Enhancing Fact-Checking: From Crowdsourced Validation to Integration with Large Language Models

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

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

Abstract

This extended abstract presents results from two recent studies [1, 2] aimed at enhancing the practical application and effectiveness of fact-checking systems. La Barbera et al. [1] detail the implementation of crowdsourcing in fact-checking, demonstrating its practical viability through experimental evaluation using a dataset of political public statements. Zeng et al. [2] build on this foundation by integrating crowdsourced data with Large Language Models, proposing the first hybrid system that combines human insights and AI capabilities.

Keywords

Misinformation, Fact Checking, Large Language Models

1. Introduction

The rapid proliferation of misinformation across digital platforms poses significant challenges to societal trust and public safety. Traditional fact-checking methods, predominantly based on experts, are unable to cope with the ever-increasing volume and speed of misinformation dissemination. This has created interest in the development of more scalable solutions that are able to enhance the accuracy and efficiency of misinformation detection methods, addressing the problem at scale. Crowdsourcing has emerged as a powerful tool in this domain [3, 4, 5, 6, 7], based on the usage of the collective wisdom of the crowd [8] applied to fact verification. While promising, the application of crowdsourcing in fact-checking requires careful consideration of factors such as task design, worker motivation, and data quality to ensure its effectiveness.

In this work, we present results from two recent studies in the field of crowdsourced fact-checking [1, 2]. Specifically, La Barbera et al. [1] explore the potential of crowdsourcing to provide a robust foundation for practical fact-checking applications at scale. Moreover, the effectiveness of crowdsourcing can be significantly enhanced by leveraging Large Language Models (LLMs), which provide powerful complementary support to human efforts. By combining LLMs with human-generated annotations, Zeng et al. [2] introduce a novel hybrid approach developed to address misinformation at scale.

IIR'24: 14th Italian Information Retrieval Workshop, September 05–06, 2024, Udine, Italy

*Corresponding author.

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

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

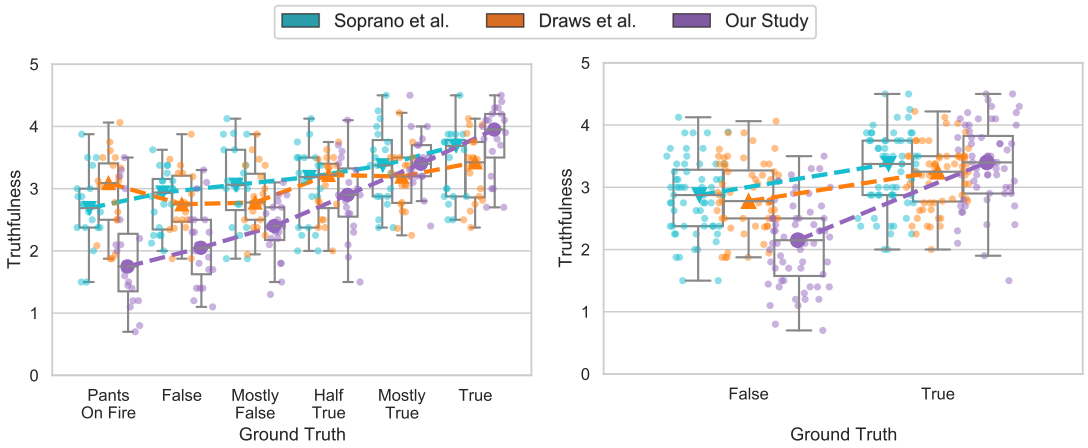

Figure 1: Worker agreement with experts on the original ground truth 6-level scale (left) and binarized 2-level scale (right). From La Barbera et al. [1].

2. Crowdsourced Fact-checking: Does It Actually Work?

Objective and Methodology. The primary objective of La Barbera et al. [1] was to develop a viable crowdsourcing approach to fact-checking. The methodology employed involved a re-design of the crowdsourcing task used in previous studies. A diverse group of workers was recruited through the Prolific platform. Participants were given tasks involving a curated dataset of political statements from the PolitiFact[1] website and were instructed to fact-check each statement using the same six-level scale used by experts.

Results. This study highlighted the effectiveness of crowdsourcing for fact-checking, demonstrating higher effectiveness when compared to previous studies [6, 9]. La Barbera et al. [1] compared workers' agreement levels across different studies. Figure 1 presents the results, with the three series replicating the findings from previous research by Soprano et al., Draws et al., and their study. For each series in the figure, the x-axis displays each statement ground truth level, while the y-axis represents the mean of assessments as supplied by the crowd. Small dots represent individual statements, with larger markers and marking the median values for statements at each truthfulness level. La Barbera et al. observed a distinct trend where the median aggregated truthfulness values consistently increased with higher ground truth levels, indicating a clear relationship between the perceived truthfulness of statements as perceived by the crowd and the actual one provided by experts. This pattern was less evident in the studies by Soprano et al. and Draws et al., where median values decreased in some cases, such as moving from Pants-On-Fire to False truthfulness levels. Further validation of our results was confirmed trough a statistically significant distinction among them: our study differed significantly from both Soprano et al. ($p < 0.01$) and Draws et al. ($p < 0.05$), whereas no significant differences were observed between the latter two studies.

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

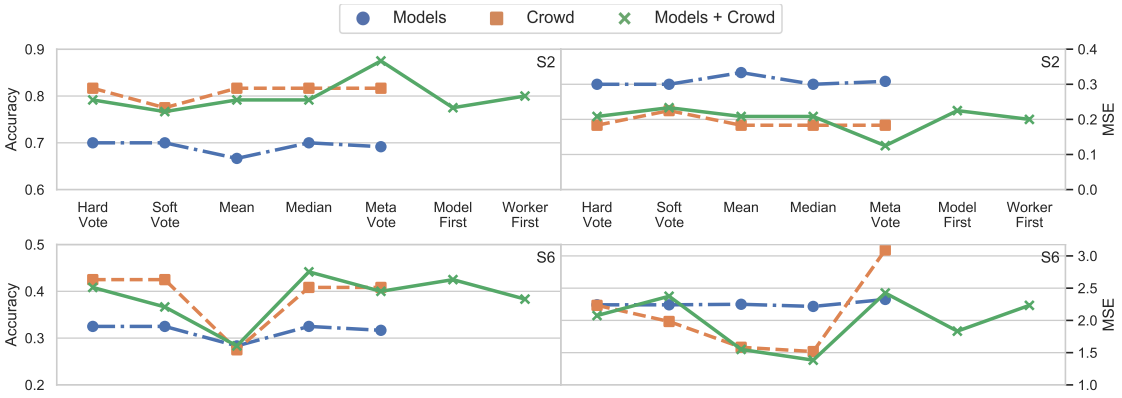

Figure 2: Accuracy and MSE for models, the crowd, and their combination. Top: S2; Bottom: S6. From Zeng et al. [2].

Discussion. The results of La Barbera et al. [1] detail a significant advancement in the application of crowdsourcing for fact-checking, achieved through an improved task design. The clear trend observed in the increasing truthfulness assessments corresponding with higher ground truth levels, particularly in our study compared to previous ones, underscores the effectiveness of the crowd. Such improvements have not only boosted the accuracy of assessments but also demonstrated the reliability of crowdsourcing as a practical and effective tool against misinformation.

The notable difference in La Barbera et al. [1]'s study compared to prior research, as statistically validated, suggests that task framing in crowdsourcing approaches can affect the outcomes of fact-checking tasks. Moreover, specific design choices can significantly enhance performance. It also highlighted that, while current results are promising, the complexity of fact-checking as a task are high. In conclusion, the presented study not only contributes to the empirical understanding and validation of crowdsourcing in fact-checking but also provides a foundational framework for future improvements in this area.

3. Combining Large Language Models and Crowdsourcing for Hybrid Human-AI Misinformation Detection

Objective and Methodology. The primary objective of Zeng et al. [2] was to enhance the effectiveness and efficiency of fact-checking by developing a hybrid system that integrates crowdsourced data with outputs from language and learning models (LLMs). The methodology involved processing the dataset of statements evaluated by crowdsourced workers in a previous study, with the addition of analysis from fine-tuned LLMs such as BERT [10], RoBERTa [11], and DeBERTa [12].

To integrate human and AI insights, several combination strategies were explored, including simple averaging, weighted averaging based on confidence scores, and more complex ensemble methods tailored to the specific characteristics of the data. The performance of the hybrid system was evaluated using standard metrics like accuracy, precision, recall, and F1-score,

supplemented by detailed error analysis to fine-tune the integration approach and ensure high-quality fact-checking.

Results. The integration of crowdsourced data with LLMs provided significant insights into the potential of hybrid systems for misinformation detection. The performance was quantitatively evaluated using two scales, S2 and S6, reflecting simpler and more complex judgment scales, respectively.

Figure 2 provide a detailed comparison of accuracy and Mean Squared Error (MSE) across individual models, crowdsourced data, and their integration. For the S2 scale, the results demonstrate consistent model performance with an accuracy around 0.7. In contrast, crowdsourced judgments maintained a higher accuracy of 0.816, surpassing individual model performances across all aggregation methods except soft-voting. For the S6 scale, the models achieved the best accuracy using hard-voting, soft-voting, and median aggregation methods (up to 0.441). The most effective hybrid combination on the S2 scale was achieved using the Meta Vote method, which delivered an outstanding accuracy of 0.875 and the lowest error rates (MSE 0.125). For the S6 scale, the median aggregation provided the best accuracy (0.441), indicating that simpler aggregation methods might be more effective for tasks with a more fine-grained scale.

Further analysis involved evaluating classification differences across various truthfulness levels using confusion matrices (not shown), which revealed that models were more consistent in classifying middle scale values like Mostly-False, Half-True, and Mostly-True but struggled with extreme categories such as Pants-On-Fire and True. Conversely, crowdsourced data showed a better capability to identify these extremes, indicating a deeper contextual understanding of the statements. The hybrid combinations provided more balanced judgments across the truthfulness spectrum of the S6 scale. While these combinations did not always outperform other methods in terms of raw accuracy or error rates, they offered a more subtle and robust classification, crucial for tasks requiring a sophisticated understanding of truthfulness.

Discussion. The integration of crowdsourced data with LLMs proposed by Zeng et al. [2] in their hybrid system for misinformation detection yields insightful results that underscore the complexity and potential of such approaches. This study not only demonstrates the viability of hybrid models in enhancing fact-checking accuracy but also reveals the complex interplay between different aggregation methods and the nature of the task.

Zeng et al. [2] suggest that hybrid models, which combine human judgment and machine intelligence, can significantly improve the reliability and accuracy of misinformation detection across various scales. The superior performance of the Meta Vote method in simpler judgment tasks (S2 scale) and the median method in more complex scenarios (S6 scale) highlights the importance of selecting appropriate aggregation strategies based on the task's specific requirements. This adaptability is crucial in real-world applications where the type and complexity of misinformation can vary greatly.

Acknowledgments. This research is partially supported by the European Union's NextGenerationEU PNRR M4.C2.1.1 – PRIN 2022 project "20227F2ZN3 MoT–The Measure of Truth: An Evaluation-Centered Machine-Human Hybrid Framework for Assessing Information Truthfulness" - 20227F2ZN3_001 – CUP G53D23002800006, and by the Strategic Plan of the University of Udine–Interdepartmental Project on Artificial Intelligence (2020-25).

References

[1] D. La Barbera, E. Maddalena, M. Soprano, K. Roitero, G. Demartini, D. Ceolin, D. Spina, S. Mizzaro, Crowdsourced Fact-checking: Does It Actually Work?, Information Processing & Management 61 (2024) 103792. doi:10.1016/j.ipm.2024.103792.

[2] X. Zeng, D. La Barbera, K. Roitero, A. Zubiaga, S. Mizzaro, Combining Large Language Models and Crowdsourcing for Hybrid Human-AI Misinformation Detection, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24, ACM, New York, NY, USA, 2024, p. 0.

[3] J. Allen, A. A. Arechar, G. Pennycook, D. G. Rand, Scaling Up Fact-Checking Using the Wisdom of Crowds, Science Advances 7 (2021) eabf4393. doi:10.1126/sciadv.abf4393.

[4] J. Allen, C. Martel, D. G. Rand, Birds of a Feather Don't Fact-Check Each Other: Partisanship and the Evaluation of News in Twitter's Birdwatch Crowdsourced Fact-Checking Program, in: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22, ACM, New York, NY, USA, 2022, pp. 1–19. doi:10.1145/3491102.3502040.

[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, ACM, New York, NY, USA, 2020, p. 439–448. doi:10.1145/3397271.3401112.

[6] 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.

[7] D. La Barbera, M. Soprano, K. Roitero, E. Maddalena, S. Mizzaro, Fact-Checking at Scale with Crowdsourcing: Experiments and Lessons Learned, in: Proceedings of the 13th Italian Information Retrieval Workshop, volume 3448, CEUR-WS.org, 2023, pp. 85–90. URL: https://ceur-ws.org/Vol-3448/paper-18.pdf.

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

[9] T. Draws, D. La Barbera, M. Soprano, K. Roitero, D. Ceolin, A. Checco, S. Mizzaro, The Effects of Crowd Worker Biases in Fact-Checking Tasks, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, ACM, Seoul, Republic of Korea, 2022, p. 2114–2124. doi:10.1145/3531146.3534629.

[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies, Volume 1 (Long and Short Papers), ACL, Minneapolis, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[11] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, 2019. doi:10.48550/arXiv.1907.11692. arXiv:1907.11692.

[12] P. He, X. Liu, J. Gao, W. Chen, DeBERTa: Decoding-enhanced BERT with Disentangled Attention, 2021. doi:10.48550/arXiv.2006.03654. arXiv:2006.03654.