

The COVID-19 Infodemic: Can the Crowd Judge Recent Misinformation Objectively?

Kevin Roitero roitero.kevin@spes.uniud.it University of Udine Udine, Italy	Michael Soprano soprano.michael@spes.uniud.it University of Udine Udine, Italy	Beatrice Portelli portelli.beatrice@spes.uniud.it University of Udine Udine, Italy	Damiano Spina damiano.spina@rmit.edu.au RMIT University Melbourne, Australia
Vincenzo Della Mea vincenzo.dellamea@uniud.it University of Udine Udine, Italy	Giuseppe Serra giuseppe.serra@uniud.it University of Udine Udine, Italy	Stefano Mizzaro mizzaro@uniud.it University of Udine Udine, Italy	Gianluca Demartini demartini@acm.org The University of Queensland, Australia

ABSTRACT

Misinformation is an ever increasing problem that is difficult to solve for the research community and has a negative impact on the society at large. Very recently, the problem has been addressed with a crowdsourcing-based approach to scale up labeling efforts: to assess the truthfulness of a statement, instead of relying on a few experts, a crowd of (non-expert) judges is exploited. We follow the same approach to study whether crowdsourcing is an effective and reliable method to assess statements truthfulness during a pandemic. We specifically target statements related to the COVID-19 health emergency, that is still ongoing at the time of the study and has arguably caused an increase of the amount of misinformation that is spreading online (a phenomenon for which the term “infodemic” has been used). By doing so, we are able to address (mis)information that is both related to a sensitive and personal issue like health and very recent as compared to when the judgment is done: two issues that have not been analyzed in related work.

In our experiment, crowd workers are asked to assess the truthfulness of statements, as well as to provide evidence for the assessments as a URL and a text justification. Besides showing that the crowd is able to accurately judge the truthfulness of the statements, we also report results on many different aspects, including: agreement among workers, the effect of different aggregation functions, of scales transformations, and of workers background / bias. We also analyze workers behavior, in terms of queries submitted, URLs found / selected, text justifications, and other behavioral data like clicks and mouse actions collected by means of an ad hoc logger.

ACM Reference Format:

Kevin Roitero, Michael Soprano, Beatrice Portelli, Damiano Spina, Vincenzo Della Mea, Giuseppe Serra, Stefano Mizzaro, and Gianluca Demartini. 2020. The COVID-19 Infodemic: Can the Crowd Judge Recent Misinformation Objectively?. In *The 29th ACM International Conference on Information and*

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-6859-9/20/10...\$15.00

<https://doi.org/10.1145/3340531.3412048>

Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3340531.3412048>

1 INTRODUCTION

“We’re concerned about the levels of rumours and misinformation that are hampering the response. [...] we’re not just fighting an epidemic; we’re fighting an infodemic. Fake news spreads faster and more easily than this virus, and is just as dangerous. That’s why we’re also working with search and media companies like Facebook, Google, Pinterest, Tencent, Twitter, TikTok, YouTube and others to counter the spread of rumours and misinformation. We call on all governments, companies and news organizations to work with us to sound the appropriate level of alarm, without fanning the flames of hysteria.”

These are the alarming words used by Dr. Tedros Adhanom Ghebreyesus, the WHO (World Health Organization) Director General during his speech at the Munich Security Conference on 15 February 2020.¹ It is telling that the WHO Director General chooses to target explicitly misinformation related problems.

Indeed, during the still ongoing COVID-19 health emergency, all of us have experienced mis- and dis-information. The research community has focused on several COVID-19 related issues [4], ranging from machine learning systems aiming to classify statements and claims on the basis of their truthfulness [23], search engines tailored to the COVID-19 related literature, as in the ongoing TREC-COVID Challenge² [26], topic-specific workshops like the NLP COVID workshop at ACL'20,³ and evaluation initiatives like the TREC Health Misinformation Track 2020.⁴ More than the academic research community, commercial social media platforms also have looked at this issue.⁵ Among all the approaches, in some very recent work, Roitero et al. [27], La Barbera et al. [17], Roitero et al. [29] have studied if crowdsourcing can be used to identify misinformation. As it is well known, *crowdsourcing* means to outsource to a large mass of unknown people (the “crowd”), by means

¹<https://www.who.int/dg/speeches/detail/munich-security-conference>

²<https://ir.nist.gov/covidSubmit/>

³<https://www.nlpcovid19workshop.org/>

⁴<https://trec-health-misinfo.github.io/>

⁵<https://www.forbes.com/sites/bernardmarr/2020/03/27/finding-the-truth-about-covid-19-how-facebook-twitter-and-instagram-are-tackling-fake-news/> and <https://spectrum.ieee.org/view-from-the-valley/artificial-intelligence/machine-learning/how-facebook-is-using-ai-to-fight-covid19-misinformation>

of an open call, a task that is usually performed by a few experts. That recent work [17, 27, 29] specifically crowdsource the task of misinformation identification, or rather assessment of the truthfulness of statements made by public figures (e.g., politicians), usually on political, economical, and societal issues. That the crowd is able to identify misinformation might sound implausible at first—isn’t the crowd the very mean to spread misinformation? However, on the basis of recent research [17, 27, 29], it appears that the crowd can provide high quality truthfulness labels, provided that adequate countermeasures and quality assurance techniques are employed.

In this paper we address the very same problem, but focusing on statements about COVID-19. This is motivated by several reasons. First, COVID-19 is of course a hot topic but, although there is a great amount of researchers efforts worldwide devoted to its study, there is no study yet using crowdsourcing to assess truthfulness of COVID-19 related statements. To the best of our knowledge, we are the first to report on crowd assessment of COVID-19 related misinformation. Second, the health domain is particularly sensitive, so it is interesting to understand if the crowdsourcing approach is adequate also in such a particular domain. Third, in the previous work [17, 27, 29] the statements judged by the crowd were not recent. This means that evidence on statement truthfulness was often available out there (on the Web), and although the experimental design prevented to easily find that evidence, it cannot be excluded that the workers did find it, or perhaps they were familiar with the particular statement because, for instance, it had been discussed in the press. By focusing on COVID-19 related statements we instead naturally target *recent* statements: in some cases the evidence might be still out there, but this will happen more rarely. Fourth, an almost ideal tool to address misinformation would be a crowd able to assess truthfulness in real time, immediately after the statement becomes public: although we are not there yet, and there is a long way to go, we find that targeting recent statements is a step forward in the right direction. Fifth, our experimental design differs in some aspects from that used in previous work, and allows us to address novel research questions.

2 BACKGROUND

2.1 COVID-19 Infodemic

The number of initiatives to apply Information Access—and, in general, Artificial Intelligence—techniques to combat the COVID-19 infodemic has been rapidly increasing (see Bullock et al. [4, p. 16] for a survey). There is significant effort on analyzing COVID-19 information on social media, and linking to data from external fact-checking organizations to quantify the spread of misinformation [7, 11, 32]. Mejova and Kalimeri [19] analyzed Facebook advertisements related to COVID-19, and found that around 5% of them contain errors or misinformation. Crowdsourcing methodologies have also been used to collect and analyze data from patients with cancer who are affected by the COVID-19 pandemic [8]. To the best of our knowledge, there is no work addressing the COVID-19 infodemic using crowdsourcing.

2.2 Crowdsourcing Truthfulness

Recent work has focused on the automatic classification of truthfulness or fact checking [2, 9, 15, 20, 22, 24]. Zubiaga and Ji [33]

investigated, using crowdsourcing, the reliability of tweets in the setting of disaster management. CLEF developed a Fact-Checking Lab [9, 22] to address the issue of ranking sentences according to some fact-checking property.

There is recent work that studies how to collect truthfulness judgments by means of crowdsourcing using fine grained scales [17, 27, 29]. Samples of statements from the PolitiFact dataset—originally published by Wang [31]—have been used to analyze the agreement of workers with labels provided by experts in the original dataset. Workers are asked to provide the truthfulness of the selected statements, by means of different fine grained scales. Roitero et al. [27] compared two fine grained scales: one in the $[0, 100]$ range and one in the $(0, +\infty)$ range, on the basis of Magnitude Estimation [21]. They found that both scales allow to collect reliable truthfulness judgments that are in agreement with the ground truth. Furthermore, they show that the scale with one hundred levels leads to slightly higher agreement levels with the expert judgments. On a larger sample of PolitiFact statements, La Barbera et al. [17] asked workers to use the original scale used by the PolitiFact experts and the scale in the $[0, 100]$ range. They found that aggregated judgments (computed using the mean function for both scales) have a high level of agreement with expert judgments. Recent work by Roitero et al. [29] found similar results in terms of external agreement and its improvement when aggregating crowdsourced judgments, using statements from two different fact-checkers: PolitiFact and RMIT ABC Fact Check (ABC). Previous work has also looked at *internal agreement*, i.e., agreement among workers [27, 29]. Roitero et al. [29] found that scales have low levels of agreement when compared with each other: correlation values for aggregated judgments on the different scales are around $\rho = 0.55 - 0.6$ for PolitiFact and $\rho = 0.35 - 0.5$ for ABC, and $\tau = 0.4$ for PolitiFact and $\tau = 0.3$ for ABC. This indicates that the same statements tend to be evaluated differently in different scales.

There is evidence of differences on the way workers provide judgments, influenced by the sources they examine, as well as the impact of worker bias. In terms of sources, La Barbera et al. [17] found that that the vast majority of workers (around 73% for both scales) use indeed the PolitiFact website to provide judgments. Differently from La Barbera et al. [17], Roitero et al. [29] used a custom search engine in order to filter out PolitiFact and ABC from the list of results. Results show that, for all the scales, Wikipedia and news websites are the most popular sources of evidence used by the workers. In terms of worker bias, La Barbera et al. [17] and Roitero et al. [29] found that worker political background has an impact on how workers provide the truthfulness scores. More in detail, they found that workers are more tolerant and moderate when judging statements from their very own political party.

3 AIMS AND RESEARCH QUESTIONS

When compared to previous work, in this paper we aim to focus on several novel aspects. With respect to La Barbera et al. [17], Roitero et al. [27, 29], we focus on claims about COVID-19, which are recent and interesting for the research community, and are arguably on a more relevant/sensitive topic to the workers. We investigate whether the health domain makes a difference in the ability of

crowd workers to identify and correctly classify (mis)information, and if the very recent nature of COVID-19 related statements has an impact as well. We focus on a single truthfulness scale, given the evidence that the scale used does not make a significant difference [17, 27, 29]. Another important difference is that we ask workers to provide a textual justification for their decision: we analyze them to better understand the process followed by workers to verify information, and we investigate if they can be exploited to derive useful information. Finally, we also exploit and analyze worker behavior.

We investigate the following specific Research Questions:

- RQ1 Are the crowd workers able to detect and objectively categorize online (mis)information related to the medical domain and more specifically to COVID-19? Which are the relationship and agreement between the crowd and the expert labels?
- RQ2 Can the crowdsourced and/or the expert judgments be transformed or aggregated in a way that it improves the ability of workers to detect and objectively categorize online (mis)information?
- RQ3 Which is the effect of workers' political bias in objectively identifying online misinformation? And the effect of workers' background and Cognitive Reflection Test (CRT) performances?
- RQ4 Which are the signals provided by the workers while performing the task that can be recorded? To what extent are these signals related to workers' accuracy? Can these signals be exploited to improve accuracy and, for instance, aggregate the labels in a more effective way?
- RQ5 Which sources of information does the crowd consider when identifying online misinformation? Are some sources more useful? Do some sources lead to more accurate and reliable assessments by the workers?

4 METHODS

In this section we present the dataset used to carry out our experiments (Section 4.1), and the crowdsourcing task design (Section 4.2). Overall, we considered one dataset annotated by experts, one crowdsourced dataset, one judgment scale (the same for the expert and the crowd judgments), and a total of 60 statements.

4.1 Dataset

We considered as primary source of information the PolitiFact dataset [31] that was built as a “benchmark dataset for fake news detection” [31] and contains over 12k statements produced by public appearances of US politicians. The statements of the datasets are labeled by expert judges on a six-level scale of truthfulness (from now on referred to as E_6): pants-on-fire, false, mostly-false, half-true, mostly-true, and true. Recently, the PolitiFact website (the source from where the statements of the PolitiFact dataset are taken) created a specific section related to the COVID-19 pandemic.⁶ For this work, we selected 10 statements for each of the six PolitiFact categories, belonging to such a COVID-19 section and with dates ranging from February 2020 to early April 2020. Table 1 contains some examples of the statements we used.

⁶<https://www.politifact.com/coronavirus/>

Table 1: Examples of COVID-19 fact-checked statements.

Statement	Source	Year	Label
“We inherited a broken test for COVID-19.”	Donald Trump	2020	pants-on-fire
“Church services cannot resume until we are all vaccinated, says Bill Gates.”	Bloggers	2020	false
“Says a 5G law passed while everyone was distracted with the coronavirus pandemic and lists 20 symptoms associated with 5G exposure.”	Facebook Post	2020	mostly-false
“Says a California surfer was alone, in the ocean, when he was arrested for violating the state’s stay-at-home order.”	Facebook Post	2020	mostly-true
“Photo shows a crowded New York City subway train during stay-at-home order.”	Viral Image	2020	true

4.2 Crowdsourcing Experimental Setup

To collect our judgments we used the crowdsourcing platform Amazon Mechanical Turk (MTurk). Each worker, upon accepting our HIT, is redirected to an external server to complete the task; we set the payment to \$1.5 for a set of 8 statements⁷. The task itself is as follows: first, a (mandatory) questionnaire is shown to the worker, to collect his/her background information such as age and political views. Then, the worker needs to provide answers to three Cognitive Reflection Test (CRT) questions, which are used to measure the personal tendency to answer with an incorrect “gut” response or engage in further thinking to find the correct answer [10].⁸ After the questionnaire and CRT phase, the worker is asked to assess the truthfulness of 8 statements: 6 from the dataset described in 4.1 (one for each of the six considered PolitiFact categories) and 2 special statements called *Gold Questions*, one clearly true and the other clearly false, manually written by the authors of this paper and used as quality checks. We used a randomization process when building the HITs to avoid all the possible source of bias, both within each HIT and considering the overall task.

To assess the truthfulness of each statement, the worker is shown: the *Statement*, the *Speaker/Source*, and the *Year* in which the statement was made. We asked the worker to provide the following information: the *truthfulness value* for the statement using the six-level scale adopted by PolitiFact, from now on referred to as C_6 (presented to the worker using a radio button containing the label description for each category as reported in the original PolitiFact website), a *URL* that s/he used as a source of information for the fact checking, and a textual *motivation* for her/his response (which can not include the URL, and should contain at least 15 words). In order to prevent the user from using PolitiFact as primary source of evidence, we implemented a custom search engine, which is based on the Bing Web Search APIs⁹ and filters out PolitiFact from the returned search results.

We logged the user behavior using a custom logger [12, 13], and we implemented in the task the following quality checks: (i) the judgments assigned to the gold questions have to be coherent (i.e., the judgment of the clearly false question should be lower than the

⁷Before deploying the task on MTurk, we investigated the average time spent to complete the task, and we related it to the minimum US hourly wage.

⁸We used the same CRT settings as Roitero et al. [29].

⁹<https://azure.microsoft.com/services/cognitive-services/bing-web-search-api/>

one assigned to true question); and (ii) the cumulative time spent to perform each judgment should be of at least 10 seconds. Note that the CRT (and the questionnaire) answers were not used for quality check, although the workers were not aware of that.

Overall, we used 60 statements in total (10 for each PolitiFact category), and each statement has been evaluated by 10 distinct workers. Thus, we deployed 100 MTurk HITs and we collected 800 judgments in total. The crowd task was launched on May 1st, 2020 and it completed on May 4th, 2020. All the data used to carry out our experiments can be downloaded at <https://github.com/KevinRoitero/crowdsourcingTruthfulness>.

5 RESULTS AND ANALYSIS

We first report some descriptive statistics about the population of workers and the data collected in our experiment (Section 5.1). Then, we address crowd accuracy (i.e., RQ1) in Section 5.2, transformation of truthfulness scales (RQ2) in Section 5.3, worker background and bias (RQ3) in Section 5.4, worker behavior (RQ4) in Section 5.5; finally, we study information sources (RQ5) in Section 5.6.

5.1 Descriptive Statistics

5.1.1 Worker Background, Abandonment, and Bias.

Questionnaire. Overall, 1113 workers resident in the United States participated in our experiment.¹⁰ In each HIT, workers were first asked to complete a demographics questionnaire with questions about their gender, age, education and political views. By analyzing the answers to the questionnaire we derived the following demographic statistics. The majority of workers are in the 26–35 age range (44%), followed by 36–50 (25%), and 19–25 (18%). The majority of workers are well educated: 47% of them have a four year college degree or a bachelor degree, 21% have a college degree, and 17% have a postgraduate or professional degree. Only about 15% of workers have a high school degree or less. Concerning political views, we had 28% of workers that identified themselves as liberals, 28% as moderate, 24% as conservative, 11% as very conservative and 9% as very liberal. Moreover, 44% of workers identified themselves as being Democrat, 31% as being Republican, and 22% as being Independent. Finally, 46% of workers agree on building a wall on the southern US border, and 42% of them disagree. Overall we can say that our sample is well balanced.

CRT Test. Analyzing the CRT scores, we found that: 31% of workers did not provide any correct answer, 34% answered correctly to 1 test question, 18% answered correctly to 2 test questions, and only 17% answered correctly to all 3 test questions.

Abandonment. When considering the abandonment ratio (measured according to the definition provided by Han et al. [12], Han et al. [13]), we found that 100 of the workers (about 9%) successfully completed the task, 991 (about 87%) abandoned (i.e., voluntarily terminated the task before completing it), and 45 (about 4%) failed (i.e., terminated the task due to failing the quality checks too many times). Most of the abandonment (80% of the 1091 workers, 85% of the 991 workers that abandoned) happened before judging the first statement (i.e., before really starting the task); about 7.52% of

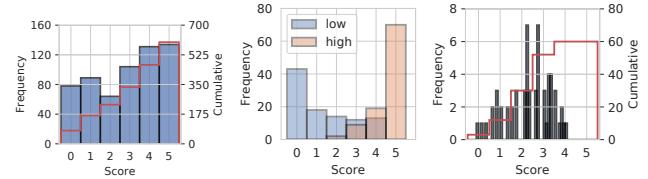


Figure 1: Distribution (in blue) and the cumulative distribution (in red) of the individual (left), gold (middle), and aggregated with mean (right).

the 1091 workers (8% of the 991 of the workers that abandoned) abandoned after the last statement (most likely once failed the quality check).

5.1.2 Crowdsourced Scores. Figure 1 shows the distribution (in blue) and the cumulative distribution (in red) of the individual (left), gold (middle), and aggregated with mean (right) scores provided by the workers for the considered PolitiFact statements.

If we focus on the distribution of the individual scores (left plot), we can see that the distribution is quite well balanced, just lightly skewed towards higher truthfulness values, represented in the right-most part of the plot. This behavior is also remarked when focusing on the red line representing the cumulative distribution, which displays almost evenly spaced steps. This is a first indication that suggests that crowd judgments are overall of a decent quality; in fact, our empirical distribution is not distant from the ideal one: since we considered 10 statements for each PolitiFact category, the perfect distribution would have been the uniform distribution.

Turning to the distribution of the gold scores (i.e., the two special statements used for quality check, shown in the middle plot), we see that the large majority of workers (i.e., 70% for High and 43% for Low) used the extreme values of the scale (i.e., pants-on-fire and true); furthermore, we see that overall the High gold question has been judged correctly more times than the Low gold question, suggesting the probably the workers found the former easier to judge than the latter.

We now turn to analyze the distribution of the scores when aggregated using the mean function (shown in the right plot). The distribution for the aggregated scores becomes roughly bell-shaped, and slightly skewed towards high truthfulness values—this behavior is consistent with the findings of Roitero et al. [29]. In the following we discuss both the external (i.e., between workers and experts) and internal (i.e., among workers) agreement of our dataset.

5.2 RQ1: Crowd Accuracy

5.2.1 External Agreement. To answer RQ1, we start by analyzing the so called external agreement, i.e., the agreement between the crowd collected labels and the experts ground truth. Figure 2 shows the agreement between the PolitiFact experts (x-axis) and the crowd judgments (y-axis). In the first plot, each point is a judgment by a worker on a statement, i.e., there is no aggregation of the workers working on the same statement. In the next plots all workers redundantly working on the same statement are aggregated using the mean (second plot), median (third plot), and majority vote (right-most plot). If we focus on the first plot (i.e., the one

¹⁰Workers provide proof that they are based in US and have the eligibility to work.

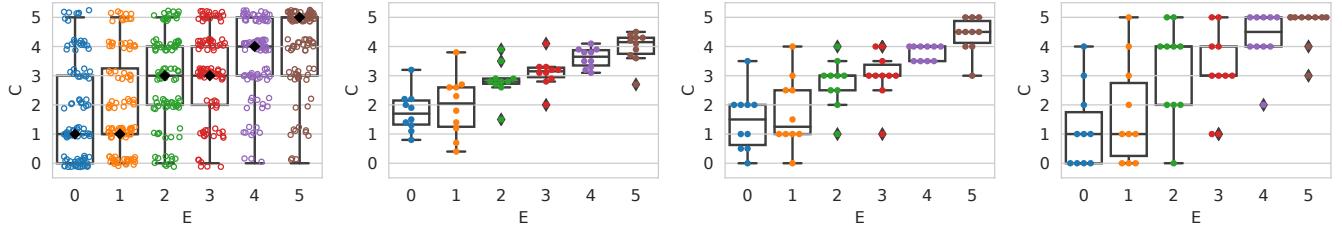


Figure 2: The agreement between the PolitiFact experts (x-axis) and the crowd judgments (y-axis). From left to right: C₆ individual judgments; C₆ aggregated with mean; C₆ aggregated with median; C₆ aggregated with majority vote.

with no aggregation function applied), we can see that, overall, the individual judgments are in agreement with the expert labels, as shown by the median values of the boxplots, which are increasing as the ground truth truthfulness level increases. Concerning the aggregated values, it is the case that for all the aggregation functions the pants-on-fire and false categories are perceived in a very similar way by the workers; this behavior was already shown in previous work [17, 29], and suggests that indeed workers have clear difficulties in distinguishing between the two categories; this is even more evident considering that the interface presented to the workers contained a textual description of the categories’ meaning in every page of the task.

If we look at the plots as a whole, we see that within each plot the median values of the boxplots are increasing when going from pants-on-fire to true (i.e., going from left to right of the x-axis of each chart). This indicates that the workers are overall in agreement with the PolitiFact ground truth, thus indicating that workers are indeed capable of recognizing and correctly classifying misinformation statements related to the COVID-19 pandemic. This is a very important and not obvious result: in fact, the crowd (i.e., the workers) is the primary source and cause of the spread of disinformation and misinformation statements across social media platforms [6]. By looking at the plots, and in particular focusing on the median values of the boxplots, it appears evident that the mean (second plot) is the aggregation function which leads to higher agreement levels, followed by the median (third plot) and the majority vote (fourth plot). Again, this behavior has already been noticed [17, 28, 29], and all the cited works used the mean as primary aggregation function.

To validate the external agreement, we measured the statistical significance between the aggregated rating for all the six PolitiFact categories; we considered both the Mann-Whitney rank test and the t-test, applying Bonferroni correction to account for multiple comparisons. Results are as follows: when considering adjacent categories (e.g., pants-on-fire and false), the difference between categories are never significant, for both tests and for all the three aggregation functions. When considering categories of distance 2 (e.g., pants-on-fire and mostly-false), the differences are never significant, apart from the median aggregation function, where there is statistical significance to the $p < .05$ level in 2/4 cases for both Mann-Whitney and t-test. When considering categories of distance 3, the differences are significant, for the mean, in 3/3 cases for the Mann-Whitney and 3/3 cases for the t-test, for the median, in 2/3 cases for the Mann-Whitney and 3/3 cases for the t-test, for the majority vote, in 0/3 cases for the Mann-Whitney and

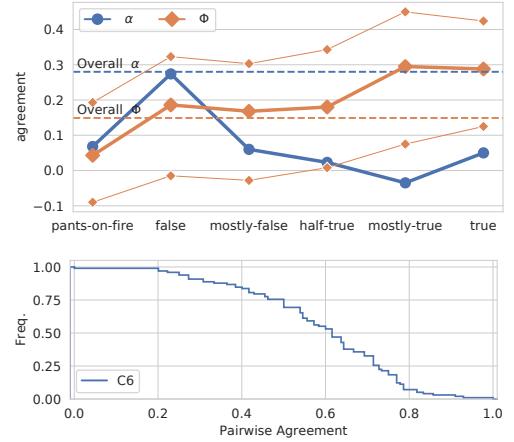


Figure 3: Workers agreement: α [16] and Φ [5] (top plot); pairwise unit agreement (bottom plot).

1/3 cases for the t-test. When considering categories of distance 4 and 5, the differences are always significant to the $p > 0.01$ level for all the aggregation functions and for all the tests, apart from the case of the majority vote function and the Mann-Whitney test, where the significance is at the $p > .05$ level. In the following we use the mean as being the most commonly used approach for this type of data [29].

5.2.2 Internal Agreement. Another standard way to address RQ1 and to analyze the quality of the work by the crowd is to compute the so called internal agreement (i.e., the agreement among the workers). Figure 3 shows in the first plot the agreement measured with α [16] and Φ [5], two popular measures often used to compute workers’ agreement in crowdsourcing tasks [18, 27–29]. The x-axis details the PolitiFact categories, while the y-axis the level of agreement measured; while α is a punctual measure, Φ allows to compute confidence intervals for the agreement measure; the plot shows the upper 97% and lower 3% confidence intervals as thinner lines. As we can see from the plot, the agreement levels measured with the two scales is very similar for the pants-on-fire, false, mostly-false, and half-true categories: note that the α measure always falls in the Φ confidence interval, and the little oscillations in the agreement value are not always indication of a real change in the agreement level, especially when considering α [5]. Having said

so, it appears that for all the two metrics the overall agreement falls in the $[0.15, 0.3]$ range, and the agreement level is similar for all the PolitiFact categories, with the exception of Φ , which shows higher agreement levels for the mostly-true and true categories. This confirms the finding, derived from Figure 2, that workers seem most effective in identifying and categorizing statements with a higher truthfulness level. This remark is also supported by [5] which shows that Φ is better in distinguishing agreement levels in crowdsourcing than α , which is more indicated as a measure of data reliability in non crowdsourced settings.

Figure 3 also shows in the second plot a measure of the agreement at the HIT level (i.e., in the set of 8 statements judged by each worker) as detailed in [18, 29]. More in detail, the plot shows the CCDF (Complementary Cumulative Distribution Function) of the relative frequencies for the agreement of the 100 HITs considered in this experiment. The plot shows that around 20% of the hits have a pairwise agreement which is very close to 1; this indicates that around 20% of the workers judged statements almost in the same way as the expert judges. Moreover, we see that 60% of the workers have a pairwise agreement greater than 0.5. Again, this result indicates a good overall agreement between crowd and expert judgments, confirming that the crowd is able to correctly identify and classify misinformation related to the COVID-19 pandemic.

5.3 RQ2: Transforming Truthfulness Scales

Given the positive results presented above, it appears that the answer to RQ1 is overall positive, even if with some exceptions. There are many remarks that can be made: first, there is a clear issue that affects the pants-on-fire and false categories, which are very often mis-classified by workers. Moreover, while PolitiFact used a six-level judgment scale, the usage of a two- (e.g., True/False) and a three-level (e.g., False / In between / True) scale is also common when assessing the truthfulness of statements [17, 29]. Finally, categories can be merged together to improve accuracy, as done for example by Tchechmedjiev et al. [30]. All these considerations lead us to RQ2, addressed in the following.

5.3.1 Merging Ground Truth Levels. For all the above reasons, we performed the following experiment: we group together the six PolitiFact categories (i.e., E_6) into three (referred to as E_3) or two (E_2) categories, which we refer respectively with 01, 23, and 45 for the three level scale, and 012 and 234 for the two level scale.

Figure 4 shows the result of such a process. As we can see from the plots, the agreement between the crowd and the expert judgments can be seen in a more neat way. As for Figure 2, the median values for all the boxplots is increasing when going towards higher truthfulness values (i.e., going from left to right within each plot); this holds for all the aggregation functions considered, and it is valid for both transformations of the E_6 scale, into two and three levels. Also in this case we computed the statistical significance between categories, applying the Bonferroni correction to account for multiple comparisons. Results are as follows. For the case of three groups, both the categories at distance one and two are always significant to the $p < 0.01$ level, for both the Mann-Whitney and the t-test, for all three aggregation functions. The same behavior holds for the case of two groups, where the categories of distance 1 are always significant to the $p < 0.01$ level.

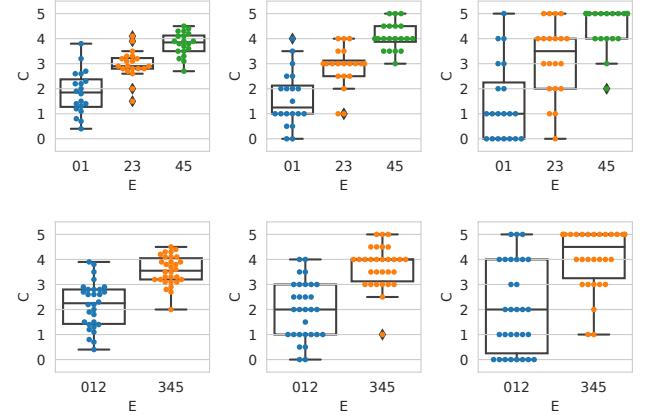


Figure 4: The agreement between the PolitiFact experts and the crowd judgments. From left to right: C_6 aggregated with mean; C_6 aggregated with median; C_6 aggregated with majority vote. First row: E_6 to E_3 ; second row: E_6 to E_2 . Compare with Figure 2.

Summarizing, we can now conclude that by merging the ground truth levels we obtained a much stronger signal: the crowd can effectively detect and classify misinformation statements related to the COVID-19 pandemic.

5.3.2 Merging Crowd Levels. Having reported the results on merging the ground truth categories we now turn to transform the crowd labels (i.e., C_6) into three (referred to as C_3) and two (C_2) categories. For the transformation process we rely on the approach detailed by Han et al. [14], that also present a complete and exhaustive discussion on the effectiveness of the scale transformation methods. This approach has many advantages [14]: we can simulate the effect of having the crowd answers in a more coarse-grained scale (rather than C_6), and thus we can simulate new data without running the whole experiment on MTurk again. As we did for the ground truth scale, we choose to select as target scales the two- and three- levels scale, driven by the same motivations. Having selected C_6 as being the source scale, and having selected the target scales as the three- and two- level ones (C_3 and C_2), we perform the following experiment. We perform all the possible cuts¹¹ from C_6 to C_3 and from C_6 to C_2 ; then, we measure the internal agreement (using α and Φ) both on the source and on the target scale, and we compare those values. In such a way, we are able to identify, among all the possible cuts, the cut which leads to the highest possible internal agreement. Also in this case, a detailed discussion on the relationships between internal agreement, effectiveness, and all the possible cuts can be found in Han et al. [14].

Figure 5 shows the results. The x-axis shows the cut performed to transform C_6 into the target scale (C_3 in the left-most plot, C_2 in the right-most plot), while the y-axis shows the internal agreement score by means of either α or Φ . As we can see by inspecting the left-most plot, (i.e., C_6 to C_3) we can see that there is, both for α

¹¹ C_6 can be transformed into C_3 in 10 different ways, and C_6 can be transformed into C_2 in 5 different ways.

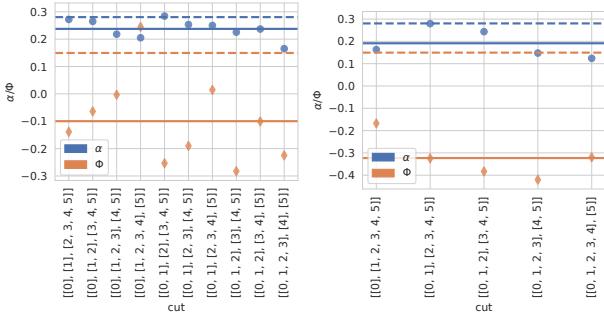


Figure 5: α and Φ cuts. From left to right: C_6 cut into three levels (C_3), C_6 cut into two levels (C_2). The cut is detailed in the x-label. The dotted line is α / Φ on C_6 , the continuous line is the average α / Φ score measured among all the cuts.

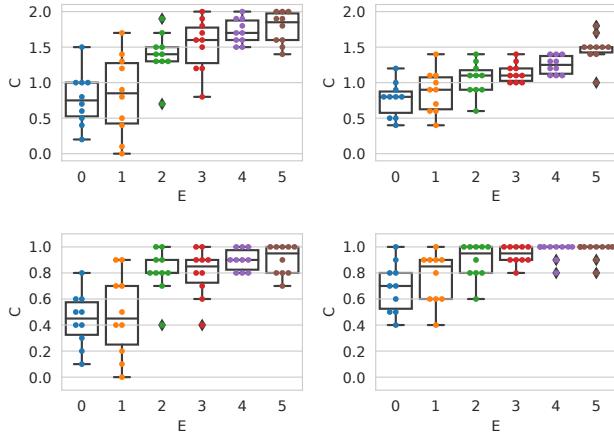


Figure 6: Comparison with E_6 . C_6 to C_3 (first row) and to C_2 (second row), then aggregated with the mean function. Best cut selected according to α (left column) and Φ (right column) (see Figure 5). Compare with Figure 2.

and Φ , a single cut which leads to higher agreement levels with the original C_6 scale. On the contrary, if we focus on the rightmost plot (i.e., C_6 to C_2), we can see that there is a single cut for α which leads to similar agreement levels as in the original C_6 scale, and there are no cuts with such a property when using Φ .

Having identified the best possible cuts for both transformations and for both agreement metrics, we now measure the external agreement between the crowd and the expert judgments, using the selected cut. Figure 6 shows such a result when considering the judgments aggregated with the mean function. As we can see from the plots, it is again the case that the median values of the boxplots is always increasing, for all the transformations. Nevertheless, inspecting the plots we can state that the overall external agreement appears to be lower than the one shown in Figure 2. Moreover, we can also state that also in the case of the transformed scales, the categories pants-on-fire and false are still not separable. Summarizing, we show that it is feasible to transform the

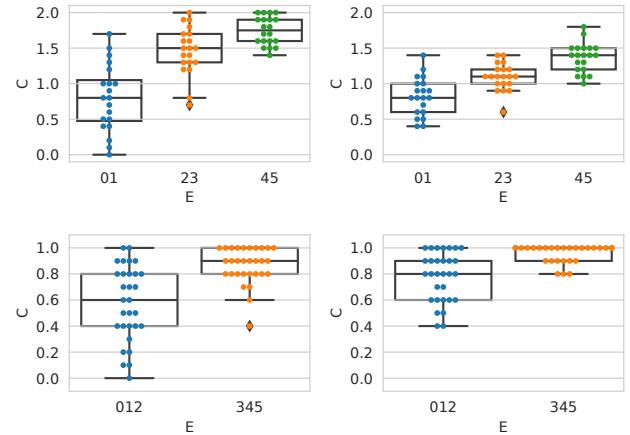


Figure 7: C_6 to C_3 (first row) and to C_2 (second row), then aggregated with the mean function. First row: E_6 to E_3 . Second row: E_6 to E_2 . Best cut selected according to α (left column) and Φ (right column) (see Figure 5). Compare with Figures 2, 4, and 6.

judgments collected on a C_6 level scale into two new scales, C_3 and C_2 , obtaining judgments with a similar internal agreement as the original ones, and with a slightly lower external agreement with the expert judgments.

5.3.3 Merging both Ground Truth and Crowd Levels. It is now natural to combine the two approaches. Figure 7 shows the comparison between C_6 transformed into C_3 and C_2 , and E_6 transformed into E_3 and E_2 . As we can see from the plots, also in this case the median values of the boxplots are increasing, especially for the E_3 case (shown in the first row). Furthermore, the external agreement with the ground truth is present, even if for the E_2 case (shown in the second row) the classes appear to be not separable. Summarizing, all these results show that it is feasible to successfully combine the aforementioned approaches, and transform into a three- and two-level scale both the crowd and the expert judgments.

5.4 RQ3: Worker Background and Bias

To address RQ3 we study if the answers to questionnaire and CRT test have any relation to worker quality.

5.4.1 Questionnaire. Table 2 (top) shows in the rows the answer to the workers political views, while on the columns the number of correctly classified statements (columns, max is 6). As we can see from the table, there is only one worker who successfully classified all 6 statements. Many workers correctly classified 1 or 2 statements (28 and 28, respectively). The next column summarizes, using Accuracy (i.e., the fraction of exactly classified statements), the quality of workers in each group. The number and fraction of correctly classified statements are however rather crude measures of worker's quality, as small misclassification errors (e.g., pants-on-fire in place of false) are as important as more striking ones (e.g., pants-on-fire in place of true). Therefore, to measure the ability of workers to

Table 2: Count of the number of workers depending on: number of statements correctly classified (columns, max is 6), vs. (top table) the answers to the Political views question (rows) and vs. (bottom table) the number of correct answers to the CRT test (rows, max is 3). The last two columns show the Accuracy (the fraction of correctly identified statements for each group) and the CEM^{ORD} score.

		Correctly classified statements						Acc	CEM ^{ORD} Mean	
		0	1	2	3	4	5	6	Sum	
Very conservative		4	3	1	0	0	0	1	9	.13
Conservative		0	9	2	3	1	0	0	15	.21
Moderate		6	6	6	7	0	1	0	26	.20
Liberal		2	8	13	4	4	2	0	33	.16
Very Liberal		0	2	6	6	2	1	0	17	.21
Sum		12	28	28	20	7	4	1	100	
		Correctly classified statements						Acc	CEM ^{ORD} Mean	
		0	1	2	3	4	5	6	Sum	
CRT	0	5	11	9	4	0	1	1	31	.14
correct	1	5	10	12	6	1	0	0	34	.22
answers	2	1	6	1	6	3	1	0	18	.21
	3	1	1	6	4	3	2	0	17	.15
Sum		12	28	28	20	7	4	1	100	

correctly classify the statements, we also compute CEM^{ORD}, an effectiveness metric recently proposed for the specific case of ordinal classification [1] (see Roitero et al. [29, §3.3] for a more detailed discussion of these issues). The last column in the table shows the average CEM^{ORD} value for the workers in each group. By looking at both Accuracy and CEM^{ORD}, it is clear that ‘Very conservative’ workers provide lower quality labels. The Bonferroni corrected two tailed t-test on CEM^{ORD} confirms that ‘Very conservative’ workers perform statistically significantly worse than both ‘Conservative’ and ‘Very liberal’ workers. The workers’ political views affect the CEM^{ORD} score, even if in a small way and mainly when considering the extremes of the scale. An initial analysis of the other answers to the questionnaire (not shown due to space limitations) does not seem to provide strong signals; a more detailed analysis is left for future work.

5.4.2 CRT Test. We now investigate the effect of the CRT test on the worker quality. Table 2 (bottom) shows the count of the number of workers depending on: number of statements correctly classified (columns, max is 6), versus the number of correct answers to the CRT test (rows, max is 3). Concerning CRT scores, we see that the minority of workers (17) answered in a correct way to all the three questions, and the majority of them answered correctly to only 1 CRT question (34) or none (31). Although there is some variation in both Accuracy and CEM^{ORD}, this is never statistically significant; it appears that the number of correct answers to the CRT test is not correlated with worker quality. We leave for future work a more detailed study of this aspect.

5.5 RQ4: Worker Behavior

We now turn to RQ4, and analyze the behavior of the workers while performing the task.

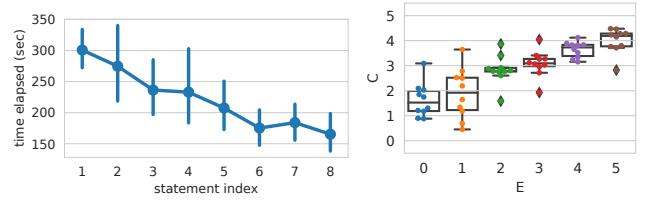


Figure 8: Position of the statement in the task vs. time elapsed, cumulative on each single statement (left). Comparison between E₆ and C₆ where the aggregation function is the weighted mean and the weights are the political views (see Table 2 top) normalized to [0.5, 1] (right).

5.5.1 Time. Figure 8 (left) shows that the amount of time spent on average by the workers on the first statements is considerably higher than on the last statements. This, combined with the fact that the quality of the assessment provided by the workers does not decrease for the last statements (CEM^{ORD} scores per position are 1: .61, 2: .60, 3: .64, 4: .58, 5: .59, 6: .54, 7: .61, 8: .62), is an indication of a learning effect: the workers learn how to assess truthfulness in a faster way.

5.5.2 Exploiting Worker Signals to Improve Quality. We have shown that, while performing their task, workers provide many signals that to some extent correlate with the quality of their work. These signals could in principle be exploited to aggregate the individual judgments in a more effective way (i.e., giving more weight to workers that possess features indicating a higher quality). For example, the relationships between worker background / bias and worker quality (Section 5.4) could be exploited to this aim.

We thus performed the following experiment: we aggregated C₆ individual scores, using as aggregation function a weighted mean, where the weights are represented by the political views, normalized to [0.5, 1]. Figure 8 (right) shows the results. We also aggregated C₆ individual scores using as aggregation function the weighted mean function where the weights are represented by the number of correct answers to CRT, normalized to [0.5, 1], which lead to similar results. Thus, it seems that leveraging quality-related behavioral signals, like questionnaire answers or CRT scores, to aggregate results does not provide a noticeable increase in the external agreement, although it does not harm. We have only scratched the surface, though, as there are many other signals, and aggregation functions, that can be tried; we leave for future work the in depth analysis of how such behavioral signals can be leveraged to improve external agreement.

5.5.3 Queries. Table 3 shows query statistics for the 100 workers which finished the task. As we can see, the higher the statement position, the lower the number of queries issued: 3.52 queries on average for the first statement, down to 2.3 for the last statement. This can indicate the attitude of workers to issue fewer queries the more time they spend on the task, probably due to fatigue, boredom, or learning effects. Nevertheless, we can see that on average, for all the statement positions each worker issues more than one query, i.e., workers often reformulate their initial query. This provides further evidence that they put effort in performing the task. The

Table 3: Statement position in the task versus: number of queries issued (top) and number of times the statement has been used as a query (bottom).

Statement Position	1	2	3	4	5	6	7	8	Sum	Mean
Number of Queries	352	280	259	255	242	238	230	230	2095	261.9
Statement as Query	22	32	31	33	34	30	29	34	245	30.6

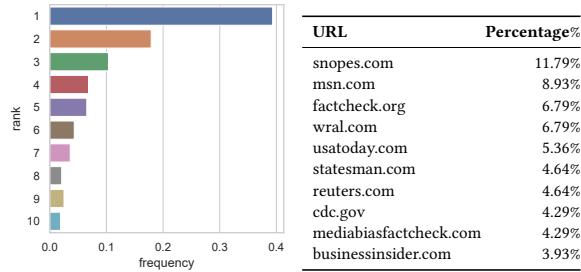


Figure 9: On the left, distribution of the ranks of the URLs selected by workers, on the right, websites from which workers chose URLs to justify their judgments.

third row of the table shows the number of times the worker used as query the whole statement. We can see that the percentage is rather low (around 13%) for all the statement positions, indicating again that workers spend effort when providing their judgments.

5.6 RQ5: Sources of Information

5.6.1 URL Analysis. Figure 9 shows on the left the distribution of the ranks of the URL selected as evidence by the worker when performing each judgment. URLs selected less than 1% times are filtered out from the results. As we can see from the plot, about 40% of workers selected the first result retrieved by our search engine, and selected the remaining positions less frequent, with an almost monotonic decreasing frequency (rank 8 makes the exception). We also found that 14% of workers inspected up to the fourth page of results (i.e., rank= 40). The breakdown on the truthfulness PolitiFact categories does not show any significant difference.

Figure 9 shows on the right part the top 10 of websites from which the workers choose the URL to justify their judgments. Websites with percentage $\leq 3.9\%$ are filtered out. As we can see from the table, there are many fact check websites among the top 10 URLs (e.g., snopes: 11.79%, factcheck 6.79%). Furthermore, medical websites are present, although in small percentage (cdc: 4.29%). This indicates that workers use various kind of sources as URLs from which they take information. Thus, it appears that they put effort in finding evidence to provide a reliable truthfulness judgment.

5.6.2 Justifications. As a final result, we analyze the textual justifications provided, their relations with the web pages at the selected URLs, and their links with worker quality. 54% of the provided justifications contain text copied from the web page at the URL selected for evidence, while 46% do not. Furthermore, 48% of the

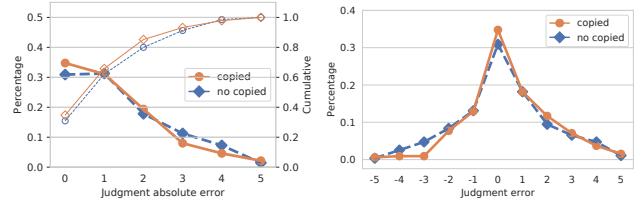


Figure 10: Effect of the origin of a justification (text copied/not copied from the URL selected) on: the absolute value of the prediction error (left; cumulative distributions shown with thinner lines and empty markers), and the prediction error (right).

justification include some “free text” (i.e., text generated and written by the worker), and 52% do not. Considering all the possible combinations, 6% of the justifications used both free text and text from web page, 42% used free text but no text from the web page, 48% used no free text but only text from web page, and finally 4% used neither free text nor text from web page, and either inserted text from a page of a different (not selected) web page or inserted part of the instructions we provided or text from the user interface.

Concerning the preferred way to provide justifications, each worker seems to have a clear attitude: 48% of the workers used only text copied from the selected web pages, 46% of the workers used only free text, 4% used both, and 2% of them consistently provided text coming from the user interface or random web pages.

We now correlate such a behavior with the workers quality. Figure 10 shows the relations between different kinds of justifications and the worker accuracy. The plots show the absolute value of the prediction error on the left, and the prediction error on the right. The lines in the plots indicate if the text inserted by the worker was copied or not from the web page selected. We did the same analysis to investigate if the worker used or not free text, and the plots were almost indistinguishable.

As we can see from the plot , statements on which workers make less errors (i.e., where $x\text{-axis} = 0$) tend to use text copied from the web page selected. On the contrary, statements on which workers make more errors (values close to 5 in the left plot, and values close to $+/- 5$ in the right plot) tend to use text not copied from the selected web page. The differences are small, but it might be an indication that workers of higher quality tend to read the text from selected web page, and report it in the justification box. To confirm this result, we computed the CEM^{ORD} scores for the two classes considering the individual judgments: the class “copied” has CEM^{ORD} = 0.62, while the class “not copied” has a lower value, CEM^{ORD} = 0.58. The behavior is consistent for what concerns the usage of free text (not shown).

By looking at the right column of Figure 10 we can see that the distribution of the prediction error is not symmetrical, as the frequency of the errors is higher on the positive side of the x-axis ($[0,5]$). These errors correspond to workers overestimating the truthfulness value of the statement (with 5 being the result of labeling a pants-on-fire statement as true). This is consistent with what observed in Sect. 5.1.2. It is also noticeable that the justifications containing text copied from the selected URL have a

lower rate of errors in the negative range, meaning that workers which directly quote the text avoid underestimating the truthfulness of the statement. These could be other useful signals to be exploited in future work to obtain more effective aggregation methods.

6 CONCLUSIONS AND FUTURE WORK

The work presented in this paper is, to the best of our knowledge, the first one investigating the ability of crowd workers to identify and correctly categorize recent health statements related to the COVID-19 pandemic. The workers performed a task consisting of judging the truthfulness of 8 statements using our customized search engine, which allows us to control worker behavior. We analyze workers background and bias, as well as workers cognitive abilities, and we correlate such information to the worker quality. We publicly release the collected data to the research community.

The answers to our research questions can be summarized as follows. We found evidence that the workers are able to detect and objectively categorize online (mis)information related to the COVID-19 pandemic (RQ1). We found that while the agreement among workers does not provide a strong signal, aggregated workers judgments show high levels of agreement with the expert labels, with the only exception of the two truthfulness categories at the lower end of the scale (pants-on-fire and false). We found that both crowdsourced and expert judgments can be transformed and aggregated to improve label quality (RQ2). We found that worker political background, self-reported in a questionnaire, is indicative of label quality (RQ3). We found several promising behavioral signals that are clearly related with worker quality (RQ4). Such signals may effectively inform new ways of aggregating crowd judgments (e.g., see [3, 25]), which we believe is a promising direction for future work. Finally, we found that workers use multiple sources of information, and they consider both fact-checking and health-related websites. We also found interesting relations between the justifications provided by the workers and the judgment quality (RQ5). Future work also includes reproducing our experiments in other crowdsourcing platforms to target other cohorts of workers.

ACKNOWLEDGMENTS

This work is partially supported by a Facebook Research award, by an Australian Research Council Discovery Project (DP190102141), and by the project HEaD – Higher Education and Development - 1619942002 / 1420AFPLO1 (Region Friuli – Venezia Giulia).

REFERENCES

- [1] Enrique Amigó, Julio Gonzalo, Stefano Mizzaro, and Jorge Carrillo-de-Albornoz. 2020. An Effectiveness Metric for Ordinal Classification: Formal Properties and Experimental Results. In *Proceedings of ACL*. 3938–3949.
- [2] Pepa Atanasova, Preslav Nakov, Lluís Márquez, Alberto Barrón-Cedeño, Georgi Karadzhov, Tsvetomila Mihaylova, Mitra Mohtarami, and James Glass. 2019. Automatic Fact-Checking Using Context and Discourse Information. *J. Data and Information Quality* 11, 3, Article 12 (2019), 27 pages.
- [3] Yukino Baba and Hisashi Kashima. 2013. Statistical Quality Estimation for General Crowdsourcing Tasks. In *Proceedings of KDD*. 554–562.
- [4] Joseph Bullock, Alexandra Luccioni, Katherine Hoffmann Pham, Cynthia Sin Nga Lam, and Miguel Luengo-Oroz. 2020. Mapping the Landscape of Artificial Intelligence Applications against COVID-19. arXiv:2003.11336
- [5] Alessandro Checco, Kevin Roitero, Eddy Maddalena, Stefano Mizzaro, and Gianluca Demartini. 2017. Let's Agree to Disagree: Fixing Agreement Measures for Crowdsourcing. In *Proceedings of HCOMP*. 11–20.
- [6] Xinran Chen, Sei-Ching Joanna Sin, Yin-Leng Theng, and Chei Sian Lee. 2015. Why Students Share Misinformation on Social Media: Motivation, Gender, and Study-level Differences. *Journal of Academic Librarianship* 41, 5 (2015), 583–592.
- [7] Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valenisse, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. 2020. The COVID-19 Social Media Infodemic. arXiv:2003.05004
- [8] Aakash Desai, Jeremy Warner, Nicole Kuderer, Mike Thompson, Corrie Painter, Gary Lyman, and Gilberto Lopes. 2020. Crowdsourcing a Crisis Response for COVID-19 in Oncology. *Nature Cancer* 1, 5 (2020), 473–476.
- [9] Tamer Elsayed, Preslav Nakov, Alberto Barrón-Cedeño, Maram Hasanain, Reem Suwaileh, Giovanni Da San Martino, and Pepa Atanasova. 2019. Overview of the CLEF-2019 CheckThat! Lab: Automatic Identification and Verification of Claims. In *Proceedings of CLEF*. 301–321.
- [10] Shane Frederick. 2005. Cognitive Reflection and Decision Making. *Journal of Economic Perspectives* 19, 4 (December 2005), 25–42.
- [11] Riccardo Gallotti, Francesco Valle, Nicola Castaldo, Pierluigi Sacco, and Manlio De Domenico. 2020. Assessing the Risks of “Infodemics” in Response to COVID-19 Epidemics. arXiv:2004.03997
- [12] Lei Han, Kevin Roitero, Ujwal Gadira, Cristina Sarasua, Alessandro Checco, Eddy Maddalena, and Gianluca Demartini. 2019. All those wasted hours: On task abandonment in crowdsourcing. In *Proceedings of WSDM*. 321–329.
- [13] Lei Han, Kevin Roitero, Ujwal Gadira, Cristina Sarasua, Alessandro Checco, Eddy Maddalena, and Gianluca Demartini. 2019. The Impact of Task Abandonment in Crowdsoourcing. *IEEE TKDE* (2019), 1–1.
- [14] Lei Han, Kevin Roitero, Eddy Maddalena, Stefano Mizzaro, and Gianluca Demartini. 2019. On Transforming Relevance Scales. In *Proceedings of CIKM*. 39–48.
- [15] Jooyeon Kim, Dongkwan Kim, and Alice Oh. 2019. Homogeneity-Based Transmissive Process to Model True and False News in Social Networks. In *Proceedings of WSDM*. 348–356.
- [16] Klaus Krippendorff. 2011. Computing Krippendorff’s Alpha-Reliability. (2011).
- [17] David La Barbera, Kevin Roitero, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2020. Crowdsourcing Truthfulness: The Impact of Judgment Scale and Assessor Bias. In *Proceedings of ECIR*. 207–214.
- [18] Eddy Maddalena, Kevin Roitero, Gianluca Demartini, and Stefano Mizzaro. 2017. Considering Assessor Agreement in IR Evaluation. In *Proceedings of ICTIR*. 75–82.
- [19] Yelena Mejova and Kyriaki Kalimeri. 2020. Advertisers Jump on Coronavirus Bandwagon: Politics, News, and Business. arXiv:2003.00923
- [20] Tsvetomila Mihaylova, Georgi Karadzhov, Pepa Atanasova, Ramy Baly, Mitra Mohtarami, and Preslav Nakov. 2019. SemEval-2019 Task 8: Fact Checking in Community Question Answering Forums. In *Proceedings of SemEval*. 860–869.
- [21] Howard R Moskowitz. 1977. Magnitude Estimation: Notes on What, How, When, and Why to Use It. *Journal of Food Quality* 1, 3 (1977), 195–227.
- [22] Preslav Nakov, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluís Márquez, Wajdi Zaghouani, Pepa Atanasova, Spas Kyuchukov, and G Da San Martino. 2018. Overview of the CLEF-2018 CheckThat! Lab on Automatic Identification and Verification of Political Claims. In *Proceedings of CLEF*. 372–387.
- [23] Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, and David Rand. 2020. Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy Nudge Intervention. *PsyArXiv* (2020). psyarxiv.com/uhbk9
- [24] Kashyap Kirithbai Popat. 2019. *Credibility Analysis of Textual Claims with Explainable Evidence*. Ph.D. Dissertation, Saarland University, Saarbrücken.
- [25] Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning From Crowds. *Journal of Machine Learning Research* 11, 43 (2010), 1297–1322.
- [26] Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen Voorhees, Lucy Lu Wang, and William R Hersch. 2020. TREC-COVID: Rationale and Structure of an Information Retrieval Shared Task for COVID-19. *Journal of the American Medical Informatics Association* (2020).
- [27] Kevin Roitero, Gianluca Demartini, Stefano Mizzaro, and Damiano Spina. 2018. How Many Truth Levels? Six? One Hundred? Even More? Validating Truthfulness of Statements via Crowdsourcing. In *Proceedings of the CIKM Workshop on Rumours and Deception in Social Media (RDSM’18)*.
- [28] Kevin Roitero, Eddy Maddalena, Gianluca Demartini, and Stefano Mizzaro. 2018. On Fine-Grained Relevance Scales. In *Proceedings of SIGIR*. 675–684.
- [29] 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 SIGIR*. 439–448.
- [30] Andon Tchechmedjiev, Pavlos Falafios, Katarina Boland, Malo Gasquet, Matthäus Zloch, Benjamin Zapilko, Stefan Dietze, and Konstantin Todorov. 2019. ClaimsKG: A Knowledge Graph of Fact-Checked Claims. In *Proceedings of ISWC*. 309–324.
- [31] William Yang Wang. 2017. “Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection. In *Proceedings of ACL*. 422–426.
- [32] Kai-Cheng Yang, Christopher Torres-Lugo, and Filippo Menczer. 2020. Prevalence of Low-Credibility Information on Twitter During the COVID-19 Outbreak. arXiv:2004.14484
- [33] Arkaitz Zubiaaga and Heng Ji. 2014. Tweet, but Verify: Epistemic Study of Information Verification on Twitter. *SNAM* 4, 1 (2014), 1–12.