SCIRE at CheckThat! 2025: Bridging Social Media, Scientific Discourse, and Scientific Literature

Notebook for the CheckThat! Lab at CLEF 2025

Parth Manish Thapliyal[1,*], Ritesh Sunil Chavan[1,†], Samridh Samridh[1,†], Chaoyuan Zuo[2] and Ritwik Banerjee[1]

[1] *Computer Science, Stony Brook University, 100 Nicolls Rd, Stony Brook, New York, United States of America*
[2] *Journalism and Communication, Nankai University, 38 Tongyan Road, Jinnan District, Tianjin, China*

Abstract

The increasing prominence of scientific discourse on social media platforms presents both unprecedented opportunities for public engagement and significant risks of misinformation. While scientific claims, references to publications, and mentions of research entities proliferate rapidly, current platforms lack robust mechanisms to validate their veracity or trace implicit sources. Manual identification and sourcing of such content is impractical at scale, and although computational methods exist for generic fact-checking or citation retrieval, they often fail to address the unique challenges of noisy, abbreviated social media language – particularly the detection of nuanced scientific discourse and the retrieval of publications from implicit, non-URL references. In this paper, we propose a unified framework tackling two critical tasks: (1) detection of **scientific web discourse**, where we identify tweets containing scientific claims, references or research entities, using a combination of natural language augmentation and supervised learning; and (2) **source retrieval for scientific claims**, employing a two-stage dense retrieval and re-ranking pipeline to link implicit mentions of sources to their actual publications from candidate pools. Our multi-stage architecture first filters and classifies scientific content, then prioritizes and resolves latent citations. Evaluations on a curated dataset provided by the CLEF-2025 CheckThat! Lab demonstrate the effectiveness of our approach, achieving significant improvements across both tasks. This work provides essential tools for automating scientific credibility assessment and aiding the verification of scientific information in online ecosystems.

Keywords

Data Augmentation, Dense Retrieval, Re-ranking, Cross-encoder, Bi-encoder, Large language model, Transformer

1. Introduction

It is increasingly evident that scientific discourse now permeates the fabric of social media [1, 2], offering transformative potential for public engagement while simultaneously amplifying the risks of misinformation [3] – including (but not limited to) acts of selective omission [4], fear-mongering [5], and use of misleading contexts [6]. For the average user, social media platforms like Twitter/X have become conduits for encountering scientific claims, references to publications, and mentions of research entities. Discerning the veracity and validity of such content remains daunting, however. Users naturally gravitate toward information confirming their existing beliefs [7, 8], and this confirmation bias, compounded by the breakneck speed of scientific information diffusion, creates fertile ground for unverified claims to proliferate as accepted facts [9]. While correcting cognitive biases is a profound societal challenge, gleaning the scientific validity of claims being shared offers a more pragmatic path forward, where progress can be objectively measured.

Comprehensive manual validation of scientific discourse is infeasible at social media scale. Computational approaches have emerged, yet existing solutions remain inadequate. Prior work in generic

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

[*] Corresponding author.

[†] These authors contributed equally.

pthapliyal@cs.stonybrook.edu (P. M. Thapliyal); riteshsunil.chavan@stonybrook.edu (R. S. Chavan); samridh.samridh@stonybrook.edu (S. Samridh); zuocy@nankai.edu.cn (C. Zuo); rbanerjee@cs.stonybrook.edu (R. Banerjee)
https://rchavan.com (R. S. Chavan); https://www.ritwikbanerjee.com (R. Banerjee)
0000-0001-5361-0193 (C. Zuo); 0000-0003-0336-0258 (R. Banerjee)

fact-checking often overlooks scientific nuance, while citation retrieval systems typically rely on explicit URLs or DOIs [10] – failing to address the prevalence of implicit, unstructured references (e.g., "a recent Nature study shows . . . "). Early studies – e.g., Levy et al. [11] and Cazalens et al. [12] – focused on coarse-grained claim detection but lacked mechanisms to prioritize content or resolve latent sources. Subsequent efforts made significant progress by ranking the "check-worthiness" of claims [13, 14, 15], and employed increasingly sophisticated models for citation matching [16, 17]. Yet, they often struggled with the shifting landscape of social media lexicon and the unique brevity seen on such platforms [18]. Critically, the synergistic tasks of detecting scientific discourse and retrieving its implicit sources remain underexplored, with no unified framework addressing both. The CLEF CheckThat! Lab's earlier datasets were pioneers in identifying significant gaps in handling scientific specificity and source ambiguity [19].

In this work, we bridge this divide by proposing an integrated pipeline for scientific information curation. Unlike prior methods constrained by narrow scope of domain or dataset, our approach leverages natural language augmentation and dense retrieval to confront the dual challenges of noisy context and implicit references. We focus on two pillars: (1) detecting scientific web discourse – identifying tweets containing claims, publication references, or research entities; and (2) retrieving sources for scientific claims – linking implicit mentions to actual publications via candidate pools. Our architecture deliberately prioritizes scalability and robustness, filtering scientific content before resolving its provenance. We use the task formulation, data, and evaluation framework provided by the fourth task of CLEF-2025 CheckThat! Lab, viz., Scientific Web Discourse [20].

2. A Quick Overview of Tasks and Dataset

Task 1 : Scientific Web Discourse Detection: This task focuses on identifying scientific content within social media posts by detecting three key elements: (1) explicit scientific claims, (2) references to research publications or studies, and (3) mentions of scientific entities like researchers or institutions. It addresses the critical challenge of distinguishing authentic scientific discourse from general online conversations, particularly in high-impact domains like COVID-19 and climate change where misinformation risks are elevated. Current research is hindered by inconsistent definitions of science-related content and a scarcity of annotated data to train detection models.

Task 2 : Scientific Claim Source Retrieval: This complementary task tackles the problem of linking implicit scientific references in social media to their source publications. Given tweets mentioning research without direct URLs or identifiers, the goal is to accurately retrieve the referenced papers from candidate pools. This addresses a key verification bottleneck in online scientific discourse, where informal mentions (e.g., "a recent Harvard study shows . . . ") lack traceable citations yet require validation against original research to combat misinformation. The absence of standardized datasets for implicit citation resolution remains a significant research gap.

Datasets: The dataset construction reflects the complexity of real-world scientific social media discourse, with tweet texts deliberately paraphrased for compliance while preserving linguistic authenticity. For the first task, the corpus comprises 1,229 training samples, 137 development samples, and 240 test instances, with each tweet annotated using binary labels across the three target categories. The evaluation employs macro-averaged F_1-scores to account for class imbalance inherent in scientific discourse detection. The second task leverages the CORD-19 publication database as its candidate pool, containing comprehensive metadata including study titles, abstracts, venues, and author information. Training instances consist of tweet-publication pairs where implicit references (e.g., "published in Nature" or "recent Stanford research") must be resolved to specific papers identified by CORD IDs. Performance assessment utilizes Mean Reciprocal Rank at 5 (MRR@5), emphasizing the practical importance of surfacing correct sources within top-ranked results for human verification workflows.

3. Methodology: Scientific Web Discourse Detection

3.1. Data Preparation and Augmentation

To address data scarcity and enhance linguistic diversity, we employed DeepSeek-R1 [21] for paraphrase-based augmentation, effectively doubling the original training set to 2,369 samples. To mitigate class imbalance, we implemented two complementary strategies: (i) preservation of all positive samples, containing at least one scientific indicator, and (ii) random undersampling of 50% of negative samples (no scientific indicators). This yielded a balanced subset of 1,663 samples for initial experiments. The full augmented corpus was reserved for cross-validation studies.

3.2. Model Architecture and Training Framework

We adopted DeBERTa-v3-large [22, 23] as our base architecture for its superior contextual representation capabilities, appending a multi-label classification head with three sigmoid-activated outputs corresponding to our target classes: (i) scientific claims, (ii) study/publication references, and (iii) scientific entity mentions. All layers remained unfrozen during fine-tuning to enable full domain adaptation.

Loss function formulation: To tackle the remaining class imbalance, the multi-label facet of this task, as well as to address learning from hard examples, we employed focal loss [24] in two sets of experiments: (1) with fixed α parameters, set to 0.7 for each one of the three classes; and (2) with dynamic weights, where α parameters are dynamically computed based on the class distribution of the fully augmented training data.

In both sets of experiments, we use a fixed $\gamma = 2$. On the other hand, the class-adaptive α weights,

$$\alpha_k = \frac{\text{total samples}}{\text{num classes} \times \text{freq}(y_k)},$$

is a scaling that inversely weights each class by its occurrence frequency, thereby increasing penalty for minority classes. For the second strand of experiments, these values are recomputed per epoch based on current batch statistics.

Training protocols: Three training paradigms were progressively explored:

1. **single-model training** on under-sampled data (1,663 samples) with fixed $\alpha = 0.7$ for each class;
2. **fully augmented data training** on the augmented corpus (2,369 samples) with dynamic α recomputation; and
3. **stratified 5-fold cross-validation** with fold-specific dynamic computation of α weights.

All runs used the Adam optimization with linear warm-up and cosine decay. The key hyperparameters included a batch size of 16 (undersampled) and 8 (fully augmented data); a learning rate $\eta = 2 \times 10^{-5}$ for fixed-α, and $\eta = 1 \times 10^{-5}$ for dynamic α; and early stopping criteria where the best checkpoint was selected by the highest macro-F_1 score on validation data.

Threshold optimization strategy: Given the multi-label nature of the task, we replaced the default 0.5 decision threshold with a systematic per-class optimization:

(1) we generated precision-recall curves for each class, and
(2) identified probability thresholds maximizing F_1 scores, done as follows:

$$\tau_k^* = \underset{\tau}{\text{argmax}} \left(\frac{2 \cdot P_k(\tau) \cdot R_k(\tau)}{P_k(\tau) + R_k(\tau)} \right),$$

where P_k and R_k denote class-k precision/recall at threshold τ.

Cross-validation ensemble: For our final architecture, we implemented a 5-model ensemble by:

1. training five independent DeBERTa-v3-large instances on mutually exclusive folds;
2. performing inference via logit averaging:

$$\hat{y}_i^{(ens)} = \sigma\left(\frac{1}{5}\sum_{f=1}^{5}\text{logits}_f^{(i)}\right) \text{; and}$$

3. applying class-specific threshold optimization to ensemble probabilities.

3.3. Evaluation framework

Model performance was assessed using macro-averaged F_1 across all classes, with complementary analysis of per-class precision/recall. All thresholds were optimized exclusively on the development set to prevent data leakage. The final benchmark was the test set available from the CLEF-2025 CheckThat! Lab's scientific web discourse task [20].

4. Methodology: Scientific Claim Source Retrieval

In this task, given a tweet containing an implicit reference to a scientific publication (without URL identifiers), we frame source retrieval as a ranking problem over candidate papers. Our solution employs a two-stage pipeline:

1. a **dense retrieval** step for efficient candidate screening from a large pool of publications; and
2. a **neural re-ranking** step for precision-guided refinement of top candidates.

4.1. Dense Retrieval

The embedding model: We used the Snowflake/snowflake-arctic-embed-l-v2.0 [25] dense retriever to encode both tweets and research papers. Instead of the entire article, only the concatenation of title and abstract were used to represent the input. On the other hand, tweets were represented by their raw text, with casing and punctuation preserved. We used cosine distance between L2-normalized embeddings as the similarity metric.

Training optimization: We split the data into 85% for training and 15% for evaluation, with fixed random seed for reproducibility. This deviation from the original training-evaluation split in the dataset was done to allow for an expansion of the training set and better focus on improving training accuracy while minimizing overfitting. It is worth noting that the final evaluation of the system was conducted on a completely separate test set

Our negative sampling strategy constructs informative triplets by combining one positive paper with nine strategically selected negatives per query: *retrieval hard negatives* (top-ranked incorrect papers from initial retrieval), *semantic hard negatives* (high-similarity non-relevant papers identified through paper-paper similarity queries), and *soft negatives* (contextually relevant papers ranked below position 5 in initial results). We employ a fixed ratio of 3:3:3 for these three types of negatives, ensuring the model learns to distinguish subtle differences between correct matches and challenging distractions at multiple relevance tiers.

4.2. Neural Re-ranking

The re-ranking phase employs a cross-encoder architecture based on `ms-marco-MiniLM-L4-v2`,[1] which processes tweet-paper pairs through concatenated input sequences formatted as `[CLS] tweet [SEP] paper_text [SEP]`. We preserve raw text casing and punctuation

[1] https://huggingface.co/cross-encoder/ms-marco-MiniLM-L4-v2

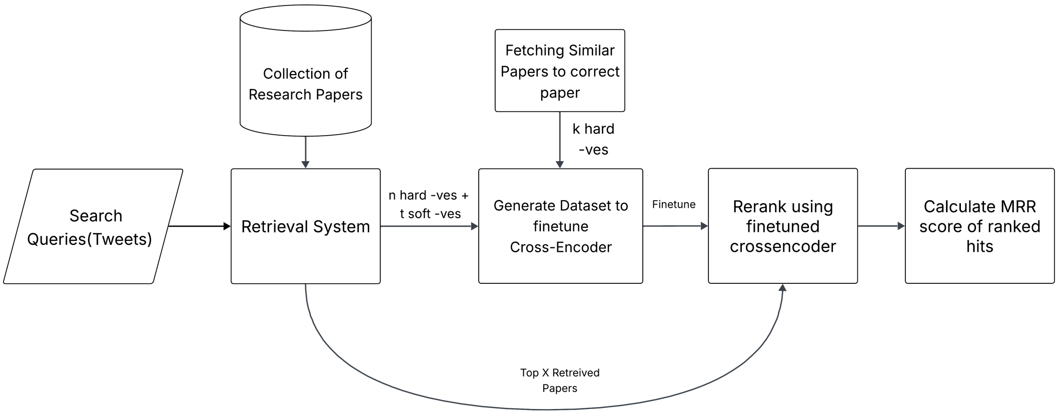
Collection of Research Papers
Fetching Similar Papers to correct paper
k hard -ves
Search Queries(Tweets)
Retrieval System
n hard -ves + t soft -ves
Generate Dataset to finetune Cross-Encoder
Finetune
Rerank using finetuned crossencoder
Calculate MRR score of ranked hits
Top X Retreived Papers

Figure 1: System architecture for the retrieval of sources for scientific claims.

Table 1
Macro-average F_1 and per-category F_1 shown for scientific web discourse detection.

Model	Macro-avg F_1	Categories		
		Scientific Claim	Study Reference	Scientific Entities
deberta_v3_large (test)	0.7917	0.7642	0.7731	0.8377
deberta_v3_large (dev)	0.9208	0.8936	0.9412	0.9275

while applying strict 512-token truncation—deliberately avoiding lemmatization or normalization to maintain linguistic authenticity. This architecture outputs a continuous relevance score $\in [0,1]$ for each candidate pair. To balance recall and MRR objectives, we re-rank only the top-20 candidates from the dense retriever—a cutoff empirically determined to optimize positional sensitivity while avoiding diminishing returns from deeper list exploration.

Baseline implementation and experimental setup: For comparative evaluation, we implemented sparse retrieval baselines (BM25/BM25+) with full preprocessing pipelines comprising the removal of function words, stemming, punctuation stripping, and then indexing complete paper texts (title + abstract). We further tested hybrid pipelines combining BM25 with TinyBERT[2] re-rankers and sentence-transformers with cross-encoders. All experiments were conducted on NVIDIA A100 GPUs using consistent evaluation protocols. The dense retriever (Snowflake-arctic) generated 1024-dimensional embeddings via SentencePiece tokenization, while the cross-encoder used WordPiece tokenization and was fine-tuned with AdamW optimization featuring 500-step linear warmup—ensuring stable convergence during relevance refinement.

5. Results and Analysis

5.1. Scientific Web Discourse

The empirical performance details for scientific web discourse detection are shown in Table 1. Our systematic experiments reveal critical insights into optimizing scientific discourse detection. The initial approach — combining DeepSeek-R1 paraphrase augmentation with strategic under-sampling of negative samples (reducing the training set to 1,663 samples), DeBERTa-v3-large fine-tuning, Focal Loss (fixed α at 0.7 and γ at 2), and per-class threshold optimization — achieved a strong macro-F1 of 0.8849. Our result with adaptive decision boundaries is a significant 5.51% improvement over using the fixed default threshold (the latter yields a macro-F1 score of 0.8298).

[2] https://huggingface.co/cross-encoder/ms-marco-TinyBERT-L2

Table 2
Initial ranking of scientific sources.

Model	Evaluation MRR	Test MRR
BM25	0.623	0.513
BM25+	0.625	0.516
sentence-transformers/multi-qa-mpnet-base-cos-v1	0.529	0.470
Lajavaness/bilingual-embedding-large	0.643	0.562
Snowflake/snowflake-arctic-embed-l-v2.0	0.685	0.575

However, expanding to the full augmented dataset (2,369 samples) with dynamically calculated α-weights (0.2832, 0.4132, and 0.3036) yielded a slight reduction in performance (0.8615 macro-F_1), suggesting that the benefits of additional data volume in this configuration are outweighed by the curated class balance with undersampling.

The most substantial gains emerged from our ensemble strategy: 5-fold cross-validation on the full augmented dataset with fold-specific dynamic α-weighting, followed by logit averaging and threshold optimization. This approach achieved a state-of-the-art macro-F_1 of 0.9208, a **3.6% absolute improvement** the best single-model result. The optimal thresholds (0.74, 0.21, 0.41) revealed *class-dependent sensitivity patterns*, with scientific claims requiring higher confidence thresholds than publication references. Once again, default thresholding degraded ensemble performance — to 0.8595, underscoring that threshold optimization remains indispensable even for sophisticated architectures.

These findings collectively confirm that model diversity through cross-validation and adaptive decision boundaries effectively mitigates variability in linguistic expression and class imbalance inherent to scientific social media discourse.

5.2. Scientific Claim Source Retrieval

Table 2 reports the results of initial ranking with sparse as well as dense encoders. After exploring various permutations of function word removal, punctuation removal, lemmatization, and stemming, we found that the best preprocessing steps consisted of function word removal, stemming, and punctuation removal. Despite earlier studies showing that removal of function words do not significantly affect retrieval with sparse encoders [26], our results indicate that affects downstream performance. In the two sparse encoders used in our experiments, BM25+ showed only marginal improvements over BM25, and hyperparameter tuning did not significantly change retrieval scores.

Among the sentence transformer bi-encoder models evaluated, `multi-qa-mpnet-base-cos-v1` — a dense representation designed for semantic search — emerged as the top performer, though notably still underperforming compared to traditional sparse retrieval methods. Other bi-encoders, including `all-MiniLM-L12-v2`, `multi-qa-distilbert-cos-v1`, and `all-distilroberta-v1`, achieved marginally lower but comparable results, confirming a consistent performance gap relative to non-neural baselines. SciBERT [27] proved particularly unsuitable for this task, likely due to its domain-specific pretraining prioritizing classification objectives over semantic alignment capabilities essential for retrieval. This performance pattern highlights a critical architectural limitation: *standard bi-encoders struggle to capture the nuanced query-document relationships required for implicit citation resolution.*

In contrast, Snowflake/`snowflake-arctic-embed-l-v2.0`, which employs separate specialized encoders for queries and documents, demonstrated superior effectiveness. This suggests that asymmetric embedding spaces better accommodate the structural disparity between conversational tweets and formal scientific text.

Results of the second stage of the retrieval pipeline — neural re-ranking — are shown in Table 3. The neural re-ranking results demonstrate a clear performance advantage for the dense retrieval pipeline. When paired with the `ms-marco-MiniLM-L4-v2` cross-encoder, the dense retriever achieved substantially higher MRR scores (Evaluation: 0.77, Test: 0.65) compared to the sparse retrieval approach with `ms-marco-TinyBERT-L2-v2` cross-encoder (Evaluation: 0.68, Test: 0.55). This **12-13% relative**

Table 3
Results of neural re-ranking with cross-encoders.

Model	Dev MRR	Test MRR
Sparse + cross-encoder/ms-marco-TinyBERT-L2-v2	0.68	0.55
Dense + cross-encoder/ms-marco-MiniLM-L4-v2	0.77	0.65

improvement across both evaluation phases highlights the effectiveness of dense embeddings in capturing semantic relationships between implicit tweet references and candidate papers. The performance gap between evaluation and test sets (approximately 0.12 MRR points for both configurations) suggests consistent generalization behavior, though further analysis would be needed to characterize the distribution shift. These results validate our architectural choice of dense retrieval coupled with moderate-sized cross-encoders for optimal precision-recall balance in scientific source resolution.

5.3. Comparative Evaluation

Compared to other systems participating in the CheckThat! Lab at CLEF 2025, our scientific discourse detection method using ensemble DeBERTa models with adaptive thresholding secured the fourth position, with 79.17% macro-F1 score. This is the aggregate score across three categories of discourse: scientific claims, reference to scientific knowledge, and mention of scientific research context. Our approach had a relatively poor performance in the first category, with 76.42% marco-F1, but a competitive macro-F1 score of 77.31% in the second category. Our approach excelled in detecting mentions of scientific research context, topping the leaderboard with a macro-F1 of 83.77%.

In the second task of source retrieval for implicit scientific claims, our approach based on Snowflake-Arctic dense retrieval and MiniLM cross-encoder re-ranking reported the MMR@5 score of 0.65, securing the 5th position (out of 30 participants) in the leaderboard.

6. Conclusion and Future Work

Our work establishes an effective framework for combating scientific misinformation on social media through two synergistic capabilities: (1) precise detection of scientific discourse in tweets via ensemble DeBERTa models with adaptive thresholding; and (2) robust source retrieval for implicit claims using Snowflake-Arctic dense retrieval and MiniLM cross-encoder re-ranking. The integration of strategic negative sampling, dynamic loss weighting, and logit ensemble methods proved critical in overcoming linguistic noise and class imbalance. This pipeline provides essential infrastructure for downstream credibility assessment, demonstrating that nuanced scientific communication patterns can be computationally modeled with high precision.

We see three promising avenues for future developments. First, extending detection to multi-modal content (images/videos containing scientific claims) would address growing visual misinformation vectors. Second, developing domain-adaptive retrieval that dynamically adjusts to emerging scientific fields could prevent vocabulary obsolescence. Third, creating real-time distillation techniques to compress our ensemble architecture would enable deployment at scale. In conjunction to these avenues, ethical frameworks for automated credibility scoring require careful development to prevent algorithmic bias while maintaining scientific rigor. These advancements would transform our pipeline from a research tool into a practical safeguard for digital scientific discourse.

Acknowledgments

This work was supported in part by a seed award from the AI Innovation Institute (AI3) at Stony Brook University (State University of New York at Stony Brook).

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT for spelling and grammar suggestions in some sections of this manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] E. Hargittai, T. Füchslin, M. S. Schäfer, How Do Young Adults Engage With Science and Research on Social Media? Some Preliminary Findings and an Agenda for Future Research, Social Media + Society 4 (2018) 1–10. doi:10.1177/2056305118797720.

[2] D. Höttecke, D. Allchin, Reconceptualizing nature-of-science education in the age of social media, Science Education 104 (2020) 641–666. doi:https://doi.org/10.1002/sce.21575.

[3] S. Iyengar, D. S. Massey, Scientific communication in a post-truth society, Proceedings of the National Academy of Sciences 116 (2019) 7656–7661. doi:10.1073/pnas.1805868115.

[4] M. J. Page, J. E. McKenzie, J. Kirkham, K. Dwan, S. Kramer, S. Green, A. Forbes, Bias due to selective inclusion and reporting of outcomes and analyses in systematic reviews of randomised trials of healthcare interventions, Cochrane Database of Systematic Reviews (2014). doi:10.1002/14651858.MR000035.pub2.

[5] H. Wolinsky, Disease mongering and drug marketing: Does the pharmaceutical industry manufacture diseases as well as drugs?, EMBO Reports 6 (2005) 612–614.

[6] C. Zuo, Q. Zhang, R. Banerjee, An Empirical Assessment of the Qualitative Aspects of Misinformation in Health News, in: A. Feldman, G. Da San Martino, C. Leberknight, P. Nakov (Eds.), Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda, Association for Computational Linguistics, Online, 2021, pp. 76–81. URL: https://aclanthology.org/2021.nlp4if-1.11/. doi:10.18653/v1/2021.nlp4if-1.11.

[7] C. Mothes, Confirmation Bias, The SAGE encyclopedia of political behavior 2 (2017) 125. doi:10.4135/9781483391144.n61.

[8] S. Galdi, B. Gawronski, L. Arcuri, M. Friese, Selective exposure in decided and undecided individuals: Differential relations to automatic associations and conscious beliefs, Personality and Social Psychology Bulletin 38 (2012) 559–569.

[9] P. Moravec, R. Minas, A. R. Dennis, Fake News on Social Media: People Believe What They Want to Believe When it Makes No Sense at All, MIS Quarterly 43 (2019) 1343–1360.

[10] C. Zuo, N. Acharya, R. Banerjee, Querying Across Genres for Medical Claims in News, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 1783–1789. doi:10.18653/v1/2020.emnlp-main.139.

[11] R. Levy, S. Gretz, B. Sznajder, S. Hummel, R. Aharonov, N. Slonim, Unsupervised corpus–wide claim detection, in: I. Habernal, I. Gurevych, K. Ashley, C. Cardie, N. Green, D. Litman, G. Petasis, C. Reed, N. Slonim, V. Walker (Eds.), Proceedings of the 4th Workshop on Argument Mining, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 79–84. doi:10.18653/v1/W17-5110.

[12] S. Cazalens, P. Lamarre, J. Leblay, I. Manolescu, X. Tannier, A content management perspective on fact-checking, in: Companion Proceedings of the The Web Conference 2018, 2018, pp. 565–574.

[13] N. Hassan, C. Li, M. Tremayne, Detecting check-worthy factual claims in presidential debates, in: Proceedings of the 24th acm international on conference on information and knowledge management, 2015, pp. 1835–1838.

[14] C. Zuo, A. Karakas, R. Banerjee, A hybrid recognition system for check-worthy claims using heuristics and supervised learning, in: CEUR workshop proceedings, volume 2125, 2018.

[15] D. Wright, I. Augenstein, Claim check-worthiness detection as positive unlabelled learning, in: T. Cohn, Y. He, Y. Liu (Eds.), Findings of the Association for Computational Linguistics: EMNLP

2020, Association for Computational Linguistics, Online, 2020, pp. 476–488. doi:10.18653/v1/2020.findings-emnlp.43.

[16] M. Färber, A. Jatowt", Citation recommendation: approaches and datasets, Int J Digit Libr 21 (2020) 371–405. doi:10.1007/s00799-020-00288-2.

[17] V. Viswanathan, G. Neubig, P. Liu, CitationIE: Leveraging the citation graph for scientific information extraction, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), 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, 2021, pp. 719–731. doi:10.18653/v1/2021.acl-long.59.

[18] D. Rousidis, E. Garoufallou, P. Balatsoukas, K. Paraskeuopoulos, S. Asderi, D. Koutsomiha, Metadata requirements for repositories in health informatics research: evidence from the analysis of social media citations, in: Metadata and Semantics Research: 7th Research Conference, MTSR 2013, Thessaloniki, Greece, November 19-22, 2013. Proceedings 7, Springer, 2013, pp. 246–257.

[19] 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.

[20] S. Hafid, Y. S. Kartal, S. Schellhammer, K. Boland, D. Dimitrov, S. Bringay, K. Todorov, S. Dietze, Overview of the CLEF-2025 CheckThat! Lab Task 4 on Scientific Web Discourse, 2025.

[21] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al., DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, arXiv preprint arXiv:2501.12948 (2025).

[22] P. He, X. Liu, J. Gao, W. Chen, DeBERTa: Decoding-enhanced BERT with Disentangled Attention, in: International Conference on Learning Representations, 2021.

[23] P. He, J. Gao, W. Chen, DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing, 2021. arXiv:2111.09543.

[24] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, Focal Loss for Dense Object Detection, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.

[25] P. Yu, L. Merrick, G. Nuti, D. Campos, Arctic-Embed 2.0: Multilingual Retrieval Without Compromise, Technical Report, Snowflake Inc., 2024.

[26] A. Trotman, A. Puurula, B. Burgess, Improvements to bm25 and language models examined, in: Proceedings of the 19th Australasian Document Computing Symposium, 2014, pp. 58–65.

[27] I. Beltagy, K. Lo, A. Cohan, SciBERT: A pretrained language model for scientific text, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), 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, 2019, pp. 3615–3620. doi:10.18653/v1/D19-1371.