# Fact-Checking at Scale with Crowdsourcing: Experiments and Lessons Learned

Discussion Paper

David La Barbera[1,*], Michael Soprano[1], Kevin Roitero[1], Eddy Maddalena[1] and Stefano Mizzaro[1]

[1]*University of Udine, Udine, Italy*

Abstract

In this paper, we present our journey in exploring the use of crowdsourcing for fact-checking. We discuss our early experiments aimed towards the identification of the best possible setting for misinformation assessment using crowdsourcing. Our results indicate that the crowd can effectively address misinformation at scale, showing some degree of correlation with experts. We also highlight the influence of worker background on the quality of truthfulness assessments.

**Keywords**

crowdsourcing and human computation, fact-checking and misinformation, truthfulness assessment

## 1. Introduction

Fact-checking is crucial as online misinformation erodes trust in traditional sources, leading to real-world consequences. Experts' capacity to keep up with social media's information overload is challenged, requiring efficient, scalable alternatives to mitigate misinformation. One promising approach in this direction is the use of crowdsourcing, which taps into the collective intelligence of a large group of people, and has been successfully used in various domains [1, 2, 3]. In the realm of fact-checking and misinformation identification, crowdsourcing potentially offers a scalable, cost-effective, and efficient solution to evaluate information truthfulness. With diverse worker backgrounds and skills, it enables a comprehensive assessment of information from multiple perspectives, while enabling timely processing of large information volumes to combat misinformation. However, there are some challenges associated with the usage of crowdsourcing for fact-checking. Ensuring qualified, reliable, and motivated workers is critical for reliable fact-checking. Worker demographics, such as gender, age, education, and cultural background, can also affect truthfulness assessments. Therefore, it is essential to understand how worker background impacts assessment quality and develop strategies to mitigate any biases or inaccuracies that may arise.

*IIR2023: 13th Italian Information Retrieval Workshop, 8th - 9th June 2023, Pisa, Italy*

*Corresponding author.

david.labarbera@uniud.it (D. L. Barbera); michael.soprano@uniud.it (M. Soprano); kevin.roitero@uniud.it (K. Roitero); eddy.maddalena@uniud.it (E. Maddalena); stefano.mizzaro@uniud.it (S. Mizzaro)

0000-0002-8215-5502 (D. L. Barbera); 0000-0002-7337-7592 (M. Soprano); 0000-0002-9191-3280 (K. Roitero); 0000-0002-5423-8669 (E. Maddalena); 0000-0002-2852-168X (S. Mizzaro)



This paper summarizes our research on using crowdsourcing for fact-checking [4, 5, 6, 7, 8]. We describe our journey from early experiments to identifying optimal settings for misinformation assessment. Our findings show that, under certain conditions, crowdsourcing is a viable tool for reliable truthfulness evaluation of publicly available information. This summary contributes to the ongoing conversation on leveraging crowdsourcing for fact-checking and for fighting online misinformation.

## 2. Crowdsourcing Truthfulness

Crowdsourcing has been effective in various information credibility research contexts, such as news quality evaluation [9]. Past studies have shown its potential for fact-checking, and recent years have seen the use of crowdsourcing-based approaches to scale manual fact-checking efforts. Examples include collecting credibility annotations on climate change [10], crowdsourcing news source quality labels [11], and tracking Twitter misinformation [12].

It is thus reasonable to hypothesize good performances of the wisdom of the crowd approach also when assessing the truthfulness of given information items. In theory, crowdsourcing offers several advantages when it comes to fact-checking and identifying misinformation. Firstly, it allows for rapid and cost-effective evaluation of a large volume of information, making it scalable. Secondly, the diversity of the worker pool can provide multiple perspectives for a comprehensive assessment. Finally, the use of crowdsourcing can enhance transparency by relying on input from multiple individuals, rather than a single expert's opinion. To experimentally verify whether a crowd of non-expert human judges can detect and objectively categorize online (mis)information thus supporting the fact-checking activity, we performed experiments with a pipeline built to leverage crowdsourcing for the misinformation assessment task. We discuss the developed experiments along with their results in the following.

### 2.1. Effect of Judgment Scale and Workers Background [4, 5]

We first conducted an experiment to evaluate the truthfulness of a subset of 120 statements made by US politicians [4]. We used the PolitiFact[1] dataset [13] to sample the statements, and we built an experiment where each statement was evaluated by 10 distinct US-based crowd workers. To assess the truthfulness of each statement, we used two different scales: a 6-level scale, which is the same scale used by PolitiFact experts, and a finer-grained 100-level scale. Each worker was asked to evaluate one statement for each of the 6 original ground truth levels, in addition to 2 custom-made statements used as gold questions for quality control. Workers were also asked to provide a URL as a source of evidence and a brief textual justification to support their provided judgments. Prior to the task, each worker filled out a demographic questionnaire. The detailed settings and results of this experiment can be found in La Barbera et al. [4].

This experiment demonstrated that crowd workers can effectively classify misinformation using the proposed truthfulness scales, as indicated by the results of the aggregation functions tested and reported by La Barbera et al. [4, Figure 4]. The arithmetic mean consistently produced

[1] https://www.politifact.com/

the most accurate final truthfulness score among workers. However, the use of a 100-level scale exhibited leniency towards values multiples of 10, suggesting that less coarse-grained scales may be preferable be workers. Additionally, political bias was observed in workers' assessments, with individuals tending to be more indulgent towards statements made by politicians of their own political affiliation. This experiment also highlighted potential systematic biases, as over 70% of the evidence sources cited by workers originated from the PolitiFact domain.

To enhance the robustness and generalizability of our findings, we also extended the previous experiment by incorporating a 3-level scale in addition to the 6 and 100-level ones along with statements from ABC News[2] as a second source for statements in addition to PolitiFact. Each worker was asked to assess six statements from PolitiFact, three from ABC, and two gold standard questions. To prevent search bias observed in the previous experiment, we excluded ABC and PolitiFact search results from worker web searches. Furthermore, to measure worker's ability to override their initial "gut" response and engage in further reflection to find the correct answer, we included a Cognitive Reflection Test (CRT) questionnaire.

In this second experiment, we confirmed and extended the findings of the first, demonstrating worker agreement with experts across various scales, with a preference for less fine-grained scales [5, Figure 2]. Our results also revealed worker bias, with political background and demographic factors affecting judgment accuracy [5, Table 5]. These findings support the use of crowdsourced truthfulness judgments to assess misinformation on a large scale. However, we found several unresolved issues concerning the statements evaluated in our experiments, as some were not recent and were made by public figures from a country other than that of the workers.

## 2.2. Impact Of Information Recency [6, 7]

To improve previous research findings, we conducted a study on the COVID-19 pandemic, a major source of daily misinformation. Our objective was to assess the feasibility and reliability of crowdsourcing as a method for evaluating the truthfulness of subjective and time-sensitive misinformation, as compared to expert judgment. We initiated our study by requesting crowd workers to evaluate the veracity of COVID-19-related statements during the pandemic and provide corroborating evidence [6]. We then carried out a longitudinal study by repeating this task multiple times with both novice and experienced workers recruited throughout the course of the pandemic [7].

Our study found that the crowd can assess the truthfulness of pandemic-related statements, in agreement with previous experiments (see [6, Figure 2]). Our results on worker behavior, agreement, scale transformations, aggregation functions, worker background and bias align with previously collected data. Our longitudinal study further confirmed the presence of worker biases observed in previous experiments (see [6, Table 2]). We also found experienced workers exhibiting different behaviors than novices; experienced workers spent more time evaluating each statement and their quality was similar to or higher than other workers. Additionally, we found that as the number of batches performed by workers increased, the average time spent on all documents decreased substantially, indicating a significant impact of time span on judgment quality for both novice and experienced workers.

[2] https://apo.org.au/collection/302996/rmit-abc-fact-check

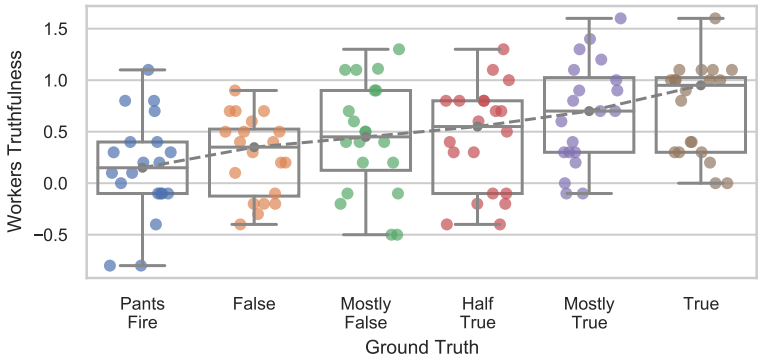


**Figure 1:** Average truthfulness score reported by workers (y-axis) compared to the experts (x-axis). Each dot is a statement. Dashed line connects the median values. Adapted from Soprano et al. [8, Figure 3].

### 2.3. Multidimensional Truthfulness [8]

Previous experiments [4, 5, 6, 7] have examined the ability of the crowd to discern the truthfulness of statements using scales with a different number of levels. However, the concept of truthfulness is complex and encompasses factors such statement clarity and precision that cannot be captured with a uni-dimensional scale. To address this issue, in Soprano et al. [8] we investigated a multidimensional notion of truthfulness.

Building on previous research [9, 14, 15, 16], we conducted an experiment where workers evaluated seven dimensions of truthfulness: Correctness, Neutrality, Comprehensibility, Precision, Completeness, Speaker's Trustworthiness, and Informativeness; we also kept a similar overall setting as for the previous experiments. Our analysis of the newly collected judgments revealed that workers were reliable compared to an expert-provided gold standard, as shown in Figure 1, providing yet another confirmation of previous findings. We also found the dimensions of truthfulness able to capture independent "facets" of information, thus suggesting that the multidimensional scale can provide a valuable basis for a more complete assessment of statement truthfulness.

## 3. Discussion and Conclusion

This paper presents our research on using crowdsourcing for fact-checking and provides a summary of our findings. While our results suggest that crowdsourcing has the potential to be a reliable way to combat misinformation at scale, more work is needed to improve its reliability for use in real settings. Future studies should explore strategies to enhance worker effectiveness by investigating additional truthfulness dimensions, worker motivation and incentives, more complex aggregation functions, and worker behavioral signals. Crowdsourcing has the potential to evaluate large volumes of information from diverse perspectives, making it a valuable resource for fact-checking.

## References


[1] J. Howe, The rise of crowdsourcing, Wired Magazine 14 (2006) 1–4. URL: https://www.wired.com/2006/06/crowds/.

[2] D. C. Brabham, Crowdsourcing as a model for problem solving: An introduction and cases, Convergence 14 (2008) 75–90. doi:10.1177/1354856507084420.

[3] D. C. Brabham, K. M. Ribisl, T. R. Kirchner, J. M. Bernhardt, Crowdsourcing applications for public health, American Journal of Preventive Medicine 46 (2014) 179–187. doi:https://doi.org/10.1016/j.amepre.2013.10.016.

[4] D. La Barbera, K. Roitero, G. Demartini, S. Mizzaro, D. Spina, Crowdsourcing Truthfulness: The Impact of Judgment Scale and Assessor Bias, in: J. M. Jose, E. Yilmaz, J. Magalhães, P. Castells, N. Ferro, M. J. Silva, F. Martins (Eds.), Advances in Information Retrieval, Springer International Publishing, Cham, 2020, pp. 207–214. doi:10.1007/978-3-030-45442-5_26, Best Short Paper Award.

[5] K. Roitero, M. Soprano, S. Fan, D. Spina, S. Mizzaro, G. Demartini, Can The Crowd Identify Misinformation Objectively? The Effects of Judgment Scale and Assessor's Background, in: Proceedings of the 43st International ACM SIGIR Conference on Research and Development in Information Retrieval., SIGIR '20, Association for Computing Machinery, 2020, p. 439–448. doi:10.1145/3397271.3401112.

[6] K. Roitero, M. Soprano, B. Portelli, D. Spina, V. Della Mea, G. Serra, S. Mizzaro, G. Demartini, The COVID-19 Infodemic: Can the Crowd Judge Recent Misinformation Objectively?, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM '20, Association for Computing Machinery, 2020, p. 1305–1314. doi:10.1145/3340531.3412048.

[7] K. Roitero, M. Soprano, B. Portelli, M. De Luise, D. Spina, V. D. Mea, G. Serra, S. Mizzaro, G. Demartini, Can The Crowd Judge Truthfulness? A Longitudinal Study On Recent Misinformation About COVID-19, Personal and Ubiquitous Computing (2021). doi:10.1007/s00779-021-01604-6.

[8] M. Soprano, K. Roitero, D. La Barbera, D. Ceolin, D. Spina, S. Mizzaro, G. Demartini, The many dimensions of truthfulness: Crowdsourcing misinformation assessments on a multidimensional scale, Information Processing & Management 58 (2021) 102710. doi:10.1016/j.ipm.2021.102710.

[9] E. Maddalena, D. Ceolin, S. Mizzaro, Multidimensional News Quality: A Comparison of Crowdsourcing and Nichesourcing, in: Proceedings of the CIKM 2018 Workshops co-located with 27th ACM International Conference on Information and Knowledge Management, 2018. URL: http://ceur-ws.org/Vol-2482/paper17.pdf.

[10] M. M. Bhuiyan, A. X. Zhang, C. M. Sehat, T. Mitra, Investigating Differences in Crowdsourced News Credibility Assessment: Raters, Tasks, and Expert Criteria, Proceedings of the ACM on Human-Computer Interaction 4 (2020). doi:10.1145/3415164.

[11] G. Pennycook, D. G. Rand, Fighting Misinformation on Social Media Using Crowdsourced Judgments of News Source Quality, Proceedings of the National Academy of Sciences 116 (2019) 2521–2526. doi:10.1073/pnas.1806781116.

[12] A. Ghenai, Y. Mejova, Catching Zika Fever: Application of Crowdsourcing and Machine Learning for Tracking Health Misinformation on Twitter, in: 2017 IEEE International

Conference on Healthcare Informatics, 2017, pp. 518–518. doi:10.1109/ICHI.2017.58.
[13] W. Y. Wang, "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection, in: R. Barzilay, M. Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, volume 4, Association for Computational Linguistics, 2017, pp. 422–426. doi:10.18653/v1/P17-2067.
[14] International Organization for Standardization, ISO/IEC 25012:2008 Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model, Technical Report, ISO, 2008. URL: https://www.iso.org/standard/35736.html.
[15] B. K. Kahn, D. M. Strong, R. Y. Wang, Information Quality Benchmarks: Product and Service Performance, Communications of the ACM 45 (2002) 184–192. doi:10.1145/505248.506007.
[16] D. Ceolin, J. Noordegraaf, L. Aroyo, Capturing the Ineffable: Collecting, Analysing, and Automating Web Document Quality Assessments, in: E. Blomqvist, P. Ciancarini, F. Poggi, F. Vitali (Eds.), Knowledge Engineer, Springer International Publishing, Cham, 2016, pp. 83–97. doi:10.1007/978-3-319-49004-5_6.