Does Evidence from Peers Help Crowd Workers in Assessing Truthfulness?

Jiechen Xu, Lei Han, Shaoyang Fan, Shazia Sadiq, and Gianluca Demartini
jiechen.xu@uq.net.au;l.han@uq.edu.au;fsysean@gmail.com;shazia@itee.uq.edu.au;demartini@acm.org
The University of Queensland, Brisbane, Australia

ABSTRACT

Misinformation has been rapidly spreading online. The current approach to deal with it is deploying expert fact-checkers that follow forensic processes to identify the veracity of statements. Unfortunately, such an approach does not scale well. To deal with this, crowdsourcing has been looked at as an opportunity to complement the work of trained journalists. In this work, we look at the effect of presenting the crowd with evidence from others while judging the veracity of statements. We implement various variants of the judgment task design to understand if and how the presented evidence may or may not affect the way of crowd workers judging truthfulness and their performance. Our results show that, in certain cases, the presented evidence may mislead crowd workers who would otherwise be more accurate if judging independently from others. Those who made correct use of the provided evidence, however, could benefit from it and generate better judgments.

CCS CONCEPTS

• **Information systems → Crowdsourcing**.

KEYWORDS

Misinformation, Crowdsourcing, Metadata, Information Credibility

ACM Reference Format:
Jiechen Xu, Lei Han, Shaoyang Fan, Shazia Sadiq, and Gianluca Demartini. 2022. Does Evidence from Peers Help Crowd Workers in Assessing Truthfulness?. In *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3487553.3524236

1 INTRODUCTION

Crowdsourcing has been recently studied as a methodology to collect truthfulness judgements at scale [11]. This is an effort to deal with the scale of misinformation online in combination with and complementary to expert fact-checkers and fully automated machine learning methods [11]. While previous work has looked at the feasibility of crowdsourcing as a means to collect reliable labels, in this work, we study the effect of presenting crowd annotators with evidence about the truthfulness of the statements they are judging. Such evidence has been provided by other peer crowd workers assessing the same statements. The hypothesis we want to test is whether crowd workers are better off making independent decisions on the truthfulness of the statements presented to them (knowing that, in such a case, there is a strong risk of bias [1]) or they could be guided in their judgements by external evidence.

To this end, in this work, we use fact-checked political statements from *PolitiFact* to design and implement a truthfulness assessment crowdsourcing task on Amazon MTurk. We deploy different variants of the task design to test different conditions in which to present external evidence from peers. We make use of a dataset made available by previous work [11] that contains URLs provided by crowd workers as evidence while judging truthfulness and we present such evidence in our task.

We focus on the following research questions (RQs):

- (RQ#1) What is the impact of peers' judgements on the judging quality? What is the best option for presenting such evidence?
- (RQ#2) What is the impact of external evidence on workers' behaviour?
- (RQ#3) What is the impact of workers' background (i.e., self-identified political standing) on the way they make use of external evidence to identify misinformation?

2 RELATED WORK

Metadata for Crowdsourced Data Annotation. Metadata is defined as structured information that describes an information source [12]. Méndez and Hooland [9] explain how metadata can be deployed to fit specific use cases. They suggest that one way to classify metadata is by how it is created (e.g., manually created by humans versus automatically generated by algorithms). Eickhoff [4] leverages human metadata in the crowdsourced relevance assessment task in which peers' assessments are presented to crowd workers. He identified a bandwagon effect among crowd workers when such human metadata is available. Other examples of human metadata in crowdsourcing annotation tasks are more veiled. Doroudi et al. [3] showed how presenting example solutions (either by experts or crowd) can benefit novice workers in developing effective task completion strategies and leads to better results. Similarly, Chang et al. [2] studied the impact of (cross-)reviewing peer-generated judgment justifications in the context of Information Retrieval evaluation. They found that human assessors can learn from the extra information provided and, thus, generate higher quality labels. In our work we look at the effect of presenting crowd workers with human metadata in a truthfulness classification task.

Crowdsourcing Truthfulness Assessments. Pennycook and Rand [10] point out that some fake stories may never be flagged as the task of judging truthfulness is usually resource intensive and the scale of the problem is massive. Roitero et al. [11] conducted a study

WWW '22 Companion, April 25–29, 2022, Virtual Event, Lyon, France.

ACM ISBN 978-1-4503-9130-6/22/04…$15.00
https://doi.org/10.1145/3487553.3524236

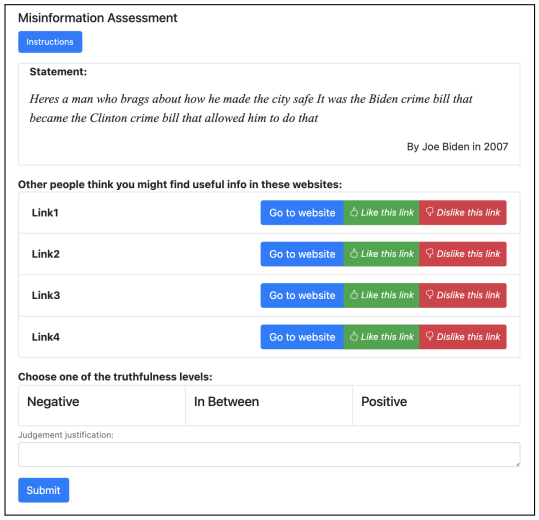

Figure 1: Interface of the truthfulness assessment task.

about crowdsourced truthfulness judgments. The result indicates that the crowd can achieve a nearly comparable result with experts' judgements after answer aggregation. The statements that crowd workers judge are usually provided by politicians that may contain ideological bias [6]. Implicit bias can influence crowd workers when they are required to make subjective judgements [1, 5]. In our work we look at how presenting evidence to crowd workers in a certain way may alleviate such bias effects.

3 METHODOLOGY

3.1 Dataset

To design and implement truthfulness assessment tasks, we use expert fact-checked political statements from *PolitiFact*. This dataset consists of short political statements mainly by US politicians [13]. Each statement has an expert editorial judgment of its truthfulness on a 6-level scale: pants-on-fire (`0`), false (1), mostly-false (2), half-true (3), mostly-true (4), and true (5). To make our study comparable against existing research, we use the same subset of *PolitiFact* as Roitero et al. [11]. This subset contains 120 statements with 20 statements at each of the 6 truthfulness levels judged by 10 distinct workers. All statements come from the two major US political parties (i.e., Democratic and Republican) from 2007 to 2015. In addition to the task design used in previous work, we provide crowd workers with peer worker-generated URLs [11] as the metadata for each statement. Therefore, each statement has at most 10 URLs.

3.2 Task Design

Following existing research [1, 11], we ask each crowd worker to judge the truthfulness of 8 statements, appearing sequentially one at a time. Among these 8 statements, two of them are gold questions (one is obviously true and the other is clearly false). To balance between the two political parties as well as among the truthfulness levels, we manually merge the original 6 levels into a 3-level scales by mapping `0` and 1 to '*Negative*', 2 and 3 to '*In Between*', and 4 and 5 to '*Positive*'. Thus, statements in each task are balanced across the 3 truthfulness levels as well as across the two political parties.

As a quality control method, we embed in our task a time-based check that requires workers spending at least 2 seconds on each statement, as in [11]. We ensure every worker can only complete one task in one condition.

Figure 1 shows the task interface, including: (i) the statement to judge (including information about the speaker and the year), (ii) the metadata in form of a URL table, (iii) buttons for truthfulness labels to click and (iv) a text area to write an optional judgement justification. Additionally, we provide an up-vote and a down-vote button next to each URL to optionally indicate if a specific URL is useful or not. For each statement, a worker must assign a truthfulness label before moving on to the next statement. We also embed a logger by JavaScript to capture their mouse clicks. Before the task, all workers are asked to complete a short demographics survey (i.e., gender, age, education level, and political leaning)[1].

3.3 Conditions

To understand if and how the presence of external evidence has an impact on crowdsourcing misinformation assessment, in our work, we consider the following variants (which we call 'conditions' in the following) of task design which we obtained by changing the way in which the metadata is shown:

- *No-URL*. In this condition, we show a traditional misinformation judgement task [11] and no metadata (i.e., URLs) is present.
- *Baseline*. We additionally present a metadata table with URLs which are in the form of a link button (e.g., 'Link 1' as shown in Fig. 1). The URLs appear in randomized order.

To investigate the effect of how the metadata is presented in the task, we also consider the following variants where we keep the same task design as 'Baseline' but using a different way to present the same URLs.

- *Order*. The URLs are sorted by popularity where the most frequently reported URL is at the top of the table. For a given URL, we also show how many peers have reported this URL.
- *Domain*. Compared to 'Baseline' showing the masked URLs via link buttons, in this condition we disclose the domain names of URLs (e.g., `www.wikipedia.org`) to the crowd worker.
- *Title*. Similar to 'Domain', instead of showing the destination of the URLs, we show the page title of the given URL.
- *Political*. In this condition, we show two metadata tables (side by side, one per party), where each table contains URLs previously provided by peer workers supporting a certain political party (i.e., either Democrats or Republicans). This allows us to investigate the potential political bias in crowd workers, that is, how crowd workers' political leanings affect the levels of trust in the information shown in the table and clash with the political background of those providing the shown information. Note that, under this condition, workers only see link buttons in randomized order, like in the 'Baseline'.

4 RESULTS

Using the task design described above, we conducted the experiments over Amazon MTurk, targeting 40 US workers per condition. Note that we allow each worker to participate in our tasks only

[1] All workers have been informed and have given consent to collect such data for our analysis, and the study has been approved by our human research ethics committee.

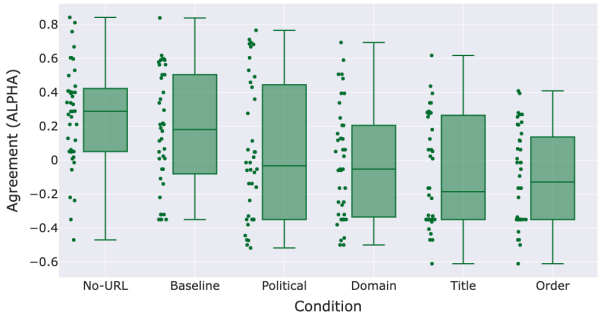

Figure 2: Quality of judgements over conditions. Conditions are sorted by the average quality in descending order.

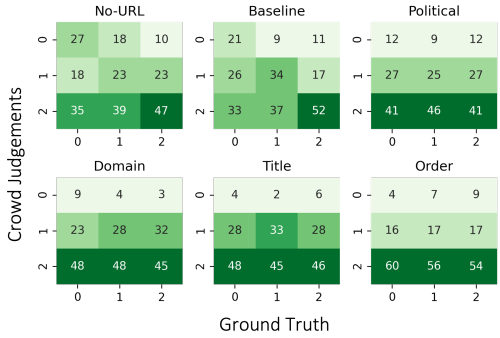

Figure 3: Confusion matrices for the numbers of judgements with a breakdown of ground-truth over conditions. Labels for row and column are workers' judgements and ground-truth, respectively. Notation: 0 – false, 1 – in-between, 2 – true.

once, and thus collected in total 1920 truthfulness judgements contributed by 240 unique workers[2].

4.1 Quality of Judgments (RQ#1)

Addressing RQ1, we first report about how the presence of evidence URLs has an impact on the judgment quality. To measure the quality of the judgments given by crowd workers, we compare their judgements with expert editorial assessments measuring agreement. We use Krippendorff's α coefficient [8] to compute the agreement, which goes from -1 (completely disagree) through 0 (equivalent to random guess) to 1 (completely agree).

Figure 2 shows the distribution of judgment quality across all conditions. Compared to 'Baseline', we do not observe statistically significant differences in the performance under the 'No-URL' or the 'Political' conditions, but the workers in the 'Domain, 'Title', and 'Order' conditions show a statistically significant lower judging quality (2-sided t-test, $p < 0.01$) as compared to the 'Baseline' condition. Because we provide more information under these three conditions (i.e., either more details about the website in 'Domain' and 'Title', or about the popularity of the URLs in 'Order'), such results may indicate that revealing this information may not contribute to quality improvements in truthfulness assessment tasks.

Figure 3 shows confusion matrices containing the number of workers and their judgments versus ground-truth expert judgements. The incorrect judgments for false statements are more than that for true statements. For example, in the 'Domain' condition (bottom-left in Fig. 3), there are 71 judgments mistakenly assigning true (48) or in-between (23) judgments to false statements, while there are 7 (i.e., 3 + 4) judgments assigning false labels to true or in-between statements.This shows how workers tend to make more mistakes when judging false statements as compared to true statements in all conditions. Compared to 'Baseline' where workers make 33 mistakes on false statements, workers in 'Domain', 'Title' and 'Order' conditions make more mistakes (i.e., 48, 48, and 60 mistakes on false statements, respectively). This also explains the overall drop in agreement under these three conditions (see Fig. 2). Moreover, we find that workers in the 'Order' condition tend to over-estimate the truthfulness of the statements (i.e., the truthfulness level that workers assigned is higher than the ground truth level, see bottom-right in Fig. 3).

[2]Based on the task time in a pilot study, we set the task reward to $1.5.

Table 1: Numbers of workers with respect to the consistency between P = {self-reported URLs} and O = {opened URLs}.

Grouping Condition	Worker Group	#Workers
$O \cap P \neq \emptyset$	Adopter	94
$O \cap P = \emptyset \wedge P \neq \emptyset$	Non-compliant	102
$O = \emptyset \wedge P = \emptyset$	Independent	3
$O \neq \emptyset \wedge P = \emptyset$	Skeptic	1

4.2 Interaction with the Provided Evidence (RQ#2)

Next, we report how crowd workers make use of the provided evidence (i.e., URLs). To this end, we consider 2 types of data:

- The URLs that workers reported as useful (i.e., by ticking up-voting buttons, see Fig. 1). We define these URLs as set P; and
- The URLs that workers have opened and visited. We define these URLs as set O.

This allows us to differentiate between workers based on how they interact with the provided evidence (e.g., if they report URLs being useful consistently with those they have opened).

Table 1 presents different worker groups with respect to the overlap between the opened URLs and the URLs voted as useful.[3] It is evident that most workers (94 + 102 = 196 out of 200) have reported some URLs being useful (i.e., $P \neq \emptyset$). Among them, only the 'adopter' group used the presented URLs to facilitate their judgements. By contrast, the useful URLs reported by the workers in the 'non-compliant' group are not to be trusted, as there is no overlap between the URLs they up-voted and those they actually opened (i.e., $O \cap P \neq \emptyset$). In fact, adopters shows a higher judging quality compared with non-compliant ones (Welch's t-test, $p < 0.01$). This shows that the provided evidence can play a positive role for crowd workers to assess misinformation, as long as they make use of this information. We also investigate the ratio of the opened and reported URLs to all the reported URLs, defined as $\kappa = \frac{|O \cap P|}{|P|} \in (0, 1]$. However, we do not observe any correlation between κ and α, nor between judgment quality and the number of URLs that are presented in the task.

[3]Note that URLs are not shown in the 'No-URL' condition (with 40 workers).

4.3 Political Impact (RQ#3)

In the condition 'Political', we split URLs over two tables corresponding to those provided by Democrats or Republicans. The number of URLs provided by Democrats, however, is more than those given by Republicans. In some extreme cases, there may be no available Republican URLs for some statements. Therefore, we distinguish between statements if they come with balanced URLs given by the two political parties, as shown below.

- *DEM-major*. The statements have at least 3 more URLs from Democrats than those given by Republicans.
- *DEM-REP*. The statements in this group have a balanced number of URLs between the two parties. That is, the difference of the number of URLs from workers supporting the two parties is less than 3 in all cases.

There is no statement having at least 3 more URLs from Republicans than those given by Democrats. We thus obtain 23 'DEM-major' statements and 97 'DEM-REP' statements. While Republican workers does not show a statistically significant difference from the Democrats on judgement quality for both statement groups, workers with different political leanings show different patterns when using the available URLs. Figure 4 shows URL preference with a breakdown over workers' self-declared political leanings. We normalize values by dividing the number of reported URLs in each group by all URLs with the same political leaning. For example, the median preference of Democratic URLs voted by Democrats in the 'DEM-REP' group is 80% (see left in Fig. 4).

Evidently, workers tend to select URLs that align to their political standings in both groups (Mann–Whitney U test, $p < 0.01$) (see left, middle, right in Fig. 4). As expected, for the balanced DEM-REP group of statements, Republican supporters select fewer Democrats URLs than Democrats workers, and the other way around (see Fig. 4 left and middle plots). These findings show that crowd workers have a strong preference for selecting URLs that align with their ideology, when political information is accessible.

5 DISCUSSION AND CONCLUSIONS

In this paper, we study how presenting peer-generated evidence has an impact on crowd workers judging statement veracity. Our results addressing RQ#1 show how providing evidence from others makes crowd workers less judgemental ending up trusting false statements. We find that when either presenting more information about the available evidence or providing workers with evidence source popularity information, their judging quality is significantly worse than those without such information. This calls for further study on how to balance between the volume and usefulness of the information provided to assist workers in recognising misinformation (i.e., distinguishing false from true statements). With respect to RQ#2, we observed that few workers leveraged the information provided via the URLs, but those who did could benefit from it in their truthfulness judgments. The judgments provided by those who indeed visited the URLs that they have up-voted are significantly better than those who just randomly report URLs as useful. This sheds light on the need to encourage workers to make better use of the provided evidence while judging statements. On the other hand, checking the overlap between the up-voted URLs and workers' behaviour (e.g., visiting the up-voted URLs) may be used as a quality control method, which is also in line with existing research showing that only certain combinations of crowd answers make sense [7]. As an answer to RQ#3, we observed that workers tend to prefer evidence which is aligned to their political thinking. These results indicate a potential for using evidence from others in supporting truthfulness judgements in crowdsourcing, but also suggest that a more in-depth investigation is required to identify the optimal way to present such information.

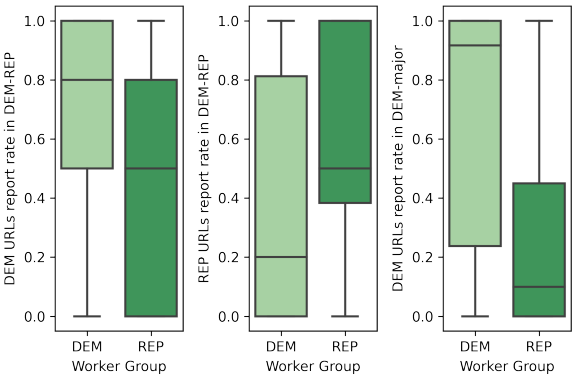

Figure 4: URLs up-vote rate over worker groups.

Acknowledgments. This work is partially supported by an ARC Discovery Project (Grant No. DP190102141), by the ARC Training Centre for Information Resilience (Grant No. IC200100022), and by a Facebook Research award.

REFERENCES

[1] David La Barbera, Kevin Roitero, Gianluca Demartini, Stefano Mizzaro, and Damiano Spina. 2020. Crowdsourcing truthfulness: the impact of judgment scale and assessor bias. In *ECIR*. Springer, 207–214.
[2] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2334–2346.
[3] Shayan Doroudi, Ece Kamar, Emma Brunskill, and Eric Horvitz. 2016. Toward a learning science for complex crowdsourcing tasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2623–2634.
[4] Carsten Eickhoff. 2018. Cognitive Biases in Crowdsourcing. In *Proceedings of WSDM* (Marina Del Rey, CA, USA). ACM, New York, NY, USA, 162–170.
[5] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. 2019. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI*. 1–12.
[6] Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the ACL (Volume 1: Long Papers)*. 1113–1122.
[7] Gabriella Kazai, Jaap Kamps, Marijn Koolen, and Natasa Milic-Frayling. 2011. Crowdsourcing for book search evaluation: impact of hit design on comparative system ranking. In *Proceedings of the 34th international ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 205–214.
[8] Klaus Krippendorff. 2007. Computing Krippendorff's alpha reliability. *Departmental papers (ASC)* (2007), 43.
[9] Eva Méndez and Seth van Hooland. 2014. Metadata typology and metadata uses. In *Handbook of metadata, semantics and ontologies*. World Scientific, 9–39.
[10] Gordon Pennycook and David G Rand. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences* 116, 7 (2019), 2521–2526.
[11] Kevin Roitero, Michael Soprano, Shaoyang Fan, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2020. Can The Crowd Identify Misinformation Objectively? The Effects of Judgment Scale and Assessor's Background. In *Proceedings of the 43rd International ACM SIGIR Conference*. 439–448.
[12] Larysa Visengeriyeva and Ziawasch Abedjan. 2020. Anatomy of metadata for data curation. *Journal of Data and Information Quality (JDIQ)* 12, 3 (2020), 1–30.
[13] William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648* (2017).