Towards Real-World Fact-Checking with Large Language Models

Iryna Gurevych[a]

[a]Ubiquitous Knowledge Processing Lab (UKP Lab),
Department of Computer Science and Hessian Center for AI (hessian.AI),
TU Darmstadt
www.ukp.tu-darmstadt.de

1 Abstract

Misinformation poses a growing threat to our society. It has a severe impact on public health by promoting fake cures or vaccine hesitancy, and it is used as a weapon during military conflicts, political elections, and crisis events to spread fear and distrust. Harmful misinformation is overwhelming human fact-checkers, who cannot keep up with the quantity of information to verify online. There is a strong potential for automated Natural Language Processing (NLP) methods to assist them in their tasks [8]. Real-world fact-checking is a complex task, and existing datasets and methods tend to make simplifying assumptions that limit their applicability to real-world, often ambiguous, claims [3, 6]. Image, video, and audio content are now dominating the misinformation space, with 80% of fact-checked claims being multimedia in 2023 [1]. When confronted with visual misinformation, human fact-checkers dedicate a significant amount of time not only to debunk the claim but also to identify accurate alternative information about the image, including its provenance, source, date, location, and motivation, a task that we refer to as image contextualization [9].

Furthermore, the core focus of current NLP research for fact-checking has been on identifying evidence and predicting the veracity of a claim. People's beliefs, however, often do not depend on the claim and the rational reasoning but on credible content that makes the claim seem more reliable, such as scientific publications [4, 5] or visual content that was manipulated or stems from unrelated contexts [1, 2, 9]. To combat misinformation, we need to show (1) "Why was the claim believed to be true?", (2) "Why is the claim false?", (3) "Why is the alternative explanation correct?" [7]. In this talk, I will zoom into two critical aspects of such misinformation supported by credible though misleading content.

Firstly, I will present our efforts to dismantle misleading narratives based on fallacious interpretations of scientific publications [4, 5]. On the one hand, we discover a strong ability of LLMs to reconstruct and, hence, explain fallacious arguments based on scientific publications. On the other hand, we make the concerning observation that LLMs tend to support false scientific claims when paired with fallacious reasoning [5]. Secondly, I will show how we can use state-of-the-art multimodal large language models to (1) detect misinformation based on visual content [2] and (2) provide strong alternative explanations for the visual content. I will conclude this talk by showing how LLMs can be used to support human fact-checkers for image contextualization [9].

References

[1] N. Dufour, A. Pathak, P. Samangouei, N. Hariri, S. Deshetti, A. Dudfield, C. Guess, P. H. Escayola, B. Tran, M. Babakar, and C. Begler. Ammeba: A large-scale survey and dataset of media-based misinformation in-the-wild. *arXiv preprint arXiv:2405.11697*, 2024. doi: 10.48550/arXiv.2405.11697. URL https://arxiv.org/abs/2405.11697.

[2] J. Geng, Y. Kementchedjhieva, P. Nakov, and I. Gurevych. Multimodal large language models to support real-world fact-checking. *arXiv preprint arXiv:2403.03627*, 2024. doi: 10.48550/arXiv.2403.03627. URL https://arxiv.org/abs/2403.03627.

[3] M. Glockner, Y. Hou, and I. Gurevych. Missing counter-evidence renders NLP fact-checking unrealistic for misinformation. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5916–5936, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.397. URL https://aclanthology.org/2022.emnlp-main.397.

[4] M. Glockner, Y. Hou, P. Nakov, and I. Gurevych. Missci: Reconstructing fallacies in misrepresented science. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4372–4405, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.240.

[5] M. Glockner, Y. Hou, P. Nakov, and I. Gurevych. Grounding fallacies misrepresenting scientific publications in evidence. *arXiv preprint arXiv:2408.12812*, 2024. doi: 10.48550/arXiv.2408.12812. URL https://arxiv.org/abs/2408.12812.

[6] M. Glockner, I. Staliūnaitė, J. Thorne, G. Vallejo, A. Vlachos, and I. Gurevych. AmbiFC: Fact-checking ambiguous claims with evidence. *Transactions of the Association for Computational Linguistics*, 12:1–18, 2024. doi: 10.1162/tacl_a_00629. URL https://aclanthology.org/2024.tacl-1.1.

[7] S. Lewandowsky, J. Cook, U. Ecker, D. Albarracín, P. Kendeou, E. J. Newman, G. Pennycook, E. Porter, D. G. Rand, D. N. Rapp, J. Reifler, J. Roozenbeek, P. Schmid, C. M. Seifert, G. M. Sinatra, B. Swire-Thomposon, S. van der Linden, T. J. Wood, and M. S. Zaragoza. The debunking handbook 2020, 2020. URL https://digitalcommons.unl.edu/scholcom/245/. Accessed: 2023-09-19.

[8] P. Nakov, D. Corney, M. Hasanain, F. Alam, T. Elsayed, A. Barrón-Cedeño, P. Papotti, S. Shaar, and G. Da San Martino. Automated fact-checking for assisting human fact-checkers. In Z.-H. Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4551–4558. International Joint Conferences

on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/619. URL https://doi.org/10.24963/ijcai.2021/619. Survey Track.

[9] J. Tonglet, M.-F. Moens, and I. Gurevych. "image, tell me your story!" predicting the original meta-context of visual misinformation. *arXiv preprint arXiv:2408.09939*, 2024. doi: 10.48550/arXiv.2408.09939. URL https://www.arxiv.org/abs/2408.09939.