Fairness and Bias in Truth Discovery Algorithms: An Experimental Analysis

Simone Lazier[*] Saravanan Thirmuruganathan[†] Hadis Anahideh[‡]

Abstract

Machine learning (ML) based approaches are increasingly being used in a number of applications with societal impact. Training ML models often require vast amounts of *labeled* data, and crowdsourcing is a dominant paradigm for obtaining labels from multiple workers. Crowd workers may sometimes provide unreliable labels, and to address this, truth discovery (TD) algorithms such as majority voting are applied to determine the consensus labels from conflicting worker responses. However, it is important to note that these consensus labels may still be biased based on sensitive attributes such as gender, race, or political affiliation. Even when sensitive attributes are not involved, the labels can be biased due to different perspectives of subjective aspects such as toxicity. In this paper, we conduct a systematic study of the bias and fairness of TD algorithms. Our findings using two existing crowd-labeled datasets, reveal that a non-trivial proportion of workers provide biased results, and using simple approaches for TD is sub-optimal. Our study also demonstrates that popular TD algorithms are not a panacea. Additionally, we quantify the impact of these unfair workers on downstream ML tasks and show that conventional methods for achieving fairness and correcting label biases are ineffective in this setting. We end the paper with a plea for the design of novel bias-aware truth discovery algorithms that can ameliorate these issues.

1 Introduction

Machine learning (ML) is becoming increasingly pervasive and is already having a significant impact on people's lives. Often, the decisions by these models impact different sub-groups in a non-uniform manner. Hence, there has been a flurry of research in designing fairness-aware machine learning algorithms. In this paper, we argue that such care must also be accorded to a crucial step in the ML process – collection of data labels. We situate our work at the intersection of two major paradigms – data-centric AI and fair ML. The former seeks to build AI systems with good-quality data while the latter ensures that the resulting system is fair. As we shall show shortly, the current process of label collection is simplistic and results in biased labels that have a disproportionate impact on the accuracy and fairness of the downstream ML models. While there has been extensive work on designing fair ML algorithms, there is limited work on developing algorithms to generate correct (and fair) labels.

Making of Data Sausage. Datasets are indispensable to the progress of the ML community allowing the design of novel models and serving as a benchmark for measuring advancements. Despite their importance, the vast majority of the work in ML is geared towards the development of new algorithms rather than data stewardship practice [18]. Recently, the fairness community has unearthed many issues with prominent benchmark datasets, such as label errors and bias [17, 19]. In this work, we systematically analyze another source of bias – labels generated through the simple aggregation of responses from biased crowd workers.

A key factor in the success of ML is the availability of large *labeled* datasets. However, creating accurate labels using a domain expert is prohibitively expensive. Hence, a common alternative is to use crowd workers for generating the labels. The workers are often hired using crowdsourcing marketplaces such as Amazon Mechanical Turk. Typically, the quality of these workers is sub-par compared to that of domain experts – albeit significantly more cost-effective. Quality control is often maintained by collecting multiple labels for each data point from various workers and implementing additional filtering mechanisms, such as pre-tests where a set of 'golden tasks' with known correct answers are given to workers before they begin labeling the main dataset [23]. It is possible that the labels of the workers conflict with one another. In this case, an aggregation method, more generally Truth Discovery algorithms (TD), is used to determine the final label. One common approach is to use majority voting. In many cases, the labels thus obtained are reasonably accurate and can be used to train ML models. A large

[*] University of Illinois Chicago. (slazie2@uic.edu)

[†] QCRI, HBKU. (sthirumuruganathan@hbku.edu.qa)

[‡] University of Illinois Chicago. (hadis@uic.edu)

number of major benchmark datasets were obtained using this approach. Even considering a single domain such as Computer Vision, a non-exhaustive list of such datasets include ImageNet [8], Places [24], MS COCO [14], ActivityNet [6], Open Street Map (OSM) dataset [11], YouTube-8M [2], and Intrinsic Images in the Wild [5]. The prevalence of crowdsourced datasets in other ML fields is equally significant. For example, as much as 10% of the papers published in Natural Language Processing uses crowdsourced dataset [21].

The pervasiveness of using crowdsourcing for obtaining labels is problematic. The use of human annotators is a double-edged sword. While humans can provide answers to challenging and subjective questions, their responses are also potentially subject to a variety of biases. With ML increasingly being utilized in sensitive applications, such as the detection of fake news, toxicity, and hate speech, the labels for these domains can be particularly subjective and involve nuanced concepts. Different workers may assign different labels to the same data point based on their lived-in experiences, which can result in the injection of subjective values and biases into the dataset. Using datasets with extensive amounts of biased labels can lead to models that are sub-par in terms of accuracy and fairness.

Summary of Our Contributions. We conduct a systematic investigation of the aggregation process for labels in crowdsourced datasets. Our study includes two publicly available diverse datasets – *Crowd Judgement* and *Twitter Toxicity*, and the results were consistent with those from our experiments on a number of other datasets in hate speech detection. Our findings indicate that biased workers are widely prevalent and cause a significant impact on the quality of the dataset labels. Surprisingly, it is possible for a worker to be accurate while also being biased , even though the dataset with the true labels might be biased itself. Hence, this precludes simplistic solutions such as removing biased workers as it could have collateral damage on the accuracy of the other items. There has been extensive work on truth discovery algorithms that seek to aggregate conflicting labels into a single 'truth'. However, we find that using such TD algorithms, which are more complex than majority voting, is not effective in solving the problem. We conduct analyses using two popular algorithms (DS and LFC) , Dawid Skene (DS) [7] and Learning From Crowds (LFC) [20], that have different probabilistic generative modeling for crowd worker bias. Nevertheless, their performance is not substantially better than that of majority voting. Furthermore, we study the impact of the biased labels on downstream ML tasks and show that each of the truth discovery algorithms suffers from similar flaws, leading to less accurate and unfair models. We also show that using fair ML algorithms does not eliminate the issue of biased labels.

We make the following key contributions:

1. We identify truth discovery algorithms for aggregating the responses of crowd workers as a major source of bias in crowdsourced datasets.
2. We conduct an extensive experimental study to examine the various facets of this bias and quantify the impact on downstream ML tasks.
3. We offer practical recommendations for practitioners and highlight promising research problems.

Paper Organization. The rest of the paper is organized as follows. In Section 2, we present the relevant background on crowdsourcing, fairness, and truth discovery algorithms. Section 3 outlines the datasets and research methodology used in our study. Our experimental analysis is structured along five main dimensions, described in detail in Sections 4-8. We conclude with some parting thoughts in Section 9.

2 Background

Crowdsourcing for ML and its Pitfalls. Over the last decade, crowdsourcing has been extensively used to advance machine learning research. Crowd workers have been used for data generation, label generation, and evaluation of ML models. The ML community has embraced crowdsourcing as a way to obtain image and linguistic annotations needed for training popular vision and language processing tasks. A key stumbling block is that the labels provided by the workers can be noisy or inaccurate. Workers can be imperfect, unmotivated, or influenced by various cognitive biases [10]. When the labels are generated for training ML models for tasks with societal impact, these biases can have serious consequences. A pioneering work identified various sources of bias that can affect ML algorithms [16]. Truth discovery algorithms have been proposed to address the former issues while fair ML algorithms have been proposed to address the latter.

Truth Discovery for Crowdsourcing. Achieving quality control in crowdsourcing involves determining the correct answers from the low-quality answers provided by crowd workers. Unfortunately, the vast majority of works use simple approaches such as majority voting, which is often insufficient. Sophisticated approaches are often based on probabilistic generative models that allow the modeling of various factors on the worker label generation. These factors could be worker-based (accuracy, bias, motivation, collusion, topic-related expertise, etc.) or task-related (difficulty, clarity, etc.). A taxonomy of truth discovery algorithms can be found in [23].

Truth Discovery - Problem Setup. Consider a set of tasks $\mathcal{T} = \{t_1, \ldots, t_n\}$ with the size of n. Let $\mathcal{W}$ be the set of annotators. and W_i be the vector of answers for task t_i collected by a subset of annotators in $\mathcal{W}$. Let m be the total number of annotators. Each task t_i could be labeled by a subset of workers. By querying from these annotators, we would like to determine the true class label y_i for any task $t_i \in \mathcal{T}$. We then denote the label of annotator j for task t_i as l_{ij}. Let $\tilde{y}_i^w$ be the answer of annotator w for task t_i and y be the unknown ground truth. Given the answers matrix of $\widetilde{Y}$, the ground truth inference of y_i for each task is identified as $\hat{y}_i$ using a truth discovery algorithm.

Algorithmic Fairness in ML. In recent years, the study of algorithmic fairness has gained increasing attention in the machine learning community. The objective of these studies is to ensure that AI systems do not discriminate against certain groups of individuals based on sensitive features such as race, gender, or age. A detailed description of the various aspects can be found in [4]. Despite the growing body of literature on fairness in machine learning, there are still a number of challenges and limitations that need to be addressed. For example, there is currently no widely accepted definition of fairness, which can lead to inconsistencies in the assessment of fairness in AI systems [16]. Additionally, many existing methods for mitigating algorithmic bias rely on adjustments to the modeling processes including preprocessing [22], in-processing [3], and post-processing [15] phases, rather than addressing the root causes of bias (e.g., data collection bias), which is especially the case in crowdsourcing settings.

3 Experimental Methodology

In this paper, we are interested a specific type of bias that results in unfair labels. For example, workers could provide incorrect labels for tasks involving members of a specific gender or race. Using TD algorithms such as majority voting perpetuates these biased labels into the dataset which in turn results in unfair models. Ideally, these unfair labels should be corrected during the label aggregation process.

In this section, we outline the experimental methodology we use to systematically study and quantify how truth discovery algorithms work in the presence of unfair workers. For ease of exposition, we have partitioned them into five research hypotheses. After describing the hypotheses, we provide more details about the experimental setting, such as the list of representative truth discovery algorithms and datasets.

Experimental Hypotheses. We conduct an extensive set of experiments to study how different TD algorithms impact the aggregation of labels from imperfect workers. Specifically, we focus on five key dimensions that expound the issue of bias in label aggregation. Sections 4-8 describe the corresponding research hypotheses in detail. In this section, we provide a high-level overview so that the reader has a holistic perspective of our experiments.

H1: Quantifying Prevalence of Unfair Workers through Analysis of Accuracy and Fairness Metrics. We begin by conducting an exploratory analysis of the accuracy and fairness metrics of individual workers. The results indicate that unfair workers are not anomalies and are instead widely prevalent.

H2: Impact of Unfair Workers on Label Flipping in Truth Discovery Algorithms. We show that unfair workers can potentially flip the labels for a non-trivial proportion of tasks. Next, we investigate the impact of unfair workers on TD algorithms to show that unfair workers can potentially flip the labels for a non-trivial proportion of tasks. Specifically, we focus on majority voting, which is a widely used method. It is worth noting that this issue affects not only majority voting but also other traditional truth discovery algorithms.

H3: Unfair Workers Reduce the Performance of TD Algorithms. Next, we investigate the performance of three representative truth discovery algorithms. Specifically, we compare the ground truth labels estimated by these methods with the true ground truth. We show that all of these algorithms have comparable (and sub-optimal) accuracy and fairness. This highlights the need for the development of new fairness-aware truth discovery algorithms, rather than relying on existing ones.

H4: Unfair Workers Negatively Impact the Downstream ML Tasks. The next hypothesis addresses a potential concern that the study of the performance of truth discovery algorithms is not merely academic. Instead, we demonstrate that training an ML model with the output of simple truth discovery algorithms result in a model that is both less accurate and less fair.

H5: The Complementarity of Fair ML and Fair TD. In the final hypothesis, we explore the idea that using fair ML algorithms cannot completely eliminate the bias introduced by traditional truth discovery algorithms. We also evaluate the effectiveness of FairTD, a fair truth discovery algorithm proposed in [13].

Truth Discovery Algorithms. There has been extensive work on techniques for aggregating conflicting responses from workers. In this paper, we compare three representative algorithms. For a detailed discussion of their comparative performance, please refer to

[23]. *Majority Voting (MV)* is a simple and widely used algorithm that chooses the response provided by the majority of the workers as the consensus. MV might work if the workers are reasonably competent and do not form any collusion. A key limitation of MV is that it weighs the response of each worker equally. However, this might be sub-optimal in a crowd setting where workers can have divergent accuracy. For example, if only one worker is an expert while others are novices, MV may fail. In such cases, it is important to model the worker accuracy so as to compute a *weighted* majority voting. Dawid-Skene (DS) [7] is a more sophisticated model that assigns different weights to workers based on the content of their confusion matrix. Worker accuracy is estimated using an iterative EM algorithm. Learning from Crowd (LFC) [20] is a probabilistic model that proposes a two-coin approach for annotators. Each worker has two biased coins with bias corresponding to sensitivity and specificity. The worker flips one of these coins based on the true label of the task and uses the outcome to determine their response. This approach has been used successfully in many domains, such as aggregating the predictions of radiologists. All three models have been widely used and have shown empirical success. In this paper, we investigate how these models perform in the presence of worker bias.

Fairness Metrics. There has been extensive work on measuring and mitigating bias in datasets and ML models. However, there is not much consensus on the appropriate fairness metrics as different metrics require that some aspect of the ML model remains comparable to all subgroups. In this paper, we use two key widely used metrics that are designed to evaluate the fairness of ML models – demographic parity and equalized odds. By treating individual crowd workers as fallible oracles (similar to ML models), we can reuse these two metrics for measuring the fairness of workers. Demographic parity (DP) is achieved when the probability of a particular prediction (such as loan disbursal) is not dependent on sensitive group membership (such as gender=Male). Equalized odds (EO) fairness metric requires that the predictions of the ML model be both independent of sensitive group membership *and* have comparable true and false positive rates. It is important to note that EO is stricter than DP. To evaluate the fairness of our models more effectively, we use two derived metrics for DP and EO – difference and ratio. A difference of 0 indicates the DP/EO has been achieved, and a ratio of 1 indicates the same. These derived metrics are useful in that they allow us to express the fairness metrics as a single scalar for easier comparison.

Datasets. We conduct our experiments using two existing diverse and representative datasets from different domains. The *Crowd Judgement* dataset [9] contains labels of the workers from a major crowdsourcing marketplace for 1000 cases from the infamous COMPAS dataset for recidivism. Groups of 20 workers evaluated the same set of 50 defendants. The second dataset *Jigsaw Toxicity* contains labels of workers on whether a Twitter comment is toxic. This dataset has been collected in [1] We used a subset of 14000 comments that were labeled for toxicity, obscenity, threats, insults, and hate.

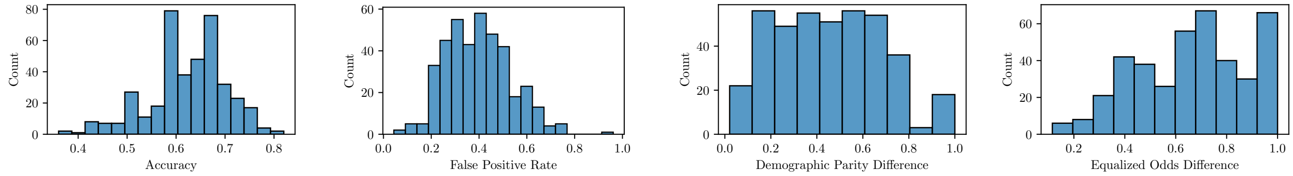
4 H1: Quantifying Prevalence of Unfair Workers through Analysis of Accuracy and Fairness Metrics

Our first set of experiments conducts an exploratory analysis of the dataset to compute the accuracy and fairness metrics of the workers. Despite conventional wisdom that workers are often imperfect and/or unfair, quantifying these factors is crucial for understanding the challenges truth discovery algorithms face.

Both the datasets (*Crowd Judgement* and *Jigsaw Toxicity*) consist of ground truth. This contrasts with typical crowdsourcing datasets where the ground truth is often unavailable. This metadata allows us to measure worker accuracy and fairness. Specifically, we focus on two accuracy-related metrics (accuracy and false positive rate) and two fairness-related metrics (demographic and equalized odds difference). We conduct three interrelated sets of experiments. In the first set of experiments, we generate histograms of the four accuracy and fairness metrics. Second, we create a scatter plot that contrasts an accuracy metric with a fairness metric. Finally, we show the accuracy-fairness distribution in an alternative tabular format as it provides unique insights.

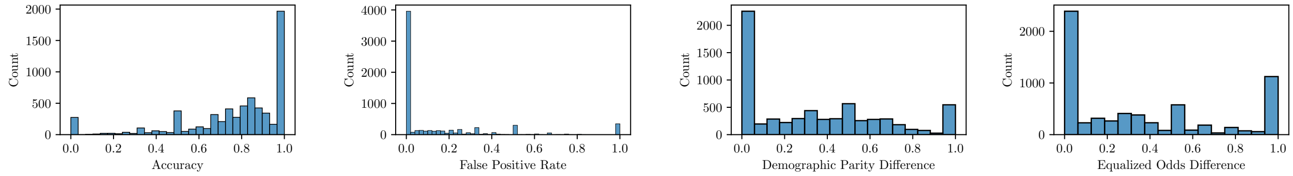
Distribution of Worker Accuracy Metrics. Figures 1a and 1b display the histogram of worker accuracy and false positive rates for the *Crowd Judgement* dataset. We can see that worker accuracy ranges approximately between 0.4 and 0.8. In other words, the workers are neither perfect (with an accuracy close to 1.0) nor incompetent (with an accuracy close to 0.0). Nevertheless, a non-trivial number of workers have a substantial false positive rate, underscoring the challenges truth discovery algorithms face. The workers are very accurate and often suffer from high false positive rates. Hence, using simple aggregation metrics such as majority voting is unlikely to be sufficient.

The corresponding histogram for *Jigsaw Toxicity* datasets is in Figures 2a and 2b. The worker accuracy distribution distinct from that of the *Crowd Judgement* dataset. For this dataset, most workers are highly accurate, with a considerable number having perfect ac-



(a) Accuracy (b) False Positive Rate (c) Demographic Parity (d) Equalized Odds

Figure 1: Histogram of Accuracy and Fairness Metrics of Workers of *Crowd Judgement* Dataset



(a) Accuracy (b) False Positive Rate (c) Demographic Parity (d) Equalized Odds

Figure 2: Histogram of Accuracy and Fairness Metrics of Workers of *Jigsaw Toxicity* Dataset

curacy, and few exhibit a high false positive rate. Even though the two datasets have very different workers in terms of accuracy, we shall show that both contain a non-trivial amount of unfair workers.

Distribution of Worker Fairness Metrics. Figures 1c and 1d shows the histogram of the demographic parity and equalized odds difference metrics of the workers for the *Crowd Judgement* dataset, respectively. Recall that a value closer to 0 indicates a fairer worker. The histograms of both these metrics show that workers exhibit a wide divergence in fairness. While there are some fair workers, the vast majority are unfair to some degree. The histograms for the *Jigsaw Toxicity* dataset, presented in Figures 2c and 2d, show that the workers are generally fairer than those of the *Crowd Judgement* dataset. This is unsurprising due to the golden tasks and clearer rubrics. Nevertheless, contain a significant number of unfair workers. We can also conclude that simply adding more workers will not solve this problem as a significant portion of these workers will be unfair too. Hence, it is imperative to design a truth discovery algorithm that cognizant of the large proportion of unfair workers for better aggregating worker labels.

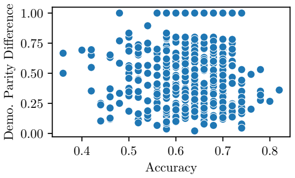
Worker Accuracy-Fairness Trade-off. A natural question that might arise is about the correlation between inaccurate and unfair workers. In other words, are inaccurate workers also unfair? If there is a high correlation between the metrics, then developing a separate class of fairness-aware truth discovery algorithms is unnecessary. We can study the relationship between these metrics by plotting an accuracy metric against a fairness metric. Figure 3a show the scatter plot of worker accuracy on the X-axis and worker demographic parity difference on the Y-axis. Figure 3b shows a similar plot for accuracy and equalized odds difference metrics. We can see values spread over the entire plot without tight clusters showing relationships between accuracy and fairness. In other words, a worker can be both accurate and unfair. The corresponding plots for the *Jigsaw toxicity* dataset in Figures 3a and 3b shows that the number of unfair workers increases as the accuracy increases. Thus, workers can be both accurate and unfair, and it is insufficient to directly use traditional truth discovery algorithms that are geared towards accuracy in the presence of unfair workers.

Table 1 shows a granular version of Figures 3 and 4 by first bucketing workers based on their accuracy and then computing the fairness metrics for all workers in this bucket. Once again, we see no strong correlation between the two metrics. For the *Crowd Judgement* dataset, there is a weak correlation where the more accurate the worker, the lower their unfairness (i.e., the metric trends towards 0). However, this effect is absent for the toxicity dataset.

5 H2: Impact of Unfair Workers on Label Flipping in Truth Discovery Algorithms

In the next set of experiments, we investigate how the presence of unfair workers impacts the labels produced by the MV algorithm, widely used for aggregating crowd worker labels in most ML datasets. We approach this investigation from two different perspectives.

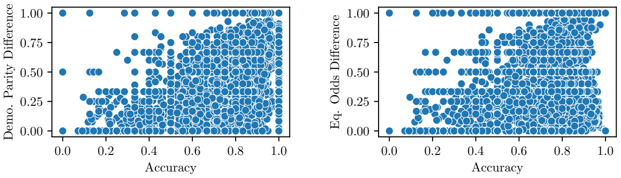
Proportion of Unfair Workers for Tasks. Intu-



(a) Accuracy vs Demographic Parity Difference

(b) Accuracy vs Equalized Odds Difference

Figure 3: *Crowd Judgement* dataset



(a) Accuracy vs Demographic Parity Difference

(b) Accuracy vs Equalized Odds Difference

Figure 4: *Twitter Toxicity* dataset

	Crowd Judgement		**Twitter Toxicity**	
Accuracy Range	**Demo. Par. Diff**	**Eq. Odds. Diff**	**Demo. Par. Diff**	**Eq. Odds. Diff**
(0.0, 0.1]	NA	NA	0.05	0.05
(0.1, 0.2]	NA	NA	0.12	0.15
(0.2, 0.3]	NA	NA	0.11	0.22
(0.3, 0.4]	0.62	0.72	0.24	0.35
(0.4, 0.5]	0.49	0.67	0.26	0.42
(0.5, 0.6]	0.45	0.69	0.3	0.5
(0.6, 0.7]	0.46	0.67	0.34	0.53
(0.7, 0.8]	0.4	0.55	0.37	0.5
(0.8, 0.9]	0.36	0.38	0.45	0.43
(0.9, 1.0]	NA	NA	0.32	0.22

Table 1: Distribution of Accuracy and Fairness metrics for both datasets.

itively, MV chooses the response provided by the majority of the workers as the consensus. Hence, if the majority of the workers assigned to a task turns out to be unfair, then the output of the task will turn out to be unfair. Our first set of experiments investigates how the proportion of workers assigned to a task changes based on the definition of fairness. We use the demographic parity and equalized odds difference metrics to determine the fairness of a worker, where a value of 0.0 indicates a lack of bias. To designate a worker as fair or unfair, we apply a threshold to these fairness metrics. Clearly, a worker with a score of 0.0 is fair while someone with a score of 1.0 is unfair. We systematically vary the minimum threshold for fairness (from $0, 0.1, \ldots 1.0$). We then count the proportion of tasks where unfair workers outnumber fair workers thereby poisoning the output of majority voting. Figure 5 plots how the proportion varies for both datasets using two fairness metrics. We can see that even with a relaxed threshold of 0.2, a large proportion of tasks are dominated by unfair workers. This observation once again shows that it is important to design fairness-aware TD algorithms.

Impact of Removing Unfair Workers. A natural rejoinder might be to eliminate unfair workers and run TD algorithms on the remainder. In our next set of experiments, we show the futility of this approach. Recall from Figures 3 and 4 that it is possible for a worker to be broadly accurate while being unfair. Figure 6 shows the impact of removing unfair workers. When we remove workers designated as unfair based on the fairness metrics used in the previous experiment, it can lead to changes in two key parameters. Firstly, removing unfair workers could potentially affect the accuracy of the tasks. This is measured using the Acc_DPD and Acc_EQD metrics for demographic parity and equalized odds, respectively. Secondly, the number of tasks with responses could also change. Recall from Figure 5 that unfair workers dominate in many tasks. So it is plausible that *all* workers are unfair for some tasks, leading to the abandonment of the task. The metrics NumQs_DPD and NumQs_EQD measure this phenomenon. From Figure 6, we can see that the overall accuracy drops when we remove unfair workers. This is due to the fact that a good chunk of them are accurate and removing them results in the collateral damage of reduced accuracy. Moreover, as we drop unfair workers, the number of abandoned tasks increases. This fact suggests that removing unfair workers is not practical for such tasks. From both these observations, we can conclude that simply dropping unfair workers is a sub-optimal solution as it results in lower accuracy. To overcome this challenge, it is crucial to design fairness-aware TD algorithms that can counteract unfair workers without compromising accuracy.

A key research challenge is that truth discovery algorithms are *unsupervised* in nature. Note that if the ground truth for the crowdsourcing tasks were known, workers would not have been needed. These algorithms often take an EM style iterative approach. They begin with an estimate (such as random initialization) of the relevant metric such as worker accuracy and improve the estimate in each iteration. While the approach for improving estimates of worker accuracy is well known, designing the corresponding algorithms for fairness is an open research problem.

6 H3: Unfair Workers Reduce the Performance of TD Algorithms.

In this section, we study how different TD algorithms perform in the presence of unfair workers. Recall that these algorithms are often designed for workers who are inaccurate. We conduct our experiment as follows. We run three representative TD algorithms over the worker responses to obtain an approximate estimate of the ground truth labels. Since both *Crowd Judgment* and *Jigsaw Toxicity* provide us with actual ground truth, we can compare this with the output of the TD algorithms. This also allows us to compute the accuracy and fairness metrics for each of these TD algorithms. Tables 2 and 3 show the result of this experiment.

At a high level, we can see that the accuracy of the three algorithms is roughly comparable, with the more complex TD algorithms based on generative modeling (DS and LFC) providing slightly more accurate estimates. We observe a similar pattern in the fairness metrics, with none of the TD algorithms able to effectively handle the unfair workers. Using advanced TD algorithms seems to exacerbate this issue, as some underlying assumptions, such as the two-coin annotator process for LFC, are violated by the presence of unfair workers. We can conclude that it is sub-optimal to directly use existing TD algorithms, and instead, it is necessary to design fairness-aware TD algorithms that explicitly model the unfairness of the workers.

7 H4: Unfair Workers Negatively Impact the Downstream ML Tasks.

Machine learning algorithms often rely on large amounts of data to learn patterns and make predictions. However, using data collected from unreliable or biased sources can have severe implications for downstream ML tasks. One potential source of biased data is crowd workers, who may have inherent biases or personal beliefs that impact the quality of the data they provide. In this scenario, it is crucial to understand the impact of unfair crowd workers on downstream machine learning tasks. This is of utmost importance as these biases can perpetuate and amplify existing societal biases, leading to discrimination and unequal treatment of certain groups.

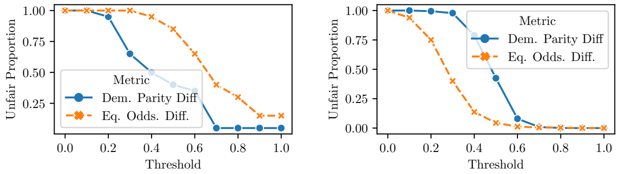
In this section, we study this impact for two representative classifiers – Logistic Regression and Random Forest. Similar results are observed for other classifiers, such as Support Vector Machines or Neural Networks. Our experiment is conducted as follows. We train an ML model $\mathcal{M}_G$, using the ground truth labels. This is the ideal model we would have obtained if the TD algorithms had been perfect. Note that the real ground truth is often unavailable in the real world. Nevertheless, we are able to conduct this experiment due to their availability for our two datasets. Next, we train a ML model $\mathcal{M}_{TD}$ based on the approximate labels that are produced by the TD algorithm. We split the dataset equally into training and test sets and repeated the experiment 10 times to compute the average accuracy and fairness metrics. By comparing the performance of the two models on the same test set, we can quantify the penalty paid by the TD algorithm.

Tables 4 and 5 presents the result for the two datasets, two classifiers, and three TD algorithms. To conserve space, we only report the delta between the metrics. Specifically, if the accuracy of the two models is a_G and a_{TD} respectively, we report $a_G - a_{TD}$. We follow the same procedure for the fairness metrics. As shown in the tables, using TD algorithms results in less accurate and more unfair models. This is not surprising as the TD algorithms are not specifically designed to optimize for fairness.

8 H5: The Complementarity of Fair ML and Fair TD.

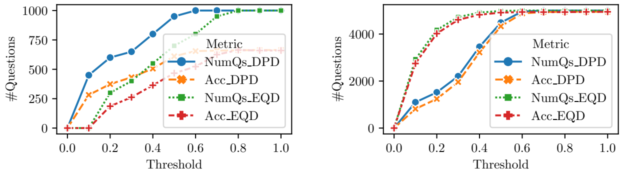
Based on the results of the previous experiments, it is clear that major TD algorithms perform sub-optimally in the presence of unfair workers. One might wonder whether the extensive amount of prior work on fair ML can mitigate the impact of unfair labels produced by TD algorithms. In other words, can the fair ML algorithms that seek to produce outputs that are fair to different subgroups overcome incorrect/unfair labels produced by TD algorithms? This section aims to provide a comprehensive comparison of the two possible approaches, fair ML and fair TD, and gain insights into the strengths and limitations of each method in achieving fairness in decision-making systems.

We compare the output of FairTD, the first and only fairness-aware truth discovery algorithm [13], with two fair ML algorithms: Exponentiated Gradient Reduction [3] and Prejudice Remover [12]. We utilize these approaches for both the Majority Voting (MV) and Dawid-Skene (DS) truth discovery strategies. The experiment's results are presented in Figure 7, where the accuracy of the considered approaches is compared with changes in the fairness constraint. The fairness constraint pertains to the allowable violation of fairness for FairTD and Exponentiated Gradient methods, and the reciprocal of the regularization term for Prejudice Remover. While



(a) *Crowd Judgement* (b) Jigsaw Toxicity

Figure 5: Proportion of tasks dominated by unfair workers for different thresholds of fairness



(a) *Crowd Judgement* (b) *Jigsaw Toxicity*

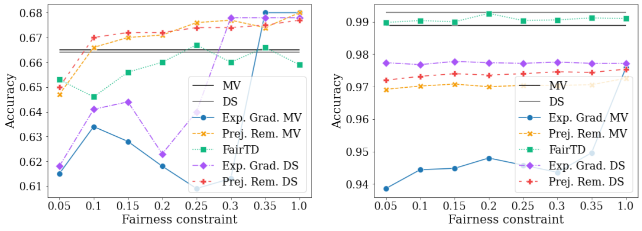
Figure 6: Impact of removing unfair workers on overall accuracy

TD Algorithm	**Accuracy**	**Demographic Parity Difference**	**Demographic Parity Ratio**	**Equalized Odds Difference**	**Equalized Odds Ratio**
Majority Voting	0.658	0.166	0.706	0.12	0.703
Dawid-Skene	0.663	0.183	0.667	0.136	0.665
Learning from Crowd	0.664	0.184	0.668	0.142	0.676

Table 2: Performance of different TD algorithms for *Crowd Judgement* dataset

increasing the fairness constraint violation led to an increase in accuracy in some cases, this relationship is not consistent across all datasets and methods tested. Furthermore, this observation may not hold true when fair workers are consistently more accurate than unfair ones. We can observe that Prejudice Remover generally outperforms Exponentiated Gradient in mitigating bias while maintaining high accuracy. FairTD, on the other hand, appears to be a robust method that provided good results in both cases, often matching or outperforming Prejudice Remover even for strict constraint values.

Based on our results, it appears that utilizing FairTD, either alone or in conjunction with Fair ML, might be a more promising approach for future research compared to relying solely on fair ML. Although Exponentiated Gradient and Prejudice Removal are valuable techniques in their own right, FairTD provides a robust and precise approach to reducing bias. Therefore, we believe that future research should concentrate on enhancing and optimizing the FairTD method to make it even more effective in addressing the challenges of TD algorithmic fairness.



(a) Fair ML vs Fair TD for *Crowd Judgement* (b) Fair ML vs Fair TD for *Jigsaw toxicity*

Figure 7: Fair ML vs Fair TD

9 Conclusion

In this paper, we investigated a key source of bias in ML datasets produced using crowdsourcing. Most of these datasets use simple majority voting to aggregate conflicting responses of crowd workers. However, this approach is sub-optimal and produces poor results in the presence of unfair workers. We conducted extensive experiments to study the impact of unfair workers on various truth discovery algorithms. and conclude that their impact on label accuracy is significant. Existing truth discovery algorithms are focused on accuracy and often cannot handle unfair workers. It is important to design fairness-aware truth discovery algorithms so that one could counteract the influence of unfair workers on the labels for ML datasets.

References

[1] *Jigsaw unintended bias in toxicity classification.* `https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification/data`. Accessed: 2023-03-20.

[2] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, *Youtube-8m: A large-scale video classification benchmark*, arXiv preprint arXiv:1609.08675, (2016).

[3] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, *A reductions approach to fair classification*, in International Conference on Machine Learning, PMLR, 2018, pp. 60–69.

[4] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*, fairmlbook.org, 2019. `http://www.fairmlbook.org`.

TD Algorithm	**Accuracy**	**Demographic Parity Difference**	**Demographic Parity Ratio**	**Equalized Odds Difference**	**Equalized Odds Ratio**
Majority Voting	0.92	0.607	0.08	0.016	0.08
Dawid-Skene	0.93	0.636	0.08	0.051	0.05
Learning from Crowd	0.95	0.64	0.08	0.06	0.06

Table 3: Performance of different TD algorithms for *Jigsaw Toxicity* dataset

Classifier	**TD Algorithm**	**Delta Accuracy**	**Delta Demo Par. Diff**	**Delta Eq. Odds. Diff**
Random Forest	Majority Voting	6.04	0.06	0.07
Random Forest	Dawid-Skene	7.63	0.04	0.06
Random Forest	Learning from Crowd	7.54	0.04	0.05
Logistic Regression	Majority Voting	2.55	0.04	0.07
Logistic Regression	Dawid-Skene	2.77	0.05	0.08
Logistic Regression	Learning from Crowd	2.86	0.05	0.07

Table 4: Performance of ML model trained with the output of fairness unaware TD algorithm for *Crowd Judgement*

Classifier	**TD Algorithm**	**Delta Accuracy**	**Delta Demo Par. Diff**	**Delta Eq. Odds. Diff**
Random Forest	Majority Voting	5.44	0.07	0.09
Random Forest	Dawid-Skene	4.63	0.05	0.07
Random Forest	Learning from Crowd	6.23	0.05	0.06
Logistic Regression	Majority Voting	4.32	0.05	0.01
Logistic Regression	Dawid-Skene	3.99	0.05	0.08
Logistic Regression	Learning from Crowd	3.96	0.04	0.07

Table 5: Performance of ML model trained with the output of fairness unaware TD algorithm for *Jigsaw Toxicity*

[5] S. Bell, K. Bala, and N. Snavely, *Intrinsic images in the wild*, ACM Transactions on Graphics (TOG), 33 (2014), pp. 1–12.

[6] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, *Activitynet: A large-scale video benchmark for human activity understanding*, in CVPR, 2015, pp. 961–970.

[7] A. P. Dawid and A. M. Skene, *Maximum likelihood estimation of observer error-rates using the em algorithm*, Journal of the Royal Statistical Society, 28 (1979), pp. 20–28.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, *Imagenet: A large-scale hierarchical image database*, in CVPR, Ieee, 2009, pp. 248–255.

[9] J. Dressel and H. Farid, *The accuracy, fairness, and limits of predicting recidivism*, Science advances, 4 (2018), p. eaao5580.

[10] C. Eickhoff, *Cognitive biases in crowdsourcing*, in WSDM, 2018, pp. 162–170.

[11] M. Haklay and P. Weber, *Openstreetmap: User-generated street maps*, IEEE Pervasive computing, 7 (2008), pp. 12–18.

[12] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, *Fairness-aware classifier with prejudice remover regularizer*, in ECML PKDD, Springer, 2012, pp. 35–50.

[13] Y. Li, H. Sun, and W. H. Wang, *Towards fair truth discovery from biased crowdsourced answers*, in KDD, 2020, pp. 599–607.

[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, *Microsoft coco: Common objects in context*, in ECCV, Springer, 2014, pp. 740–755.

[15] P. K. Lohia, K. N. Ramamurthy, M. Bhide, D. Saha, K. R. Varshney, and R. Puri, *Bias mitigation post-processing for individual and group fairness*, in ICASSP, IEEE, 2019, pp. 2847–2851.

[16] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, *A survey on bias and fairness in machine learning*, ACM Computing Surveys (CSUR), 54 (2021), pp. 1–35.

[17] C. G. Northcutt, A. Athalye, and J. Mueller, *Pervasive label errors in test sets destabilize machine learning benchmarks*, arXiv preprint arXiv:2103.14749, (2021).

[18] K. Peng, A. Mathur, and A. Narayanan, *Mitigating dataset harms requires stewardship: Lessons from 1000 papers*, arXiv preprint arXiv:2108.02922, (2021).

[19] V. U. Prabhu and A. Birhane, *Large image datasets: A pyrrhic win for computer vision?*, arXiv preprint arXiv:2006.16923, (2020).

[20] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, *Learning from crowds.*, JMLR, 11 (2010).

[21] B. Shmueli, J. Fell, S. Ray, and L.-W. Ku, *Beyond fair pay: Ethical implications of nlp crowdsourcing*, arXiv preprint arXiv:2104.10097, (2021).

[22] B. H. Zhang, B. Lemoine, and M. Mitchell, *Mitigating unwanted biases with adversarial learning*, in Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 2018, pp. 335–340.

[23] Y. Zheng, G. Li, Y. Li, C. Shan, and R. Cheng, *Truth inference in crowdsourcing: Is the problem solved?*, PVLDB Endowment, 10 (2017), pp. 541–552.

[24] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, *Places: A 10 million image database for scene recognition*, IEEE Transactions on PAMI, 40 (2017), pp. 1452–1464.