*Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 26–31
Marseille, 20 June 2022


# Bias Discovery Within Human Raters: A Case Study of the Jigsaw Dataset

**Marta Marchiori Manerba, Riccardo Guidotti, Lucia Passaro, Salvatore Ruggieri**
University of Pisa
{marta.marchiori@phd.,riccardo.guidotti@,lucia.passaro@,salvatore.ruggieri@}unipi.it

**Abstract**

Understanding and quantifying the bias introduced by human annotation of data is a crucial problem for trustworthy supervised learning. Recently, a perspectivist trend has emerged in the NLP community, focusing on the inadequacy of previous aggregation schemes, which suppose the existence of a single ground truth. This assumption is particularly problematic for sensitive tasks involving subjective human judgments, such as toxicity detection. To address these issues, we propose a preliminary approach for bias discovery within human raters by exploring individual ratings for specific sensitive topics annotated in the texts. Our analysis's object focuses on the Jigsaw dataset, a collection of comments aiming at challenging online toxicity identification.

**Keywords:** NLP Perspectivism, Human Raters, Individual Annotations, Fairness, Bias, Toxicity Detection

## 1. Introduction

At every stage of a supervised learning process, biases can arise and be introduced in the pipeline, ultimately leading to possible harm (Suresh and Guttag, 2019; Dixon et al., 2018). The role of the datasets used to train these supervised models is crucial, as they may reinforce such biases and propagate them. There might be multiple reasons why a dataset is biased, e.g., due to skewed sampling strategies or to the prevalence of a particular demographic group disproportionately associated with a class outcome (Ntoutsi et al., 2020), ultimately establishing conditions of privilege and discrimination. (Sap et al., 2019; Davidson et al., 2019; Ball-Burack et al., 2021), for example, show that annotators tend to label as toxic messages in Afro-American English more frequently than when annotating other messages, which could lead to the training of a system reproducing the same kind of racial dialect bias. The phenomenon's complexity is not limited to algorithms but is deeply rooted and bound in historical, cultural, and social perceptions. Therefore, it is very relevant to investigate the impact of annotators' social and cultural backgrounds on the produced labelled data. It is clear that when the labelling is performed on subjective tasks, such as the online toxicity detection, it becomes even more relevant to explore agreement reports and preserve individual and divergent opinions. Having access to the disaggregated data annotations and being aware of the dataset's intended use can inform both models' outcome assessment and comprehension, including facilitating bias detection (Suresh and Guttag, 2019).

Given these evident socio-technical challenges, significant trust problems emerge, mainly regarding the robustness and quality of datasets and the related trustworthiness of models trained on these collections and their automated decisions. Recently, a perspectivist trend has emerged in the NLP community, focusing on datasets collecting human judgments, especially for sensitive tasks involving subjective decisions such as toxicity detection. The main issue concerns the inadequacy of previous aggregation schemes, which assume the existence of a single ground truth and reduce the final label through the standard approaches of disagreement resolution, primarily through majority voting. (Basile, 2020) propose a new paradigm to maintain multiple perspectives naturally arising from raters having different cultural backgrounds. The authors pursue the goal of granting significance to divergent opinions, equally important and correct, according to individual sensitivities. They stress the importance of publishing disaggregated dataset versions and the positive impact of these collections for developing more inclusive yet accurate, fairness-aware measures and automated decisions. (Röttger et al., 2021) critically discuss two annotation approaches: the descriptive-perspectivist paradigm versus the prescriptive-reductionist one. Among other recommendations, the authors suggest that dataset collectors should intentionally choose and pursue one of the two paradigms according to the intended usage for that particular collection.

In line with the perspectivist approach, this work aims to value disagreement and investigate a different way to weight annotations. Specifically, we propose a preliminary approach for bias discovery within human raters[1] by exploring individual ratings for specific sensitive topics annotated in the texts. Although investigating biases within human annotators was already explored in (Sap et al., 2019; Davidson et al., 2019; Ball-Burack

[1] In this contribution, we use the terms *rater* and *annotator* interchangeably.

et al., 2021; Sap et al., 2021), our proposed method, compared to these previous works, is not limited to a single bias ground (e.g., race or gender). Indeed, thanks to the nature of the dataset under examination, the sensitive identities taken into account are more diverse, embracing, for example, sexual orientations, disabilities, religions, etc. Finally, to preserve the role of different perspectives, as performed in the work of (Wich et al., 2020), our assessment focuses on the disaggregated dataset, hence on individual annotations, and not only on the harmonized ground truth. Our analysis focuses on the Jigsaw dataset (Jigsaw, 2018), a collection of comments aiming at challenging online toxicity identification. The dataset is manually annotated to investigate unintended model bias for a broad spectrum of sensitive demographic identities.

Starting from the description of the Jigsaw dataset in Section 2, we report in Section 3 the fairness approach adopted and the preliminary analysis results in Section 4. Finally, in Section 5, we present the takeaways.

## 2. Dataset Description

This section briefly describes the object of our analysis, i.e., the *Jigsaw Unintended Bias in Toxicity Classification*[2] dataset, published within a Kaggle competition[3] which took place in 2018 (Jigsaw, 2018). Aiming to explore unintended model bias through a broad spectrum of online dialogues, the dataset collects contents from the *Civil Comments* platform that allowed to start conversations and post comments on news sites. Curated by Jigsaw[4], a Google unit dealing with disinformation, toxicity, censorship, and extremisms, the collection gathers posts ranging from 2015 to 2017 annotated by a degree of toxicity by human raters through the crowd rating platform *Figure Eight*.[5] The structure of the dataset, including its cascade annotation, allows for sheding some light on the impact of the sociocultural characteristics of the raters, especially when dealing with sensitive tasks involving subjective decisions such as toxicity detection.

**Annotation Process and Labeling Schema.** Comments in the dataset were annotated to identify toxicity. Specifically, by toxicity, the curators mean extremely rude, offensive, humiliating, or/and harmful content. The dataset presents several levels of annotation, which we will describe in detail in the next paragraphs. The toxicity is registered across a range of other labels: VERY TOXIC, TOXIC, HARD TO SAY and NOT TOXIC. A comment is considered toxic if the toxicity value assigned by the aggregations of individual raters annotations is greater than or equal to $0.5$. Toxic comments were further labelled with the type of abusiveness: TOXICITY, SEVERE TOXICITY, OBSCENE, THREAT, INSULT, IDENTITY ATTACK, SEXUAL EXPLICIT. The dataset was divided by the curators in training $(1,804,874)$, public $(97,320)$ and private test $(97,320)$ sets, for a total of $1,999,514$ instances. To enclose several different perspectives, every comment was annotated by up to 10 raters[6] since the dataset creators acknowledged the subjectivity of the task. Interestingly, some comments were annotated by more than 10 raters, up to even thousands.[7] For a subset of the dataset, annotators were also asked to indicate whether the text mentioned demographic identities, such as specific races or genders. To ensure that the comments in the subset had identity mentions, data were filtered as follows. The curators started with a random sample of around $250,000$ comments. Then, through model predictions and word matching, they found approximately other $250,000$ instances, which most likely contained references to the sensitive identities within the texts. The collection resulting from the union of the two subsets, was then manually labeled by the raters. Identities appearing in more than $500$ comments were found relevant, including: *male, female, homosexual (gay or lesbian), christian, jewish, muslim, black, white, psychiatric or mental illness*. Others were detected but occurred less frequently. In addition to the aggregated dataset, the curators also published two additional sheets useful to investigate raters' behaviour (Table 1). The first sheet reports the individual raters annotations of the sensitive identities for a total of $2,597,365$ annotations, collected for $405,159$ unique comments labeled by $4,592$ different raters. The second sheet collects the judgments related to the toxicity degrees, amounting to $15,855,266$ individual annotations for $1,804,874$ unique comments, i.e., the aggregated training set size, labeled by $8,899$ different raters. Both sheets thus contain comments repeated as many times as different annotators were asked to label them (this is why these tables are larger than the dataset for model training).

| | **Identities** | **Toxicity** |
|---|---|---|
| Comments | 405,159 | 1,804,874 |
| Raters | 4,592 | 8,899 |

Table 1: Comments and raters for the individual annotations regarding *(i)* sensitive identities and *(ii)* degrees of toxicity, respectively.

**Disaggregated Data.** We explore the impact of raters' bias by analysing the dataset of individual judgements related to toxicity. As reported above, the dataset consists of $15,855,266$ individual annotations, often reporting the same, repeated comments labeled by different raters. Specifically, it has $1,804,874$ unique comments, i.e., the aggregated training set size,

[2]Jigsaw Unintended Bias in Toxicity Classification.

[3]Competition overview.

[4]Jigsaw.

[5]The platform was acquired by Appen.

[6]The attributes gathering this information are "toxicity annotator count" and "identity annotator count".

[7]They motivate this choice with very vague reasons: *"due to sampling and strategies used to enforce rater accuracy"*.

labeled by $8,899$ different raters. We add the "target annotator" column. This new attribute has a binary label indicating whether the annotator considered the comment toxic. This information is derived from the individual annotations that collect toxicity grades: specifically, if at least one of the judgments is affirmative, the comment is considered toxic by some annotator. Notably, the value of this attribute may differ from the assigned label in the released training dataset. We also retrieve from the training dataset other important and informative columns, including "target" and all columns reporting the sensitive identities. Henceforth, when we reference the dataset, we address this disaggregated version.

## 3. Raters Bias Discovery

This section describes the metrics we adopt to detect biases of human raters and the fairness assessment approach we propose.

### 3.1. Fairness Metrics

To detect potential biases in individual raters, we choose to adopt the *Bias AUCs* evaluation metrics proposed by (Borkan et al., 2019).[8] They are defined as the ROC-AUC computed over specific subsets of the data. The use of these metrics within the competition and the work proposing them (Borkan et al., 2019) focuses on assessing unintended biases of models on the test set. Since our aim is to determine if they can also capture bias in humans, we want to propose their application in a new context, different from the purposes for which they were originally developed. Exploring the comments for which multiple annotations are available in the training dataset, we intend to use the label of the aggregated dataset version as ground truth and the judgment of the individual rater as prediction, thus discovering and evaluating biases within human raters annotations. To compute the ROC-AUC, we sorted the data according to a comment toxicity score, ranging from $0$ to $1$. Such as score is derived from the individual rater's annotations and computed as the number of toxicities identified (i.e. labelled with $1$ by the rater), divided by the number of toxicity types (i.e. $7$). According to (Borkan et al., 2019), we formalize the following metrics:

**Definition 1 (Bias AUCs)** *We define the Bias AUCs measures as:*

$$\begin{aligned} \textit{Sub}_s &= \mathrm{AUC}\left(D_s^- + D_s^+\right) \\ \textit{BPSN}_s &= \mathrm{AUC}\left(D^+ + D_s^-\right) \\ \textit{BNSP}_s &= \mathrm{AUC}\left(D^- + D_s^+\right) \end{aligned}$$

*where $s$ is a subgroup, $D^+$ are the toxic comments, $D^-$ the non-toxic comments, $D_s^+$ the toxic comments in the identity subgroup, and $D_s^-$ the non-toxic comments in the identity subgroup.*

[8]The Kaggle competition proposed the same metrics.

We specify that in the formulas the $+$ symbol operates a concatenation between different subsets of the dataset. The three metrics are calculated separately on these subsets for each sensitive identity. More in detail, in our setting, *Subgroup AUC* ($\mathrm{Sub}_s$) is calculated for toxic and non toxic comments that contain the sensitive identity $s$. A low score indicates that the annotator deviates from the ground truth of the dataset by differently identifying toxic and non-toxic comments containing that identity. *BPSN (Background Positive, Subgroup Negative) AUC* ($\mathrm{BPSN}_s$) instead is computed for non-toxic comments that contain the sensitive identity $s$ and toxic comments that do not contain it. A low score means that the annotator exchanges non-toxic comments containing the identity for toxic ones that do not (consistently with the ground truth of the dataset). Finally, *BNSP (Background Negative, Subgroup Positive) AUC* ($\mathrm{BNSP}_s$) is calculated for toxic comments that contain the sensitive identity $s$ and non-toxic comments that do not contain it. Obtaining a low score means that the annotator exchanges toxic comments mentioning the identity for non-toxic ones that do not (always according to the ground truth of the dataset).
Since our goal is to analyze the annotators w.r.t. the metrics above, we decided to average them by defining the Average Bias AUC that aggregates the individual *Bias AUCs* scores.

**Definition 2 (Average Bias AUC)** *We define the Average Bias AUC as*

$$\textit{Avg Bias AUC}_s = \frac{\textit{Sub}_s + \textit{BPSN}_s + \textit{BNSP}_s}{3}$$

The intuition is that, given a certain sensitive identity $s$, we will have a high Avg Bias AUC if the rater is not biased w.r.t. a certain background or subgroup; we will have a low value on the other hand.

### 3.2. Methodology

This section illustrates the process followed to perform the raters' bias assessment. We applied this methodology for the Jigsaw dataset but it can be easily replicated on other datasets having disaggregated annotations.
To assess biases, we followed the definitions reported in Section 3.1, computing the three metrics, i.e., *Subgroup AUC, BPSN* and *BNSP*. We recall that each measure is computed on different data subsets and for each identity subgroup present in the comments annotated by each rater. Regarding the identities detected in the comments, we adopt the ground truth of the aggregated training dataset because the focus of this analysis is on variation in toxicity judgment. Further investigations on disagreement concerning individual identity annotations will be conducted as future work. As ground truth for the toxicity, we binarize the target score from the aggregated training set. Concerning the predictions, we deploy the individual toxicity judgment of each annotator, as explained in the previous Section 2. After

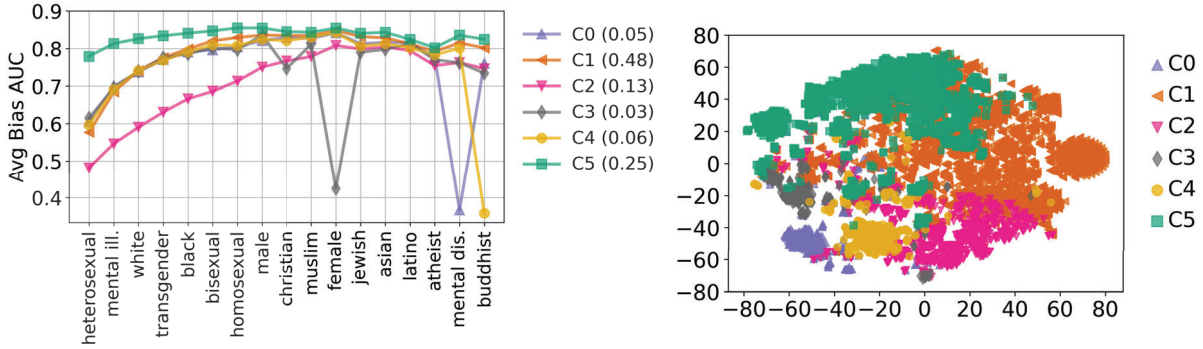


Figure 1: Left: Avg Bias AUC for each cluster centroid (between brackets, the population percentage). Right: t-SNE visualization of clusters in two dimensions.

that, we aggregate the three metrics according to Definition 2, resulting in a score for each rater for each sensitive identity.

To identify recurrent and recognizable groups of raters achieving similar identity scores, we then apply the KMeans clustering algorithm (Macqueen, 1967). We choose the $k$ value for the number of clusters by evaluating the SSE score curve observed varying $k$. We adopt KMeans since we conduct a preliminary analysis, but more advanced clustering techniques could be used as alternatives. Finally, to visualize the clusters, we adopt the t-distributed Stochastic Neighbor Embedding (t-SNE) visualization (Van der Maaten and Hinton, 2008).

## 4. Preliminary Results and Discussion

This section reports the results of the preliminary analysis conducted.[9] As a first step, we focus on evaluating only raters who annotated at least 10 comments, aiming at finding as many identities in the texts as feasible. Thus, starting from the dataset containing $15,855,266$ annotations generated by $8,899$ raters, we filter for $15,847,581$ annotations for a total of $8,034$ raters.

Following the stages defined in the previous section, we calculate the metrics for all the 24 identities available. We then remove the values *other gender, other sexual orientation, other religion, other race or ethnicity, other disability*. Finally, we only keep the identities for which the missing values are lower than 30%, resulting in 17 residual identities. For the remaining identities, we fill the missing values with the average values of each identity.[10] We then apply the KMeans algorithm (Macqueen, 1967) on the data frame resulting from the process, i.e., having as columns the sensitive identities and as rows the annotators. The value of each cell is derived from the aggregated metrics as given in Definition 2. We identify 6 clusters, i.e., 6 different trends in annotators' rating behaviour. We test $k$ in a range from 2 to 100, finding that for $k = 6$ the SSE does not decrease significantly. We report in Figure 1 the Avg Bias AUC for each of the cluster centroids, along side the percentage of the population size of each group. A cluster centroid is the most representative point of a group. Technically, it is calculated by averaging the identities of the *Avg Bias AUC* scores within that cluster. If an identity obtains a low value for this aggregated metric the cluster of annotators demonstrate a biased behaviour.

Generally, we recognize the utility of clustering annotators and display the metric calculated for subgroups. In fact, this setting contributes to the identification of critical disparities in accuracy that may be symptomatic of bias, demonstrated by a propensity to assign toxicity judgments in conjunction with particular identities.Noting the percentage of clusters population as a first aspect, we observe that the most populated are in order `clusters 1, 5,` and `2` (respectively of $0.48, 0.25$ and $0.13$ percent). The remaining clusters have a population between $0.06$ and $0.03$ percent. `Cluster 5` proves to be the best for all identities w.r.t. the *Avg Bias AUC* scores obtained. For *heterosexual*, there is a disparity for the Avg Bias AUC metric between $0.2$ and $0.3$ points compared to the other clusters. Almost all clusters show an increasing trend for the central identities in the chart. It is interesting to focus on `clusters 0, 3` and `4`, whose performance is good, tending to approach `cluster 1`, the best after `5`. Differently from the other clusters, `0, 3` and `4` register a significant drop for the *mental disability, female* and *buddhist* identities, respectively. These results highlight that groups of annotators register a divergent rating behaviour for specific identities, demonstrating a different sensibility w.r.t. the ground truth. The line that shows poor agreement for all identities, i.e., that deviates towards low levels on average, is with reference to `cluster 2`. This cluster represents the $0.13$ percent of the annotators, i.e., $1,044$ on a total of $8,034$.

In Figure 1, we report the t-SNE visualization (Van der

[9] https://github.com/MartaMarchiori/Bias-Discovery-In-Human-Raters.

[10] We also tried replacing the missing values with the maximum and minimum values, and the results did not change. Thus, the replacement of the missing values is not affected by choice of this aggregation function.

Maaten and Hinton, 2008) in two dimensions. The plot highlights similar aspects of the previous one. Clusters achieving low scores for some identities are located in the lower-left area of the visualisation and have a rounded compact shape. In fact, `clusters 0, 3` and `4`, that differ a lot for some identities, create aggregations that depart from the central mass. `Cluster 5`, characterised by the highest scores, is located at the top and has an elongated shape, which implies a larger variability within it. The most populated cluster, i.e., `cluster 1`, is relatively scattered.

More individual annotations and comments dealing with sensitive topics would be needed for each rater to allow for a more appropriate assessment. However, we acknowledge the difficulty in real datasets to collect and organize this kind of data balancing minorities' frequency. More precisely, the distributions reflect online discourse, both in terms of identity presence and their unequal division within abusive versus non-abusive samples.

## 5. Conclusion and Future Work

In this paper we have proposed a preliminary approach for bias discovery of human raters by exploring individual ratings for specific sensitive topics annotated in the texts. Our analysis's object consisted of the Jigsaw dataset, a collection of comments aiming at challenging online toxicity identification. We measured the biases of raters through the *Bias AUCs* metrics. By dividing the annotators' behaviour into clusters, we assessed disparate treatments that occurred for particular sensitive identities, such as specific religions or disabilities. Therefore, the main inference drawn concerns the different levels of agreements registered by the clusters of annotators w.r.t. the ground truth evaluated separately for each diverse sensitive identity. Most trends show close alignment and consistency, except for isolated entities by a few clusters.

A first validation of our method would be to compare the resulting annotators groups identified through our clustering approach with other unsupervised strategies for annotator community grouping and analysis, such as the one presented by (Wich et al., 2020). An interesting experimental extension would consists of applying the proposed methodology to other datasets concerning toxicity detection.[11] It would require explicit sensitive identities mentioned in the texts and the disaggregated versions of individual annotations. The variety of sensitive identities do not constitute a limitation. In fact, the analysis could retrieve meaningful insights even by comparing a few (one or more) identities, e.g., comments grouped for the target of the text, addressing for example females or males. Instead, different thresholds regarding the number of comments needed for each sensitive identity and the least amount of annotations for each rater should be tested and evaluated on a case-by-case basis, i.e., depending on the size of datasets.

In addition, adopting the perspectivist's view would certainly be a good practice to ask data collectors and organizers for disaggregated versions of other similar sensitive tasks, encouraging a more responsible documentation process. One dimension to be explored further is to analyze the content of comments for which the datasets have multiple conflicting annotations. It would be helpful to detect a potential correlation between a given topic and a strong rater's disagreement to qualify the content of the comments that triggered the most controversy among annotators. Furthermore, adopting metrics to identify biases that don't need ground truth could release the analysis from the assumption of the robustness of a gold standard. Finally, having obtained a measure of bias for each rater, a critical experiment would be to construct an alternative version of the dataset that aggregates the annotations differently. Specifically, the annotation of a rater with a high bias score would have less weight for that specific sensitive identity than the judgment of a rater with a lower bias. A comparison between a classifier trained on the original and the weighted data could be an indicative test, focusing the analysis on the unintended bias of the models according to the metrics introduced by (Borkan et al., 2019).

**Acknowledgements.** This research was partially supported by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215. The contents reflects only the author's view and the Commission is not responsible for any use that may be made of the information it contains.

## 6. Bibliographical References

Ball-Burack, A., Lee, M. S. A., Cobbe, J., and Singh, J. (2021). Differential tweetment: Mitigating racial dialect bias in harmful tweet detection. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 116–128.

Basile, V. (2020). It's the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks. In *2020 AIxIA Discussion Papers Workshop, AIxIA 2020 DP*, volume 2776, pages 31–40. CEUR-WS.

Borkan, D., Dixon, L., Sorensen, J., Thain, N., and Vasserman, L. (2019). Nuanced metrics for measuring unintended bias with real data for text classification. In Sihem Amer-Yahia, et al., editors, *Companion of The 2019 World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 491–500. ACM.

Davidson, T., Bhattacharya, D., and Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. *CoRR*, abs/1905.12516.

[11] Examples could be the dataset proposed by (Sap et al., 2020), called Social Bias Inference Corpus or other collections published within the Perspectivist Data Manifesto.

Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernández, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., Broelemann, K., Kasneci, G., Tiropanis, T., and Staab, S. (2020). Bias in data-driven artificial intelligence systems - an introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3).

Röttger, P., Vidgen, B., Hovy, D., and Pierrehumbert, J. B. (2021). Two contrasting data annotation paradigms for subjective NLP tasks. *CoRR*, abs/2112.07475.

Sap, M., Card, D., Gabriel, S., Choi, Y., and Smith, N. A. (2019). The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678.

Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A., and Choi, Y. (2020). Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490.

Sap, M., Swayamdipta, S., Vianna, L., Zhou, X., Choi, Y., and Smith, N. A. (2021). Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *CoRR*, abs/2111.07997.

Suresh, H. and Guttag, J. V. (2019). A framework for understanding unintended consequences of machine learning. *CoRR*, abs/1901.10002.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Wich, M., Kuwatly, H. A., and Groh, G. (2020). Investigating annotator bias with a graph-based approach. In Seyi Akiwowo, et al., editors, *Proceedings of the Fourth Workshop on Online Abuse and Harms, WOAH 2020, Online, November 20, 2020*, pages 191–199. Association for Computational Linguistics.

## 7. Language Resource References

Jigsaw. (2018). *Jigsaw Unintended Bias in Toxicity Classification*. distributed via Kaggle.