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

Christoph Hube, Besnik Fetahu, Ujwal Gadiraju
L3S Research Center, Leibniz Universtät 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?

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

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

[1] http://www.people-press.org/2007/01/24/global-warming-a-divide-on-causes-and-solutions

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*[2]. 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 [3]. 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 [50]. 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).

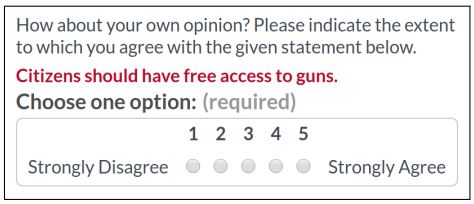

Figure 2: Example main statement to gather workers stances on the topic of '*Gun Control*'.

[2]https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues
[3]https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view

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

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

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.

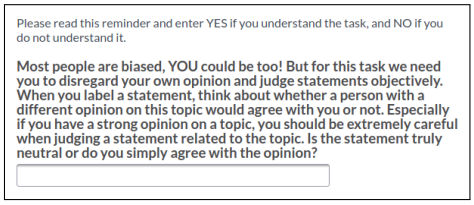

Figure 3: Message snippets serve as reminders to create awareness of potential biases in the *AwaRe* condition.

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 FigureEight[4], a primary crowdsourcing platform, and acquired responses from 120 workers. Each crowdsourcing job contained a task

[4]http://www.figure-eight.com/

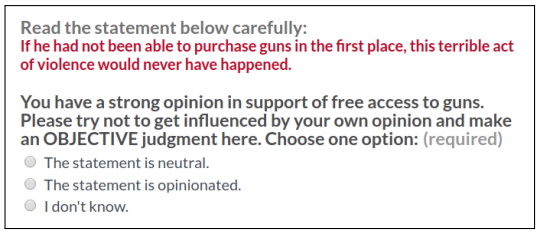

Figure 4: Example personalized instruction to workers who strongly agree that citizen should have free access to guns, on statements related to '*Gun Control*' in *PerNu* condition.

description including a brief explanation of "neutral" and "opinionated" statements. We also provided some labeling examples for both classes. To ensure reliability of responses, we restricted participation of workers on the platform to 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 check questions 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 to avoid potential learning effects. Workers were allowed 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.

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*⟶*neut*).
- A contra-statement labeled as neutral (*con*⟶*neut*).
- A neutral statement labeled as opinionated (*neut*⟶*op*).

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*⟶*neut* misclassification rate should be comparatively higher than the *con*⟶*neut* misclassification rate. For (strong) opposers we should observe an opposing trend, where the *con*⟶*neut* misclassification rate should be higher than the *pro*⟶*neut* misclassification rate.

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

$$
\begin{aligned}
bias_{w_x} &= \frac{\left(m_{pro}(w_x) - \frac{\sum_{w_i \in w} m_{pro}(w_i)}{|w|}\right)}{\sigma} \\
&\quad - \frac{\left(m_{con}(w_x) - \frac{\sum_{w_i \in w} m_{con}(w_i)}{|w|}\right)}{\sigma} \\
&= zscore(m_{pro}(w_x)) - zscore(m_{con}(w_x))
\end{aligned}
\tag{1}
$$

where $m_{pro}(w_x)$ is the *pro*⟶*neut* misclassification rate, $m_{con}(w_x)$ is the *con*⟶*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*⟶*neut* and *con*⟶*neut*. The *neut*⟶*op* misclassification rate has no direct relation to worker bias since we cannot attribute a pro or contra bias to it.

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; $p = 2.3e\text{-}12$, $F(3, 10709)=19.13$ 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 α [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 α. 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.

	Baseline	SoPro	AwaRe	PerNu
# workers	102	106	93	72
# judgments	3060	3180	2790	2160
# misclassifications	618	565	424	502
Misclassification Rate	0.20	0.18	0.15	0.23
Fleiss' Kappa	0.33	0.43	0.49	0.33
Krippendorff's α	0.34	0.44	0.38	0.33
TCT (in mins)	7.00	7.37	9.13	8.88

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 (*neut→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 *con→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.

Figure 5: Misclassification rates for all conditions and all worker categories: ssup = *strong supporters*, sup = *supporters*, und = *undecided*, opp = *opposers*, sopp = *strong opposers*. The misclassification types are: pro as neutral, neutral as opinionated, and contra as neutral.

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.

	Baseline			**SoPro**			**AwaRe**			**PerNu**		
	pro→neut	con→neut	bias	pro→neut	con→neut	bias	pro→neut	con→neut	bias	pro→neut	con→neut	bias
strong supporter	1.26	-1.36	2.62	-0.01	-1.08	1.07	-1.72	-1.20	-0.52	1.16	-0.64	1.80
supporter	-1.00	-0.08	-0.93	-0.30	-0.79	0.49	0.21	-0.50	0.71	0.13	-0.40	0.53
undecided	1.13	-0.18	1.30	-1.21	0.98	-2.19	1.23	1.40	-1.17	0.64	1.90	-1.26
opposer	-0.39	-0.15	-0.25	-0.32	-0.53	0.20	0.63	-0.67	1.29	-0.15	0.05	-0.20
strong opposer	-0.99	1.76	-2.75	1.83	1.41	0.42	-0.34	0.96	-1.31	-1.79	-0.91	-0.88
total bias			7.85			4.37			4.00			4.47

Table 4: Misclassification rates for random worker samples from different combinations of worker categories. *all* = all workers, *strong* = strong supporters/strong opposers, $\overline{strong}$ = 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.

		Baseline		**SoPro**		**AwaRe**		**PerNu**	
		pro→neut	con→neut	pro→neut	con→neut	pro→neut	con→neut	pro→neut	con→neut
N=3	*all*	0.272	0.042	0.255	0.033	**0.223**	0.018	0.349	**0.114**
	strong	0.353	0.141	0.267	0.060	0.236	0.035	0.420	0.0348
	$\overline{strong}$	**0.267**	**0.037**	**0.249**	**0.028**	0.251	**0.015**	**0.329**	0.115
N=5	*all*	0.242	0.017	**0.224**	0.011	**0.192**	0.005	0.329	0.068
	strong	0.326	0.121	0.228	0.015	0.210	0.016	0.405	**0.023**
	$\overline{strong}$	**0.237**	**0.014**	**0.224**	**0.009**	0.230	**0.003**	**0.308**	0.069

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 meaure; $p = 0.0002$, $F(3, 1784) = 6.57$. 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 → 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 → 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→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→neut* misclassification rate increases to 1.83 for *strong opposers*, leading to a drop in bias for *strong opposers* (0.42). This *sopp* 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.

PerNu. 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 '$\overline{strong}$' 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 $\overline{strong}$. 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.

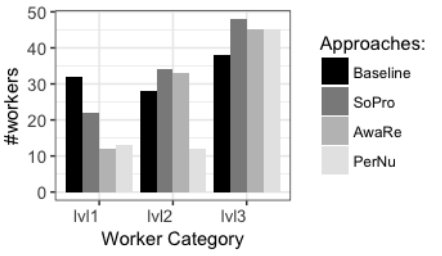

Figure 6: No. of workers per worker Level for all conditions.

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.

	Baseline			**SoPro**			**AwaRe**			**PerNu**		
	Level 1	Level 2	Level 3	Level 1	Level 2	Level 3	Level 1	Level 2	Level 3	Level 1	Level 2	Level 3
total	0.21	0.19	0.19	0.20	0.21	0.15	0.21	0.14	0.15	0.21	0.20	0.25
ssup pro	0.29	0.40	0.32	0.29	0.44	0.26	0.14	0.24	0.29	0.35	0.47	0.37
ssup con	0.11	0.09	0.05	0.02	0.13	0.07	0.03	0.04	0.03	0.17	0.17	0.25
ssup bias	0.97	**2.45**	0.52	1.04	**2.37**	-0.60	-0.04	**0.48**	-0.58	-1.32	-0.84	**1.11**
sopp pro	0.23	0.21	0.33	0.46	0.28	0.38	0.43	0.19	0.33	0.33	0.50	0.21
sopp con	0.25	0.21	0.14	0.11	0.28	0.08	0.21	0.06	0.20	0.17	0.00	0.24
sopp bias	-2.55	**-2.97**	-1.21	1.49	**-2.57**	2.0	0.86	-1.22	**-1.65**	**-1.39**	1.71	-0.62

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 asses 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 *ssup*s or *sopp*s 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

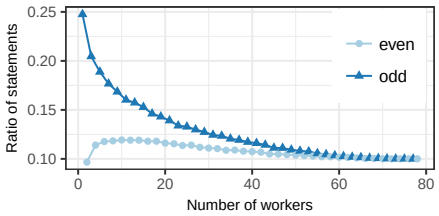

Figure 7: Ratio of statements whose labels are provided by a majority of biased workers after k workers (random worker samples from the *Baseline* condition, averaged over $10K$ iterations). 'k' is split between even and odd for readability. In case of a tie there is no majority of biased workers.

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, *"opinonated"* 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 needs to be extremely high, i.e, more than 40. Contrary, in the case of *even* number of judgments, this ratio is lower due to the fact that often we end up with a tie, and thus cannot employ the majority voting scheme. Therefore, in such cases we may end up with a large portion of statements without a clear aggregated label.

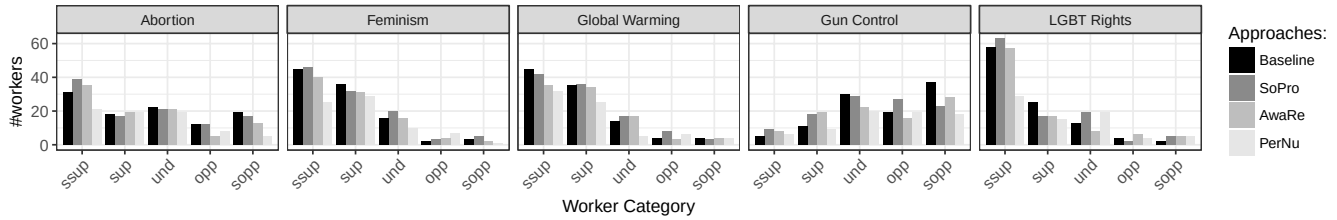

Figure 8: No. of workers for each topic, condition, and worker category.

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 *ssup*s. For *sopp*s the *SoPro* approach receives the highest reduction and might therefore be the preferred approach for situations with a large number of *sopp*s. 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 *ssup*s and *sup*s for all topics except Gun Control, where the tendency is more towards *sopp*s and *opp*s. This suggests that workers on FigureEight 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

[1] Alan Aipe and Ujwal Gadiraju. 2018. SimilarHITs: Revealing the Role of Task Similarity in Microtask Crowdsourcing. In *Proceedings of the 29th on Hypertext and Social Media*. ACM, 115–122.

[2] Ricardo Baeza-Yates. 2018. Bias on the web. *Commun. ACM* 61, 6 (2018), 54–61.

[3] W Lance Bennett. 2016. *News: The politics of illusion*. University of Chicago Press.

[4] Maxwell T Boykoff and Jules M Boykoff. 2004. Balance as bias: global warming and the US prestige press. *Global environmental change* 14, 2 (2004), 125–136.

[5] Róger Brown. 1960. Gilman. *The Pronouns of the Power and Solidarity* (1960).

[6] Carrie J Cai, Shamsi T Iqbal, and Jaime Teevan. 2016. Chain reactions: The impact of order on microtask chains. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 3143–3154.

[7] Lydia B Chilton, John J Horton, Robert C Miller, and Shiri Azenkot. 2010. Task search in a human computation market. In *Proceedings of the ACM SIGKDD workshop on human computation*. ACM, 1–9.

[8] Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, Panagiotis G. Ipeirotis, and Philippe Cudré-Mauroux. 2015. The Dynamics of Micro-Task Crowdsourcing: The Case of Amazon MTurk. In *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015*. 238–247. https://doi.org/10.1145/2736277.2741685

[9] Carsten Eickhoff. 2018. Cognitive Biases in Crowdsourcing. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 162–170.

[10] Carsten Eickhoff and Arjen P de Vries. 2013. Increasing cheat robustness of crowdsourcing tasks. *Information retrieval* 16, 2 (2013), 121–137.

[11] Boi Faltings, Radu Jurca, Pearl Pu, and Bao Duy Tran. 2014. Incentives to counter bias in human computation. In *Second AAAI conference on human computation and crowdsourcing*.

[12] Gavan J Fitzsimons, J Wesley Hutchinson, Patti Williams, Joseph W Alba, Tanya L Chartrand, Joel Huber, Frank R Kardes, Geeta Menon, Priya Raghubir, J Edward Russo, et al. 2002. Non-conscious influences on consumer choice. *Marketing Letters* 13, 3 (2002), 269–279.

[13] Roger Fowler. 2013. *Language in the News: Discourse and Ideology in the Press*. Routledge.

[14] Liye Fu, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Tie-breaker: Using language models to quantify gender bias in sports journalism. *arXiv preprint arXiv:1607.03895* (2016).

[15] Ujwal Gadiraju, Alessandro Checco, Neha Gupta, and Gianluca Demartini. 2017. Modus operandi of crowd workers: The invisible role of microtask work environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 49.

[16] Ujwal Gadiraju, Besnik Fetahu, Ricardo Kawase, Patrick Siehndel, and Stefan Dietze. 2017. Using worker self-assessments for competence-based pre-selection in crowdsourcing microtasks. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 4 (2017), 30.

[17] Ujwal Gadiraju, Ricardo Kawase, and Stefan Dietze. 2014. A taxonomy of microtasks on the web. In *25th ACM Conference on Hypertext and Social Media, HT '14, Santiago, Chile, September 1-4, 2014*. 218–223. https://doi.org/10.1145/2631775.2631819

[18] Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. 2015. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1631–1640.

[19] Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. 2017. Clarity is a worthwhile quality: On the role of task clarity in microtask crowdsourcing. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*. ACM, 5–14.

[20] Stephan Greene and Philip Resnik. 2009. More than words: Syntactic packaging and implicit sentiment. In *Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics*. Association for Computational Linguistics, 503–511.

[21] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics* (1979), 65–70.

[22] David S Holmes. 1968. Dimensions of projection. *Psychological bulletin* 69, 4 (1968), 248.

[23] David S Holmes. 1978. Projection as a defense mechanism. *Psychological Bulletin* 85, 4 (1978), 677.

[24] Tobias Hossfeld, Christian Keimel, Matthias Hirth, Bruno Gardlo, Julian Habigt, Klaus Diepold, and Phuoc Tran-Gia. 2014. Best practices for QoE crowdtesting: QoE assessment with crowdsourcing. *IEEE Transactions on Multimedia* 16, 2 (2014), 541–558.

[25] Christoph Hube and Besnik Fetahu. 2018. Detecting Biased Statements in Wikipedia. In *Companion of the The Web Conference 2018 on The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 1779–1786.

[26] Christoph Hube and Besnik Fetahu. 2018. Neural Based Statement Classification for Biased Language. *arXiv preprint arXiv:1811.05740* (2018).

[27] Nguyen Quoc Viet Hung, Nguyen Thanh Tam, Lam Ngoc Tran, and Karl Aberer. 2013. An evaluation of aggregation techniques in crowdsourcing. In *International Conference on Web Information Systems Engineering*. Springer, 1–15.

[28] Panagiotis G Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*. ACM, 64–67.

[29] Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 1113–1122.

[30] Ayush Jain, Akash Das Sarma, Aditya Parameswaran, and Jennifer Widom. 2017. Understanding workers, developing effective tasks, and enhancing marketplace dynamics: a study of a large crowdsourcing marketplace. *Proceedings of the VLDB Endowment* 10, 7 (2017), 829–840.

[31] Ling Jiang, Christian Wagner, and Bonnie Nardi. 2015. Not Just in it for the Money: A Qualitative Investigation of Workers' Perceived Benefits of Micro-task Crowdsourcing. In *System Sciences (HICSS), 2015 48th Hawaii International Conference on*. IEEE, 773–782.

[32] Ece Kamar, Ashish Kapoor, and Eric Horvitz. 2015. Identifying and accounting for task-dependent bias in crowdsourcing. In *Third AAAI Conference on Human Computation and Crowdsourcing*.

[33] David R. Karger, Sewoong Oh, and Devavrat Shah. 2011. Iterative Learning for Reliable Crowdsourcing Systems. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*. 1953–1961. http://papers.nips.cc/paper/4396-iterative-learning-for-reliable-crowdsourcing-systems

[34] Nicolas Kaufmann, Thimo Schulze, and Daniel Veit. 2011. More than fun and money. Worker Motivation in Crowdsourcing-A Study on Mechanical Turk.. In *AMCIS*, Vol. 11. 1–11.

[35] Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 1301–1318.

[36] Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability. (2011).

[37] Qiang Liu, Jian Peng, and Alexander T. Ihler. 2012. Variational Inference for Crowdsourcing. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake*

Tahoe, Nevada, United States. 701–709. http://papers.nips.cc/paper/4627-variational-inference-for-crowdsourcing

[38] John Lyons. 1970. New horizons in linguistics, Volume. (1970).

[39] Catherine C Marshall and Frank M Shipman. 2013. Experiences surveying the crowd: Reflections on methods, participation, and reliability. In *Proceedings of the 5th Annual ACM Web Science Conference*. ACM, 234–243.

[40] Nolan Miller, Paul Resnick, and Richard Zeckhauser. 2005. Eliciting informative feedback: The peer-prediction method. *Management Science* 51, 9 (2005), 1359–1373.

[41] Edward Newell and Derek Ruths. 2016. How one microtask affects another. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 3155–3166.

[42] Dražen Prelec. 2004. A Bayesian truth serum for subjective data. *science* 306, 5695 (2004), 462–466.

[43] Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Anna Jerebko, Charles Florin, Gerardo Hermosillo Valadez, Luca Bogoni, and Linda Moy. 2009. Supervised Learning from Multiple Experts: Whom to Trust when Everyone Lies a Bit. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09)*. ACM, New York, NY, USA, 889–896. https://doi.org/10.1145/1553374.1553488

[44] Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic Models for Analyzing and Detecting Biased Language.. In *ACL (1)*. 1650–1659.

[45] Jakob Rogstadius, Vassilis Kostakos, Aniket Kittur, Boris Smus, Jim Laredo, and Maja Vukovic. 2011. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. *ICWSM* 11 (2011), 17–21.

[46] Suzanne Romaine et al. 2000. *Language in society: An introduction to sociolinguistics*. Oxford University Press.

[47] Dietram A Scheufele. 1999. Framing as a theory of media effects. *Journal of communication* 49, 1 (1999), 103–122.

[48] Gün R Semin and Klaus Fiedler. 1988. The cognitive functions of linguistic categories in describing persons: Social cognition and language. *Journal of personality and Social Psychology* 54, 4 (1988), 558.

[49] Aaron D Shaw, John J Horton, and Daniel L Chen. 2011. Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. ACM, 275–284.

[50] Anselm L Strauss. 1987. *Qualitative analysis for social scientists*. Cambridge University Press.

[51] James Surowiecki. 2005. *The wisdom of crowds*. Anchor.

[52] Fabian L. Wauthier and Michael I. Jordan. 2011. Bayesian Bias Mitigation for Crowdsourcing. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*. 1800–1808. http://papers.nips.cc/paper/4311-bayesian-bias-mitigation-for-crowdsourcing

[53] Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational linguistics* 30, 3 (2004), 277–308.

[54] Tae Yano, Philip Resnik, and Noah A Smith. 2010. Shedding (a thousand points of) light on biased language. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, 152–158.

[55] Honglei Zhuang and Joel Young. 2015. Leveraging In-Batch Annotation Bias for Crowdsourced Active Learning. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM 2015, Shanghai, China, February 2-6, 2015*. 243–252. https://doi.org/10.1145/2684822.2685301