Understanding and Mitigating Worker Biases in the Crowdsourced Collection of Subjective Judgments

Christoph Hube, Besnik Fetahu, Ujwal Gadiraju
L3S Research Center, Leibniz Universität Hannover
Hannover, Germany
{hube, fetahu, gadiraju}@L3S.de

ABSTRACT
Crowdsourced data acquired from tasks that comprise a subjective component (e.g. opinion detection, sentiment analysis) is potentially affected by the inherent bias of crowd workers who contribute to the tasks. This can lead to biased and noisy ground-truth data, propagating the undesirable bias and noise when used in turn to train machine learning models or evaluate systems. In this work, we aim to understand the influence of workers’ own opinions on their performance in the subjective task of bias detection. We analyze the influence of workers’ opinions on their annotations corresponding to different topics. Our findings reveal that workers with strong opinions tend to produce biased annotations. We show that such bias can be mitigated to improve the overall quality of the data collected. Experienced crowd workers also fail to distance themselves from their own opinions to provide unbiased annotations.

1 INTRODUCTION
Microtask crowdsourcing provides remarkable opportunities to acquire human input at scale for a variety of purposes [35] including the creation of ground-truth data and the evaluation of systems. A survey of crowdsourcing tasks on Amazon’s Mechanical Turk [8] revealed that one of the most popular tasks is that of interpretation and analysis (IA) [17]. In many scenarios, such interpretation tasks may be prone to biases of workers. These biases are subject to various factors, such as cultural background of workers, personal opinion on a topic, ideological, or other group memberships of a person.

Such factors are well studied from the language point of view and the use of language to express statements on a given subject at hand [5, 13, 38, 46]. For instance, sociolinguistic studies show gender biases in English language in terms of authority (e.g. how a person is addressed, title + firstname + lastname) [5, 38] or in terms of over-lexicalization [46] (e.g. young married woman). Language bias and bias in language use occurs in various contexts, e.g. journalism [14]. Subjective language [53] can be seen as a subproblem of language bias (e.g. framing, opinions etc.), which often is presented through subtle linguistic cues that carry an implicit sentiment [20, 48] and often are deliberately used in order to convey a specific stance towards a subject. Thus, differentiating between neutrally phrased and opinionated statements is subject to the worker’s ideological memberships.

Studies [3, 4] show that the political or ideological stance of a person can influence the perception and interpretation of facts. In interpretation tasks such as distinguishing between opinions and facts, worker awareness of possible biases that may be introduced due to their personal or ideological stances is crucial in providing noise free judgments. For example, surveys\(^1\) show that only 23% of the U.S population who identify politically with the Republican party believe that humans have an influence in climate change.

Several natural language understanding tasks that rely on crowdsourced labeling are prone to worker biases. For instance, Yano et al. [54] showed that in determining biased language in text corresponding to the news genre, a pivotal quality concern is the actual political stances of the workers. Here, the perceived bias of labelers was found to vary depending on their political stance. Other examples of ground-truth acquisition through crowdsourcing where workers biases may lead to subjective judgments include opinion detection, sentiment analysis etc. In general, the ability to mitigate biased judgments from workers is crucial in reducing noisy labels and creating higher quality data. To this end, we address the following research questions:

RQ#1: How does a worker’s personal opinion influence their performance on tasks including a subjective component? 

RQ#2: How can worker bias stemming from strong personal opinions be mitigated within subjective tasks?

RQ#3: How does a worker’s experience influence their capability to distance themselves from their opinion?

---

Based on the aforementioned observations in prior works, that suggest an influence of personal stances in subjective labeling tasks we construct the following hypotheses:

**H#1:** Workers are more likely to make a misclassification if such a classification is in line with their personal opinion.

**H#2:** Experienced workers are relatively less susceptible to exhibiting bias.

The main contributions of our work in this paper are:

- A novel measure for worker bias in subjective tasks based on misclassification rates and workers’ opinions.
- Novel techniques for mitigating systemic worker bias stemming from personal opinions.
- Revealing the impact of such systemic worker bias on aggregated ground-truth labels.

## 2 RELATED LITERATURE

### Bias in Crowdsourcing Data Acquisition

Recent works have explored task related factors such as complexity and clarity that can influence and arguably bias the nature of task-related outcomes [19]. Work environments (i.e., the hardware and software affordances at the disposal of workers) have also shown to influence and bias task related outcomes such as completion time and work quality [15]. Eickhoff studied the prevalence of cognitive biases (ambiguity effect, anchoring, bandwagon and decoy effect) as a source of noise in crowdsourced data curation, annotation and evaluation [9]. Gadiraju et al. showed that some crowd workers exhibit inflated self-assessments due to a cognitive bias [16]. Crowdsourcing tasks are often susceptible to participation biases. This can be further exacerbated by incentive schemes [10]. Other demographic attributes can also become a source of biased judgments. It has also been found that American and Indian workers differed in their perceptions of non-monetary benefits of participation. Indian workers valued self-improvement benefits, whereas American workers valued emotional benefits [31]. Newell and Ruths showed that intertask effects could be a source of systematic bias in crowdsourced tasks [41]. Other works revealed a significant impact of task order on task outcomes [1, 6]. Zhuang and Young [55] explore the impact of in-batch annotation bias, where items in a batch influence the labelling outcome of other items within the batch.

These prior works have explored biases from various standpoints; task framing and design, demographic attributes, platforms for participation and so forth. In contrast, we aim to analyze and mitigate the bias in subjective labeling tasks stemming from personal opinions of workers using the example task of bias detection.

### Subjective Annotations through Crowdsourcing

For many tasks such as detecting subjective statements in text (i.e., text pieces reflecting opinions), or biased and framing issues that are often encountered in political discourse [13, 47], the quality of the ground-truth is crucial. Yano et al. [54] showed the impact of crowd worker biases in annotating statements (without their context) where the labels corresponded to the political biases, e.g. very liberal, very conservative, no bias, etc. Their study shows that crowd workers who identify themselves as moderates perceive less bias, whereas conservatives perceive more bias in both ends of the spectrum (very liberal and very conservative). In a similar study, Iyyer et al. [29] showed the impact of the workers in annotating statements with their corresponding political ideology. In nearly 30% of the cases, it was found that workers annotate statements with the presence of a bias, however, without necessarily being clear in the political leaning (e.g. liberal or conservative). While it is difficult to understand the exact factors that influence workers in such cases, possible reasons may be their lack of domain knowledge, i.e., with respect to the stances with which different political ideologies are represented on a given topic, or it may be due to the political leanings of the workers themselves. Such aspects remain largely unexplored and given their prevalence they represent an important family of quality control concerns in ground-truth generation through crowdsourcing.

In this work, we take a step towards addressing these unresolved quality concerns of crowdsourcing for such subjective tasks by disentangling bias induced through strong personal opinions or stances.

### Mitigation of Bias

In large batches that consist of several similar tasks, Ipeirotis et al. showed that it is possible to use statistical methods and eliminate systematic bias [28]. The authors relied on synthetic experiments to do so. In other related work, Faltings et al. propose a game theoretic incentive scheme to counter the anchoring effect bias among workers [11]. Wauthier and Jordan [52] propose a machine learning model, which accounts for bias in a labelling task, where the labels are obtained through crowdsourcing. Here, the task is to predict labels, where consensus among the labellers is missing. Our work addresses the case where complete agreement among labellers, may still lead to a biased label. We explore various approaches to mitigate such undesirable bias, stemming from personal stances of workers.

Kamar et al. introduced and evaluated probabilistic models for identifying and correcting task-dependent bias [32]. Other lines of work [33, 37], rightly assume different expertise among the crowdsourcing workers, and thus propose models that improve over the majority voting label aggregation scheme. Such approaches are suitable for cases where there is disagreement among the workers. However in subjective tasks, the presence of varying ideological backgrounds of workers means that it is possible to observe biased labels with complete agreement among the workers, rendering such models inapplicable.
Raykar et al. [43] introduce an approach for combining labels provided by multiple types of annotators (experts and novices) to obtain a final high quality label. In contrast to their work, we aim to mitigate the effects of worker bias during the annotation process directly via interventions.

3 METHOD AND EXPERIMENTAL SETUP
In our study of crowd worker bias we focus on the task of labeling biased statements, a task that has found prominence in recent times to create ground truth data and evaluate methods for bias detection in text ([25, 26, 44]). We chose this task as an experimental lens due to its inherent susceptibility to worker subjectivity. In this task, workers are presented with statements pertaining to controversial topics and asked to decide whether the statement is “neutral” or “opinionated”. All statements revolve around a set of specific controversial topics wherein workers can be assumed to have diverging opinions. During the course of the task, we ask workers for their own opinion on each of the topics. Given this information, we define a measure of worker bias and investigate different approaches to mitigate potential bias.

Statement Extraction
We chose 5 controversial and widely discussed topics from US politics (Abortion, Feminism, Global Warming, Gun Control, and LGBT Rights) from Wikipedia’s List of controversial issues. We chose these popular and controversial topics so that a majority of crowd workers (from USA) could arguably have some basic understanding of the topic and an opinion.

For each of the chosen topics we selected a main statement that reflects the central pro/contra aspect of the controversy, e.g. “Abortion should be legal”. We extracted biased statements from the English Wikipedia using the approach introduced by [44] and [25] for articles that cover the given topics, e.g. LGBT rights by country or territory for LGBT rights. The approach relies on “POV” tags in comments for Wikipedia article revisions, which are added by Wikipedia editors for statements violating the NPOV principle. In this context Wikipedia provides an explanation of opinionated statements. By extracting statements that have been removed or modified for POV reasons, we obtained a set of biased statements for each topic. The strength of the opinionated words in a statement has been addressed in [25], but is beyond the scope of our work.

Authors of this paper acted as experts to validate that all statements in the final set contain explicit bias according to Wikipedia’s definition. Where necessary, we modified the statements briefly to make them comprehensible out of context. We removed phrases that where irrelevant or confusing (for example, we removed the phrase “resulting or caused by its death” from the statement “An abortion is the murder of a human baby embryo or fetus from the uterus resulting or caused by its death.”) and replaced very specific words to make the statements clearer and easier to understand (for example, we replaced “misandry” with “hate against men”).

We split the resulting set of biased statements into pro statements that support the main statement for this topic and contra statements that oppose the main statement. Additionally, we extracted neutral statements from the latest versions of the articles. We followed the process of open coding to ensure that the statements were reliably identified as pro, contra and neutral. We iteratively coded the resulting statements as either ‘pro’, ‘contra’, or ‘neutral’ until unanimous agreement was reached on each statement, thereby forming the ground truth for our experimental tasks.

Crowdsourcing Task Design
We manually selected 6 of the extracted statements for each topic; 2 pro, 2 contra, and 2 neutral statements. Our final statement set contains 30 extracted statements and 5 main statements. Table 1 shows the main statements and an extracted example statement for each topic.

Workers were asked to label each of the 30 extracted statements as either “neutral” or “opinionated”. We also provided a third option, “I don’t know”, which workers were encouraged to select in case they were not sure (see Figure 1).

Figure 1: Example statement labeling task corresponding to the topic of ‘Feminism’.

We also gathered each worker’s opinion corresponding to each topic from the statement group. We presented the main statement for each topic, and gathered responses from workers on a 5-point Likert scale ranging from 1: Strongly Disagree to 5: Strongly Agree (see Figure 2).

Figure 2: Example main statement to gather workers stances on the topic of ‘Gun Control’.

Table 1: The controversial topics chosen for this study together with the main statements and one example statement each from the corresponding Wikipedia articles.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Main Statement</th>
<th>Example Statement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abortion</td>
<td>Abortion should be legal.</td>
<td>An abortion is the murder of a human baby embryo or fetus from the uterus.</td>
</tr>
<tr>
<td>Feminism</td>
<td>Women have to fight for equal rights.</td>
<td>Feminists impose pressure on traditional women by denigrating the role of a traditional housewife.</td>
</tr>
<tr>
<td>Global Warming</td>
<td>Global warming is a real problem caused by humanity.</td>
<td>The global warming theory is perpetuated only for financial and ideological reasons.</td>
</tr>
<tr>
<td>Gun Control</td>
<td>Citizens should have free access to guns.</td>
<td>In some countries such as the United States, gun control may be legislated at either a federal level or a local state level.</td>
</tr>
<tr>
<td>LGBT Rights</td>
<td>Homosexual couples should have the same rights as heterosexual couples.</td>
<td>There are many inspiring activists who fight for gay rights.</td>
</tr>
</tbody>
</table>

Study Design
In our study we analyze worker behavior under different conditions with the goal of mitigating worker bias. We consider the following variations.

**Standard Bias Labeling Task (Baseline).** In this condition, we consider the standard bias labeling task as introduced in Section 3, without an explicit method or attempt for bias mitigation. Although it is now common practice to deploy crowdsourcing jobs with quality control mechanisms embedded in them [18, 35], it is still uncommon to control for biases stemming from worker opinions. Thus, we consider the more typical setting which is devoid of any form of bias control as a baseline condition for further comparisons.

**Social Projection (SoPro).** Two popular methods to induce honest reporting in the absence of a ground-truth are the Bayesian truth serum method (BTS) [42] and the peer-prediction method [40]. In a related study, Shaw et al. found that when workers think about the responses that other workers give then they work more objectively [49]. We draw inspiration from such truth-inducing methods as well as from the theory of social projection [22, 23] and aim to analyze the effect of social projection on mitigating biases stemming from worker opinions. In this condition workers are asked to label statements according to how they believe the majority of other workers would label them. We modified the task title and descriptions to adequately describe this condition. Apart from these minor changes, the task was identical to the baseline condition.

**Awareness Reminder (AwaRe).** Recent work has reflected on the importance of creating an awareness of existing biases in order to alleviate the biases [2]. We aim to analyze the impact of creating an awareness of biases stemming from personal opinions among workers, on their capability of being objective. In this condition we encouraged workers to reflect on the controversial nature of the topics in the task, and the potential bias that could be induced by their personal opinions on their judgments. We explore whether workers who are explicitly made aware of the subjective component in the task, go on to be more careful while making judgments.

We appended a message in the task description to create awareness among workers and presented 6 reminders at random intervals within the task bearing the identical message. To ensure that workers read or acknowledge the reminders, we created an interaction where workers were asked to type “YES” if they understood what was expected of them and “NO” otherwise. Figure 3 depicts the message snippet that serves as a reminder.

**Personalized Nudges (PerNu).** Similar to the AwaRe condition, here we investigate whether workers can deliberately influence the results of the task by distancing themselves from their personal opinions. In this condition, we first gather responses from workers on the main statements pertaining to each of the topics as shown in Figure 2. Using this knowledge of worker stances on a given topic (gathered on a 5-point Likert scale), we present personalized instructions to workers alongside each statement that is to be labeled. For example, if a worker strongly agrees with a main statement that ‘Citizens should have free access to guns’, then the worker receives a personalized instruction drawing attention to his potential bias while judging all statements related to ‘Gun Control’, as shown in the Figure 4. Note that the personalized instructions are phrased according to the degree of agreement or disagreement of the workers with the main statement.

Experimental Setup
For each task variation we deployed a job on FigureEight4, a primary crowdsourcing platform, and acquired responses from 120 workers. Each crowdsourcing job contained a task

4http://www.figure-eight.com/
Mitigating Worker Biases in Crowdsourced Collection of Subjective Judgments

CHI’19, May 2019, Glasgow, Scotland, UK

We first compute the misclassification rates for all types of misclassifications and all worker categories. The misclassification rate for a specific misclassification type is defined as the fraction of the number of misclassifications for a statement type ("pro", "contra", or "neutral") and the number of all judgments for statements of the same type. To assure that the bias measure is robust across different task variations, we normalize the misclassification rates for each worker category by computing the z-scores of each value.

According to hypothesis $H\#1$, due to the bias stemming from a worker’s personal opinions a (strong) supporter of topic $t$ is more likely to label a pro statement of topic $t$ as neutral, while a (strong) opposer of topic $t$ is more likely to label a contra statement of topic $t$ as neutral.

If hypothesis $H\#1$ holds, then (strong) supporters should be more likely to misclassify pro statements as being neutral compared to contra statements, i.e. the $pro\rightarrow neut$ misclassification rate should be comparatively higher than the $con\rightarrow neut$ misclassification rate. For (strong) opposers we should observe an opposing trend, where the $con\rightarrow neut$ misclassification rate should be higher than the $pro\rightarrow neut$ misclassification rate.

To test $H\#1$, we define bias for a worker category as the difference between the normalized $pro\rightarrow neut$ and the normalized $con\rightarrow neut$ values for this category. The following equation presents our measure for computing worker bias:

$$bias_{w_x} = \frac{m_{pro}(w_x) - \frac{\sum_{w_i \in w} m_{pro}(w_i)}{|w|}}{\frac{\sum_{w_i \in w} m_{con}(w_i)}{|w|}}\frac{\sigma}{\sigma}$$

where $m_{pro}(w_x)$ is the $pro\rightarrow neut$ misclassification rate, $m_{con}(w_x)$ is the $con\rightarrow neut$ misclassification rate for worker category $w_x$, and $w$ is the set of worker categories.

Relatively high positive values show that workers are more likely to regard a pro statement as neutral compared to a contra statement and therefore indicate pro bias. At the same time, relatively low negative values show that workers are more likely to regard a contra statement as neutral compared to a pro statement and therefore indicate contra bias. If $H\#1$ holds, we should observe a tendency towards pro bias for (strong) supporters and a tendency towards contra bias for (strong) opposers.

Note that a high misclassification rate alone does not necessarily indicate worker bias. It is possible that workers of a specific category generally perform badly in labeling opinionated statements. We therefore consider the misclassification rates for both $pro\rightarrow neut$ and $con\rightarrow neut$. The $neut\rightarrow op$ misclassification rate has no direct relation to worker bias since we cannot attribute a pro or contra bias to it.

**Measuring Worker Bias**

For each topic, we split workers into the 5 worker categories: strong opposer, opposer, undecided, supporter, strong supporter.

This categorization is based on the worker opinions of the main statement corresponding to a topic (gathered on a 5-point Likert scale), with strong opposer referring to ‘Strongly Disagree’ and strong supporter to ‘Strongly Agree’. We refer to the workers of the category strong opposer as strong opposers, and likewise for the other categories.

To measure worker bias, we focus on the misclassifications, i.e. the worker labels that do not coincide with the given ground-truth classes. We argue that incorrectly labeled statements can serve as indicators of worker bias. Given our task design, there are three different forms of misclassifications:

- A pro-statement labeled as neutral ($pro\rightarrow neut$).
- A contra-statement labeled as neutral ($con\rightarrow neut$).
- A neutral statement labeled as opinionated ($neut\rightarrow op$).

We also provided some labeling examples for both classes. To ensure reliability of responses, we restricted participation of workers on the platform to workers from Level 1 or above (2, 3). FigureEight workers are awarded level badges based on their accuracy across several test questions across hundreds of tasks of different types. Level 3 workers are the workers of the highest quality, followed by Level 2 and Level 1. In a multiple choice question, workers were asked to provide their FigureEight contributor level (1, 2 or 3). We also included two attention checks to filter out inattentive workers [39]. All job units (statements to label, main statements for opinion, attention checks, and contributor level question) appear in a random order to control for ordering effects. Workers who participated in one condition were not allowed to complete tasks in any other condition or to submit their responses only after completing the full set of units. Since the chosen topics focus on USA politics, we restricted participation on the platform to workers from the USA to avoid effects of domain knowledge. We compensated each worker at a fixed hourly rate of 7.5 USD based on our estimates of task completion time.
4 RESULTS AND ANALYSIS
In this section, we present the results of our study. We analyze and compare worker performance and bias in the 4 different variations described earlier.

Worker Categories
For each condition, we first filtered out workers who did not pass at least one of the two attention check questions. In case of the AwaRe condition, we additionally filtered out workers who did not enter ‘YES’ in response to all the reminder snippets. This leaves us with 102 workers in the Baseline condition, 106 workers in the SoPro, 93 workers in the AwaRe, and 72 workers in the PerNu condition.

Worker distributions across categories for each condition and each topic are provided in the appendix (Figure 8).

Worker Performance
Table 2 shows the overall results for each condition. On average, workers perform well across all conditions with average misclassification rates of 0.20 (Baseline), 0.18 (SoPro), 0.15 (AwaRe), and 0.23 (PerNu). We found a significant difference in worker performance between the conditions; \( p = 2.3 \times 10^{-12}, F(3, 356) = 19.13 \) using a one-way ANOVA. Post-hoc Tukey-HSD test revealed a sig. diff. between Baseline and AwaRe \( (p = 0.001) \) with a small effect size (Hedge’s \( g = 0.12 \)).

We measure inter-worker agreement using Fleiss’ Kappa and Krippendorff’s \( \alpha \) [36]. For both measures, the agreement values of all the other conditions are higher compared to the Baseline with the highest agreement observed in the AwaRe condition using Fleiss’ Kappa and the SoPro condition according to Krippendorff’s \( \alpha \). The generally low to moderate inter-worker agreement is consistent with expected agreement in similar tasks [24].

Table 2: Worker performance, agreement, and average task completion time (TCT) across all conditions.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Baseline</th>
<th>SoPro</th>
<th>AwaRe</th>
<th>PerNu</th>
</tr>
</thead>
<tbody>
<tr>
<td># workers</td>
<td>102</td>
<td>106</td>
<td>93</td>
<td>72</td>
</tr>
<tr>
<td># judgments</td>
<td>3060</td>
<td>3180</td>
<td>2790</td>
<td>2160</td>
</tr>
<tr>
<td># misclassifications</td>
<td>618</td>
<td>565</td>
<td>424</td>
<td>502</td>
</tr>
<tr>
<td>Misclassification Rate</td>
<td>0.20</td>
<td>0.18</td>
<td>0.15</td>
<td>0.23</td>
</tr>
<tr>
<td>Fleiss’ Kappa</td>
<td>0.33</td>
<td>0.43</td>
<td>0.49</td>
<td>0.33</td>
</tr>
<tr>
<td>Krippendorff’s ( \alpha )</td>
<td>0.34</td>
<td>0.44</td>
<td>0.38</td>
<td>0.33</td>
</tr>
<tr>
<td>TCT (in mins)</td>
<td>7.00</td>
<td>7.37</td>
<td>9.13</td>
<td>8.88</td>
</tr>
</tbody>
</table>

The average task completion time (TCT) for workers in the Baseline and SoPro is ~7 mins. In case of the AwaRe and the PerNu condition we observe higher task completion times of ~9 mins, which can partly be attributed to the additional information snippets that we confront workers with in both conditions. A one-way ANOVA showed a significant difference in TCT across interventions; \( p = 0.012, F(3,356)=3.73 \). Post-hoc Tukey-HSD test revealed a significant difference between TCT w.r.t. Baseline and AwaRe \( (p = 0.025) \) with a medium effect size (Hedge’s \( g = 0.44 \)).

Figure 5 illustrates the misclassification rates per worker category for each misclassification type and each condition. We refer to the different types of misclassifications as introduced in Section 3. Workers selected the “I don’t know” option in only 2.8% of cases. We did not consider these to be misclassifications. Using Welch’s T-test, we found that workers in all categories and across all conditions label a pro statement as being neutral significantly more often than they label a contra statement as neutral; \( t(1424) = 18.781, p < .001 \). We also found a large effect size; Hedge’s \( g = 0.70 \).

The rate of neutral statements being misclassified as opinionated appears to be consistent among different worker categories in the Baseline condition. In the SoPro and PerNu conditions (strong) supporters exhibit a lower misclassification rate for \( \text{neut} \rightarrow \text{op} \), while in the AwaRe condition (strong) opposers exhibit a lower misclassification rate. Across all conditions we did not find correlation between worker categories and the accuracy of judging neutral statements.

Worker Bias
As stated in Section 3, we measure worker bias as the difference between the normalized misclassification rates for pro and contra statements. The normalized misclassification rates and the resulting bias for all worker categories are presented in Table 3. High positive values indicate pro bias and low negative values indicate contra bias. A bias value close to 0 indicates that the group of workers is not biased to either the pro or contra side. For the sake of convenience while making comparisons across conditions, we use the notion of total bias, which is the sum of the absolute bias values in each condition.

First, we will focus on the Baseline condition to analyze results in the absence of a bias mitigation approach. We found that the misclassification rates for pro statements are similarly high for all worker categories, with the largest rate for strong supporters (1.26). In case of contra statements we found a larger gap between strong supporters (-1.36) and strong opposers (-1.76), meaning that strong opposers are significantly more likely to misclassify a contra statement as neutral compared to strong supporters. We conducted a one-way ANOVA to investigate the effect of the worker category on the \( \text{con} \rightarrow \text{neut} \) misclassification rate. We found a significant difference between the 5 worker categories at the \( p < 0.05 \) level; \( F(4,509) = 2.85 \). Post-hoc comparisons using the Tukey-HSD test revealed a significant difference between strong supporters and strong opposers at the \( p < 0.05 \) level with a medium effect size; Hedge’s \( g = 0.54 \).

The bias measure shows that strong supporters exhibit pro bias (2.62) and strong opposers exhibit contra bias (-2.75). We do not observe such tendencies in case of the other worker categories.
Table 3: Normalized misclassification rates (z-score) and worker bias for all worker categories and all conditions. For the bias column, positive values indicate pro bias, negative values indicate contra bias. Total bias is the sum of the absolute bias values. The introduced mitigation approaches achieve lower (total) bias values compared to the baseline.

Table 4: Misclassification rates for random worker samples from different combinations of worker categories. all = all workers, strong = strong supporters/strong opposers, strong = without strong supporters/strong opposers. Sample sizes N = 3 and N = 5. Lowest misclassification rates are highlighted for each condition. Including only workers with a strong opinion leads to higher misclassification rates.

Bias Mitigation

As depicted in Table 3, we see that the total bias for all three mitigation approaches is reduced compared to Baseline with AwaRe achieving the lowest score. A one-way ANOVA revealed a significant effect of our interventions on the total bias measure: $p = 0.0002$, $F(3,1784) = 6.57$. Post-hoc Tukey-HSD test revealed a significant difference between Baseline and AwaRe ($p = 0.036$) with a small effect size (Hedge’s $g = 0.19$), significant difference between AwaRe and PerNu ($p = 0.001$) with a slightly larger effect size (Hedge’s $g = 0.30$), and significant difference between SoPro and PerNu ($p = 0.022$) with a small effect size (Hedge’s $g = 0.19$).

Using Wilcoxon signed-rank tests we found that in the case of con $\rightarrow$ neut, AwaRe performs significantly better against Baseline ($p < .05$, effect size: Hedge’s $g=0.24$) and PerNu ($p < .01$, effect size: Hedge’s $g=0.72$). In the case of pro $\rightarrow$ neut the misclassification rates do not show any significant difference, apart from AwaRe being significantly better than PerNu ($p < .01$, effect size: Hedge’s $g=0.45$). To control for Type-I error inflation in our multiple comparisons, we used the Holm-Bonferroni correction for family-wise error rate (FWER) [21], at the significance Level of $\alpha < .05$.

SoPro. In this condition we found that the normalized pro $\rightarrow$ neut misclassification rate for strong supporters drops to -0.01 leading to a decrease in bias compared to the Baseline. Additionally, we found that the pro $\rightarrow$ neut misclassification rate increases to 1.83 for strong opposers, leading to a drop in bias for strong opposers (0.42). This supp bias value is closest to 0 for all conditions. Interestingly, we observe a change in bias for the undecided worker category from 1.30 to -2.19.
**Aware.** For this condition, we found that the pro→neut misclassification rate for strong supporters drops further to -1.72 when compared to the Baseline and SoPro conditions. This leads to a small bias that is closer to 0 when compared to the other conditions. For strong opposers we see a bias drop compared to the Baseline, from -2.75 to -1.31. The consequent total bias in the Aware condition was found to be the lowest across all conditions.

**PerNa.** In this case we note that we obtain the highest total bias score amongst our proposed approaches. We still see a non-significant drop in bias for both strong supporters and strong opposers compared to the Baseline.

**Impact of Worker Categories on Resulting Quality**

An important element in the creation of high-quality ground-truth using crowdsourcing is a diversity of opinion that can manifest from acquiring multiple independent judgments from workers [51]. One of the simplest methods used for aggregating multiple judgments in crowdsourced tasks is majority voting [27]. In the absence of gold-standard data, and especially for subjective tasks, majority voting or a variation of the algorithm is arguably a popular aggregation technique. Thus, to analyze the potential impact of worker categories on the resulting quality of aggregated judgments, we consider majority voting.

Consider a typical microtask crowdsourcing platform; task completion is generally driven by a self-selection process where workers pick and complete tasks they wish to [7]. Various factors ranging from worker motivation [34, 45] to marketplace dynamics such as task availability [8, 30], dictate which workers end up self-selecting and completing a given task from the available group of workers at any given point in time. Based on our findings pertaining to worker categories, we know that strong supporters and strong opposers correspond to the most systemic bias (see Baseline condition in Table 3). To measure the impact of workers from different categories on the average quality of aggregated judgments, we carry out simulations consisting of randomly selected workers from all categories. To this end, considering all worker categories, we ran 10,000 simulations of acquiring judgments from randomly teamed worker combinations with N=3 and N=5 for each of the 30 statements. Requesters often use 3 or 5 workers to gather redundant judgments and ensure quality. This is also recommended practice on FigureEight. Thus, this setting replicates standard crowdsourcing task configurations of obtaining multiple judgments from workers and assigning a label after aggregation.

To investigate the impact of strong supporters and strong opposers on the resulting quality, we consider three grouping strategies. ‘all’ considers the set of all workers (here N=3 or N=5 workers are picked at random from the entire pool of all workers), ‘strong’ is the set of workers who exhibited a strong bias; strong supporters for pro→neut and strong opposers for con→neut (here N=3 or N=5 workers are picked at random from the subset of strong supporters and strong opposers), and ‘strong’ is the set of workers present in all after filtering out all workers from the strong subset. Table 4 presents the average pro→neut and con→neut misclassification rates for worker groups across the 10,000 runs in each of the conditions.

When randomly selecting worker samples from the full set of workers (all), we see that the total misclassification rates drop as compared to the average misclassification rates per worker in Figure 5. This shows that groups of workers achieve higher accuracy, even when workers with strong opinions are included.

In the Baseline scenario, the misclassification rate of the strong worker group exhibits higher misclassification rate when compared to strong. We assess the significance of the misclassification rate through the Kruskal-Wallis test, which yields a significant difference with $p < .01$. This result is intuitive as the presence of workers in the end of both extremes (ssup and sopp), adds to the amount of biased judgments collected for a given subjective task.

**Effects of Worker Level**

Workers on the FigureEight platform can earn three different Level badges based on their accuracy and experience over time. In the FigureEight job settings, Level 1 is described as “All qualified contributors”, Level 2 as a “Smaller group of more experienced, higher accuracy contributors”, and Level 3 as the “Smallest group of most experienced, highest accuracy contributors”. We acquired self-reported worker Levels in our study. This allows us the opportunity to analyze potential correlations between worker performance/bias and the worker experience as represented by the worker level.

![Figure 6: No. of workers per worker Level for all conditions.](image)

Figure 6 shows the distributions of workers of each Level across the different conditions. In all conditions, Level 3 workers are the largest group of workers. There were more Level 1 workers than Level 2 workers in the Baseline condition, while for the other two conditions the number of Level 2 workers is higher.

Table 5 shows the average misclassification rates for the different approaches and the corresponding worker levels. Overall, the results vary. While we see high bias values for
Table 5: Misclassification rates and bias for each worker level and specific worker categories and orientations. The total shows the overall misclassification rate for workers of the given level. For each condition, we highlight the highest pos. bias score for ssup and the highest neg. bias score for sopp across the worker levels. The results show that no single level group clearly outperforms the other level groups across conditions.

<table>
<thead>
<tr>
<th></th>
<th>Baseline Level 1</th>
<th>Baseline Level 2</th>
<th>Baseline Level 3</th>
<th>SoPro Level 1</th>
<th>SoPro Level 2</th>
<th>SoPro Level 3</th>
<th>AwaRe Level 1</th>
<th>AwaRe Level 2</th>
<th>AwaRe Level 3</th>
<th>PerNu Level 1</th>
<th>PerNu Level 2</th>
<th>PerNu Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>sopp pro</td>
<td>0.29</td>
<td>0.40</td>
<td>0.32</td>
<td>0.29</td>
<td>0.44</td>
<td>0.26</td>
<td>0.14</td>
<td>0.24</td>
<td>0.29</td>
<td>0.35</td>
<td>0.47</td>
<td>0.37</td>
</tr>
<tr>
<td>sopp con</td>
<td>0.11</td>
<td>0.09</td>
<td>0.05</td>
<td>0.02</td>
<td>0.13</td>
<td>0.07</td>
<td>0.03</td>
<td>0.04</td>
<td>0.03</td>
<td>0.17</td>
<td>0.17</td>
<td>0.25</td>
</tr>
<tr>
<td>sopp bias</td>
<td>0.97</td>
<td>2.45</td>
<td>0.52</td>
<td>1.04</td>
<td>2.37</td>
<td>-0.60</td>
<td>-0.04</td>
<td>0.48</td>
<td>-0.58</td>
<td>-1.32</td>
<td>-0.84</td>
<td>1.11</td>
</tr>
<tr>
<td>sopp pro</td>
<td>0.23</td>
<td>0.21</td>
<td>0.33</td>
<td>0.46</td>
<td>0.28</td>
<td>0.38</td>
<td>0.43</td>
<td>0.19</td>
<td>0.33</td>
<td>0.33</td>
<td>0.50</td>
<td>0.21</td>
</tr>
<tr>
<td>sopp con</td>
<td>0.25</td>
<td>0.21</td>
<td>0.14</td>
<td>0.11</td>
<td>0.28</td>
<td>0.08</td>
<td>0.21</td>
<td>0.06</td>
<td>0.20</td>
<td>0.17</td>
<td>0.00</td>
<td>0.24</td>
</tr>
<tr>
<td>sopp bias</td>
<td>-2.55</td>
<td>-2.97</td>
<td>-1.21</td>
<td>1.49</td>
<td>-2.57</td>
<td>2.0</td>
<td>0.86</td>
<td>-1.22</td>
<td>-1.65</td>
<td>-1.39</td>
<td>1.71</td>
<td>-0.62</td>
</tr>
</tbody>
</table>

Level 2 workers for Baseline and SoPro, the bias values for AwaRe and PerNu are mixed. We computed a non-parametric Kruskal-Wallis test to assess the correlation between the worker level and their misclassification rates for pro and con statements. In this case, we do not distinguish between ssup and sopp. The test revealed that none of the bias differences between worker levels are significant. As a consequence, we do not control for worker level in our bias analysis.

Implication of Strong Supporters and Strong Opposers on Resulting Quality

Our findings show that the strong supporters and strong opposers are most susceptible to systemic bias due to their strong opinions. Let us consider the impact of a ssup or sopp contributing to a task where multiple judgments are aggregated using majority voting. In such a setting, ssup or sopp can bias a task outcome if there is a majority of either ssups or soppss in the cohort of workers annotating the same statement. To quantify the possible implication, we draw random samples of k workers (k = 1...102) from the Baseline condition and assess the fraction of resulting biased outcomes for each k, averaged across 10,000 iterations. Our findings are presented in Figure 7. We note that if 3 judgments are collected for each statement, over 17% of the statements end up with a biased label. With 5 judgments, over 15% end up with a biased label. This converges to around 10% around k=60. Requesters seldom gather so many judgments on a single statement, especially in large-scale jobs where costs are an important trade-off. This shows that the presence of ssup or sopp can be undesirable if their susceptibility to their opinions and the resulting bias is not mitigated.

5 DISCUSSION

Intuitively, workers belonging to the extreme categories (ssup and sopp) exhibit systematic bias stemming from their personal opinions. We found evidence of this, where ssup and sopp provided biased judgments inline with their stance on a given topic. Table 3 shows that ssup and sopp have the highest biased scores as a consequence of their unbalanced misclassification rates for pro and contra statements. This finding supports hypothesis H#1.

Impact on Resulting Ground-Truth Quality. Our findings suggest that negative effects on the created annotations due to biased workers can be effectively canceled out by increasing redundancy. By including non-biased workers in the ground-truth creation, misclassifications stemming from personal opinions can be averted. Importantly, the biggest threat to introducing systemic bias is having a group with a large majority of biased workers contributing to a task.

Implications of Bias. If we opt for a majority voting label aggregation scheme, even for fairly simple tasks with binary outcomes (i.e., "opinionated" or "neutral") the amount of judgments needed to overcome bias is very high. Figure 7 shows that if we consider odd numbers of judgments (less than 5), more than 20% of task units end up with a group of workers who are susceptible towards their strong stance (ssup or sopp). To reduce the amount of statements which end up with a majority of workers in the extreme categories, the number of judgments needed to achieve a tie is split between even and odd for readability. In case of a tie there is no majority of biased workers.
Bias Mitigation. To avoid biased judgments, we aim to mitigate bias through social projection (SoPro) or by making workers aware of their possible inclinations towards a topic (AwaRe and PerNu).

Our analysis results show that all three approaches reduce the total bias and average bias for workers with extreme opinions when compared to the Baseline. Additionally, SoPro and AwaRe lead to an improvement in general worker performance by reducing the overall number of misclassifications, therefore increasing the quality of resulting ground-truth labels in general. We achieve the highest total bias reduction rate of 49\% with the AwaRe approach as well as the highest bias reduction for ssups. For sopps the SoPro approach receives the highest reduction and might therefore be the preferred approach for situations with a large number of sopps. Another trade-off is the task completion time (TCT). The average TCT for AwaRe is significantly higher in our analysis compared to the Baseline, while the increase for SoPro is not significant.

We note that the most sophisticated approach PerNu provides significantly worse results for both bias and general worker performance. This behaviour can possibly be explained through the theory of central and peripheral persuasion from marketing research [12]. The AwaRe approach falls into the category of peripheral persuasion, where general reminder snippets increase worker awareness regarding the potential bias entailing the task. In contrast, PerNu can be seen as a central persuasion technique, where personalized instructions are provided at the statement level, thus, actively and directly informing workers while they make their judgments. According to [12], peripheral persuasion techniques are most suitable in inflicting attitude change when compared to central persuasion ones. We will investigate this further in our future work.

Worker Level Effects. Our results across different worker levels show that there is no significant difference in bias scores between workers with varying experience. This shows that filtering by levels is not a reliable strategy for mitigating worker bias. As a consequence, due to the lack of support we reject hypothesis H\#2. We note that Level 1 and 2 workers appear to be more receptive of treatment interventions provided by the different bias mitigating approaches. However, further qualitative studies are needed to establish this.

6 CONCLUSIONS

Systematic bias stemming from worker opinions can be a major problem in subjective labeling tasks. We showed that crowdsourced ground-truth annotations are susceptible to potentially biased workers who tend to produce systematically biased and noisy labels.

Our results show that judgments of workers who have extreme personal stances (i.e. strong supporters or strong opposers) pertaining to a particular topic, show a significant tendency to be influenced by their opinions. We found that performance or experience indicators like worker levels, do not play a significant role in reducing misclassification rates, making such indicators unreliable in mitigating systemic biases stemming from opinions in subjective tasks.

To mitigate such aforementioned worker bias, we proposed interventions based on social projection and making workers aware of their personal stances and potential biases, thus encouraging them to set aside their personal opinions during the course of task completion. Our approaches, SoPro and AwaRe provide significant improvement in terms of both worker bias reduction and general worker performance. Finally, we found that the PerNu approach, which actively provides the worker with personalized bias-related feedback during the task completion, does not provide any improvement over the other bias mitigation approaches.

A WORKER DISTRIBUTIONS ACROSS CATEGORIES

Figure 8 shows the distributions of workers across worker categories. There is a tendency towards ssups and sups for all topics except Gun Control, where the tendency is more towards sopps and opps. This suggests that workers on Figure Eight tend to be more liberal in their views on average, with Abortion, Feminism, Global Warming, and LGBT Rights being traditionally supported by liberals in the US. Our findings are inline with similar observations in [54], where the task was to assess how biased a statement is w.r.t liberal vs. conservative bias.

Acknowledgments. This work is partially supported by the ERC Advanced Grant ALEXANDRIA (grant no. 339233), DESIR (grant no. 731081), AFEL (grant no. 687916), DISKOW (grant no. 60171990) and SimpleML (grant no. 01IS18054).
REFERENCES


