourse Detection**: Given a social media post (tweet), detect if it contains (1) a scientific claim, (2) a reference to a scientific study/publication, or (3) mentions of scientific entities, e.g. a university or scientist.
- **Task 4b - Claim-source Retrieval**: Given an implicit reference to a scientific paper, i.e., a social media post (tweet) that mentions a research publication without a URL (e.g., "Stanford study shows that vaccines don't work"), retrieve the mentioned paper from a pool of candidate papers.

In the remainder of the paper, we introduce the datasets used in our two tasks (Section 2), describe system submissions and results (Section 3), present related work (Section 4), and conclude with final remarks (Section 5).

2. Dataset

2.1. Task 4a: Scientific Web Discourse Detection

The dataset for task **4a** is an extension of the SciTweets corpus [15] and consists of 1,606 posts from X (formerly Twitter) annotated with the different forms of science-related online discourse, which are scientific claims (Category 1), scientific references (Category 2), and references to science contexts or entities (Category 3), following the definitions by Hafid et al. [15]:

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

Category 1 - Scientific knowledge (scientifically verifiable claims): Does the text include a claim or a question that could be scientifically verified? i.e., claims that can only be verified with the help of scientific publications or research data used in the scientific process (see posts 2 and 5 in Table 1).

Category 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 posts 3 and 5 in Table 1).

Category 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 posts 3, 4, and 5 in Table 1).

Not science-related: Texts that do not fall under either of the three previous categories. (see post 1 in Table 1)

Table 2 shows the number of posts per category and per split.

Table 1
Task 4a: Dataset examples

X post	Cat 1	Cat 2	Cat 3
McDonald's breakfast stop then the gym			
65% of cats born with blue eyes are deaf.	■		
@user Please read this research analysis https://www.apa.org/pubs/journals/releases/psp-pspp0000147.pdf.		■	■
How is University of Chicago shaping the future of science? Find out on April 6			■
A fifth of US high school students use tobacco, finds survey http://www.bmj.com/content/349/bmj.g6885	■	■	■

Table 2
Task 4a: Number of X posts per category and per data split

Split	Category 1	Category 2	Category 3	Total
Train	333	224	306	1,229
Dev	26	26	34	137
Test	121	56	97	240
Total	480	306	437	1,606

2.2. Task 4b: Scientific Claim Source Retrieval

The dataset for task **4b** consists of two sets, a query set and a collection set. The query set contains 15,699 X (Twitter) posts with implicit references to scientific papers from the CORD-19 corpus [17]. The collection set contains metadata, such as title, abstract, and affiliations of the 7,718 CORD-19 scientific papers, which the posts from the query set implicitly refer to. Table 3 shows the number of posts (query set) and number of publications (collection set) per split. Note that the collection set is identical for all three splits. Tables 4 and 5 show two exemplary posts with implicit references and the corresponding CORD-19 publications. For instance, given the first post in Table 4, the participants are asked to retrieve the correct publication, which is the first publication in Table 5. See [16] for more details on how the data was sampled and annotated.

3. Results and Overview of the Systems

3.1. Task 4a: Scientific Web Discourse Detection

Task **4a** is a multilabel classification task and was evaluated through the macro-averaged F1-score. The **baseline** is a DeBERTaV3-base model trained on the train set for 10 epochs with a learning rate of $2e^{-5}$ and a batch size of 16. For the final test set predictions, we used the checkpoint with the best dev set performance, resulting in a test set macro F1-score of 0.7668 (rank 7).

Table 3
Task 4b: Number of X posts and CORD-19 publications per data split

Split	**X Posts** (Query Set)	**Publications** (Collection Set)
Train	12,853	7,718
Dev	1,400	
Test	1,446	
Total	15,699	7,718

In total, ten teams participated in task **4a**. Table 6 provides an overview of the different approaches and their performances for those teams that submitted a paper description of their work. The F1-score and rank indicate the performance and position on the final test set leaderboard. Most teams relied on transformer-based models such as DeBERTa-v3, SciBERT, and Twitter-Roberta, while some also used LLMs. In addition, different techniques such as LLM-based data augmentation, ensemble methods, and other optimizations were employed.

Team **ClimateSense** [18] fine-tuned a twitter-roberta-base-2022-154m model and identified the best-performing checkpoint, for each category, based on the dev set performance. Using the embeddings of these checkpoints, traditional classifiers were trained for each category (Nearest centroid classifiers for categories 1 and 3, and a Naive Bayes classifier for category 2), ranking first on the overall leaderboard with an F1-score of 0.7998. They also experimented with SetFit [19], and training classifiers on top of existing sentence encoders, but found the approaches to have lower performance compared to their final system.

Team **VerbaNexAI** [20] fine-tuned a DeBERTa-v3-base model using hyperparameters, i.e., learning rate and number of epochs, found with 5-fold cross-validation. To improve performance, they adjusted the binary cross-entropy loss with class weights and employed threshold-tuning. For their final submission, ranking second, they used a soft-voting ensemble of their two strongest models (one with class weights and one without).

Team **SBU-SCIRE** [21] augmented the training data to 2,369 samples with paraphrases using DeepSeek-R1. They trained five DeBERTa-v3-large models on the augmented dataset using 5-fold cross-validation with a focal loss [22] to address class imbalance and focus on hard examples. For inference on the test set, they averaged the logits of the five models before applying optimized class-specific thresholds to the sigmoid probabilities.

Team **DS@GT** [23] trained different transformer-based models, such as DeBERTa-v3-base and -large, and used zero-shot and few-shot classification with GPT-4o and GPT-4o mini. While the transformer-based models performed better for categories one and three, the LLM-based approaches outperformed them on category two. Thus, for their final submission, they combined DeBERTa-v3-base (categories one and three) with GPT-4o mini using few-shot with five examples based on semantic similarity (category two).

Team **TurQUaz** [24] employed various LLMs, such as Gemma3 (12B), Qwen3 (8B), DeepSeek-R1 (8B), in different collaborative settings. Specifically, they investigated three settings: (1) a single debate, where two LLMs argue in favor of or against a specific classification (e.g., whether a post contains a scientific claim) and a third model acts as a judge, (2) a team debate, following the same approach as the single debate but with multiple LLMs collaborating on each side, and (3) a council debate, where multiple LLMs argue together to reach a consensus. For their final submission, they chose the council debate, outperforming the two other settings. While the approach, overall, did not improve upon the baseline, it ranked first in identifying scientific references (category two).

Team **JU_NLP** [25] generated embeddings using SciBERT and Twitter-RoBERTa models to capture both scientific and social media discourse characteristics of posts. The embeddings were concatenated and used to train a two-layer classification head.

Overall, fine-tuning existing PLMs such as DeBERTa-v3 or twitter-roberta-base-2022-154m performs

Table 4
Task 4b: Dataset examples - X Posts

X post	CORD Id
Peer-reviewed in the New England Journal of Medicine regarding Delta (B.1.617.2): •Pfizer is 90% effective •AstraZeneca is 70% effective. This falls in line with vaccine efficacy of other variants. Yes, the vaccines ARE indeed effective against Delta.	5g02ykhi
Published in the journal Antiviral Research, the study from Monash University showed that a single dose of Ivermectin could stop the coronavirus growing in cell culture – effectively eradicating all genetic material of the virus within two days.	ivy95jpw

Table 5
Task 4b: Dataset examples - publications

CORD Id	Title	Date	Venue + Authors	Selected Abstract Text
5g02ykhi	Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant	21-07-2021	New England Journal of Medicine Jamie Lopez Bernal, Nick Andrews, Charlotte Gower, Eileen Gallagher, Ruth Simmons, Simon Thelwall, Julia Stowe, Elise Tessier, [...]	[...] With the BNT162b2 vaccine, the effectiveness of two doses was [...] 88.0% among those with the delta variant. With the ChAdOx1 nCoV-19 vaccine, the effectiveness of two doses was [...] 67.0% among those with the delta variant. [...]
ivy95jpw	The FDA-approved drug ivermectin inhibits the replication of SARS-CoV-2 in vitro	03-04-2020	Antiviral Research Caly, Leon; Druce, Julian D.; Catton, Mike G.; Jans, David A.; Wagstaff, Kylie M.	[...] We report here that Ivermectin [...] is an inhibitor of the causative virus (SARS-CoV-2), with a single addition to Vero-hSLAM cells 2 h post infection with SARS-CoV-2 able to effect 5000-fold reduction in viral RNA at 48 h. [...]

best in terms of macro-avg F1-score, with some LLM approaches outperforming them in the identification of scientific references (category two).

3.2. Task 4b: Scientific Claim Source Retrieval

Task **4b** is a retrieval task and was evaluated by the MRR@5 (Mean Reciprocal Rank) score. BM25 ranking using the title and abstract of the papers and the text of the X posts serves as the **baseline** with an MMR@5 of 0.43. The best-performing team reached an MMR@5 of 0.68.
In total, 30 teams participated in task **4b**. Table 7 provides an overview of the different approaches and their performance for teams that submitted a paper description of their work. Most teams relied on a combination of retrieval methods (dense, sparse, or both) and re-ranking models. Retrieval methods included both lexical and semantic methods. LLMs such as ChatGPT, LLaMa, and Gemma were mainly used as re-rankers, but did not always outperform fine-tuned transformer-based models. Additionally, some teams experimented with data augmentation and style transfer techniques.

Team **AIRwaves** [26] employed a two-stage pipeline using neural representation learning for can-

Table 6
Task 4a: Overview of the approaches

Team	Models					Misc.			Perf.	
	DeBERTa-v3	SciBERT	Twitter-RoBERTa	LLMs	Others	Data Augmentation	Ensemble	Other Optimizations	Macro-avg. F1-Score	Rank
ClimateSense [18]			■		■		■		0.7998	1
VerbaNexAI [20]	■						■	■	0.7983	2
SBU-SCIRE [21]	■					■		■	0.7917	4
DS@GT [23]	■			■	■		■		0.7685	6
DeBERTa-v3 Baseline	■								0.7668	7
TurQUaz [24]				■			■		0.7615	8
JU_NLP [25]		■	■				■		0.7347	9

didate generation with a fine-tuned E5-large model, followed by neural re-ranking with a SciBERT cross-encoder to re-order the top predictions. Incorporating one additional BM-25-mined hard negative example per query improved the performance significantly.

Team **Deep Retrieval** [27] combined lexical BM25-based retrieval with a semantic search-based approach using an INF-Retriever-v1 retrieval model to generate candidates, which were then re-ranked with a BAAI/bge-reranker-v2-gemma cross-encoder. While the semantic retrieval outperformed the lexical BM-25 approach, combining and re-ranking the generated candidates from both approaches yielded the best performance.

Team **ATOM** [28] explored different retriever and re-ranking models. For their final submission, they used a GTR-T5-Large model to retrieve candidates, followed by the MXBAI-base-v2 re-ranker. The team further experimented with enriching the collection set with full-texts, but did not find it to improve performance compared to using the provided abstracts.

Team **SBU-SCIRE** [21] used a Snowflake/snowflake-arctic-embed-l-v2.0 model for dense retrieval, followed by an ms-marco-MiniLM-L4-v2 cross-encoder for re-ranking. To improve the training effectiveness of the dense retriever, they used a strategically sampled set of nine hard negative examples per query.

Team **SeRRa** [29] used a multi-step pipeline including dense retrieval for candidate generation with a Sentence-BERT model, re-ranking using a SciBERT-based binary classification model, and a final ranking through pairwise comparisons of the top 10 re-ranked documents with the input claim using a ModernBERT model. In addition, they evaluated the effect of hard negative sampling for the final re-ranking step and found a significant improvement in performance.

Team **Claim2Source** [30] systematically evaluated the impact of seven different style transfer techniques applied to both claims and source documents using a LLaMa 3.3-70B-Instruct model. The style transfer techniques involved three strategies for rewriting claims, such as converting informal posts into more formal language, and four strategies for transforming publications, e.g., writing a concise social media post based on a scientific abstract. They observed that making claims more formal tended to help retrieval, but rewriting the titles and abstracts of the publications usually degraded the performance. Still, for their final approach, they relied on dense retrieval with a GritLM-7B model without applying any style transfer.

Team **DS@GT** [31] explored eight different two-stage retrieval and re-ranking pipelines and investigated the effect of rewriting posts in formal language using ChatGPT. They found that appending a formal paraphrase to the original post slightly improved the re-ranking performance. Their final approach combined BM25-retrieval with a T5 model for reranking.

Table 7
Task 4b: Overview of the approaches

Team	Models				Misc.		Perf.	
	Dense Retrieval	Sparse Retrieval	Re-ranking	LLMs	Data Augmentation	Style transfer	MRR@5	Rank
AIRwaves [26]	■		■				0.67	2
Deep Retrieval [27]	■	■	■	■			0.66	3
ATOM [28]	■		■				0.66	4
SBU-SCIRE [21]	■		■				0.65	5
SeRRa [29]	■		■				0.61	8
Claim2Source [30]	■	■		■		■	0.59	12
DS@GT [31]		■	■	■	■	■	0.58	16
BM25 Baseline		■					0.43	28

Overall, most teams followed a two-stage approach, combining a dense retrieval step with neural re-ranking. Strategically sampling hard negative samples led to performance gains, while results from applying style transfer techniques were mixed.

4. Related Work

4.1. Task 4a: Scientific Web Discourse Detection

Scientific web discourse has been studied by various disciplines, including social sciences, measuring the engagement with scientific publications on social media [32, 33, 34], as well as NLP research, detecting and verifying scientific claims [35, 36, 37, 38, 5]. To facilitate such research, high-quality datasets with robust definitions are crucial. While existing resources often target specific domains [37, 35], or generate synthetic claims [39], we follow the domain-agnostic definitions of [15] and extend their SciTweets corpus to provide 1,606 X posts in total.

From a computational perspective, task **4a** includes the detection of claims, which has been studied in prior work [35, 40, 41, 37]. For example, in a previous iteration of the CheckThat! lab, the goal was to identify relevant claims in tweets [41]. More closely related to scientific claims, Wuhrl and Klinger [35] detect claims in tweets from the biomedical domain. In contrast, our task involves the detection of scientific claims independent of the scientific field. Furthermore, to the best of our knowledge, the two other subtasks of task **4a**, the detection of scientific references and scientific contexts, have not been addressed in prior work.

4.2. Task 4b: Scientific Claim Source Retrieval

The task of scientific claim source retrieval is closely related to the problem of evidence retrieval in automated fact-checking [5], as explored by previous research [42, 39, 43, 44]. For example, the FEVER shared task asked participants to retrieve evidence from Wikipedia for human-authored claims [42]. In the previous edition of the CheckThat! lab 2024, a similar task was introduced: given a rumour, evidence tweets from authority Twitter accounts should be retrieved [43]. However, existing works are different because they use synthetic claims [42], claims that originate from different sources, such as scientific publications [39], or evidence from sources other than scientific publications. Furthermore, previous evidence retrieval tasks focus on retrieving any relevant evidence useful to fact-check a claim, whereas we are interested in finding the one piece of evidence that is most likely the basis for the stated claim. The original source that a claim is based on has been shown to be one of the primary pieces of evidence used by fact-checkers [45]. Another related line of work is the retrieval of publications referred to by news articles [46, 47, 48]. While similar to the implicit references in our dataset, we assume the references in news articles are more formal and have a larger context.

5. Conclusion and Future Work

We presented an overview of Task 4 of the CheckThat! lab at CLEF 2025, which comprised two subtasks: identifying and distinguishing between different forms of scientific web discourse (task **4a**), and retrieving the scientific publication given a social media post with an implicit reference (task **4b**).

For task **4a**, most teams relied on fine-tuning transformer-based models, with some exploring LLMs for data augmentation as well as for zero- and few-shot classification. In task **4b**, two-stage retrieval pipelines were used by many teams, including the use of various LLMs for candidate generation and reranking. The highest-ranked team for task **4a**, team **ClimateSense** [18] fine-tuned a twitter-roberta-base-2022-154m model and, for each category, used the best-performing checkpoint to extract embeddings for training a traditional classifier using a weighted loss function. Team **AIRwaves** [26], the top-ranked team with a system description paper and second overall in task **4b**, implemented a two-stage pipeline including candidate generation with a fine-tuned E5-large model and reranking the top predictions with a SciBERT-based cross-encoder. While these teams achieved an F1-score of 0.80 and an MRR@5 score of 0.67, there is still room for improvement across both tasks.

In total, 40 teams submitted their predictions, and 13 submitted system description papers, attracting considerable interest. In future iterations of these tasks, we plan to expand the languages covered (e.g., including French and German), include additional online discourse platforms such as Telegram or Bluesky, and incorporate more realistic scenarios (e.g., moving beyond the COVID-19 domain in task **4b**).

6. Acknowledgments

This work has been funded by the AI4Sci grant (co-funded by MESRI (France, grant UM-211745), BMBF (Germany, grant 01IS21086), and the French National Research Agency (ANR)), as well as the Leibniz Association as part of the Leibniz Collaborative Excellence funding programme (grant no K490/2022).

7. Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, Grammarly to perform the following tasks: "Grammar and spelling check", "Paraphrase and reword". After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] S. Dunwoody, Science journalism: Prospects in the digital age, in: Routledge handbook of public communication of science and technology, Routledge, 2021, pp. 14–32.
[2] M. Brüggemann, I. Lörcher, S. Walter, Post-normal science communication: exploring the blurring boundaries of science and journalism, Journal of Science Communication 19 (2020) A02.
[3] A. Barrón-Cedeño, F. Alam, T. Chakraborty, T. Elsayed, P. Nakov, P. Przybyła, J. M. Struß, F. Haouari, M. Hasanain, F. Ruggeri, et al., The clef-2024 checkthat! lab: Check-worthiness, subjectivity, persuasion, roles, authorities, and adversarial robustness, in: European Conference on Information Retrieval, Springer, 2024, pp. 449–458.

[4] G. Wang, K. Harwood, L. Chillrud, A. Ananthram, M. Subbiah, K. Mckeown, Check-covid: Fact-checking covid-19 news claims with scientific evidence, in: Findings of the Association for Computational Linguistics: ACL 2023, 2023, pp. 14114–14127.
[5] P. Nakov, D. Corney, M. Hasanain, F. Alam, T. Elsayed, A. Barron-Cedeno, P. Papotti, S. Shaar, G. Da San Martino, et al., Automated fact-checking for assisting human fact-checkers, in: IJCAI, International Joint Conferences on Artificial Intelligence, 2021, pp. 4551–4558.
[6] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, science 359 (2018) 1146–1151.
[7] K. Garimella, G. D. F. Morales, A. Gionis, M. Mathioudakis, Quantifying controversy on social media, ACM Transactions on Social Computing 1 (2018) 1–27.
[8] Y. M. Rocha, G. A. de Moura, G. A. Desidério, C. H. de Oliveira, F. D. Lourenço, L. D. de Figueiredo Nicolete, The impact of fake news on social media and its influence on health during the covid-19 pandemic: A systematic review, Journal of Public Health (2021) 1–10.
[9] S. Hafid, S. Schellhammer, Y. S. Kartal, T. Papastergiou, S. Dietze, S. Bringay, K. Todorov, An in-depth analysis of the linguistic characteristics of science claims on the web and their impact on fact-checking, ACM Trans. Web (2025). URL: https://doi.org/10.1145/3746170. doi:10.1145/3746170, just Accepted.
[10] R. Banerjee, A. H. Kelkar, A. C. Logan, N. S. Majhail, N. Pemmaraju, The democratization of scientific conferences: Twitter in the era of covid-19 and beyond, Current hematologic malignancy reports 16 (2021) 132–139.
[11] S. Kreps, D. Kriner, Model uncertainty, political contestation, and public trust in science: evidence from the covid-19 pandemic. sci. adv. 6, eabd4563, 2020.
[12] T. August, D. Card, G. Hsieh, N. A. Smith, K. Reinecke, 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, 2020, pp. 1–12.
[13] E. Chandrasekharan, M. Samory, S. Jhaver, H. Charvat, A. Bruckman, C. Lampe, J. Eisenstein, E. Gilbert, 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 (2018) 1–25.
[14] F. Alam, J. M. Struß, T. Chakraborty, S. Dietze, S. Hafid, K. Korre, A. Muti, P. Nakov, F. Ruggeri, S. Schellhammer, V. Setty, M. Sundriyal, K. Todorov, V. Venktesh, Overview of the CLEF-2025 CheckThat! Lab: Subjectivity, fact-checking, claim normalization, and retrieval, in: J. Carrillo-de Albornoz, J. Gonzalo, L. Plaza, A. García Seco de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.
[15] S. Hafid, S. Schellhammer, S. Bringay, K. Todorov, S. Dietze, Scitweets-a dataset and annotation framework for detecting scientific online discourse, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, 2022, pp. 3988–3992.
[16] S. Hafid, Y. S. Kartal, S. Schellhammer, V. Jacot, S. Bringay, S. Dietze, K. Todorov, Disambiguation of implicit scientific references on x, Proceedings of the 36th ACM Conference on Hypertext and Social Media (2025).
[17] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Burdick, D. Eide, K. Funk, Y. Katsis, R. M. Kinney, Y. Li, Z. Liu, W. Merrill, P. Mooney, D. A. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A. D. Wade, K. Wang, N. X. R. Wang, C. Wilhelm, B. Xie, D. M. Raymond, D. S. Weld, O. Etzioni, S. Kohlmeier, CORD-19: The COVID-19 open research dataset, in: Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020, Association for Computational Linguistics, Online, 2020. URL: https://www.aclweb.org/anthology/2020.nlpcovid19-acl.1.
[18] G. Burel, P. Lisena, E. Daga, R. Troncy, H. Alani, ClimateSense at CheckThat! 2025: Combining Fine-tuned Large Language Models and Conventional Machine Learning Models for Subjectivity and Scientific Web Discourse Analysis, in: [49], 2025.
[19] L. Tunstall, N. Reimers, U. E. S. Jo, L. Bates, D. Korat, M. Wasserblat, O. Pereg, Efficient few-shot learning without prompts, ArXiv abs/2209.11055 (2022). URL: https://api.semanticscholar.org/CorpusID:252439001.
[20] M. Sosa, J. Serrano, J. C. Martinez Santos, E. Puertas, VerbaNexAI Lab at CheckThat! 2025:

Fine-Tuning DeBERTa for Multi-Label Scientific Discourse Detection in Tweets, in: [49], 2025.
[21] P. Thapliyal, R. Chavan, S. Samridh, C. Zuo, R. Banerjee, SBU-SCIRE at CheckThat! 2025: Bridging Social Media, Scientific Discourse, and Scientific Literature, in: [49], 2025.
[22] T.-Y. Lin, P. Goyal, R. B. Girshick, K. He, P. Dollár, Focal loss for dense object detection, 2017 IEEE International Conference on Computer Vision (ICCV) (2017) 2999–3007. URL: https://api.semanticscholar.org/CorpusID:47252984.
[23] A. Parikh, H. Truong, J. Schofield, M. Heil, DS@GT at CheckThat! 2025: Ensemble Methods for Detection of Scientific Discourse on Social Media, in: [49], 2025.
[24] T. Saraç, S. Mergen, M. Kutlu, TurQUaz at CheckThat! 2025: Debating Large Language Models for Scientific Web Discourse Detection, in: [49], 2025.
[25] A. Majumdar, D. Das, P. Pal, JU_NLP at CheckThat! 2025: Leveraging Hybrid Embeddings for Multi-Label Classification in Scientific Social Media Discourse, in: [49], 2025.
[26] C. Ashbaugh, L. Baumgärtner, T. Greß, N. Sidorov, D. Werner, AIRwaves at CheckThat! 2025: Retrieving Scientific Sources for Implicit Claims on Social Media with Dual Encoders and Neural Re-Ranking, in: [49], 2025.
[27] P. J. Sager, A. Kamaraj, B. F. Grewe, T. Stadelmann, Deep Retrieval at CheckThat! 2025: Identifying Scientific Papers from Implicit Social Media Mentions via Hybrid Retrieval and Re-Ranking, in: [49], 2025.
[28] M. Staudinger, A. El-Ebshihy, W. Kusa, F. Piroi, A. Hanbury, ATOM at CheckThat! 2025: Retrieve the Implicit - Scientific Evidence Retrieval, in: [49], 2025.
[29] G. Marchetti, G. Rocha, H. L. Cardoso, Team SeRRa at CheckThat! 2025: Sequential Re-Ranking in a Scientific Claim Source Retrieval Pipeline, in: [49], 2025.
[30] T. Schreieder, M. Färber, Claim2Source at CheckThat! 2025: Zero-Shot Style Transfer for Scientific Claim-Source Retrieval, in: [49], 2025.
[31] J. Schofield, S. Tian, H. Truong, M. Heil, DS@GT at CheckThat! 2025: Exploring Retrieval and Reranking Pipelines for Scientific Claim Source Retrieval on Social Media Discourse, in: [49], 2025.
[32] J. Carlson, K. Harris, Quantifying and contextualizing the impact of biorxiv preprints through automated social media audience segmentation, PLoS Biology 18 (2020) e3000860.
[33] R. Haunschild, L. Bornmann, D. Potnis, I. Tahamtan, Investigating dissemination of scientific information on twitter: A study of topic networks in opioid publications, Quantitative Science Studies (2021) 1–56.
[34] A. A. Díaz-Faes, T. D. Bowman, R. Costas, Towards a second generation of 'social media metrics': Characterizing twitter communities of attention around science, PloS one 14 (2019) e0216408.
[35] A. Wuhrl, R. Klinger, Claim detection in biomedical twitter posts, in: Workshop on Biomedical Natural Language Processing, 2021. URL: https://api.semanticscholar.org/CorpusID:233387769.
[36] I. Srba, B. Pecher, M. Tomlein, R. Moro, E. Stefancova, J. Simko, M. Bielikova, 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, 2022, pp. 2949–2959.
[37] P. Smeros, C. Castillo, K. Aberer, SciClops: Detecting and Contextualizing Scientific Claims for Assisting Manual Fact-Checking, Proceedings of the 30th ACM International Conference on Information & Knowledge Management (2021) 1692–1702. URL: http://arxiv.org/abs/2110.13090. doi:10.1145/3459637.3482475, arXiv: 2110.13090.
[38] N. Hassan, G. Zhang, F. Arslan, J. Caraballo, D. Jimenez, S. Gawsane, S. Hasan, M. Joseph, A. Kulkarni, A. K. Nayak, V. Sable, C. Li, M. Tremayne, ClaimBuster: the first-ever end-to-end fact-checking system, Proceedings of the VLDB Endowment 10 (2017) 1945–1948. URL: https://dl.acm.org/doi/10.14778/3137765.3137815. doi:10.14778/3137765.3137815.
[39] D. Wadden, K. Lo, L. L. Wang, S. Lin, M. van Zuylen, A. Cohan, H. Hajishirzi, Fact or fiction: Verifying scientific claims, in: Conference on Empirical Methods in Natural Language Processing, 2020. URL: https://api.semanticscholar.org/CorpusID:216867133.
[40] M. Sundriyal, M. S. Akhtar, T. Chakraborty, Overview of the claimscan-2023: Uncovering truth in social media through claim detection and identification of claim spans, Proceedings of the

15th Annual Meeting of the Forum for Information Retrieval Evaluation (2023). URL: https://api.semanticscholar.org/CorpusID:264820358.

[41] P. Nakov, A. Barrón-Cedeño, G. D. S. Martino, F. Alam, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouani, C. Li, S. Shaar, H. Mubarak, A. Nikolov, Y. S. Kartal, Overview of the clef-2022 checkthat! lab task 1 on identifying relevant claims in tweets, in: Conference and Labs of the Evaluation Forum, 2022. URL: https://api.semanticscholar.org/CorpusID:251472020.

[42] J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, A. Mittal, The fact extraction and verification (fever) shared task, ArXiv abs/1811.10971 (2018). URL: https://api.semanticscholar.org/CorpusID:53645946.

[43] F. Haouari, T. Elsayed, R. Suwaileh, Overview of the clef-2024 checkthat! lab task 5 on rumor verification using evidence from authorities, in: Conference and Labs of the Evaluation Forum, 2024. URL: https://api.semanticscholar.org/CorpusID:271771424.

[44] J. Chen, G. Kim, A. Sriram, G. Durrett, E. Choi, Complex claim verification with evidence retrieved in the wild, ArXiv abs/2305.11859 (2023). URL: https://api.semanticscholar.org/CorpusID:258822852.

[45] M. Glockner, Y. Hou, I. Gurevych, Missing counter-evidence renders nlp fact-checking unrealistic for misinformation, ArXiv abs/2210.13865 (2022). URL: https://api.semanticscholar.org/CorpusID:253107194.

[46] J. Wang, B. Yu, News2pubmed: A browser extension for linking health news to medical literature, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 2605–2609.

[47] K. Kousha, M. Thelwall, An automatic method to identify citations to journals in news stories: A case study of uk newspapers citing web of science journals, Journal of Data and Information Science 4 (2019) 73–95.

[48] J. Ravenscroft, A. Clare, M. Liakata, Harrigt: Linking news articles to scientific literature, in: Proceedings of ACL, 2018, p. 19.

[49] G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 2025.