Information Processing and Management 61 (2024) 103792

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Crowdsourced Fact-checking: Does It Actually Work?

David La Barbera [a,*], Eddy Maddalena [a], Michael Soprano [a], Kevin Roitero [a], Gianluca Demartini [b], Davide Ceolin [c], Damiano Spina [d], Stefano Mizzaro [a]

[a] University Of Udine, Via Delle Scienze 206, Udine, Italy
[b] The University of Queensland, St Lucia Queensland 4072, Brisbane, Australia
[c] Centrum Wiskunde & Informatica (CWI), Science Park 123, Amsterdam, The Netherlands
[d] RMIT University, 124 La Trobe St, Melbourne, Australia

ARTICLE INFO

Keywords:
Crowdsourcing
Misinformation
Truthfulness classification

ABSTRACT

There is an important ongoing effort aimed to tackle misinformation and to perform reliable fact-checking by employing human assessors at scale, with a crowdsourcing-based approach. Previous studies on the feasibility of employing crowdsourcing for the task of misinformation detection have provided inconsistent results: some of them seem to confirm the effectiveness of crowdsourcing for assessing the truthfulness of statements and claims, whereas others fail to reach an effectiveness level higher than automatic machine learning approaches, which are still unsatisfactory. In this paper, we aim at addressing such inconsistency and understand if truthfulness assessment can indeed be crowdsourced effectively. To do so, we build on top of previous studies; we select some of those reporting low effectiveness levels, we highlight their potential limitations, and we then reproduce their work attempting to improve their setup to address those limitations. We employ various approaches, data quality levels, and agreement measures to assess the reliability of crowd workers when assessing the truthfulness of (mis)information. Furthermore, we explore different worker features and compare the results obtained with different crowds. According to our findings, crowdsourcing can be used as an effective methodology to tackle misinformation at scale. When compared to previous studies, our results indicate that a significantly higher agreement between crowd workers and experts can be obtained by using a different, higher-quality, crowdsourcing platform and by improving the design of the crowdsourcing task. Also, we find differences concerning task and worker features and how workers provide truthfulness assessments.

1. Introduction

Given the ever-increasing amount of misinformation spread online daily, in recent years there has been an increased interest in designing reliable techniques for fact-checking. Fact-checking is a complex activity with a workflow that involves several

* Corresponding author.
Linkedin: david-la-barbera-8a1a646a, David La Barbera (D.L. Barbera), Linkedin: eddy-maddalena (E. Maddalena), Linkedin: michaelsoprano, Michael Soprano (M. Soprano), Linkedin: gianlucademartini, Gianluca Demartini (G. Demartini), Linkedin: davideceolin, Davide Ceolin (D. Ceolin), Linkedin: damianospina, Damiano Spina (D. Spina), Linkedin: stefano-mizzaro-1234082, Stefano Mizzaro (S. Mizzaro).
E-mail addresses: david.labarbera@uniud.it (D.L. Barbera), eddy.maddalena@uniud.it (E. Maddalena), michael.soprano@uniud.it (M. Soprano), kevin.roitero@uniud.it (K. Roitero), demartini@acm.org (G. Demartini), davide.ceolin@cwi.nl (D. Ceolin), damiano.spina@rmit.edu.au (D. Spina), stefano.mizzaro@uniud.it (S. Mizzaro).
URLs: http://www.eddymaddalena.net (E. Maddalena), http://www.michaelsoprano.com (M. Soprano), http://www.kevinroitero.com (K. Roitero), http://www.gianlucademartini.net (G. Demartini), http://www.cwi.nl/en/people/davide-ceolin/ (D. Ceolin), http://www.damianospina.com (D. Spina), https://users.dimi.uniud.it/~stefano.mizzaro/ (S. Mizzaro).

https://doi.org/10.1016/j.ipm.2024.103792
Received 21 July 2023; Received in revised form 8 April 2024; Accepted 19 May 2024
Available online 31 May 2024
0306-4573/© 2024 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

steps (Graves, 2017; Mena, 2019; Spina et al., 2023; Vlachos & Riedel, 2014): check-worthiness (i.e., understanding if a given statement is of interest to a possibly large audience), evidence retrieval (i.e., retrieving the evidence pro and/or against the statement), truthfulness or veracity assessment, discussion among the assessors to reach a consensus, and assignment and publication of the final verdict for the statement inspected.

In this paper, we focus on one of the steps involved in the fact-checking workflow: truthfulness assessment, which is an instance of a classification problem. The common approaches used for truthfulness assessment are either (i) expert assessment, i.e., journalists or fact-checkers who use a rigorous approach to assess the truthfulness of a piece of information, or (ii) automated classification systems, which employ Artificial Intelligence techniques to process a high amount of information in a limited time and with a low cost. However, these two approaches are rather the ends of a spectrum, and each of them presents some limitations (Demartini, Mizzaro, & Spina, 2020; La Barbera, Roitero & Mizzaro, 2022). Therefore, some researchers have proposed leveraging the "wisdom of the crowd" (Howe, 2006) for truthfulness assessment, and many proposals have been made in this direction (Allen, Arechar, Pennycook, & Rand, 2021; La Barbera, Roitero, Demartini, Mizzaro, & Spina, 2020; Saeed, Traub, Nicolas, Demartini, & Papotti, 2022; Sethi, 2017). The results of this line of research, however, appear inconclusive or even contradictory: while Allen et al. (2021) and Zhao and Naaman (2023) claim that crowdsourcing is effective, other researchers – although finding some signals – do not seem to reach an effectiveness level higher than that of completely automatic systems (Allen, Martel, & Rand, 2022; Roitero, Soprano, Fan et al., 2020; Saeed et al., 2022). The explanation for this lower effectiveness might be due to different aspects: the use of a different and possibly low-quality crowd, the selection of a different dataset with statements to be fact-checked that are difficult to assess, the choice of a too fine- or coarse-grained truthfulness scale, a worse task design, or a combination of the aforementioned.

Some researchers have proposed hybrid approaches, which combine automated classification and manual annotation (Demartini et al., 2020; La Barbera, Roitero, Mizzaro, 2022; Qu et al., 2022). These hybrid approaches aim to leverage the advantages of automated classification systems while still incorporating the expertise of human annotators. Once again, these attempts seem to be inconclusive (Qu et al., 2022): the reached effectiveness is still unsatisfactory and it is also still unclear if crowd-based manual classification can effectively complement automated classification.

Therefore, by analyzing the literature one is left wondering whether crowdsourcing can be a useful tool for assessing the truthfulness of statements. Our overall aim in this paper is to understand if effective fact-checking decisions can indeed be obtained from the crowd as claimed by some, but not by others. To this end, we select two studies by Draws et al. (2022) and Soprano et al. (2021) who obtained sub-optimal results from the crowd, and we reproduce and improve their approach: we not only aim at replicating their results, but we also highlight some limitations of these studies and try to improve them addressing such limitations. It can be said that we are therefore having a somehow optimistic attitude in this work: we conjecture that crowdsourcing can indeed be useful in truthfulness assessment, and we try to make it work in cases where it had not worked earlier. We remark that our aim is not to identify which component (i.e., task design, crowd composition, etc.) contributes the most to increase the overall effectiveness both for the crowd and for the model, leaving this question as future work.

The contributions of this work are two-fold. On the one hand, our study corroborates that crowdsourcing, under certain conditions, is a reliable mechanism for scaling truthfulness assessments in fact-checking. On the other hand, our improved experimental design, building upon previous work, informs how the crowdsourcing task needs to be designed and how the pool of workers needs to be defined for cost-effective fact-checking with the crowd.

This paper is organized as follows: in Section 2 we analyze existing literature on crowdsourcing and automated approaches for truthfulness classification. In Section 3 we explore aims and motivations for the current work, investigating inconsistencies and limitations of the current approaches before listing our research questions. In Section 4 we outline our experimental design, describing the performed crowdsourcing task and the effectiveness and agreement metrics used. In Section 5 we discuss our results. The paper concludes with Section 6, where we summarize our findings and discuss possible directions for future research.

2. Background

In Section 2.1 we discuss some notable work that used crowdsourcing-based approaches to detect misinformation, while Section 2.2 briefly summarizes the state of the art for Deep Learning and hybrid Human-In-The-Loop (HITL) approaches.

2.1. Crowdsourcing approaches

In the literature, many researchers addressed the misinformation detection problem using crowdsourcing approaches. La Barbera et al. (2020) found that worker bias has an effect when assessing a given statement. Later, this line of research has been further developed by Roitero, Soprano, Fan et al. (2020), who used several truthfulness scales to find similar agreement levels. Soprano et al. (2021) used the same setting but with a multidimensional truthfulness scale finding that the assessments provided by the workers over seven dimensions of truthfulness are reliable and that such dimensions measure different aspects of truthfulness. Then, Draws et al. (2022) performed an analysis on the work done by Soprano et al. (2021) to identify potential and systematic biases. They performed a novel crowdsourcing task over a new set of statements to test their hypotheses, finding that workers generally tend to overestimate truthfulness. They also found evidence of cognitive biases. Bias-aware aggregation methods has been proposed to mitigate cognitive biases such as confirmation bias (Gemalmaz & Yin, 2021). Recently, there has also been interest in enhancing label accuracy and reliability in crowdsourced data labeling by leveraging label correlations and neighboring instances' noisy labels (Jiang, Zhang, Tao, & Li, 2022; Li, Jiang, & Xue, 2023).

Sethi (2017) used the crowd to verify the validity of alternative facts using fundamental argumentation ideas in a graph-theoretic framework, while Allen et al. (2021) analyzed articles' headlines, finding that a politically balanced crowd correlates with experts' decisions. Zhao and Naaman (2023) compared the contributions of crowd workers and professional fact-checkers, finding that while the crowd is faster and performs similarly to experts in terms of accuracy, objectivity, and clarity, it often relies on existing professional knowledge to evaluate a piece of information. Saeed et al. (2022) compared crowd and experts using Twitter Birdwatch data, finding that they use different sources of information to fact-check, and confirming the good scalability and efficiency of the crowd as compared to fact-checkers. The effectiveness of Birdwatch as a fact-checking tool was investigated by Allen et al. (2022), finding that users are more likely to negatively evaluate tweets from those with whom they disagree politically.

2.2. Automatic and hybrid approaches

We focus on the evaluation of truthfulness classification performed by relying on human-based approaches only and in particular on crowdsourced ones. However, it is crucial to acknowledge the existence of fully- and semi-automated fact-checking techniques that have gained significant attention from researchers (Das, Liu, Kovatchev, & Lease, 2023; Demartini et al., 2020; Spina et al., 2023). Therefore, we briefly discuss and present the main results and findings from the existing literature that relies on those methodologies.

The use of automatic methods is particularly relevant, as proven by literature reviews (Ahmed, Aljabouh, Donepudi, & Choi, 2021; Collins, Hoang, Nguyen, & Hwang, 2021; Hu, Wei, Zhao, & Wu, 2022; Manzoor, Singla, & Nikita, 2019), and by public challenges that exist, such as the CLEF CheckThat! Lab (Nakov et al., 2022, 2021). In this respect, Aphiwongsophon and Chongstitvatana (2018), Hu et al. (2022), and Tanvir, Mahir, Akhter, and Huq (2019) have compared a variety of well-known machine learning algorithms to demonstrate their effectiveness on different datasets, finding that the accuracy of automatic methods can vary a lot. Others focused on developing more complex automatic approaches: Hakak et al. (2021) extracted significant features influencing misinformation classification to develop an ensemble model to achieve high accuracy over a test dataset, and La Barbera, Roitero, Mackenzie et al. (2022) took advantage of evidence retrieval to develop a composite deep learning pipeline for misinformation detection.

Another line of research consists in trying to combine crowdsourcing and automatic methods to develop Human-In-The-Loop (HITL) approaches to misinformation detection. As discussed by Ximenes and Ramalho (2022), some recent initiatives during COVID-19 can be seen as preliminary cases on how to deal with misinformation by means of interdisciplinary solutions possibly taking a Human-Centered Artificial Intelligence approach. Some practical HITL implementations exist. Dong, Sarker, and Qian (2022) integrated user feedback into the loop of learning and inference to recognize misinformation in a decentralized manner. Yang, Vega-Oliveros, Seibt, and Rocha (2021) proposed a new pipeline to group and summarize similar claims using both human assessors and automatic methods. Godel et al. (2021) investigated the effectiveness of a scalable model for real-time crowdsourced fact-checking, finding that machine learning models using the crowd data perform better than simple aggregation methods, but worse than the experts.

Other researchers investigate the connections between automatic and crowdsourcing-based methods; for instance, Qu et al. (2022) look at combining the two approaches using confidence scores to understand if indicating the precision of a performed classification is useful (i.e., if crowd workers and the used BERT transformer model can assess the correctness of a classification). Their results suggest that the two approaches make different mistakes, and that the confidence scores do not appear to be a reliable indicator of classification effectiveness.

3. Aims and motivations

In Section 3.1 we detail the inconsistencies in the effectiveness reported in the literature, while in Section 3.2 we outline the existing limitations. Then, in Section 3.3 we list the aims and research questions of this work.

3.1. Inconsistencies in reported effectiveness

When considering the literature detailed in Section 2, the inconsistencies among the published results become evident. The majority of the published works (Allen et al., 2021, 2022; Saeed et al., 2022; Zhao & Naaman, 2023) agree that the crowd can correctly identify misinformation, but with substantially different degrees of accuracy (we focus here on binary accuracy to provide comparable figures). For example, Draws et al. (2022), Roitero, Soprano, Fan et al. (2020), and Soprano et al. (2021) achieve a binary accuracy of 0.627, 0.571, and 0.580 respectively.

When considering only the automatic methods listed by Hu et al. (2022, Table 6) that are tested on data from the same dataset (i.e., PolitiFact), the reported crowd accuracies are comparable only with the lower end of the reported approaches, with the automatic methods reporting scores in the 0.531–0.904 range. On the other hand, other studies that employ the crowd have achieved higher accuracy. Both La Barbera et al. (2020) (accuracy of 0.841 on the same PolitiFact source of data) and Allen et al. (2021) (accuracy of 0.826) are directly comparable to the best automatic methods reported by Hu et al. (2022): 0.874 from Zhou, Wu, and Zafarani (2020) and 0.904 from Shu, Cui, Wang, Lee, and Liu (2019).

3.2. Existing limitations

When considering crowd-based effectiveness from previous works that achieved sub-optimal results (Draws et al., 2022; Roitero, Soprano, Fan et al., 2020; Soprano et al., 2021) we can draw multiple hypotheses to explain the poor performances. First, Draws et al.

Table 1
The seven dimensions of truthfulness, as defined by Soprano et al. (2021, Appendix A).

Dimension	Description
Correctness	The statement is expressed in an accurate way, as opposed to being incorrect and/or reporting mistaken information.
Neutrality	The statement is expressed in a neutral/objective way, as opposed to subjective/biased.
Comprehensibility	The statement is comprehensible/understandable/readable as opposed to difficult to understand.
Precision	The information provided in the statement is precise/specific, as opposed to vague.
Completeness	The information reported in the statement is complete as opposed to telling only a part of the story.
Speaker's Trustworthiness	The speaker is generally trustworthy/reliable as opposed to untrustworthy/unreliable/malicious.
Informativeness	The statement allows us to derive useful information as opposed to simply stating well known facts and/or tautologies.

(2022), Roitero, Soprano, Fan et al. (2020), and Soprano et al. (2021) use Amazon Mechanical Turk[1] as the crowdsourcing platform to recruit workers. Allen et al. (2022) and Saeed et al. (2022), on the other hand, rely on more specialized workers from Twitter Birdwatch platform, now Community Notes.[2] This could impact the effectiveness of the results and motivate the differences in terms of agreement with the experts detailed in Section 3.1, since the low quality of the data collected using Amazon Mechanical Turk has been shown in recent studies. For instance, Peer, Rothschild, Gordon, Evernden, and Damer (2022) show that other specialized platforms such as Prolific[3] allow to collect data of higher quality. More generally, Chmielewski and Kucker (2020) and Kennedy et al. (2020) show that the quality of crowdsourced results is platform-dependent.

La Barbera et al. (2020), on the other hand, achieved better results using workers from Amazon Mechanical Turk. This could be due to various reasons; for instance, the crowdsourcing tasks run by Draws et al. (2022) and Soprano et al. (2021) could have been perceived as too long, since the workers were required to evaluate 11 statements after answering multiple questionnaires. Furthermore, for each of the 11 statements workers had to provide assessments not only for the overall truthfulness, but also for their confidence in its evaluation and seven additional dimensions of truthfulness, thus leading to a total of $9 * 11 = 99$ different assessments. To better understand this issue, it is useful to clarify how Soprano et al. (2021) chose the seven dimensions used, at a later time, also by Draws et al. (2022).

Ceolin, Noordegraaf, and Aroyo (2016) proposed to decompose the notion of truthfulness into the aforementioned seven dimensions of truthfulness, reported in Table 1. They derived five of them (namely, Correctness, Completeness, Precision, Comprehensibility, and Neutrality) from the ISO 25012 Model (International Organization for Standardization, 2008), focused on information quality characteristics, while the remaining two capture other aspects of a considered piece of information: Speaker's Trustworthiness was introduced by Kahn, Strong, and Wang (2002), while the formulation of Informativeness has been directly proposed by Ceolin et al. Then, Ceolin et al. (2016) used them to perform user studies on web documents about vaccination debates while, later, Maddalena, Ceolin, and Mizzaro (2018) performed a crowdsourcing experiment to understand if the crowd can perform information quality assessments similarly to the experts. Finally, Draws et al. (2022) and Soprano et al. (2021) used the dimensions decomposition proposed by Ceolin et al. (2016) to perform their crowdsourcing-based truthfulness assessment tasks. Even though Soprano et al. find that the dimensions capture orthogonal aspects of truthfulness, thus being worth crowdsourcing, their combination did not lead to particular improvements while trying to predict the ground truth of a given statement. This suggests that while each dimension provides unique insights, their collective impact does not necessarily enhance predictive performance of the overall truthfulness. One possible explanation is the presence of redundancy or inter-correlation among some of those dimensions, which may not add novel information when combined. Therefore, future crowdsourcing tasks could potentially rely on a lower number of dimensions of truthfulness; by focusing on a carefully chosen subset, the aim is to reduce complexity and potential redundancy while maintaining or even improving the effectiveness of truthfulness assessments.

As for other factors, the task layout could have had an impact on the workers' effectiveness since a lot of information was displayed concurrently. It might also be the case that the definition of some of the dimensions could have had some influence; indeed, some workers perceive the descriptions reported in Table 1 as unclear or confusing. Lastly, the statements used to collect truthfulness assessments were rather outdated (spanning from 2007 to 2015). As found in the longitudinal study by Roitero, Soprano et al. (2021), the time-span between the statements and the truthfulness judgments can have a major effect on the workers' effectiveness.

3.3. Research questions

We can conclude that some of the previous studies obtained effectiveness levels that could be below optimal, in light of the inconsistencies found (Section 3.1) and the discussed limitations (Section 3.2). Thus, in this paper, we attempt to reproduce those results with more appropriate task design and experimental settings.

We propose a novel setting that aims to improve those of the selected previous studies by Draws et al. (2022) and Soprano et al. (2021) (referred to in the following as Soprano et al. and Draws et al., respectively) and we analyze whether the proposed changes allow obtaining assessments that lead to better results in terms of the agreement among workers and with experts. In particular, we study whether a different and higher-quality task design and a different population of crowd workers address the limitations

[1] https://www.mturk.com/.
[2] https://github.com/twitter/communitynotes.
[3] https://www.prolific.co/.

of Soprano et al. and Draws et al. studies. Then, we recruit a crowd of workers from the, allegedly high quality, Prolific platform and we analyze multiple crowdsourcing-based studies (La Barbera et al., 2020; Roitero, Soprano, Fan et al., 2020; Roitero, Soprano, Portelli et al., 2020; Soprano et al., 2021) to develop an improved task design. Lastly, we focus on a particular aspect of the selected previous studies, i.e., the seven dimensions of truthfulness, and we study whether there is any change in how the workers provide their assessments. The research questions can be summarized as follows:

RQ1 What level of effectiveness can be achieved by relying on crowdsourcing for truthfulness assessment? Can we improve the results of the two selected previous studies, after addressing the shortcomings in their task design and worker population?
RQ2 Does the new study design impact the overall effectiveness of the crowd in terms of both task components and worker features?
RQ3 Do the workers provide assessments for the seven dimensions of truthfulness consistently between the previous studies selected and our updated setting? Can the workers' self reported confidence be considered as a reliable measure in the new study design?

4. Methodology

This section is structured as follows: in Section 4.1 we describe the data used, in Section 4.2 we describe the motivations and changes made to the task originally designed by Soprano et al. (2021), while in Section 4.3 we describe how we performed scale transformations to allow direct comparisons of assessments collected on different scales. In Section 4.4 we describe how we define, measure, and compare task effectiveness. Throughout this section we highlight and motivate the differences from the previous studies. We release the whole dataset at: https://osf.io/j7as8/.

4.1. Dataset

We use statements from the PolitiFact website[4] for the new crowdsourcing task, as done by Draws et al. (2022) and Soprano et al. (2021). This organization has been fact-checking statements made by US politicians, political organizations, other public figures, and posted on social media since 2007. To date, the website has recorded more than 24,000 fact-checks and continues to be updated regularly. The statements are labeled by expert judges on a six-level truthfulness scale: Pants-On-Fire, False, Mostly-False, Half-True, Mostly-True, and True. Following the approach of PolitiFact, we adopted the same six-level truthfulness scale to ensure consistency and comparability in our analysis and specifically to allow for better and more direct comparisons between assessments from expert fact checkers and workers.

For our crowdsourcing task, we sample a set of 120 statements, 20 for each of the 6 ground truth levels, different from those used by Soprano et al. and Draws et al. for our crowdsourcing task. Particularly, we use recent statements from 2022 to minimize the issue of crowd workers already possessing knowledge of the information they need to verify – a phenomenon previously observed by Zhao and Naaman (2023); more generally, other researchers use up-to-date statements within their experiments (Draws et al., 2022; La Barbera et al., 2020; Roitero, Soprano, Fan et al., 2020; Soprano et al., 2021). Furthermore, PolitiFact provides a verdict and a collection of links to the sources (mostly articles) used by expert fact-checkers to assess truthfulness of each statement on the six-level scale. However, we discovered that various links provided in the website's fact-checking pages are either broken or lead to non-existent resources. This issue is more common in older statements, possibly because the linked resources have been removed or transferred. This is not negligible and further convinced us to rely on more recent statements with verifiable fact-checks.

Still differently from Soprano et al. and Draws et al., we decided to not include statements from the RMIT ABC Fact Check[5] dataset. The main motivation for this decision is that by removing 3 statements (one for each of the three RMIT ABC Fact Check ground truth levels) we shorten the task, thus reducing the required workload.

4.2. The crowdsourcing task

We update the task design proposed by Soprano et al. (2021, Section 4.3) and then reproduced by Draws et al. (2022, Section 4.1) to address the limitations of the previous studies described in Section 3.2. We relied on the Crowd_Frame (Soprano, Roitero, Bombassei De Bona, & Mizzaro, 2022) framework and used the original configuration files used by Soprano et al. and Draws et al. for their tasks, provided upon inquiry. We study the previous crowdsourcing tasks (Draws et al., 2022; La Barbera et al., 2020; Roitero, Soprano, Fan et al., 2020; Soprano et al., 2021) by considering their workers' feedback and by running an ad-hoc user study.

Soprano et al. task design featured four questionnaires. The first one was used to collect demographic information about the workers, while the remaining three were Cognitive Reflection Tests (CRT) (Frederick, 2005). Draws et al. expanded on Soprano et al. design by incorporating two new questionnaires after the demographic one and before the three CRTs, resulting in a total of six questionnaires. These two new questionnaires, proposed by De Vries, Piotrowski, and de Vreese (2023), measure two variables, namely the digital skill level and knowledge of the workers, which could be considered as proxies for their effectiveness. We further

[4] https://www.politifact.com/.
[5] https://apo.org.au/collection/302996/rmit-abc-fact-check.

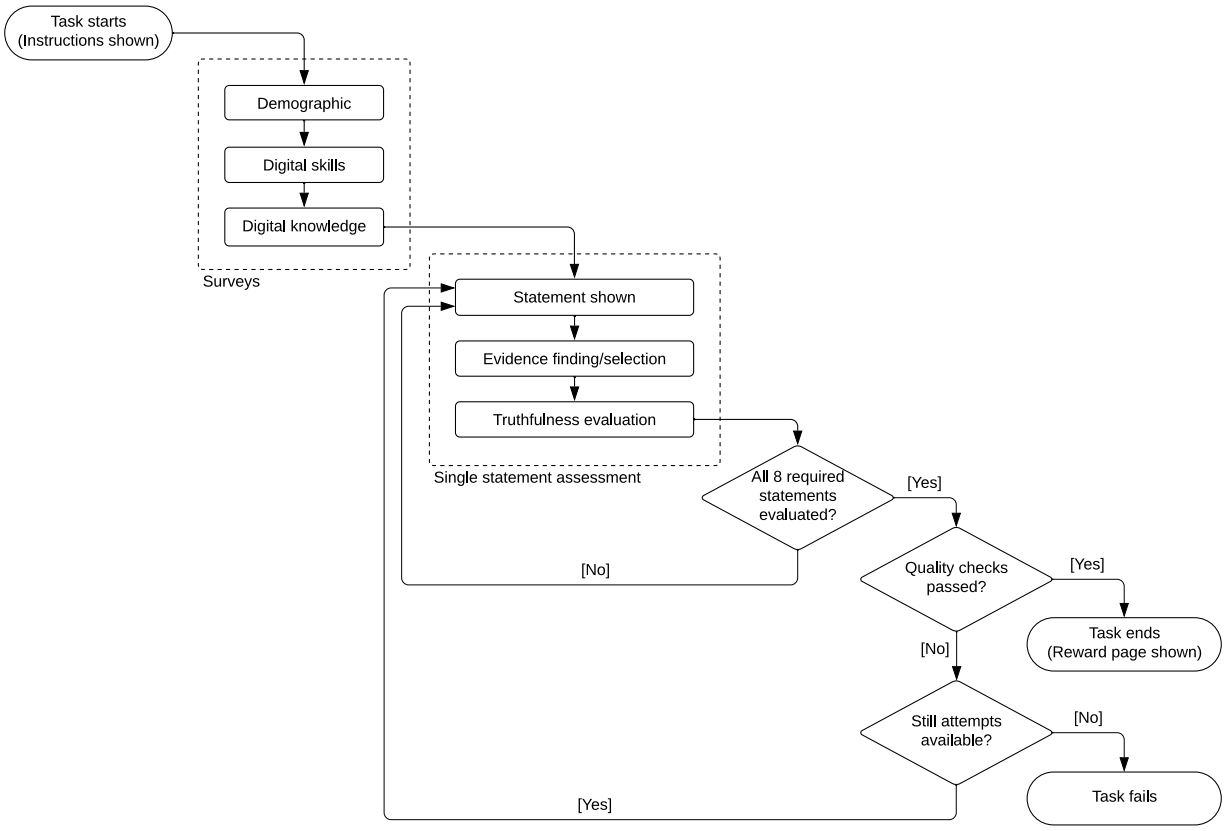

Fig. 1. Flowchart showing the steps required to carry out the task.

updated Draws et al. design by removing the three CRTs, thus lowering the number of questionnaires to three. The rationale behind this choice is that Draws et al. shown no statistical significant correlation between the workers' cognitive abilities measured using CRTs and their effectiveness in performing the fact-checking task.

In light of the considerations about the seven dimensions of truthfulness described in Section 3.2, we renamed some of the previously used terms since they were perceived as unclear and confusing by the workers. Specifically, we rename Neutrality to Unbiasedness, Correctness to Accuracy, and Overall Truthfulness to simply Truthfulness.

We have also replaced the five-level Likert scale used to collect assessments for the truthfulness dimension in the previous tasks. Instead, we employ the six-level scale used by PolitiFact to improve the assessments' consistency and obtain assessments that are directly comparable with the ground truth. This scale has also been used by La Barbera et al. (2020), Roitero, Soprano, Fan et al. (2020) and Roitero, Soprano, Portelli et al. (2020).

Another modification involves how the assessment of each statement is performed by the workers. Soprano et al. and Draws et al. presented together to each worker the multidimensional assessment scale, the customized search engine, and the statement text and attributes. We deduced that the workers were overwhelmed by the amount of information shown, and that this might have influenced their performances. Thus, we updated the task assessment interface design, which is now split into five sequential different steps. First, the workers must read the statement, then they must search for evidence using the customized search engine. Once an URL is selected, the workers need to provide the assessment for each dimension of truthfulness. Lastly, they have to express the perceived confidence level of their own assessment. We also rewrote the instructions to be more concise and favor clarity, and implemented many minor graphical tweaks and edits to improve the overall style of the interface. Fig. 1 shows the flowchart that outlines the steps of the task after the modifications described above. The process is split into two main phases: initially, the worker completes three general surveys. Subsequently, the worker assesses the truthfulness of the eight required statements. By convenience, the arrows of the second block represent the flow of a linear task execution. However, workers were allowed to revisit and revise earlier statements before submitting their final submission.

Soprano et al. and Draws et al. set a HIT reward of 2 USD including the set of 11 assessments, computed on the basis of the time needed to finish the task and the U.S. Minimum Salary Wage of 7.25 USD per hour. Overall, 180 statements in total were employed. Each statement has been evaluated by 10 distinct crowd workers. Thus, 200 HITs were published and 2200 assessments were collected. They published their tasks using the Amazon Mechanical Turk crowdsourcing platform. Since we deploy the updated version of the task on the Prolific platform we must adapt workers' payment to a different payment model: the task requester

indicates an estimate of the time needed to complete the task and proposes the payment amount accordingly. We thus set a completion time of 20 min and a payment of 3 GBP (i.e., around 3.60 USD),[6] for a 9 GBP (around 10.80 USD) hourly rate. The completion time was estimated by performing in-house timed tests and by looking at the time spent by the worker recruited by Soprano et al. and Draws et al.. Overall, the amount of statements that each worker had to judge was lower than Soprano et al. and Draws et al. tasks, since we used PolitiFact statements only (see Section 4.1). We published 200 HITs, where each HIT involved assessing six statements plus two additional ones used as gold questions. We recruited only US-based workers and we collected a total of 1600 assessments.

4.3. Scales transformation

We perform a preprocessing step on the assessments collected by Draws et al. (2022) and Soprano et al. (2021) to allow direct comparison with those collected with our new crowdsourcing task. Indeed, while the adoption of the 6-level truthfulness scale, described in Section 4.2, simplifies the comparison with the PolitiFact ground truth, it complicates the one with the previous studies since they used a 5-level scale with discrete values in the $[-2, +2]$ range.

Thus, we define a linear scale transformation to compute effectiveness measures. For each assessment performed using the 5-level scale, we map the values collected in the $[0, 5]$ range by multiplying each value by 1.25. Since we employ a linear scale transformation, the bias introduced by this operation is minimal (Han, Roitero, Maddalena, Mizzaro & and Demartini, 2019). We detail each metric used as an effectiveness measure in the following.

4.4. Effectiveness and agreement measures

In our crowdsourced fact-checking evaluation, we carefully selected a set of effectiveness measures, each one serving a unique purpose in the assessment process and aligning with those commonly employed in the literature. These measures encompass both internal and external agreement metrics used in previous studies (Draws et al., 2022; La Barbera et al., 2020; Roitero, Soprano, Fan et al., 2020; Soprano et al., 2021). This selection ensures that our methodology is consistent with established practices and provides a comprehensive evaluation framework.

Furthermore, it is essential to acknowledge the complexities inherent in the nature of ordinal categorical scales used in such assessments. As discussed by Roitero, Soprano, Fan et al. (2020, Section 3.3), the truthfulness scales are not simple nominal scales with independent categories; they are ordinal, indicating a hierarchy in the error magnitude for misclassifications. For example, the error in misclassifying a True statement as Mostly-True is smaller than classifying it as Half-True. This complexity justifies why, although some proposals exist (Amigo, Gonzalo, Mizzaro, & Carrillo-de Albornoz, 2020), there is no single universally accepted measure to capture the error magnitudes in such scales (Roitero, Soprano, Fan et al., 2020). Therefore, our approach incorporates a mix of classification measures, such as accuracy (which does not account for the magnitude of an error), and regression measures, such as mean squared error (that assume an arbitrary equidistant error between truthfulness levels).

In more detail, we align with the work of Roitero, Soprano, Fan et al. (2020) and report five measures to reflect a necessary compromise in dealing with the subtleties of ordinal scales. The accuracy metric, detailed in Section 4.4.1, is essential for comparing the crowd classifications with expert judgments, offering a direct measure of the crowd's ability to match expert evaluations. In Section 4.4.2, we discuss the worker's external agreement, where the Krippendorff's α reliability coefficient is employed to measure the congruence between workers' and experts' assessments, providing insight into the external validity of the crowdsourced data. In Sections 4.4.3 and 4.4.4, we make use of pairwise agreement and Mean Squared Error (MSE), respectively. Pairwise agreement offers a perspective on the consistency of crowd judgments against expert labels, highlighting the uniformity in worker evaluations. MSE, on the other hand, quantifies the variance between workers' assessments and the ground truth, providing a numerical measure of the accuracy of their judgments. Lastly, Section 4.4.5 covers the worker's internal agreement using Krippendorff's α. This measure focuses on the intrinsic agreement within the crowd, i.e, the coherence among workers' assessments independently of the expert judgments.

4.4.1. Accuracy

The first effectiveness measure employed is the accuracy of the classifications performed by the crowd when compared with the experts' ground truth. Given that only the new task is performed using a 6-level scale, and that the scale transformation detailed in Section 4.3 does not map the values to the 6-level labels (i.e., we still have 5 distinct values, just with different values), we compute accuracy over a coarse-grained version of scales, in this case, the 2-level scale. To compute the overall accuracy of the workers, for each task we compute the aggregated score (using, as aggregation functions, mean, mode, and median) between the 10 workers' assessments for each statement. Then, we map them to a 2-level scale, using the values 0 and 2.5 as cutting points for the 5-level and for the 6-level scale respectively. The aggregated values less than or equal to the reported cutting points will be mapped to False, and the others True.

We also binarize (i.e., collapse into two levels) the ground truth as done by Roitero, Soprano, Fan et al. (2020): we map the first three levels of the scale (Pants-On-Fire, False, Mostly-False) as False, and the last three levels (Half-True, Mostly-True, and True) to True. In this case, we assume that the labels from both the starting scales are equally distributed, thus it is sufficient to use the mid value as the cutting point.

[6] Amazon Mechanical Turk uses USD as currency for payments, while Prolific uses GPB.

4.4.2. Worker's external agreement – Agreement with the experts

As a second effectiveness measure, we use Krippendorff's α reliability coefficient (Krippendorff, 2008) to measure the agreement between workers and experts under the assumption that if the crowd agrees with the expert labels the effectiveness of the assessments is high.

Since Krippendorff's α is dependent on the scale used to collect assessments (Checco, Roitero, Maddalena, Mizzaro, & Demartini, 2017), the scale transformation detailed in Section 4.3 is a prerequisite for comparing directly the collected assessments of each task against the external agreement. We thus compute Krippendorff's α per task in two different ways: either by considering each worker's raw assessment against the experts' ground truth, and by computing how much each worker agrees with the experts.

4.4.3. Pairwise agreement

As a third effectiveness measure, we use the pairwise agreement, computed as the portion of cases in which the annotators agree, divided by the total number of observations (Maddalena, Roitero, Demartini, & Mizzaro, 2017). Given a pair of statements judged differently by the experts, we analyze each worker's assessments in relation to the ground truth as provided by the experts, looking for consistency.

Thus, the pairwise agreement score as defined by Maddalena et al. (2017) is high when the crowd judges truthfulness consistently to the experts. We compute the pairwise agreement for each worker over the judged 6 statements judged and compare their scores with the ground truth. We recall that for each task, a worker evaluates a statement for each of the 6-level of the ground truths. Workers' assessments may either align with expert evaluations or differ from them. In cases where a worker evaluates two statements similarly and cannot discern the difference in labels, a "tie" occurs. We consider both inclusive and exclusive methods of calculating pairwise agreement, with ties included or excluded in the count of agreement pairs, respectively.

It should be noted that the two variants of the pairwise agreement discussed above differ in the degree to which they weigh the worker effectiveness. Specifically, the variant that excludes ties penalizes workers who are unable to discern a difference in the assessments, whereas the other does not capture this distinction. As there is no evident preference for a given variant for our analysis, we report both values.

4.4.4. Mean Squared Error (MSE)

As a fourth effectiveness measure, we compute the Mean Squared Error (MSE), which quantifies the average of the squared errors between the truthfulness assessments provided by the workers and the ground truth. We measure MSE by computing the average error:

(i) per task, computing the average error of the workers;
(ii) per ground truth level, considering for each of the statements with given ground truth; and
(iii) per statement, by considering the assessments given by the 10 workers for each of the statements.

Note that MSE relies on the assumption that assessments are equally spaced (i.e., the difference between Pants-On-Fire and False is half the difference between Mostly-False and Half-True), a reasonable assumption coherent with previous works (Draws et al., 2022; La Barbera et al., 2020; Roitero, Soprano, Fan et al., 2020; Soprano et al., 2021).

4.4.5. Worker's internal agreement – Agreement within workers

We now inspect an additional measure which gives us a better understanding of our results. Specifically, we compute the Krippendorff's α agreement score by exclusively considering the assessments provided by workers while disregarding those of experts; that is, the internal agreement among workers.

This measure is slightly different from the ones defined above, as the observed agreement between workers can be high even when they agree on a wrong assessment. Nevertheless, internal agreement is a valuable intrinsic measure in crowdsourcing tasks. Roitero, Soprano, Fan et al. (2020) and Soprano et al. (2021) provide a detailed discussion of the theoretical motivations.

5. Results

In Section 5.1 we report some descriptive statistics for the performed task. In Section 5.2 we compare the three considered tasks using our effectiveness measures (i.e., we focus on RQ1). In Section 5.3 we investigate the impact of some worker's and task features within our updated design (RQ2). Finally, in Section 5.4 we investigate how workers provide assessments for the seven truthfulness dimensions and their confidence (RQ3).

5.1. Workers' demographics

Henceforth, we will refer to our updated task design, detailed in Section 4.2, as Our Study for simplicity. For a direct comparison with previous task designs, see Draws et al. (2022, Section 5.1) and Soprano et al. (2021, Section 4.4). We calculate the abandonment rate using the definition by Han, Roitero, Gadiraju et al. (2019). In total, 200 out of 273 workers (73.26%) successfully completed the task. Among the remaining 73 workers, 55 (20.15%) voluntarily abandoned the task before completing a try, while 18 (6.59%) failed, leading to the termination of the task due to quality check failures. We observe lower abandonment and failure rates among workers compared to the studies conducted by Soprano et al. and Draws et al.. We recall that Our Study involves workers recruited from

Table 2
Summary of the answers provided by the workers to the demographic questionnaire.

Question	Data
Age range	0.5% 0–18, 10.5% 19–25, 29.5% 26–35, 36% 36–50, 23% 51–80, 0.5% 81+
Education level	2% high school incomplete or less, 12.5% high school graduate, 22% some college, 39% four year college degree/bachelor, 4% some postgraduate or professional schooling, 20.5% postgraduate or professional degree
Family income before taxes	2.5% less than $10k, 4% $10k to less than $20k, 7.5% $20k to less than $30k, 7% $30k to less than $40k, 7% $40k to less than $50k, 24% $50k to less than $75k, 16.5% $75k to less than $100k, 19.5% $100k to less than $150k, 12% $150k or more.
Political views	3.5% very conservative, 19.5% conservative, 22.5% moderate, 34% liberal, 18.5% very liberal, 2% choose not to answer
Consideration in today's politics	22% republican, 47.5% democrat, 28% independent, 2.5% something else
Opinion on U.S. southern border	29.5% agree, 53% disagree, 17.5% no opinion either way
Opinion on U.S. climate change policies	84% agree, 8.5% disagree, 7.5% no opinion either way

Table 3
Summary of the demographic data provided by the Prolific platform.

Question	Data
Gender	60% male, 39.5% female, 0.5% consent revoked
Ethnicity	73.5% white, 15% black, 4.5% asian, 3.5% mixed, 2% other, 1% data expired, 0.5% consent revoked
Country of residence	100% United States
Country of birth	92.5% United States, 1% Nigeria, 0.5% Germany, 0.5% Italy, 0.5% United Kingdom, 0.5% Kenya, 0.5% Canada, 0.5% Ghana, 0.5% Russian Federation, 0.5% Ireland, 0.5% consent revoked, 2% data expired
Nationality	97.5% United States, 1% Brasil, 0.5% Philippines, 0.5% Sweden, 0.5% consent revoked
Native language	100% English
Language fluency	98% English, 0.5% Japanese, 0.5% Italian, 0.5% Spanish, 0.5% consent revoked
Student status	76% no, 16% yes, 0.5% consent revoked, 7.5% data expired
Employment status	53.5% full-time, 8.5% unemployed (and job seeking), 7% not in paid work (e.g., homemaker, retired, or disabled) 6% part-time, 3% other, 0.5% consent revoked, 21.5% data expired

the Prolific platform, whereas Soprano et al. and Draws et al. rely on Amazon Mechanical Turk. Here, we report the demographic statistics based on the 200 workers who successfully completed Our Study.

We start by analyzing the answers provided to the demographic questionnaire,[7] summarized in Table 2. Such answers are explicitly provided by the workers participating in this study; therefore, there is no guarantee that they are entirely genuine and fully reflective of objective reality.

The majority of workers (72/200, 36%) are between 36 and 50 years old. In terms of education, 39% (78/200) obtained a four-year college/bachelor's degree or higher. Concerning total family income before taxes in the previous year, 24% of workers (48/200) earned between $50k and $75k. Regarding political views, 34% (68/200) identify as liberal, and in today's politics, 47.5% (95/200) identify as Democrats. A majority of workers (106/200, 53%) disagreed with building a wall on the U.S. southern border. Finally, the vast majority of workers (168/200, 84%) believed that the government should increase environmental regulations to prevent climate change.

Additionally, the Prolific platform provides task requesters with additional demographic information about each worker, summarized in Table 3. However, certain data is either marked as expired or obfuscated due to workers revoking consent. The majority of workers (120/200, 60%) identify as male. In terms of ethnicity, the majority (147/200, 73.5%) are white. All workers reside in the USA, with the vast majority (185/200, 92.5%) born there, and almost all (195/200, 97.5%) hold full citizenship. All workers have English as their native language, with a small percentage (3/200, 1.5%) also being fluent in other languages. The majority (156/200, 76%) are not students. Approximately half of the workers (107/200, 53%) are employed full-time.

Overall, our sample is well-balanced in terms of demographic characteristics, aligning with the findings from previous studies (Draws et al., 2022; La Barbera et al., 2020; Roitero, Soprano, Fan et al., 2020; Roitero, Soprano et al., 2021; Roitero, Soprano, Portelli et al., 2020; Soprano et al., 2021), with only a few exceptions.

5.2. RQ1: *Effectiveness*

In this section, we report each effectiveness measure computed as detailed in Section 4.4 and summarized in Table 4. Section 5.2.1 describes the accuracy values, while Section 5.2.2 describes the external agreement with experts. Then, Section 5.2.3 discusses the pairwise agreement and Section 5.2.4 addresses the mean squared error. Lastly, Section 5.3.2 analyzes the internal agreement among workers.

[7] The whole demographic questionnaire is available in the configuration files published in the repository (see Section 4).

Table 4
The agreement and effectiveness measures computed for the considered tasks.

Measure	Soprano et al.	Draws et al.	Our Study
2-level accuracy	0.571	0.580	0.817
External agreement α	0.204	0.065	0.681
Pairwise external agreement (ties)	0.662	0.652	0.829
Pairwise external agreement (no ties)	0.620	0.619	0.812
MSE	2.497	2.691	1.484
Internal agreement α	0.089	0.040	0.219

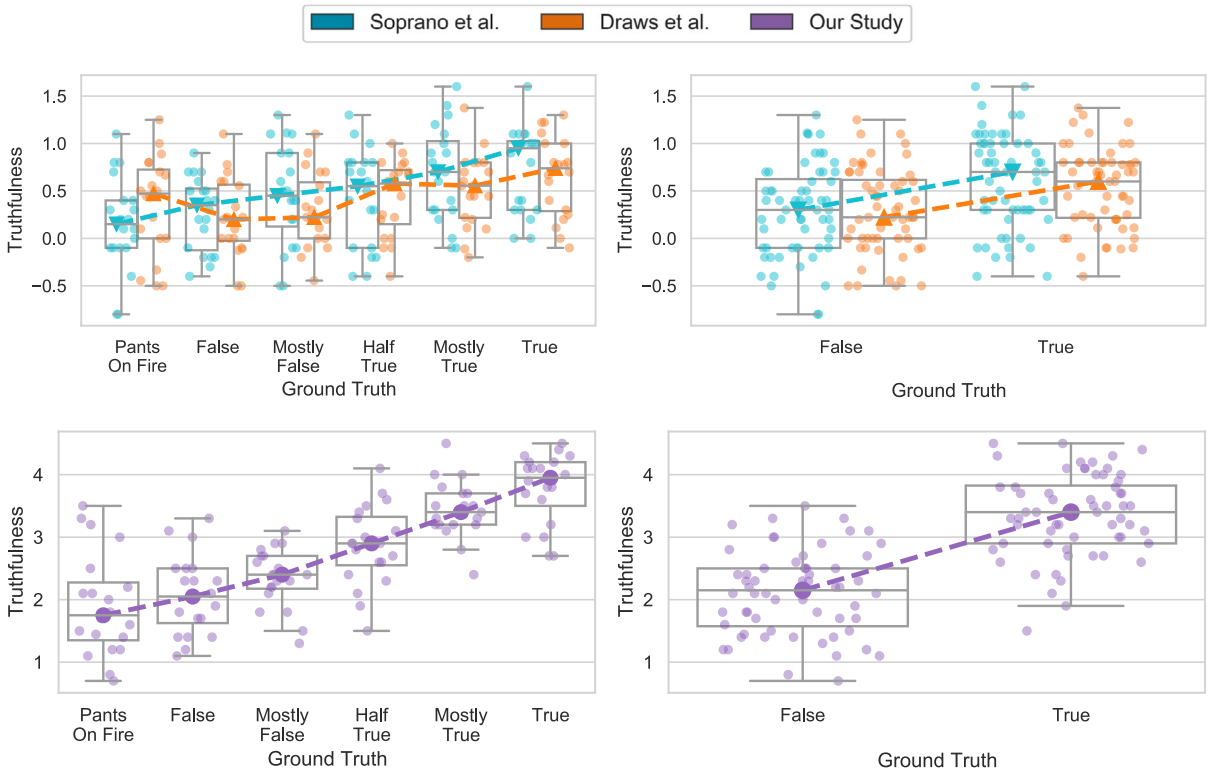

Fig. 2. Worker external agreement for the three considered studies. Top: Draws et al. (2022) and Soprano et al. (2021), Bottom: Our Study. Left: original ground truth 6-level scale; right: binarized 2-level (for interpretation of the references to color in this figure legend, please refer to the web version of this article).

5.2.1. Accuracy

Soprano et al. and Draws et al. achieved similar binary accuracy levels, respectively 0.571 and 0.580. In Our Study, we clearly outperform these scores, obtaining an accuracy of 0.817 on the binarized data. Such a value is a first confirmation of the good performance achieved by the crowd recruited for Our Study.

Thus, the crowd can judge misinformation with a higher accuracy than the previous ones, while being consistent with other crowdsourcing-based studies as those by Allen et al. (2021) and La Barbera et al. (2020). Furthermore, the obtained accuracy value is comparable with those obtained with state-of-the-art automatic methods for truthfulness assessment (Hu et al., 2022) (see also Section 3.1).

5.2.2. Worker's external agreement – Agreement with the experts

We compare workers' agreement for each study, using the assessments collected according to each original truthfulness scale (i.e., 5-level for Soprano et al. and Draws et al., 6-level for Our Study). Fig. 2 shows the results; the plots on the top row reproduce the results obtained by Soprano et al. and Draws et al., while those on the bottom row refer to Our Study. For each plot, the x-axis shows each ground truth level, while the y-axis shows the assessments aggregated using the mean. Each small dot refers to a single statement, whereas the larger circles and triangles highlight the median values of the statements in each ground truth level, also represented by the internal line of the box. The left column plots show the original ground truth levels in ascending order of

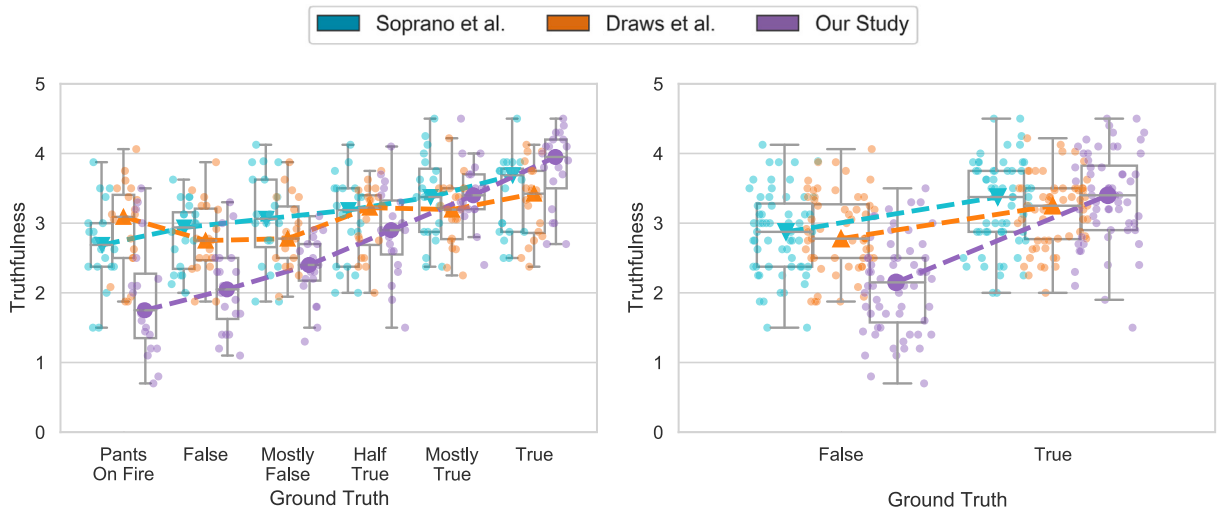

Fig. 3. Worker external agreement for the PolitiFact statements over the three studies with the 5-level scale of the original studies mapped into a 6-level scale. Left: original ground truth 6-level scale; right: binarized 2-level. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

truthfulness from left to right, while the right one shows the binarized levels after the assessment aggregation step.[8] The results obtained for Our Study have a higher effectiveness compared to those of Soprano et al. and Draws et al.. Let us first consider the left column. A clear trend can be seen for Our Study: the median aggregated truthfulness values increase steadily when moving from lower to higher ground truth values. The trend is less evident in Soprano et al. study. When considering the study of Draws et al., the median values also decreases, as happens for example when moving from Pants-On-Fire to False. This behavior can be seen more clearly by considering the binarized ground truth levels reported in the right column plots.

We now use the scale transformation as detailed in Section 4.3 to directly compare each study. Fig. 3 shows the results. The left column plots show, similarly to Fig. 2, the original ground truth levels, while the right column plots the binarized ones. As it can be seen, the higher effectiveness of the results obtained for Our Study becomes more evident. The boxplots for Our Study report a trend that increases more than the others, for both the 6-level ground truth levels and for the binarized ones, as shown by the dashed lines linking each boxplot median value to the subsequent ones. In particular, the boxplot for the Mostly-False ground truth value shows a lower median value compared to the one of Pants-On-Fire and False, meaning that it is perceived as more false by the workers while providing their assessments, as hinted by looking at Fig. 3.

More remarks can be drawn by inspecting Fig. 3. When comparing the median values of Our Study with those of Soprano et al. and Draws et al., one can see that the differences are higher for false statements than for true ones. Furthermore, the workers recruited for Our Study tend to assign lower truthfulness levels to false statements than those of Soprano et al. and Draws et al. studies. This is less evident when considering the boxplots for ground truth levels closer to True, as the median values show lower differences compared to those observed for ground truth levels closer to False. This can be further seen when considering the ground truth levels that are adjacent in the original scale. For instance, the difference between the Mostly-True and True levels median values is 0.55. When considering the Pants-On-Fire and False levels, it is of 0.33. This may thus suggest that the truthfulness levels are not coherently understood, or maybe even that fine-grained scales should be avoided. We leave a detailed analysis for future work.

We then check the above results for statistical significance. To this aim, we apply the Kruskal–Wallis Test (Kruskal & Allen Wallis, 1952) and Dunn's Test (Jean Dunn, 1964) to evaluate the statistical significance of differences across three experimental conditions: Our Study, Soprano et al., and Draws et al.. The Kruskal–Wallis Test, a non-parametric method suitable for comparing more than two independent samples not following a normal distribution, indicates a statistical significant difference among these three experiments at the 0.01 significance level. This finding justifies the testing for further pairwise comparisons between the three studies, for which we employ the Dunn's Test with Bonferroni correction to adjust for multiple comparisons (Bland & Altman, 1995; Sedgwick, 2012). The results from Dunn's Test revealed that Our Study is statistically significantly different from Soprano et al. at the 0.01 level, and from Draws et al. at the 0.05 level. However, no significant difference is found between Draws et al. and Soprano et al. ($p > 0.05$). These findings suggest significant differences of the experimental conditions on the outcomes, highlighting the effectiveness of Our Study in differing significantly from both Soprano et al. and Draws et al., while the latter two appear not to be significantly different from each other.

[8] We use the following color code in the remaining figures of this work: light blue for Soprano et al. (2021), orange for Draws et al. (2022), and purple for Our Study.

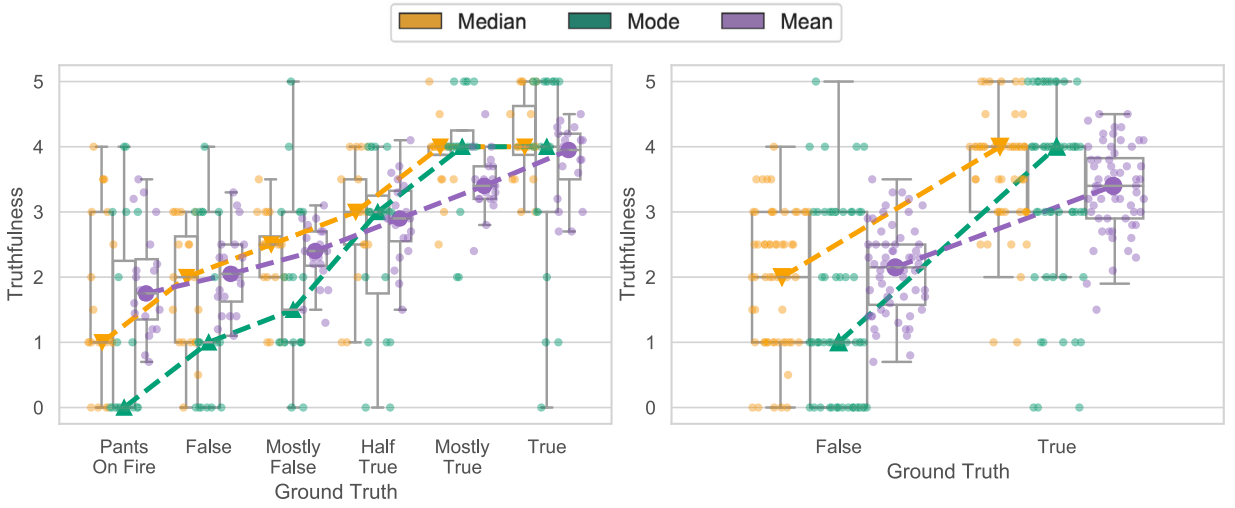

Fig. 4. Worker external agreement for the PolitiFact statements over the three aggregation metrics considered for Our Study: in purple for the mean, yellow for the median, and green for the mode. Left: original ground truth 6-level scale; right: binarized 2-level. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In the previous analysis, we have presented our results considering the mean as aggregation function (i.e., the function used to aggregate the truthfulness levels given by the workers on each statement). To make our discussion more comprehensive, we perform a comparative analysis of aggregation metrics for Our Study considering also the median and the mode (i.e., majority voting) as aggregation functions. Results are summarized in Fig. 4. Similarly to the previous Figs. 2 and 3, on the left part of the plot we show the original ground truth levels, while on the right column the binarized ones. In purple we show the results when the mean is used as aggregation function (these results correspond to those in Fig. 3), while in yellow and green we show the results when we use the median and mode, respectively. These plots show that the mean function consistently exhibits lower variance among statements for the same ground truth level, as evidenced by a reduced interquartile range in the boxplots both for the six and in the two-level scale. This suggests that the mean may provide a more stable measure of central tendency when performing misinformation detection by means of crowdsourcing as compared to other aggregation metrics such as median and mode, that show both wider interquartile ranges and longer whiskers.

To further compare the three aggregation functions, we also undertake a set of statistical tests to examine the presence of statistically significant differences among mean, median, and mode. To this end, we mirror the previous methodology applying the Kruskal–Wallis and Dunn's Tests to assess the statistical significance of differences across the aggregation functions. We apply these tests on both the absolute and squared errors between the workers' scores and the ground truth, to determine whether one aggregation function yields more or fewer errors than the others. For both absolute and squared errors, both the Kruskal–Wallis and Dunn's Tests yield a p-value greater than 0.05, suggesting the absence of statistically significant differences in error rates among the three aggregation functions. We performed additional testing through the t-test and Tukey HSD, which resulted in consistent outcomes.

Coupling these findings with existing literature (La Barbera et al., 2020; Roitero, Soprano, Fan et al., 2020; Soprano et al., 2021) that consistently indicates higher quality results from using the mean as an aggregation function, we decided to use the mean as the main aggregation function in our analysis.

Finally, we discuss workers' external agreement. Initially, we measure the agreement between the aggregated assessments provided by the workers and the ground truth provided by the experts. We obtain $\alpha = 0.204$ for Soprano et al., $\alpha = 0.065$ for Draws et al., and $\alpha = 0.681$ for Our Study. Thus, we find a lower external agreement score with experts for workers from Draws et al. and Soprano et al. studies than the ones from Our Study. Then, we investigate how the external agreement varies as the number of workers assessing each statement decreases. Given that each statement is assessed by 10 crowd workers, we randomly sample workers for each possible cardinality (i.e., from 1 to 10). To avoid biases and estimate the mean value, we perform bootstrap repeating the agreement computation 100 times. Fig. 5 shows the result. The x-axis shows each cardinality of workers sampled from those who evaluated a given statement, while the y-axis shows the corresponding α. As we can see, for each study considered the external agreement increases for lower cardinalities and then stabilizes for samples ranging from 8 to 10 workers. When considering Our Study, noticeably higher α scores can be seen for each cardinality. Moreover, Our Study is the one showing the highest external agreement variability for the 10 considered cardinalities. In fact, the boxplots show increasing median α scores. However, such values are always higher for Our Study than in Soprano et al. and Draws et al. ones.

5.2.3. Pairwise agreement

Table 4 shows the pairwise agreement with experts when considering ties or when excluding them, across all the three studies considered. The pairwise agreement scores confirm again the higher effectiveness obtained for Our Study (0.829 with ties, 0.812

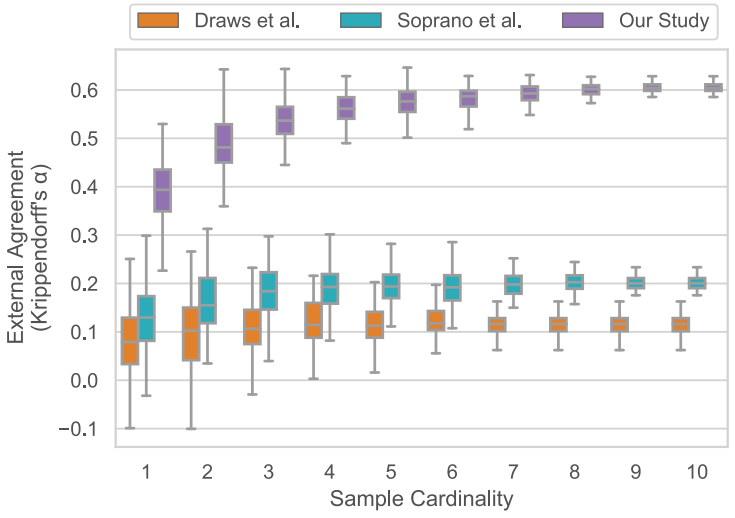

Fig. 5. Worker agreement with experts when considering samples of workers from 1 to 10. Each sample is bootstrapped 100 times. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

without). Such scores are higher than those obtained by the previous studies: Soprano et al. obtained 0.662 and 0.620, while Draws et al. 0.652 and 0.619.

5.2.4. Mean Squared Error (MSE)

We compute the Mean Squared Error (MSE) for each study considered when aggregating the scores provided by each worker when assessing a statement. Table 4 shows the results. Our Study has a lower and thus better MSE value (i.e., 1.484) compared to the studies of Draws et al. and Soprano et al.

When computing variants of MSE (e.g., by considering the raw assessments provided by each worker, or by considering each ground truth level separately), that we do not report, we obtain very similar results.

5.3. RQ2: Task and workers' features

The results reported in Section 5.2 show the higher effectiveness of Our Study compared with the ones of Soprano et al. and Draws et al., consistently across several measures. We now focus on understanding whether the setting of Our Study can improve Soprano et al. and Draws et al. with respect to other features.

Section 5.3.1 investigates the consistency between two of the effectiveness measures employed, while Section 5.3.2 discusses the internal agreement among workers. Finally, Section 5.3.3 addresses notable features of the new crowd of workers.

5.3.1. Agreement scores consistency

We start by investigating the consistency between two measures of external agreement (i.e., agreement of the workers with the experts), that is Krippendorff's α (Section 5.2.2) and pairwise agreement (Section 5.2.3), both computed at the worker level. In such a way, we measure differences on the two considered metrics when investigating workers agreement with the experts. The results are summarized in Fig. 6. The top row shows the pairwise agreement without ties, while in the bottom row, the ties are included. Each column is a study.

Fig. 6 shows a positive statistically significant correlation between Krippendorff's α and the pairwise agreement for each study, regardless of whether ties are considered in the measure or not. Specifically, the effectiveness of assessments collected by Our Study is higher than the one observed for Draws et al. and Soprano et al. studies, especially when comparing the plots shown in Fig. 6 bottom row. This confirms the higher effectiveness of Our Study. Note that by excluding ties we obtain a more stringent evaluation criterion since each tie is considered as an error made by the worker.

5.3.2. Worker's internal agreement – Agreement within workers

Considering worker's internal agreement, Table 4 shows that the workers from Our Study have a higher internal agreement ($\alpha = 0.219$) when compared with those of previous studies ($\alpha = 0.089$ for Soprano et al. and $\alpha = 0.040$ for Draws et al.). Then, we study the correlation between workers' internal and external agreements. Fig. 7 shows the results for the three studies considered. The x-axis shows the workers' internal agreement, while the y-axis shows the corresponding external agreement α. Each dot is a worker.

The three considered studies show a consistently positive and statistically significant correlation between the two measures, which implies that workers with higher internal agreement have also higher external agreement. Furthermore, we can see a stronger

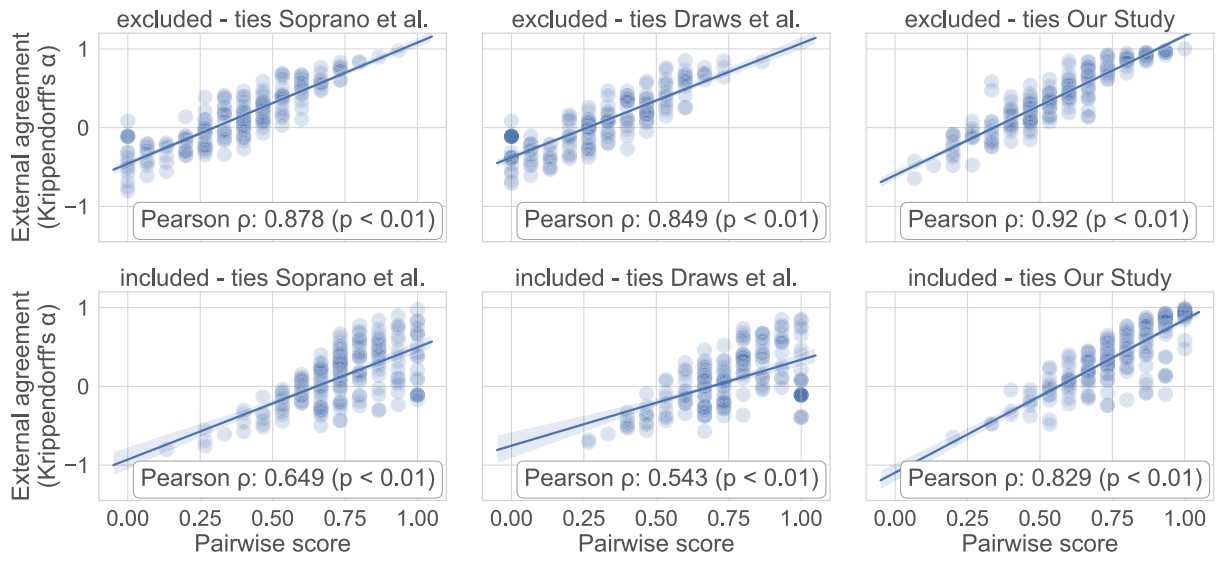

Fig. 6. Correlation between two measures of external agreement with experts: external pairwise agreement vs. external agreement α. The first row shows pairwise agreement with ties excluded, the second row with ties included. Each column is a study.

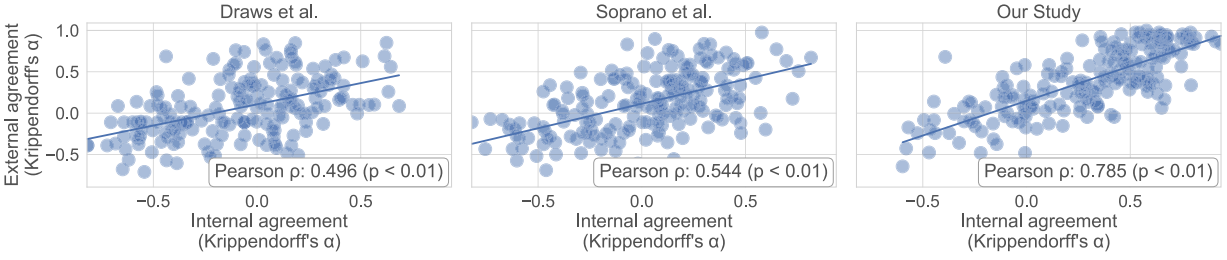

Fig. 7. Correlation between internal and external agreement.

correlation in Our Study ($\rho = 0.785$). This means that the internal agreement in Our Study provides a stronger signal of the expected overall quality than the studies of Soprano et al. and Draws et al.. In other words, the internal agreement metric in Our Study highlights a stronger indication of overall annotation quality compared to the studies by Soprano et al. and Draws et al.. Specifically, in our context, a higher degree of agreement among workers correlates more significantly with their effectiveness, meaning their consensus is more closely aligned with expert assessments. This finding is especially relevant in practical applications where internal agreement is readily available, offering a reliable measure of data quality. In fact, unlike external agreement, which requires expert input – often unavailable or limited in real-world settings – internal agreement can always be assessed using methods like Krippendorff's α. This not only validates the reliability of our internal agreement metric but also highlights its practicality and applicability in scenarios where expert judgments are not easy or feasible to obtain.

5.3.3. Effect of worker features

We now investigate some notable worker features and their effects. First, we inspect workers' self-reported age, derived from the demographic data provided by Prolific, and their relation with the external agreement measured using Krippendorf's α for each worker. We break down the correlations using each worker's self-reported gender. Fig. 8 shows the results. The x-axis shows the workers' age, while the y-axis is the corresponding external agreement. As shown, workers who identify themselves as females show a consistent external agreement across all workers' ages, while male workers show an increasing external agreement as their age increases.

Then, we analyze the questionnaire answers (which are either on categorical or ordinal scale) to test if the worker's political background has an impact on the assessments provided, as reported by Allen et al. (2021, 2022), Draws et al. (2022) and La Barbera et al. (2020). The results in Our Study are consistent with previous findings: workers who identify as Democrat or Independent achieve a higher level of binary accuracy (0.758 and 0.761 respectively) than the ones who identify as Republican (binary accuracy of 0.654). Nevertheless, each of these groups of workers achieves a lower accuracy compared to the whole crowd from Our Study. This result confirms the importance of diversity in the background and political beliefs of workers when a statement is evaluated.

We also observe statistically significant correlations ($p < 0.005$) between the self-reported workers' confidence and the truthfulness assessments of Soprano et al. and Draws et al. studies (0.53 and 0.58 respectively). These correlations are not significant

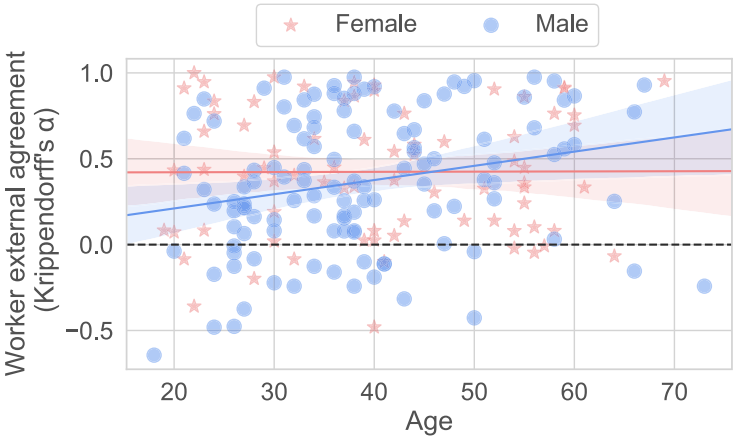

Fig. 8. Correlation for Our Study between worker's age (x-axis) and worker's external agreement α (y-axis), breakdown on gender.

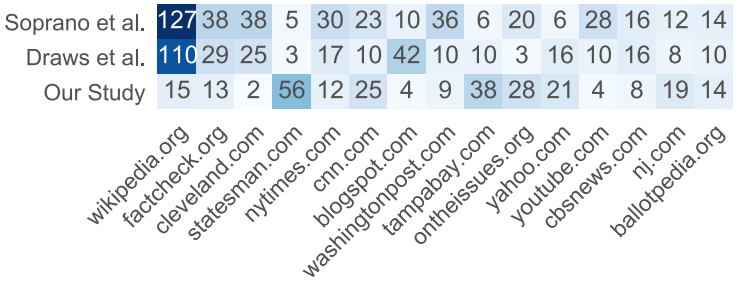

Fig. 9. Top 15 sources reported as evidence from the workers.

in Our Study. Such a finding provides evidence of higher workers' confidence when judging statements considered as True, and lower confidence when judging statements considered False. In other words, in Our Study the perceived truthfulness of a statement does not influence the confidence of the workers who judged it. We plan to further investigate this phenomenon in future work. As a last insight, we analyze the top-15 sources of evidence chosen by the workers to support their assessments for each study considered. Fig. 9 shows the result; interestingly, the workers recruited by Soprano et al. and Draws et al. relied heavily on Wikipedia as a source of evidence, while those of Our Study prefer newspaper websites and, in general, a more distributed set of websites.

5.4. RQ3: *Effect of truthfulness dimensions and confidence*

Finally, we explore how the workers provide assessments for the seven dimensions of truthfulness across the three studies. First, for each study, we use Pearson's ρ and Kendall's τ to compute the correlation between each dimension pair, including Ground Truth and worker's Confidence, as done by Soprano et al. (2021). Our results show statistically significant correlations between each considered pair, with the only exception of a non-significant correlation between Ground Truth and the worker's Confidence, in each study. To identify possible differences in these correlations between the previous experiments and Our Study we compute the difference between correlation values (both for Pearson's ρ and Kendall's τ) for each dimension pair. A positive difference indicates a higher correlation value for Our Study, while a negative difference indicates a lower correlation value for Our Study. Fig. 10 shows the results. The two heatmaps on the left (separated from the right part by the dimension names at the center of the figure) show the correlation differences between Soprano et al. and Our Study, while those on the right show correlation differences between Draws et al. and Our Study. The heatmaps on the top (separated from the bottom by the dimension names written vertically) show the differences between studies computed using Pearson's ρ, while those on the bottom use Kendall's τ. Thus, each cell shows the correlation difference between two studies on two given dimensions. We emphasize the distinction between Ground Truth and Confidence from the seven dimensions of truthfulness by increasing the white space between these elements and using bold typography for their labels. This visual adjustment highlights that they are not part of the seven dimensions of truthfulness.

The figure shows a positive difference in correlation values between multiple dimension pairs: Accuracy/Speaker's Trustworthiness, and Accuracy/Neutrality, and Comprehensibility/Completeness, being particularly evident between the first pair. A positive difference in correlation values indicates that workers from Our Study perceive, for instance, statements with a higher ground truth (the most true ones) as more accurate, and said by more trustworthy speakers. More generally, the workers perceive the dimensions for which there is a positive difference between correlation values with the ground truth as more related with the

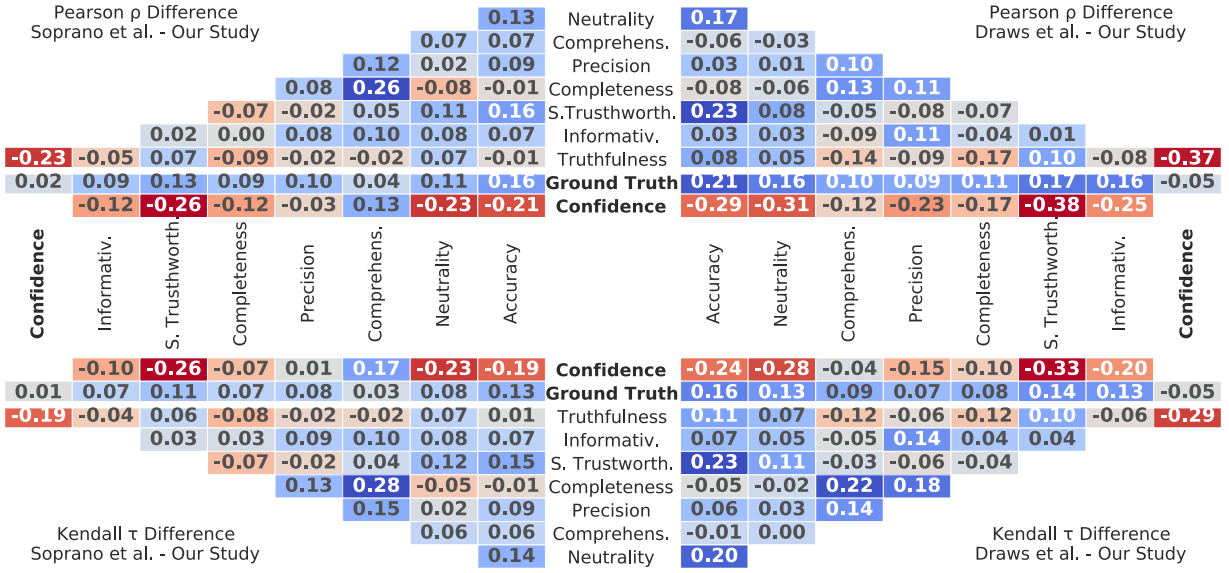

Fig. 10. Difference between correlation values computed using Pearson's ρ (top) and Kendall's τ (bottom), for each dimension considered in Soprano et al. and Our Study (left), and Draws et al. and Our Study (right).

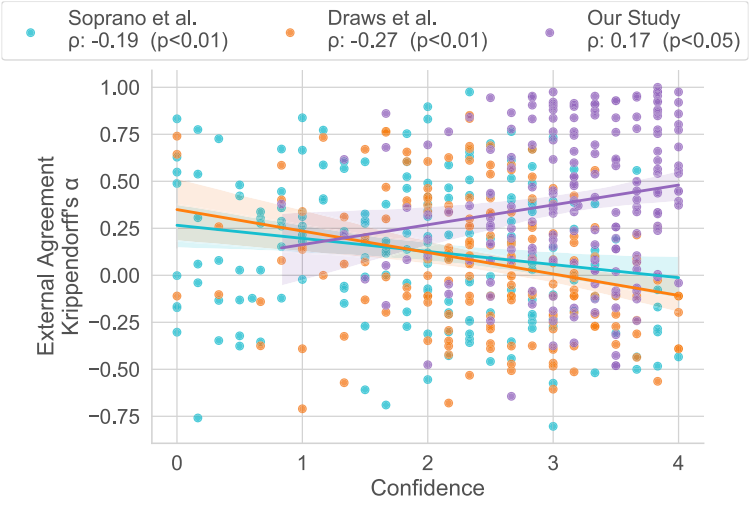

Fig. 11. Correlation between mean worker Confidence (x-axis) and mean worker's external agreement computed using Krippendorff's α (y-axis) for each considered study. Each dot is a statement. Legend annotated with statistical significance for Pearson's ρ.

underlying ground truth of the statement. Moreover, these dimensions act in Our Study as a good quality indicator of the worker's performed judgment: these dimensions could be the ones mainly considered by the workers when performing fact-checking tasks.

For most of the dimensions pairs shown in Fig. 10, however, the correlation differences are very close to zero between Our Study and the previous ones; this means that the assessments provided for Our Study are consistent with those provided in Soprano et al. and Draws et al. studies. Finally, when comparing the worker's self-reported Confidence with other dimensions for each study (first row starting from the center in each plot), we see multiple values with a negative correlation difference. This suggests some differences between the self-reported worker's Confidence from each experiment, and a lower correlation between the Confidence and each dimension.

To further investigate workers' self-reported Confidence behavior, we look at its relation with the worker's external agreement (i.e., agreement of the workers with the experts). Fig. 11 shows the results. The x-axis reports the mean confidence values, while the y-axis shows the average workers' external agreement. Each dot is a statement. As can be seen, both Soprano et al. and Draws et al. studies have a similar and statistically significant negative correlation value. Thus, the higher the worker's Confidence, the lower their agreement with the experts. This result remarks the low quality of these workers, whose confidence cannot be trusted

as a quality indicator of the performed truthfulness assessments. Differently, in Our Study there is evidence of a small positive and statistically significant ($p < 0.05$) correlation between external agreement (i.e., agreement of the workers with the experts) and Confidence. Thus, workers with higher external agreement show a higher Confidence. Therefore, higher-quality workers provide Confidence scores that are more reliable, that can act as a proxy of quality over the performed assessments.

6. Conclusions and future work

This research aims to shed light on contradictory results found in the literature and understand whether a crowd of workers can effectively assess the truthfulness of statements, one of the key steps in fact-checking. We pursue that by building on the work of Draws et al. (2022) and Soprano et al. (2021) to reproduce and improve their methodology. In the following, we provide a synthesis of our findings for each Research Question detailed in Section 3.3, detailing their implications and contributions to the field:

RQ1 By reproducing Draws et al. (2022) and Soprano et al. (2021) experimental design – but improving task design and recruiting a new pool of workers using a different crowdsourcing platform – we have been able to improve their suboptimal results. We thus found additional evidence that supports what others reported (Allen et al., 2021; Zhao & Naaman, 2023): in specific conditions, crowdsourcing can achieve comparable effectiveness when compared to experts in assessing the truthfulness of statements. This positive result highlights the need for researchers and practitioners to take additional care when (i) designing a crowdsourcing task with the purpose of collecting truthfulness labels, and (ii) selecting the pool of workers to employ.

RQ2 Concerning the impact of the new study design both in terms of task features and workers recruited on the overall effectiveness, we found some differences: (i) workers who identify as female show a more consistent external agreement (i.e., agreement with the experts) for each age range than those who identify as male, (ii) the perceived truthfulness of a statement does not influence the workers' self-reported Confidence, and (iii) workers prefer a more distributed set of websites to provide evidence.

RQ3 Focusing on how workers provide assessments using the seven dimensions of truthfulness, we found that only for a subset of them workers provide assessments in a different fashion. Indeed, the majority of dimensions have a strong correlation with Truthfulness. Thus, we derive that future studies can employ only a subset of the dimensions which have been originally proposed, and in particular, they should consider statement Accuracy and Speaker's Trustworthiness. Finally, by inspecting the correlation between dimensions we found that workers with a higher quality show a higher degree of Confidence in their judgments, meaning that in such cases Confidence might act as a proxy measure for label quality.

In sum, our study corroborates the reliability of using crowdsourcing as a mechanism to scale truthfulness assessments for fact-checking. Our results provide insights on how non-experts can effectively engage in misinformation management processes (Demartini et al., 2020). It is worth noting that the aim is not to replace experts, but rather make the fact-checking process more cost-effective (Spina et al., 2023). We made several contributions to the field of crowdsourced truthfulness assessment, which we summarize as follows.

- We introduced a novel task design and experimental settings that improved previous studies (Draws et al., 2022; Soprano et al., 2021), achieving higher accuracy and agreement with expert assessments.
- Our findings highlighted the importance of diverse and balanced worker demographics, as well as the choice of the crowdsourcing platform, in enhancing the quality of the collected truthfulness assessments.
- We demonstrated the utility of considering multiple dimensions of truthfulness in the assessment process, providing insights into how these dimensions correlate with each other and with the overall ground truth.
- We found that worker confidence in our improved task design is a reliable indicator of worker assessment quality, contrasting with previous studies where confidence did not show a significant correlation with external agreement.

There are other aspects of this complex task that are left for future work. Given the differences in the reported confidence, results obtained by Qu et al. (2022) need to be re-discussed, as Confidence might indeed act as a proxy for the judgments provided by the workers. Another line of future work consists of replicating the crowdsourcing task in other scenarios that include languages other than English, geo-political contexts beyond the US, other types of media other than text, and those that incorporate different task designs, particularly employing other evaluation techniques like magnitude estimation, a method which assigns numerical values to represent the perceived intensity of an effect, a technique that has demonstrated effectiveness in various domains including relevance assessment (Roitero, Maddalena, Mizzaro and Scholer, 2021).

Finally, given the available data, we can only speculate on the relative importance of another aspect we have focused on, namely the selection of high quality workers. Our initial impression is that the quality of workers holds greater significance. To confirm this, additional ablation studies are needed.

CRediT authorship contribution statement

David La Barbera: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Data curation, Conceptualization. **Eddy Maddalena:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation, Conceptualization. **Michael Soprano:** Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Data curation, Conceptualization. **Kevin Roitero:** Writing – review & editing,

Writing – original draft, Validation, Software, Methodology, Data curation, Conceptualization. **Gianluca Demartini:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Conceptualization. **Davide Ceolin:** Writing – review & editing, Visualization, Validation, Supervision, Conceptualization. **Damiano Spina:** Writing – review & editing, Writing – original draft, Supervision, Funding acquisition, Conceptualization. **Stefano Mizzaro:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Conceptualization.

Data availability

Instructions at the beginning of Section 4.

Disclosure and acknowledgments

This research is partially supported by the Australian Research Council, Australia (DE200100064, CE200100005, IC200100022). Damiano Spina is the recipient of an Australian Research Council (ARC) DECRA Research Fellowship (DE200100064), and Associate Investigator of the ARC Centre of Excellence for Automated Decision-Making and Society (CE200100005), and a research collaborator of RMIT ABC Fact Check. Gianluca Demartini is a Chief Investigator of the ARC Training Centre for Information Resilience (IC200100022).

This research is also supported by the Swiss National Science Foundation (SNSF) under contract number CRSII5_205975 and 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, by the Strategic Plan of the University of Udine–Interdepartment Project on Artificial Intelligence (2020-25), by the Netherlands eScience Center project "The Eye of the Beholder" (project nr. 027.020.G15), and it is part of the AI, Media & Democracy Lab (Dutch Research Council project number: NWA.1332.20.009). For more information about the lab and its further activities, visit https://www.aim4dem.nl/.

Any opinions, findings, and conclusions expressed in this article are those of the authors and do not necessarily reflect those of the sponsors.

References

Ahmed, A. A. A., Aljabouh, A., Donepudi, P. K., & Choi, M. S. (2021). Detecting Fake News Using Machine Learning: A Systematic Literature Review. *Psychology and Education Journal*, *58*(1), http://dx.doi.org/10.17762/pae.v58i1.1046.

Allen, J., Arechar, A. A., Pennycook, G., & Rand, D. G. (2021). Scaling Up Fact-Checking Using the Wisdom of Crowds. *Science Advances*, *7*(36), eabf4393. http://dx.doi.org/10.1126/sciadv.abf4393.

Allen, J., Martel, C., & Rand, D. G. (2022). 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* (pp. 1–19). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3491102.3502040.

Amigo, E., Gonzalo, J., Mizzaro, S., & Carrillo-de Albornoz, J. (2020). An Effectiveness Metric for Ordinal Classification: Formal Properties and Experimental Results. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of The Association for Computational Linguistics* (pp. 3938–3949). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.acl-main.363.

Aphiwongsophon, S., & Chongstitvatana, P. (2018). Detecting Fake News with Machine Learning Method. In *15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology* ECTI-CON, (pp. 528–531). Chiang Rai, Thailand: IEEE, http://dx.doi.org/10.1109/ECTICon.2018.8620051.

Bland, M. J., & Altman, D. G. (1995). Multiple Significance Tests: The Bonferroni Method. *BMJ*, *310*(6973), 170. http://dx.doi.org/10.1136/bmj.310.6973.170.

Ceolin, D., Noordegraaf, J., & Aroyo, L. (2016). Capturing the Ineffable: Collecting, Analysing, and Automating Web Document Quality Assessments. In E. Blomqvist, P. Ciancarini, F. Poggi, & F. Vitali (Eds.), *Knowledge Engineer* (pp. 83–97). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-319-49004-5_6.

Checco, A., Roitero, K., Maddalena, E., Mizzaro, S., & Demartini, G. (2017). Let's Agree to Disagree: Fixing Agreement Measures for Crowdsourcing. In S. Dow, & A. T. Kalai (Eds.), *Proceedings of the Fifth AAAI Conference on Human Computation and Crowdsourcing* (pp. 11–20). Québec City, Québec, Canada: AAAI Press, URL: https://aaai.org/ocs/index.php/HCOMP/HCOMP17/paper/view/15927.

Chmielewski, M., & Kucker, S. C. (2020). An MTurk Crisis? Shifts in Data Quality and the Impact on Study Results. *Social Psychological and Personality Science*, *11*(4), 464–473. http://dx.doi.org/10.1177/1948550619875149.

Collins, B., Hoang, D. T., Nguyen, N. T., & Hwang, D. (2021). Trends in Combating Fake News On Social Media – A Survey. *Journal of Information and Telecommunication*, *5*(2), 247–266. http://dx.doi.org/10.1080/24751839.2020.1847379.

Das, A., Liu, H., Kovatchev, V., & Lease, M. (2023). The State of Human-centered NLP Technology for Fact-checking. *Information Processing & Management*, *60*(2), Article 103219. http://dx.doi.org/10.1016/j.ipm.2022.103219.

De Vries, D. A., Piotrowski, J., & de Vreese, C. (2023). DigiIQ - Digital Competence across the Lifespan. URL: https://osf.io/dfvqb/.

Demartini, G., Mizzaro, S., & Spina, D. (2020). Human-in-the-loop Artificial Intelligence for Fighting Online Misinformation: Challenges and Opportunities. *IEEE Data Engineering Bulletin*, *43*(3), 65–74, URL: http://sites.computer.org/debull/A20sept/p65.pdf.

Dong, X., Sarker, S., & Qian, L. (2022). Integrating Human-in-the-loop into Swarm Learning for Decentralized Fake News Detection. In *IDSTA*, *International Conference on Intelligent Data Science Technologies and Applications* (pp. 46–53). San Antonio, TX, USA: IEEE, http://dx.doi.org/10.1109/IDSTA55301.2022.9923043.

Draws, T., La Barbera, D., Soprano, M., Roitero, K., Ceolin, D., Checco, A., et al. (2022). The Effects of Crowd Worker Biases in Fact-Checking Tasks. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2114–2124). Seoul, Republic of Korea: Association for Computing Machinery, http://dx.doi.org/10.1145/3531146.3534629.

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, *19*(4), 25–42. http://dx.doi.org/10.1257/089533005775196732.

Gemalmaz, M. A., & Yin, M. (2021). Accounting for Confirmation Bias in Crowdsourced Label Aggregation. In Z.-H. Zhou (Ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence* (pp. 1729–1735). International Joint Conferences on Artificial Intelligence Organization, http://dx.doi.org/10.24963/ijcai.2021/238.

Godel, W., Sanderson, Z., Aslett, K., Nagler, J., Bonneau, R., Persily, N., et al. (2021). Moderating With The Mob: Evaluating The Efficacy of Real-Time Crowdsourced Fact-Checking. *Journal of Online Trust and Safety*, *1*(1), http://dx.doi.org/10.54501/jots.v1i1.15.

Graves, L. (2017). Anatomy of a Fact Check: Objective Practice and the Contested Epistemology of Fact Checking. *Communication, Culture and Critique*, *10*(3), 518–537. http://dx.doi.org/10.1111/cccr.12163.

Hakak, S., Alazab, M., Khan, S., Gadekallu, T. R., Maddikunta, P. K. R., & Khan, W. Z. (2021). An Ensemble Machine Learning Approach Through Effective Feature Extraction to Classify Fake News. *Future Generation Computer Systems*, *117*, 47–58. http://dx.doi.org/10.1016/j.future.2020.11.022.

Han, T. L., Roitero, K., Gadiraju, U., Sarasua, C., Checco, A., Maddalena, E., et al. (2019). The Impact of Task Abandonment in Crowdsourcing. *IEEE Transactions on Knowledge & Data Engineering*, *1*(1), http://dx.doi.org/10.1109/TKDE.2019.2948168, 1–1.

Han, L., Roitero, K., Maddalena, E., Mizzaro, S., & Demartini, G. (2019). On Transforming Relevance Scales. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (pp. 39–48). Beijing, China: Association for Computing Machinery, http://dx.doi.org/10.1145/3357384.3357988.

Howe, J. (2006). The Rise of Crowdsourcing. *Wired Magazine*, *14*(6), 1–4, URL: https://www.wired.com/2006/06/crowds/.

Hu, L., Wei, S., Zhao, Z., & Wu, B. (2022). Deep Learning For Fake News Detection: A Comprehensive Survey. *AI Open*, *3*, 133–155. http://dx.doi.org/10.1016/j.aiopen.2022.09.001.

International Organization for Standardization (2008). *ISO/IEC 25012:2008 Software Engineering — Software Product Quality Requirements and Evaluation (SQuaRE) — Data Quality Model*: *Technical report*, ISO, URL: https://www.iso.org/standard/35736.html.

Jean Dunn, O. (1964). Multiple Comparisons Using Rank Sums. *Technometrics*, *6*(3), 241–252. http://dx.doi.org/10.1080/00401706.1964.10490181.

Jiang, L., Zhang, H., Tao, F., & Li, C. (2022). Learning From Crowds With Multiple Noisy Label Distribution Propagation. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(11), 6558–6568. http://dx.doi.org/10.1109/TNNLS.2021.3082496.

Kahn, B. K., Strong, D. M., & Wang, R. Y. (2002). Information Quality Benchmarks: Product and Service Performance. *Communications of the ACM*, *45*(4), 184–192. http://dx.doi.org/10.1145/505248.506007.

Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P. D., Jewell, R., & Winter, N. J. G. (2020). The Shape of and Solutions to the MTurk Quality Crisis. *Political Science Research and Methods*, *8*(4), 614–629. http://dx.doi.org/10.1017/psrm.2020.6.

Krippendorff, K. (2008). Computing Krippendorff's Alpha-Reliability. *UPENN Libraries*, *1*, 43, URL: https://repository.upenn.edu/asc_papers/43.

Kruskal, W. H., & Allen Wallis, W. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, *47*(260), 583–621. http://dx.doi.org/10.1080/01621459.1952.10483441.

La Barbera, D., Roitero, K., Demartini, G., Mizzaro, S., & Spina, D. (2020). 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* (pp. 207–214). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-45442-5_26.

La Barbera, D., Roitero, K., Mackenzie, J., Spina, D., Demartini, G., & Mizzaro, S. (2022). BUM at CheckThat!-2022: A Composite Deep Learning Approach to Fake News Detection using Evidence Retrieval. In G. Faggioli, N. Ferro, A. Hanbury, & M. Potthast (Eds.), *CEUR Workshop Proceedings*: *Vol. 3180*, *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of The Evaluation Forum* (pp. 564–572). Bologna, Italy: CEUR-WS.org, URL: https://ceur-ws.org/Vol-3180/.

La Barbera, D., Roitero, K., & Mizzaro, S. (2022). A Hybrid Human-In-The-Loop Framework for Fact Checking. In *NL4AI '22*, *Proceedings of the 6th Workshop on Natural Language for Artificial Intelligence* (pp. 1–10). Udine, Italy: CEUR-WS.org, URL: https://ceur-ws.org/Vol-3287/paper4.pdf.

Li, H., Jiang, L., & Xue, S. (2023). Neighborhood Weighted Voting-Based Noise Correction for Crowdsourcing. *ACM Transactions on Knowledge Discovery from Data*, *17*(7), http://dx.doi.org/10.1145/3586998.

Maddalena, E., Ceolin, D., & Mizzaro, S. (2018). 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* (pp. 1–5). URL: http://ceur-ws.org/Vol-2482/paper17.pdf.

Maddalena, E., Roitero, K., Demartini, G., & Mizzaro, S. (2017). Considering Assessor Agreement in IR Evaluation. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval* (pp. 75–82). Association for Computing Machinery, http://dx.doi.org/10.1145/3121050.3121060.

Manzoor, S. I., Singla, J., & Nikita (2019). Fake News Detection Using Machine Learning approaches: A systematic Review. In *ICOEI*, *Proceedings of the 3rd International Conference on Trends in Electronics and Informatics* (pp. 230–234). Tirunelveli, India: IEEE, http://dx.doi.org/10.1109/ICOEI.2019.8862770.

Mena, P. (2019). Principles and Boundaries of Fact-checking: Journalists' Perceptions. *Journalism Practice*, *13*(6), 657–672. http://dx.doi.org/10.1080/17512786.2018.1547655.

Nakov, P., Barrón-Cedeño, A., da San Martino, G., Alam, F., Struß, J. M., Mandl, T., et al. (2022). Overview of the CLEF–2022 CheckThat! Lab on Fighting the COVID-19 Infodemic and Fake News Detection. In A. Barrón-Cedeño, G. Da San Martino, M. Degli Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, N. Ferro (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction* (pp. 495–520). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-031-13643-6_29.

Nakov, P., Da San Martino, G., Elsayed, T., Barrón-Cedeño, A., Míguez, R., Shaar, S., et al. (2021). The CLEF-2021 CheckThat! Lab on Detecting Check-Worthy Claims, Previously Fact-Checked Claims, and Fake News. In D. Hiemstra, M.-F. Moens, J. Mothe, R. Perego, M. Potthast, F. Sebastiani (Eds.), *Advances in Information Retrieval* (pp. 639–649). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-72240-1_75.

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data Quality of Platforms and Panels for Online Behavioral Research. *Behavior Research Methods*, *54*(4), 1643–1662. http://dx.doi.org/10.3758/s13428-021-01694-3.

Qu, Y., Roitero, K., Barbera, D. L., Spina, D., Mizzaro, S., & Demartini, G. (2022). Combining Human and Machine Confidence in Truthfulness Assessment. *Journal of Data and Information Quality*, *15*(1), http://dx.doi.org/10.1145/3546916.

Roitero, K., Maddalena, E., Mizzaro, S., & Scholer, F. (2021). On the Effect of Relevance Scales in Crowdsourcing Relevance Assessments for Information Retrieval Evaluation. *Information Processing & Management*, *58*(6), Article 102688. http://dx.doi.org/10.1016/j.ipm.2021.102688.

Roitero, K., Soprano, M., Fan, S., Spina, D., Mizzaro, S., & Demartini, G. (2020). Can The Crowd Identify Misinformation Objectively? The Effects of Judgment Scale and Assessor's Background. In *SIGIR '20*, *Proceedings of the 43st International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 439–448). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3397271.3401112.

Roitero, K., Soprano, M., Portelli, B., De Luise, M., Spina, D., Della Mea, V., et al. (2021). Can the Crowd Judge Truthfulness? A Longitudinal Study on Recent Misinformation about COVID-19. *Personal and Ubiquitous Computing*, *27*, 59–89. http://dx.doi.org/10.1007/s00779-021-01604-6.

Roitero, K., Soprano, M., Portelli, B., Spina, D., Della Mea, V., Serra, G., et al. (2020). The COVID-19 Infodemic: Can the Crowd Judge Recent Misinformation Objectively? In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (pp. 1305–1314). Virtual Event, Ireland: Association for Computing Machinery, http://dx.doi.org/10.1145/3340531.3412048.

Saeed, M., Traub, N., Nicolas, M., Demartini, G., & Papotti, P. (2022). Crowdsourced Fact-Checking at Twitter: How Does the Crowd Compare With Experts? In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management* (pp. 1736–1746). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3511808.3557279.

Sedgwick, P. (2012). Multiple Significance Tests: The Bonferroni Correction. *The BMJ*, *344*, http://dx.doi.org/10.1136/bmj.e509.

Sethi, R. J. (2017). Crowdsourcing the Verification of Fake News and Alternative Facts. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media* (pp. 315–316). Prague, Czech Republic: Association for Computing Machinery, http://dx.doi.org/10.1145/3078714.3078746.

Shu, K., Cui, L., Wang, S., Lee, D., & Liu, H. (2019). DEFEND: Explainable Fake News Detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 395–405). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3292500.3330935.

Soprano, M., Roitero, K., Bombassei De Bona, F., & Mizzaro, S. (2022). Crowd_Frame: A Simple and Complete Framework to Deploy Complex Crowdsourcing Tasks Off-the-Shelf. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining* (pp. 1605–1608). Virtual Event, AZ, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3488560.3502182.

Soprano, M., Roitero, K., La Barbera, D., Ceolin, D., Spina, D., Mizzaro, S., et al. (2021). The Many Dimensions of Truthfulness: Crowdsourcing Misinformation Assessments on a Multidimensional Scale. *Information Processing & Management*, *58*(6), Article 102710. http://dx.doi.org/10.1016/j.ipm.2021.102710.

Spina, D., Sanderson, M., Angus, D., Demartini, G., McKay, D., Saling, L. L., et al. (2023). Human-AI Cooperation to Tackle Misinformation and Polarization. *Communications of the ACM*, *66*(7), 40–45. http://dx.doi.org/10.1145/3588431.

Tanvir, A. A., Mahir, E. M., Akhter, S., & Huq, M. R. (2019). Detecting Fake News using Machine Learning and Deep Learning Algorithms. In *ICSCC*, *Proceedings of the 7th International Conference on Smart Computing & Communications* (pp. 1–5). New York, USA: IEEE, http://dx.doi.org/10.1109/ICSCC.2019.8843612.

Vlachos, A., & Riedel, S. (2014). Fact Checking: Task Definition and Dataset Construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science* (pp. 18–22). Baltimore, MD, USA: Association for Computational Linguistics, http://dx.doi.org/10.3115/v1/W14-2508.

Ximenes, B. H., & Ramalho, G. (2022). The Best of Both Worlds: Mixed Systems with ML and Humans in the Loop to Combat Fake Information. In J. Y. C. Chen, G. Fragomeni, H. Degen, & S. Ntoa (Eds.), *HCI International 2022 – Late Breaking Papers: Interacting With eXtended Reality and Artificial Intelligence* (pp. 583–597). Cham: Springer Nature Switzerland, http://dx.doi.org/10.1007/978-3-031-21707-4_42.

Yang, J., Vega-Oliveros, D., Seibt, T., & Rocha, A. (2021). Scalable Fact-checking with Human-in-the-Loop. In *WIFS*, *2021 IEEE International Workshop on Information Forensics and Security* (pp. 1–6). Montpellier, France: IEEE, http://dx.doi.org/10.1109/WIFS53200.2021.9648388.

Zhao, A., & Naaman, M. (2023). Variety, Velocity, Veracity, and Viability: Evaluating the Contributions of Crowdsourced and Professional Fact-checking. http://dx.doi.org/10.31235/osf.io/yfxd3, URL: osf.io/preprints/socarxiv/yfxd3.

Zhou, X., Wu, J., & Zafarani, R. (2020). SAFE: Similarity-Aware Multi-modal Fake News Detection. In H. W. Lauw, R. C.-W. Wong, A. Ntoulas, E.-P. Lim, S.-K. Ng, & S. J. Pan (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 354–367). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-47436-2_27.