LimitBias! Measuring Worker Biases in the Crowdsourced Collection of Subjective Judgments

Christoph Hube, Besnik Fetahu and Ujwal Gadiraju

L3S Research Center, Leibniz Universität Hannover
Appelstrasse 4, Hannover 30167, Germany
{hube, fetahu, gadiraju}@L3S.de

Abstract. Crowdsourcing results acquired for tasks that comprise a subjective component (e.g. opinion detection, sentiment analysis) are affected by the inherent bias of the crowd workers. This leads to weaker and noisy ground-truth data. In this work we propose an approach for measuring crowd worker bias. We explore worker bias through the example task of *bias detection* where we compare the worker's opinions with their annotations for specific topics. This is a first important step towards mitigating crowd worker bias in subjective tasks.

1 Introduction

Crowdsourcing is one of the most common means in obtaining ground-truth data for training automated models for a large variety of tasks [9, 7, 6]. In many cases, the annotations are affected by the subjective nature of tasks (e.g. opinion detection, sentiment analysis) or the biases of the workers themselves. For instance, for tasks like determining the political leaning or biased language in a piece of text the annotations, how we perceive something as liberal/conservative or biased is subject to several factors like *framing* and *epistemological* biases in language, social and cultural background of workers, etc. [3].

In this work, we aim at understanding and mitigating the worker biases in crowdsourced annotation tasks that are of subjective nature (e.g., political leaning of a statement, biased language etc.). In particular, we are interested in the case where for a given set of strict annotation rules how does the workers' bias influence their annotation quality. Furthermore, having this setting in mind, how can we mitigate such worker biases in subjective tasks. We propose an approach for measuring crowd worker bias based on the example task of labeling statements as either biased or neutral. In addition to the main task we ask workers for their personal opinion on each statement's topic. This additional information allows us to measure correlations between a worker's opinion and their choice of labeling. In future work we will introduce methods for mitigating the measured bias.

2 Related Work

2.1 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 [5]. Work envi-

ronments (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 [4]. In other closely related work, Eickhoff has studied the prevalence of cognitive biases as a source of noise in crowdsourced data curation, annotation and evaluation [1]. Eickhoff studied the effect size of common cognitive biases such as the ambiguity effect, anchoring, bandwagon and decoy effect in a typical relevance judgment task framework. Crowdsourcing tasks are often susceptible to participation biases. This can be further exacerbated by incentive schemes [2]. 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 [8].

In this work, we aim to disentangle the potential sources of worker bias using the example task of *bias detection*. This will be a first holistic approach towards bias management in crowdsourcing.

2.2 The Case of Subjective Annotations

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 [9, 3], the quality of the ground-truth is crucial.

Yano et al. [10] show 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*). Interestingly, the distribution of workers is heavily biased towards *moderates*. This raises several issues. First, how can we ensure a balanced representation of workers, where for subjective tasks a balanced representation is crucial. Second, which judgments are more reliable having in mind that more conservative workers tend to perceive statements as more biased in both ends of the political spectrum.

In a similar study to, Iyyer et al. [7] 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, respectively the stances with which different political ideologies are represented on a given topic, or it may be the political leanings of the workers themselves. Such aspects remain largely unexplored and given their prevalence they represent significant quality concerns in ground-truth generation through crowdsourcing.

In this work, we aim at addressing these unresolved quality concerns of crowdsourcing for subjective tasks by disentangling all the possible bias factors.

3 Measuring crowd worker Bias

In Section 1 we introduced the problem of measuring crowd worker bias for a crowdsourcing task including a subjective component. In this section we propose an approach for measuring crowd worker bias for the example task of labeling statements as either biased or neutral. The same approach can be used for other tasks as stated in Section 1.

For the example task we use statements from datasets of subjective and opinionated statements that have been extracted from Wikipedia [6] or ideological books [7]. We first create a set of 10 statement groups with each group containing statements for one controversial topic from a list of widely discussed controversial topics, e.g. *abortion*, *capital punishment*, *feminism*. Each statement group contains one main statement that reflects the central pro/against aspect of the controversy, e.g. "Abortion should be legal". In our task design we use the main statement to determine the worker's opinion on the given topic. Furthermore each group contains 4 additional opinionated statements from the dataset that follow the group's topic, two statements that support the main statement and two against it.

To accurately measure worker bias, we divide the task into two subtasks. In the first subtask we show the worker the opinionated statements. The worker has to label each statement as either "biased" or "neutral". We explain the concepts of biased and neutral wording to the workers and give them a guideline when to label a statement as biased or neutral. We give multiple examples for both classes. We additionally provide a third "I don't know" option. The task design for the first subtask is depicted in Figure 1. A similar task design has been used to create a ground truth for the problem of bias detection [6].

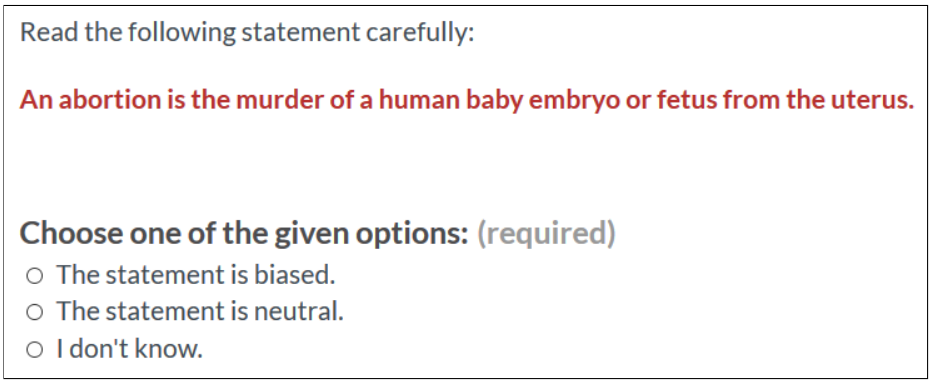

Fig. 1. Main task example.

In the second subtask we ask the worker's opinion for each topic from the statement group. We show the worker the main statement from each group and 5 options on a Likert scale reaching from "I strongly agree" to "I strongly disagree". The task design for the second subtask is depicted in Figure 2.

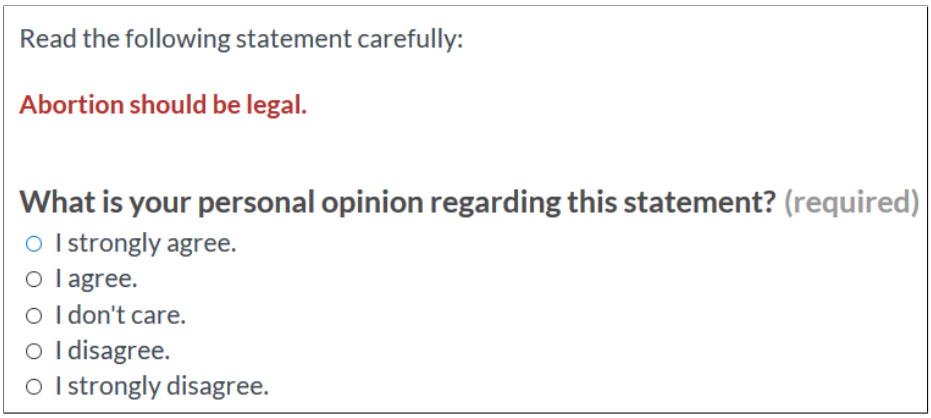

Fig. 2. Opinion example.

Our hypothesis is that workers who agree with a statement are more likely to label it as neutral, i.e. a worker who agrees that abortion should be illegal is more likely to label the statement "An abortion is the murder of a human baby embryo or fetus from the uterus" as neutral. As stated in Section 1 this behavior can negatively influence the crowdsourcing results of this task since crowd workers should label according to the given guidelines and not to personal opinion.

4 Future Work

We introduced an approach for measuring crowd worker bias for crowdsourcing tasks including a subjective component. For future work we are planning to develop methods for mitigating the measured bias. Possible approaches could include balancing judgments between workers of different opinions, making workers aware of their biases (meta-cognition), and discounting strongly biased crowdworkers. Furthermore we want to analyze the influence of task design on worker bias.

Acknowledgments This work is funded by the ERC Advanced Grant ALEXANDRIA (grant no. 339233), DESIR (grant no. 31081), and H2020 AFEL project (grant no. 687916).

References

1. Carsten Eickhoff. Cognitive biases in crowdsourcing. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 162–170. ACM, 2018.
2. Carsten Eickhoff and Arjen P de Vries. Increasing cheat robustness of crowdsourcing tasks. *Information retrieval*, 16(2):121–137, 2013.
3. Roger Fowler. *Language in the News: Discourse and Ideology in the Press*. Routledge, 2013.

4. Ujwal Gadiraju, Alessandro Checco, Neha Gupta, and Gianluca Demartini. 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):49, 2017.
5. Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. 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*, pages 5–14. ACM, 2017.
6. Christoph Hube and Besnik Fetahu. Detecting biased statements in wikipedia. In *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*, pages 1779–1786, 2018.
7. Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1113–1122, 2014.
8. Ling Jiang, Christian Wagner, and Bonnie Nardi. 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*, pages 773–782. IEEE, 2015.
9. Dietram A Scheufele. Framing as a theory of media effects. *Journal of communication*, 49(1):103–122, 1999.
10. Tae Yano, Philip Resnik, and Noah A Smith. 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*, pages 152–158. Association for Computational Linguistics, 2010.