

What Does it Take to Effectively Communicate Fact Checking Results?

Amelie Wühlrl

IT University of Copenhagen
Copenhagen, Denmark
amwy@itu.dk

Introduction. While automatic fact verification has grown into an established research field in the last years, work on the follow-up step, i.e., conveying fact checking results to users, is still lacking. Presumably, this is because it requires a multidisciplinary approach, as NLP methods to generate such result briefs with a meaningful impact may depend on the consideration of complex psychological processes. In this extended abstract, we discuss the setting and the tasks, questions and considerations connected to it.

Automatic fact verification. Fact verification has become an established task in Natural Language Processing (NLP), both for general domain as well as scientific and medical claims (Guo, Schlichtkrull, and Vlachos 2022; Vladika and Matthes 2023). A full fact checking pipeline consists of multiple components. Guo, Schlichtkrull, and Vlachos (2022) highlight four tasks: First, claim detection needs to determine which elements in the discourse are check-worthy. Claims are considered check-worthy if they are factual and thus theoretically verifiable and of interest to society or specific stakeholders (Majer and Šnajder 2024). Evidence retrieval constitutes the second part in the pipeline. Here, the objective is to identify one or multiple pieces of evidence – primarily text, but also other evidence modalities such as images (Chen, Tang, and Thomas 2024) – that are relevant to the claim. Once we obtain a claim–evidence pair, verdict prediction infers the relation between the claim and the evidence document by predicting if the evidence *supports* or *refutes* the claim. In an effort to make fact checking results more transparent, trustworthy and therefore convincing, the final step in a fact checking pipeline is justification or explanation production. This includes explainability or interpretability approaches such as identifying parts in the evidence that were most influential for verdict prediction or justifying a verdict by generating an evidence summary discussing how the evidence led to a verdict. All four components constitute active areas of research within NLP, however, the step which follows the fact check in a real-world setup – conveying the results to users – is underexplored.

How to convey fact checking results to users in a meaningful way? For the sake of this argument, let us assume we have a system capable of reliably verifying claims that users are sharing online. Beyond generating fact checking justifications and explanations such as straight-forward ev-

idence summaries, it is crucial to carefully evaluate what needs to be considered when communicating verification results. Conveying fact checking results or explanations to social media users can only have a meaningful impact and a chance to deconstruct misinformation, if we take into consideration the psychological processes that shape human behavior, cognition and communication phenomena. In the case of fact verification, this could be processes such as cognitive biases, information overload, or other psychological and behavioral inclinations of the audience.

Which psychological & behavioral inclinations should we consider? To this end, we have to turn to research fields such as science communication and psychology which have an advanced understanding of how to effectively communicate a fact check (pui Sally Chan et al. 2017; Soprano et al. 2024). With that in mind, we may leverage NLP to (a) customize the level of detail of an explanation as to not overwhelm users with information they are unable or unwilling to parse; or (b) adapt the focus or framing of an explanation by taking into account a user’s background and knowledge as reflected in previous interactions. Moreover, we can explore customizing the presentation of fact checking results in order to appeal to their psychological, behavioral or communicative characteristics. For instance, regulatory focus theory (Higgins 1998) could inform the way a model phrases results such that the explanation appeals to a user’s motivation. Similarly, when the focus is on medical claims, leveraging the transtheoretical model of behavior change (Prochaska and Velicer 1997), we could explore detecting a patient’s mental stage, and optimize the explanation that de-constructs false information to resonate with their mental state. Alternatively, cognitive appraisal theory (Smith and Ellsworth 1985; Scherer 2005) allows us to detect specific components in text that are eliciting a ‘misinformation reaction’, which an explanation could then target. Similarly, it could help estimating check-worthiness, by detecting claims or other triggers in online discourse that have the potential to evoke an emotional response and therefore accelerate how quickly a claim spreads. Exploring how psychology and communication science can inform fact checking explanation generation to ultimately make the results more impactful, is a crucial research path for future work at the intersection of NLP, psychology and sociology. As we dis-

cussed, this line of work could even extend into earlier steps in the verification pipeline, such as determining which is the most check-worthy claim for a specific user or the community as a whole.

What are the triggers of distorted public discourse?

Zooming out a bit further to misinformation at the level of media discourse, we could investigate if targeted response to misinformation may even enable counteracting the unsolicited amplification of topics. During this process “a particular phenomenon or point of view may acquire disproportionate importance or prevalence as a result of being selected for media coverage”¹. Unfortunately, to a certain extent fact checking may even add to amplifying a topic. If a news agency publishes a fact check verifying a claim in a debate, this inevitably draws new attention not only to the claim, but also to the debate as a whole. To address and counteract this process, we need to identify the characteristics of events, political topics, social media claims, or scientific findings that have the potential to be over-amplified – a question which social scientists may already be investigating. Further, we require a detailed understanding of *why* a particular piece of news or information under investigation in a fact check resonates with the people who share or engage with it – a question which psychologists may be able to provide insights on. From the NLP side, this information could for example inform the development of early warning systems for media amplification and discourse distortion. It further involves a detailed understanding of how information disseminates online and in other spaces of public discourse, based on which journalists, politicians and social media users can work on re-framing narratives.

Investigating the downsides of personalization. At the same time, we have to consider the trade offs related to this type of personalization. As a prerequisite for this strategy, we need to acquire substantial knowledge about the user which is costly and, importantly, involves mining potentially sensitive data. In most cases, we would probably need to infer or generalize certain characteristics, which requires careful considerations regarding potential model-induced biases. Moreover, strong personalization of how the fact checking results are presented runs the risk of only being appealing and accessible to a specific user. If such posts are shared, they might lose their effectiveness. This motivates a two-fold strategy which consists of addressing the results from both a general perspective *and* a personalized one.

Outlook. The challenges we outlined in this extended abstract clearly require expertise going beyond NLP. While NLP practitioners and researchers can develop the tools and methods to perform a fact check and generate a summary of the results, we have to work together with researchers from psychology and sociology to understand the processes that shape how individuals perceive and interact with the generated content. Therefore, this paper is motivated by the need to start an exchange between these scientific fields, to define

and explore how AI, and specifically NLP, can support counteracting misinformation in an effective and human-centric way. This extended abstract serves as a starting point for a joint discussion and an outlook on which directions to explore.

References

- Chen, T.-C.; Tang, C.-W.; and Thomas, C. 2024. Meta-SumPerceiver: Multimodal Multi-Document Evidence Summarization for Fact-Checking. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8742–8757. Bangkok, Thailand: Association for Computational Linguistics.
- Guo, Z.; Schlichtkrull, M.; and Vlachos, A. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, 10: 178–206.
- Higgins, T. E. 1998. Promotion and prevention: Regulatory focus as a motivational principle. *Advances in experimental social psychology*, 30.
- Majer, L.; and Šnajder, J. 2024. Claim Check-Worthiness Detection: How Well do LLMs Grasp Annotation Guidelines? In Schlichtkrull, M.; Chen, Y.; Whitehouse, C.; Deng, Z.; Akhtar, M.; Aly, R.; Guo, Z.; Christodoulopoulos, C.; Cocarascu, O.; Mittal, A.; Thorne, J.; and Vlachos, A., eds., *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, 245–263. Miami, Florida, USA: Association for Computational Linguistics.
- Prochaska, J. O.; and Velicer, W. F. 1997. The transtheoretical model of health behavior change. *American journal of health promotion : AJHP*, 12(1): 38–48. Place: United States.
- pui Sally Chan, M.; Jones, C. R.; Jamieson, K. H.; and Albarracín, D. 2017. Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation. *Psychological Science*, 28(11): 1531–1546. PMID: 28895452.
- Scherer, K. R. 2005. What are emotions? And how can they be measured? *Social science information*, 44(4): 695–729.
- Smith, C. A.; and Ellsworth, P. C. 1985. Patterns of cognitive appraisal in emotion. *Journal of personality and social psychology*, 48(4): 813.
- Soprano, M.; Roitero, K.; La Barbera, D.; Ceolin, D.; Spina, D.; Demartini, G.; and Mizzaro, S. 2024. Cognitive Biases in Fact-Checking and Their Countermeasures: A Review. *Inf. Process. Manage.*, 61(3).
- Vladika, J.; and Matthes, F. 2023. Scientific Fact-Checking: A Survey of Resources and Approaches. In *Findings of the Association for Computational Linguistics: ACL 2023*, 6215–6230. Toronto, Canada: Association for Computational Linguistics.

¹<https://www.oxfordreference.com/display/10.1093/acref/9780199646241.001.0001/acref-9780199646241-e-67>