Investigating "Who" in the Crowdsourcing of News Credibility

Md Momen Bhuiyan
Virginia Tech
momen@vt.edu

Amy X. Zhang
University of Washington
axz@cs.uw.edu

Connie Moon Sehat
Hacks/Hackers
connie@hackshackers.com

Tanushree Mitra
Virginia Tech
tmitra@vt.edu

ABSTRACT

Concerns about the spread of misinformation online via news articles have led to the development of many tools and processes involving human annotation of their credibility. However, much is still unknown about how different people judge news credibility or the quality or reliability of news credibility ratings from populations of varying expertise. In this work, we consider credibility ratings from two "crowd" populations: 1) students within journalism or media programs, and 2) crowd workers on UpWork, and compare them with the ratings of two sets of experts: journalists and climate scientists, on a set of 50 climate-science articles. We find that both groups' credibility ratings have higher correlation to journalism experts compared to the science experts, with 10-15 raters to achieve convergence. We also find that raters' gender and political leaning impact their ratings. Among article genre of news/opinion/analysis and article source leaning of left/center/right, crowd ratings were more similar to experts respectively with opinion and strong left sources.

KEYWORDS

credibility, science news, crowdsourcing, misinformation

ACM Reference Format:
Md Momen Bhuiyan, Amy X. Zhang, Connie Moon Sehat, and Tanushree Mitra. 2020. Investigating "Who" in the Crowdsourcing of News Credibility. In *Proceedings of Computation+Journalism Symposium (C+J'20).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/xxx

1 INTRODUCTION

Misinformation—or information that is false or misleading—can quickly reach thousands to millions of readers via online social and search platforms, helped by inattentive or malicious sharers and algorithms optimized for engagement. In recent years, platforms and third party organizations have developed tools and processes for people to label the credibility of news articles to slow the spread of misinformation.

Some initiatives include Facebook's fact-checking program and Climate Feedback's use of domain experts. However, expert feedback is hard to scale. Other initiatives such as TruthSquad, FactcheckEU, and WikiTribune have pursued a lower-barrier crowdsourced approach, which sometimes run into issues with quality; workarounds include final judgments by experts or delegating primary research to experts and secondary tasks to the crowd [3]. Efforts to automate fact-checking still require human judgment and advances in understanding the crowd labeling of data [1].

In this work, we delve more deeply into the notion of "crowd" and "expert" by examining the article credibility ratings of two populations with different backgrounds — journalism students and UpWork workers— and compare their ratings with those of two different forms of expertise: journalistic and scientific. We also looked at how personal traits and article genre may have related to the ratings.

Our articles set of 50 stories about climate science were annotated by 49 students, 26 Upwork workers, 3 science and 3 journalism experts. Analyses reveal that crowd annotators' perception of the credibility of the articles has higher correlation to journalism experts' ratings than science ones. Among personal attributes, less educated and non-Democrat characteristics lead to higher error. Genre-wise crowd groups are more accurate with the experts on opinion articles and articles from strong left-leaning sources.

From this work, we gain a deeper understanding of the conditions under which crowdsourced annotations might serve as a proxy for reliable expert knowledge, specifically learning more about "who" in terms of the annotation crowd and in addition, how article genre may play a role.

2 RELATED WORK

Much has been made about the "wisdom of crowds" but it is still unclear whether crowdsourcing can be an effective strategy for assessing misinformation at larger scales. Partly this has to do with the limits of crowds on certain topics. It is accepted that collective wisdom can be better than an individual's judgment, including those of individual experts [17]. However, there are situations in which the collective is a lot worse because they do not have enough *relevant* information, suggesting a baseline expertise in the crowd is necessary [16]. Traits related to crowd diversity and their ability to preserve some amount of independent decision-making have been shown important, along with size; in addition to the suitability of the raters themselves, task difficulty also plays a part [9, 13, 18]. The key question is not *whether* crowdsourcing is a viable approach but exactly how—what set of parameters unlocks "wisdom of select crowds" [9]?

Evaluating news may be an area where a general audience may not perform as well as experts. For instance, research has found that most Americans do only slightly better than chance at distinguishing factual from opinion news statements [11], and half are unfamiliar with "op-ed" [6]. This is concerning as opinion pieces have different journalistic standards compared to news articles. Some segments of the population are also potentially worse at assessing news. Research has found that conservative leaning, older, and highly engaged with political news profiles were more likely to engage with "fake news" in the US, and a partisan distrust of certain kinds of climate science and conservatives [7, 10, 19].

The strategy that our study focuses upon is one that centers a reader's assessment of "credibility." Credibility has been defined by Flanagin and Metzger as equating with the believability of a source or message, and is made up of two primary dimensions: *trustworthiness* and *expertise* [5]. Domain experts may better grasp underlying truth values and journalists may be better at fact-checking, less experienced readers may yet be able to assess news credibility. Research shows that attributes about articles' content (eg., emotional tone) and context (eg., citation to reputable sources) can signal expert judgment of credibility [20].

We build upon prior work by examining how different populations with differing backgrounds, demographics, and levels and types of expertise perform on the task of assessing article credibility. We also analyze crowds' performance on articles divided by genre and sources' political leaning.

3 STUDY DESIGN

Three main questions guide this study, the first two focusing on suitability aspects of raters and the latter of tasks:

- RQ1: How well do different crowd raters as well as different expert raters agree with themselves and with each other on article credibility ratings?
- RQ2: How do personal characteristics of age, education, gender, and political leaning affect credibility ratings?
- RQ3: How do characteristics of news articles, such as article genre (news, opinion, analysis) and political bias of the source, affect credibility ratings?

Raters

We gathered credibility ratings on climate science articles from four different groups, including two "crowd" groups consisting of: 1) 49 journalism and media students, as well as 2) 26 Upwork crowdworkers, and two "expert" groups comprising: 3) three climate scientists, and 4) three journalists. Individuals in the crowd groups were paid $150 for completing all tasks on time, while experts were paid $300.

Students: The first group was canvassed through the *Credibility Coalition* network, which has worked directly with nonprofits and journalism schools to build up a cohort motivated to combat misinformation. They are predominantly pursuing higher education in the U.S. and tend politically liberal. They actively recruited, eg. with campus Republican clubs, to achieve more demographic balance for the study.

Upwork: In addition, we also used the *Upwork* platform for freelancers to gather a broader sample of participants. We restricted participants to the U.S. Participants were admitted on first-come basis until demographic balance became an issue (e.g. politically liberal respondents were declined once more conservatives were needed for balance).

Scientists: Three science experts were referred by science organizations (Climate Feedback, AAAS, National Academy of Sciences), all possessing a PhD in a climate-related field.

Journalists: Our three journalism experts, reached through personal networks, each possess at least seven years of professional journalism experience in the U.S.

	#	α	Avg. Cred. (Std. Dev.)
Student	49	0.44	3.49 (1.32)
Upwork	26	0.48	3.34 (1.33)
Expert[Science]	3	0.75	3.21 (1.27)
Expert[Journalism]	3	0.83	3.60 (1.42)

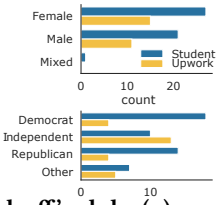

Table 1: Inter-rater reliability using Krippendorff's alpha (α) within all 4 rater groups on the question of credibility across 50 articles, along with average credibility rating. The figures show user distribution by gender and political party.

Rater Task

Before participation, crowd raters filled out a demographic survey followed by committing to an Annotator Code of Conduct, outlining their duties to be as accurate and diligent as possible, provided in their informed consent. Once eligible, they received reading and rating tasks which included their credibility perception per article on a 5-point Likert scale, ranging from *very low* (1) to *very high* (5). Instructions to fill out the seven question survey across a 7-10 day period (estimated at 10 hours) were provided in a handbook with a recommended limit of 10-15 minutes per article.

Articles

We selected 50 articles related to climate and environment issues that have a high degree of consensus among domain experts. Articles were written in English and represented a range of liberal to conservative positions or attitudes towards climate problems. To gather articles, we used the Buzzsumo social media research tool in late 2018 to find the most popular articles over the previous year with the keywords of "climate change," "global warming," "environment," and "pollution." Among the top results, our team selected a set of

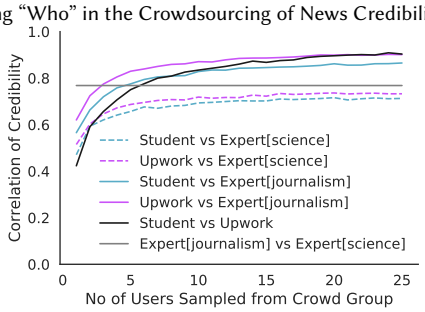

Figure 1: Correlation of credibility ratings among all pairs in four groups: 2 crowd and 2 expert groups. In each crowd group, we sample the number of raters from 1–25. For expert groups, we take all 3 ratings. Then we compute the Spearman ρ between the mean responses from each group on all 50 articles. The plot shows average ρ after 100 resamplings.

articles with varying amounts of scientific reference. According to our journalism experts, the articles contain 30 news, 8 analysis and 8 opinion articles (discussed later).

4 RESULTS

RQ1: Comparison Within and Between Rater Groups

Table 1 presents the inter-rater reliability (IRR) and average credibility ratings within each of our two crowd groups—`Student` and `Upwork`—and two expert groups—`Science` and `Journalism`. Overall, we see that the experts had much higher IRR within each group than the crowd raters, with the journalists most aligned at 0.83. On average, `Science` had the lowest scores while `Journalism` had the highest.

We also compute the correlation of credibility ratings among all combinations of groups using Spearman's ρ. Figure 1 shows the pairwise correlation between rater groups when we vary the number of raters from 1 to 25 in `Student` or `Upwork`. We randomly sample 100 times from each group and then average the result; no individual rater has undue weight (example of approach in prior work [12]). With only 3 raters in each group of experts, we simply average them per group. We find that the correlation between the two expert groups is 0.77. Correlation between the two crowd groups starts off low at about 0.4 with only 1 rater, but becomes high ($\rho = 0.9$) with 15 or more raters within each group. This suggests that when averaging across 15 or more raters, both rater populations begin performing about equivalently.

When we dive into the correlation of each crowd group to each expert group, differences emerge. First, we notice that `Upwork` has slightly higher correlation with both sets of experts than `Student`. The gap, while small in both cases, is nonetheless robust in the case of journalists (0.04, $t = 2.31$, $p < 0.02$) averaging across 1–25 raters. In the case of scientists, the gap was 0.02 ($t = 1.59$, $p < 0.11$). Second, we note that it takes about 15 crowd raters to achieve about 0.85 correlation with journalists. However, crowd raters get only about 0.7 correlation with scientists using 15 raters, and ratings do not improve at 25 raters. The difference between correlation with scientists versus journalists is a major one, with about 0.13 for `Student` and 0.15 for `Upwork`.

RQ2: Personal Factors Affecting Credibility Ratings

To determine how crowd raters' personal characteristics, such as their age and gender, relate to how well they agreed with experts, we perform an OLS regression on the error in our crowd raters' credibility rating when compared to experts' average rating. In Table 2, we present 6 models, where ratings from just `Student`, just `Upwork`, and then `Student` and `Upwork` *combined* are compared against ratings from `Science` and then `Journalism`. To meet minimum sampling requirements, we recoded their education into three groups: combined "High School", "Some College No Degree" and "Some College" into one and "4 Year College" with "Community College/Vocational Training" into another. We divided raters into "18-25", "26-30" and "31+" age groups.

Among our variables, consistent across all models, crowd raters with a non-Democrat political leaning made greater errors in their assessment. In addition, males made less error compared to females; the difference is small but significant in all the models except one. Among age groups, people aged 26–30 made less error compared to those aged 18–25; however those values are only significant in the omnibus models. Other age ranges had no significant results. On the other hand, crowd raters with a four-year college or community college degree made more errors compared to those with a graduate degree. Surprisingly, raters with a high school degree or some college experience made fewer errors compared to those with a graduate degree in one of our models (`Student+Upwork` compared with `Journalism`). This may be because the majority of our crowd raters in the `Student` group are assumed to still be in college, and perform relatively well due to exposure to journalism and media studies.

RQ3: Rating Performance According to Article Type

Finally, we investigate how our four groups performed when assessing different types of articles. Given the difficulty cited earlier that Americans have with factual and opinion statements, one area of interest is article *genre*. Our journalism experts additionally rated the genre of articles in our dataset alongside the crowd raters; in addition to the genres of "news" and "opinion," we added "analysis" (understood as a close examination of a complex news event by a specialist, including reporters [14]) for experts. Across news and opinion, the journalism experts had an IRR of 0.97; but when adding analysis as a third category, the IRR went down to 0.71.[1] In the

[1]Separately, we wondered whether our crowd raters could label genre. When asked to consider just news vs. opinion, IRR was lower at 0.43 for `Student` and 0.49 for `Upwork` but the majority assessment of each crowd group was

	Expert[Science]						Expert[Journalism]					
	Student		Upwork		Stud.+Upwork		Student		Upwork		Stud.+Upwork	
	β(sig.)	Err.	β(sig.)	Err.	β(sig.)	Err.	β(sig.)	Err.	β(sig.)	Err.	β(sig.)	Err.
Intercept	0.13*	(0.06)	-0.04	(0.07)	0.12***	(0.04)	0.27***	(0.06)	0.11	(0.07)	0.26***	(0.04)
Edu[4Y&CColl]	0.13***	(0.04)	0.05*	(0.02)	0.07***	(0.02)	0.11***	(0.04)	0.04*	(0.02)	0.06***	(0.02)
Edu[HS&SColl]	-0.00	(0.04)	0.01	(0.04)	-0.03	(0.02)	-0.03	(0.04)	0.01	(0.04)	-0.05***	(0.02)
Gender[Male]	-0.02	(0.01)	-0.04*	(0.02)	-0.03***	(0.01)	-0.03*	(0.01)	-0.04***	(0.02)	-0.03***	(0.01)
Age[26-30]	-0.05	(0.04)	-0.01	(0.03)	-0.06***	(0.02)	-0.06	(0.04)	-0.03	(0.03)	-0.06***	(0.02)
Pol[Indep.]	0.06***	(0.02)	0.11***	(0.02)	0.06***	(0.01)	0.07***	(0.02)	0.12***	(0.02)	0.07***	(0.01)
Pol[Other]	0.04	(0.02)	0.15***	(0.03)	0.08***	(0.01)	0.06**	(0.02)	0.15***	(0.03)	0.08***	(0.01)
Pol[Repub.]	0.08***	(0.02)	0.13***	(0.04)	0.10***	(0.01)	0.10***	(0.02)	0.14***	(0.04)	0.11***	(0.01)

Table 2: OLS regression on error in credibility rating compared to experts' average rating after recoding and non-significant rows omitted. The reference for education, gender, age and political leaning are: Graduate degree, Female, 18-25 and Democrat. Numbers in green are negative coefficients with significant p-values contributing to less error; numbers in red are vice-versa.

Figure 2: The two figures show the average credibility rating and standard deviation by each of the four rater groups broken down by article type, along with Mann-Whitney U-test results between pairs where red lines indicate indistinguishable pairs.

end, 48 articles had a majority category across the 3 experts, with 30 as *news*, 8 as *opinion*, and 8 as *analysis*.

Another area of interest is the *political leaning* of article sources. Using Media Bias/Fact Check, a site that classifies media sources on a political bias spectrum (prior example of approach [4]). we recoded their 7 categories into three higher-level ones of strong left, center, and strong right resulting respectively in 6, 24, and 15 articles for our set (5 omitted).

From an article source perspective, articles from both stronger right/left sources have higher IRR than those in the center. This suggests that annotators might have used the leaning of sources as shortcuts to identify credibility [15].

In addition, we examined how closely annotators evaluated credibility related to experts based on news genre and political leaning. Through a Mann-Whitney U-test, we looked at the average credibility and the values that support an assumption that the groups can be treated equally (null hypothesis, see Figure 2). Among the news-related categories, our test suggests that crowd groups were better at rating credibility on opinion articles given indistinguishable differences with experts (e.g., student-science: $U = 4270$, $p = 0.22$). Our test also shows that `Upwork` ratings are very close to those of science experts in all categories (slightly more distinctive on analysis articles at $p = 0.08$); however there are a number of factors that could explain this, and more analysis is needed. `Student` ratings on opinion and news articles are also similar to the journalism experts, which make sense given that many of the students were recruited through journalism networks.

Along political lines, ratings of both crowd groups are indistinguishable from experts on articles from strong left sources; also remarkable is the higher rating of credibility by science compared to journalism experts. For articles from center sources, `Upwork` ratings are similar to the science experts; the higher credibility rating from journalism experts may come from a professional experience which aligns with the center sources. For strong right article sources, Upwork workers rate them closer to journalism than science experts, who rated this set of articles least credible of all.

5 DISCUSSION AND CONCLUSION

For RQ1 and RQ2, we considered the suitability of the raters. Under RQ1, we found that crowd raters overall have lower

100% aligned with experts. Most articles labeled "analysis" by journalists were labeled "news" by the crowd groups.

internal agreement compared to experts, and need about 10–15 raters before they begin to plateau in their correlation with experts. `Upwork` was slightly more aligned with experts than `Student`, though this difference becomes negligible around 15 raters. `Science` and `Journalism` experts have relatively high agreement within and between each other. Crowd raters overall were more aligned with Journalism than Science experts, who tended to rate articles lower.

Under RQ2, we found that non-Democrats had consistently positive association with errors compared to experts. This result is not surprising given the politicization of climate science despite scientific consensus. Also noteworthy are cases in which articles from right-leaning sources were perceived to be generally credible, such as a straight-forward reporting of a "glitter ban" call by some scientists. More granular investigation into articles may provide further insight into how annotators' perceived distinct cases; we leave this for future research. Overall, our findings support the hypothesis that a distributed demographic (eg, gender, politics, education) can correlate with experts after 10–15 raters.

Delving into the intricacies of the articles was the focus of RQ3, which started to explore task suitability. On the *genre* of articles–news/opinion/analysis–both groups of crowd annotators' credibility ratings are very similar to experts' ratings for opinion articles. Along political lines, crowd groups provide indistinguishable ratings on articles from left-leaning sources. These results suggest that the crowd may have the ability to replace experts' annotations in certain categories of articles. However, the conditions are not yet well understood.

With genre, it may be the case that some difficulties for raters arise from the lack of genre labeling in US mainstream media, excepting some amount of opinion columns [8]. Without being well labeled or well understood, non-opinion news categories may require readers to rely on other structural aspects when the topic is more difficult to understand. Future studies may focus on understanding such alternate signals and their relationship to experts' credibility perception. Some differences may also stem from how journalists versus scientists consider news. The category of *analysis* as do article sources from the center of the political spectrum are representative of professional journalism; these differences may relate to perceptions of *credibility* and specialized fields of knowledge — the summarized perspectives of lay persons versus more technically accurate recounting, and the complicated interaction of *trustworthiness* and *expertise* [5].

Disagreement among raters is neither always bad nor always about their capacities, but at times about suitability of the task [2]. Deeper understanding the parameters of task suitability in relationship with the expertise in question is needed to better leverage the capacities of crowdsourcing.

ACKNOWLEDGMENTS

This paper would not be possible without the valuable support of the Credibility Coalition, with special thanks to Caio Almeida, An Xiao Mina, Jennifer 8. Lee, Rick Weiss, Kara Laney, and especially Dwight Knell. Bhuiyan and Mitra were partly supported through NSF grant IIS-1755547.

REFERENCES

[1] Bill Adair, Chengkai Li, Jun Yang, and Cong Yu. 2019. Automated Pop-Up Fact-Checking: Challenges & Progress. In *Proc. C+J Symposium.*
[2] Lora Aroyo and Chris Welty. 2015. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Magazine* 36, 1 (2015), 15–24.
[3] Mevan Babakar. 2018. Crowdsourced Factchecking.
[4] Alexandre Bovet and Hernán A Makse. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature communications* 10, 1 (2019), 7.
[5] Andrew J Flanagin and Miriam J Metzger. 2008. Digital media and youth: Unparalleled opportunity and unprecedented responsibility. *Digital media, youth, and credibility* (2008), 5–27.
[6] American Press Institute & The AP-NORC Center for Public Affairs Research. 2018. Americans and the news media: What they do–and don't–understand about each other. *The Media Insight Project* (2018).
[7] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on Twitter during the 2016 U.S. presidential election. *Science* 363, 6425 (Jan 2019), 374–378.
[8] Rebecca Iannucci and Bill Adair. 2017. Reporters' Lab Study Results: Effective News Labeling and Media Literacy.
[9] Albert Mannes, Jack Soll, and Richard Larrick. 2014. The Wisdom of Select Crowds. *Journal of personality and social psychology* (2014).
[10] Aaron M. McCright, Katherine Dentzman, Meghan Charters, and Thomas Dietz. 2013. The influence of political ideology on trust in science. *Environmental Research Letters* 8, 4 (Nov 2013), 044029.
[11] Amy Mitchell, Jeffrey Gottfried, Michael Barthel, and Nami Sumida. 2018. *Can Americans Tell Factual From Opinion Statements in the News?*
[12] Tanushree Mitra and Eric Gilbert. 2015. CREDBANK: A Large-Scale Social Media Corpus with Associated Credibility Annotations. In *Proc. ICWSM'15.*
[13] Tanushree Mitra, Clayton J Hutto, and Eric Gilbert. 2015. Comparing person-and process-centric strategies for obtaining quality data on amazon mechanical turk. In *Proc. CHI'15.* ACM, 1345–1354.
[14] Art Silverblatt, Donald C. Miller, Julie Smith, and Nikole Brown. 2014. *Media Literacy: Keys to Interpreting Media Messages, 4th Edition: Keys to Interpreting Media Messages.* ABC-CLIO.
[15] S Shyam Sundar. 2008. The MAIN model: A heuristic approach to understanding technology effects on credibility. *Digital media, youth, and credibility* 73100 (2008).
[16] Cass R. Sunstein. 2006. When Crowds Aren't Wise. *Harvard Business Review* (Sep 2006). https://hbr.org/2006/09/when-crowds-arent-wise
[17] James Surowiecki. 2004. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations.* Doubleday.
[18] Christian Wagner and Ayoung Suh. 2014. The Wisdom of Crowds: Impact of Collective Size and Expertise Transfer on Collective Performance. (Jan 2014), 594–603. https://doi.org/10.1109/HICSS.2014.80
[19] Lorraine Whitmarsh. 2011. Scepticism and uncertainty about climate change: Dimensions, determinants and change over time. *Global Environmental Change* 21, 2 (May 2011), 690–700.
[20] Amy X Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, et al. 2018. A structured response to misinformation: Defining and annotating credibility indicators in news articles. In *Companion Proceedings of The Web Conference 2018.* 603–612.