Proceedings of the Eighth AAAI Conference on Human Computation and Crowdsourcing (HCOMP-20)

Does Exposure to Diverse Perspectives Mitigate Biases in Crowdwork? An Explorative Study

Xiaoni Duan
Purdue University
duan79@purdue.edu

Chien-Ju Ho
Washington University in St. Louis
chienju.ho@wustl.edu

Ming Yin
Purdue University
mingyin@purdue.edu

Abstract

Earlier research has shown the promise of enabling worker interactions in crowdwork to mitigate worker biases and improve the quality of crowdwork. In this study, we focus on one characteristic of the interacting workers that may influence the effectiveness of worker interactions in enhancing crowdwork—the *diversity of perspectives* that the interacting workers bring together—and we explore whether and how interactions between a set of workers holding different perspectives can help mitigate biases in crowdwork. Through two sets of randomized experiments, we find that whether interactions between workers with different perspectives can help mitigate biases in crowdwork depends on task properties. We also find no conclusive evidence in our experimental settings suggesting that interactions among workers with diverse perspectives reduce biases in crowdwork to a larger extent compared to interactions among workers with similar perspectives.

Introduction

Crowdsourcing has become a ubiquitous paradigm for obtaining data from people to enhance machine intelligence. Recent studies, however, reveal considerable concerns on the quality of human-annotated datasets as humans are notorious for being prone to *biases*, which may result in systematic deviations between the data collected from them and the ideal (Hube, Fetahu, and Gadiraju 2019; Otterbacher et al. 2019). Such biases may come from multiple sources, including the "blind spots" in worker's knowledge, as well as worker's political viewpoints and cultural background.

Among various approaches that researchers have developed to combat biases in the crowdsourced data, it is recently shown that enabling *interactions* between crowd workers working on the same task can decrease worker biases and result in data of higher quality (Drapeau et al. 2016; Chang, Amershi, and Kamar 2017; Tang, Ho, and Yin 2019). Despite its promise, systematic understandings of how the designs of such worker interactions affect their effectiveness in mitigating biases in crowdwork are largely lacking. In this paper, we focus on one specific aspect in designing worker interactions—the diversity of perspectives that the interacting workers bring together. In the social science literature, there are mixed empirical evidence showing that long-term, repeated interactions to opposing views could result in either more (Bail et al. 2018) or less (Guilbeault, Becker, and Centola 2018) biased belief. In crowdsourcing contexts, where interactions between workers are often short-term and even one-off, does enabling workers with different perspectives to interact with each other help mitigate biases in crowdwork, and does it bring about higher levels of bias reduction in crowdsourced data compared to having workers interact with others holding similar perspectives?

As an initial attempt to answer this question, we conducted two sets of randomized experiments on Amazon Mechanical Turk. In the first experiment, we asked workers to work on an objective task of differentiating two styles of artificially-generated face images. Different workers were "trained" to own the skills of recognizing different discriminating attributes of the two styles, thus workers' "perspectives" are operationalized as their skills. In the second experiment, we asked workers to evaluate whether a statement containing political messages is a factual or an opinion statement, and worker's perspectives are reflected through the political values that they hold. In both experiments, we allowed interactions among a subset of workers—they formed coworker pairs to work on the same tasks together, so that they can see each other's answers to the tasks and engage in discussions, which were set to be 2-minute maximum per task to keep the microtask nature of crowdwork. Since both experiments involve binary classification tasks, we use workers' accuracy in the tasks as a proxy to measure biases in the crowdwork. Our results suggest that depending on task type and difficulty, interactions between workers with diverse perspectives may or may not mitigate biases in crowdwork, and they do not reduce biases to a larger extent than interactions between workers holding similar perspectives.

Experiment 1: Exposure to Diverse Skills

We begin with an experiment on an *objective* task of image classification in which different workers were trained to own different specialized skills, and we aim to understand whether and how interactions between workers with diverse skills help mitigate biases in crowdwork.

Figure 1: Interface of the face image style classification task. **(Top)** We first showed distinct sets of artificial face images to workers to train them into one of the two "types" in detecting the styles of GAN-generated face images; **(Bottom)** trained workers were then asked to determine between two sets of face images, which set belongs to a target style.

Experimental Task

In this experiment, we asked workers to classify the styles of face images that were artificially generated by a generative adversarial network (GAN). Specifically, we adopted the facial attribute editing tool developed by He et al. (2019) to generate artificial face images that differ on two key attributes: the *skin color* and the *hair color*. We created face images of two different "styles": one style was with relatively pale skin *and* dark hair, while the other style was with relatively tanned skin color *and* light hair color. In our experiment, we first "trained" workers to recognize GAN-generated faces of these two styles by showing them example images from both styles. Then, we asked the trained workers to complete a sequence of style classification tasks where in each task, the worker was presented with two sets of GAN-generated face images and was asked to identify which set belongs to a target style. Figure 1 (bottom) shows an example of the interface of our style classification task.

Importantly, by showing different workers the same text instructions but distinct sets of example images that have more salient differences on skin color or hair color during training, we created different "perspectives" among workers in solving this task (Figure 1 top). As a result, workers may produce biased work in this task due to their varying specialities or different "blind spots" in their knowledge.

Experimental Procedure

Our experiment was separated into two phases. Phase 1 was the training phase, in which workers were recruited from MTurk to learn about two styles of GAN-generated face images by inspecting the example images from both styles (Figure 1 top). We randomly trained each worker into one of the two types (i.e., "sensitive to skin color" or "sensitive to hair color"). As discussed above, the example images that a worker saw was determined by the type she was trained into. After carefully reviewing the example images, each worker completed a same sequence of 12 style classification tasks.

Upon the completion of Phase 1, we randomly divided all trained workers into two groups, the *independent* group and the *interaction* group, and then recruited all these trained workers to complete an additional sequence of 4 style classification tasks in Phase 2. At the beginning of Phase 2, we reminded workers the differences between the two styles of face images by showing them *exactly the same* sets of example images that they saw in Phase 1. For workers in the independent group, they were instructed to complete the sequence of style classification tasks *on their own*. In contrast, workers in the interaction group were randomly matched with a co-worker in the same group, and the pair of workers completed the same sequence of tasks *together*—In each task, a worker was first asked to submit her independent answer; then, the worker could discuss this task with her co-worker for up to 2 minutes, during which she could see her co-worker's answer, and she was asked to explain why she believed her answer was correct; finally, each worker in the pair needed to submit her final answer to the task separately. Since workers in the interaction group were randomly paired up with each other, interactions between workers in Phase 2 can occur either between workers with similar perspectives or between workers with different perspectives. We revealed *no* explicit information about this possible difference in perspectives to interacting workers, though workers may figure this out by themselves through discussions.

We designed a pool of 16 tasks with varying levels of difficulty: the *easy/difficult* tasks contained two sets of face images with large/small differences on both skin color and hair color, while the *intermediate* tasks contained two sets of face images with large differences on only one attribute (either skin color or hair color). The 4 style classification tasks that each worker completed in Phase 2 (1 easy, 1 difficult, and 2 intermediate tasks) were randomly sampled from this pool.

Our experiment was open to U.S. workers only, and each worker was allowed to participate once. Each worker received a flat payment of $0.35 in Phase 1. In Phase 2, besides the base payment of $0.50, we also offered a $0.20 bonus in each task if the worker's final answer was correct.

Experimental Results

In total, 1,062 workers participated in our Phase 1, among whom 392 workers provided valid data in Phase 2 (independent group: 116 workers with 58 workers for each type; interaction group: 276 workers). From workers in the interaction group, we got 78 pairs of workers with the same perspective and 60 pairs of workers with different perspectives. Examining worker's performance on the 12 tasks in Phase 1, we confirmed that workers of different types focused on different attributes to determine the style of face images.

Figure 2 (left) compares worker's average performance across *all* style classification tasks in Phase 2 between workers who worked on their own, workers who interacted with others sharing the same perspective, and workers who inter-

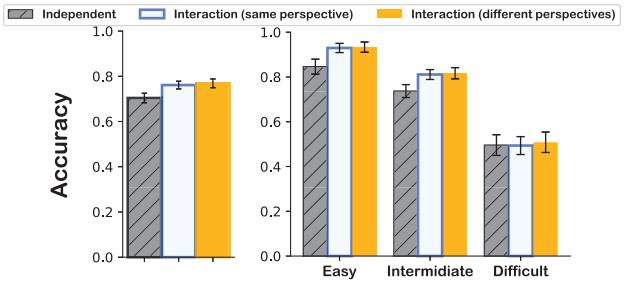

Figure 2: Comparing worker's performance in Experiment 1 between independent workers, workers who interacted with others with the same perspective, and workers who interacted with others with different perspectives, across all tasks (**left**) and broken down by task difficulty (**right**).

acted with others who had different perspectives. We used workers' accuracy in the style classification tasks to quantify their biases in these tasks; the higher the accuracy is, the smaller the bias. Visually, while it appears that allowing workers with different perspectives to interact with each other decreases the biases in crowdwork as compared to the case when they work alone, it does not seem to result in a higher level of bias reduction compared to having workers with the same perspective interact. Indeed, a one-way analysis of variance (ANOVA) suggests a significant difference in worker's accuracy in the Phase 2 style classification tasks between independent workers and workers who participated in interactions ($p = 0.040$). However, post-hoc Tukey HSD tests reveal that while interacting with workers holding different perspectives marginally reduces worker biases compared to independent work ($p = 0.057$), it does not reduce the biases to a larger extent than interactions between workers with the same perspective ($p = 0.956$).

We then break down the comparison by task difficulty and Figure 2 (right) shows the results. Interestingly, we find that compared to independent work, allowing workers with different perspectives to interact only shows marginal benefits in bias reduction on easy tasks and intermediate tasks, but *not* difficult tasks (post-hoc Tukey HSD tests, independent vs. interaction between different perspectives: $p = 0.056, 0.085, 0.9$ for easy, intermediate, and difficult tasks). Still, we find no difference in the amount of biases reduced by interactions between workers with similar or different perspectives, regardless of the task difficulty.

Experiment 2: Exposure to Diverse Values

To see whether results of our first experiment are limited by our design of imposing perspectives on workers, we conducted a second experiment using a different type of task, in which workers will naturally be influenced by their own subjective belief when completing the task.

Experimental Tasks

In our second experiment, we asked workers to complete a sequence of statement evaluation tasks. In each task, we presented the worker with a news statement (e.g., "Immigrants who are in the U.S. illegally have some rights under the Constituion"), and we asked her to determine whether the statement was a *factual* or an *opinion* statement. News statements used in this experiment were selected from a recent survey conducted by Pew Research Center (Mitchell 2018), thus we had the ground-truth label for each statement. The Pew Research Center survey results revealed that people tend to label both factual and opinion statements as factual when they align with their political side (Mitchell 2018). Thus, in this experiment, we considered each worker's political value as a natural characterization of the worker's perspectives.

Experimental Procedure

Our second experiment was again divided into two phases. Phase 1 was conducted to recruit a set of workers from MTurk and measure their political values. We adopted the political typology quiz developed by Pew Research Center (Doherty, Kiley, and Johnson 2017) to categorize worker's political attitudes as leaning liberal or conservative. The procedure of Phase 2 is completely analogous to that in Experiment 1. Again, as interacting workers were randomly paired up, they may have the same or different political values. This experiment was again open to U.S. MTurk workers only, restricting each worker to participate in once. The payment structure was the same as that used in Experiment 1.

Experimental Results

We recruited a total of 1,504 workers through our Phase 1 experiment (988 liberal workers, 516 conservative workers). In Phase 2, we obtained valid data from 331 workers (independent group: 101 workers with 58 leaning liberal and 43 leaning conservative; interaction group: 230 workers), and for the interaction group, we got 68 pairs of workers with the same political value (37 liberal pairs, 31 conservative pairs) and 47 pairs with different political values.

We first compare workers' average performance on all Phase 2 statement evaluation tasks across independent workers, workers who interacted with another worker with similar political value and workers who interacted with another worker with different political values in Figure 3 (left). Again, we used worker's average accuracy in the tasks to quantify the amount of bias in the crowdwork produced, with higher accuracy indicating smaller bias. Inspecting Figure 3 (left), however, we find interactions between workers with diverse political values does not reduce biases in crowdwork compared to either independent work or interactions between workers with similar political values (one-way ANOVA: $p = 0.193$). Repeating the comparison separately for workers holding liberal views and workers holding conservative views in Figure 3 (right), we still observe no significant differences in the amount of biases produced in the data between workers who completed the work independently and workers who interacted with another worker who had the same or different political view as themselves (one-way ANOVA within liberal workers: $p = 0.109$, within conservative workers: $p = 0.961$). In other words, for the statement evaluation tasks, having workers exposed to different

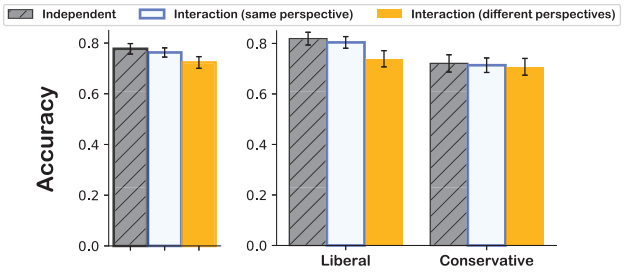

Figure 3: Comparing worker's performance in Experiment 2 between independent workers, workers who interacted with others with the same perspective, and workers who interacted with others with different perspectives, across all tasks (**left**) and broken down by worker's own value (**right**).

political values does not help mitigate worker's biases in the task, regardless of what the worker's own political value is.

Conclusions and Discussions

In this paper, we explored the effects of allowing interactions between crowd workers with diverse perspectives on mitigating biases in crowdwork. We do not find conclusive evidence that interactions between workers holding diverse perspectives lead to higher levels of bias reduction in the data generated by the crowd compared to interactions between workers with similar perspectives. We also observed that whether the interactions between workers with diverse perspectives can help mitigate biases in crowdwork depends on the difficulty of the task, as well as the type of task.

We provide a few reasons for why we get limited evidence showing the advantages of allowing interactions among workers with diverse perspectives in mitigating biases in crowdwork. First, we found that when interacting with other workers with a different perspective than themselves, workers may fail to understand the perspective of their co-workers. For example, from the chat messages we recorded in Experiment 1, we observed that the interactions between workers with different skills could go into two different ways—some workers could *successfully* understand the information shared by other workers with different perspectives, while some other workers *failed* to do so.

Moreover, the difficulty for workers to fully understand and appreciate the value of each other's perspectives during the interactions may be further increased by how interactions were structured in our experiment—workers had a 2-minute maximum discussion time on each task, and we did not provide any accuracy feedback to the interacting workers. The short amount of time for each discussion period, though well reflects the microtask nature of crowdwork, may keep workers from fully elaborating and deliberating on each other's viewpoints. Indeed, we found that the average number of chat messages in one task was 3.38 and 2.72 in our first and second experiment, respectively, suggesting that workers may lack the sufficient time needed to build a common ground with their co-workers who had different perspectives than themselves. In addition, the absence of accuracy feedback implies that when workers could not understand each other's perspective, this impression of incomprehensibility may get reinforced over multiple runs of interactions without workers seeing the value of the different perspective.

As a practical lesson, we have learned from our study that reaping the benefits of diversity in microtask-based crowdsourcing context to mitigate biases in crowdwork is not an easy task. Our observations in this study clearly suggest the needs for providing scaffolding for mutual understanding when workers with diverse perspectives interact with each other. On the other hand, both of our experiments involve binary classification tasks and workers' biases in the tasks are measured via their accuracy. For tasks that can not be represented as binary classifications, it would be critical to define proper measurements to quantify biases, and our results may not generalize to those cases. We hope the explorative results that we report in this study could open more discussions on how to better design worker interactions to fully release the potential of diversity in mitigating biases in crowdwork.

References

Bail, C. A.; Argyle, L. P.; Brown, T. W.; Bumpus, J. P.; Chen, H.; Hunzaker, M. F.; Lee, J.; Mann, M.; Merhout, F.; and Volfovsky, A. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences* 115(37):9216–9221.

Chang, J. C.; Amershi, S.; and Kamar, E. 2017. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI)*.

Doherty, C.; Kiley, J.; and Johnson, B. 2017. Political typology reveals deep fissures on the right and left: Conservative republican groups divided on immigration,'openness'. *Pew Research Center*.

Drapeau, R.; Chilton, L. B.; Bragg, J.; and Weld, D. S. 2016. Microtalk: Using argumentation to improve crowdsourcing accuracy. In *Fourth AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*.

Guilbeault, D.; Becker, J.; and Centola, D. 2018. Social learning and partisan bias in the interpretation of climate trends. *Proceedings of the National Academy of Sciences* 115(39):9714–9719.

He, Z.; Zuo, W.; Kan, M.; Shan, S.; and Chen, X. 2019. Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing* 28(11):5464–5478.

Hube, C.; Fetahu, B.; and Gadiraju, U. 2019. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12.

Mitchell, A. 2018. *Distinguishing between factual and opinion statements in the news*. Pew Research Center.

Otterbacher, J.; Barlas, P.; Kleanthous, S.; and Kyriakou, K. 2019. How do we talk about other people? group (un) fairness in natural language image descriptions. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, 106–114.

Tang, W.; Ho, C.-J.; and Yin, M. 2019. Leveraging peer communication to enhance crowdsourcing. In *The World Wide Web Conference*, 1794–1805.