Correcting Annotator Bias in Training Data: Population-Aligned Instance Replication (PAIR)

Anonymous ACL submission

Abstract

Models trained on crowdsourced labels may not reflect broader population views, because those who work as annotators do not represent the population. We propose Population-Aligned Instance Replication (PAIR), a method to address bias caused non non-representative annotator pools. Using a simulation study of offensive language and hate speech, we create two types of annotators with different labeling tendencies and generate datasets with varying proportions of the types. Models trained on unbalanced annotator pools show poor calibration compared to those trained on representative data. By duplicating labels from underrepresented annotator groups to match population proportions, PAIR reduces bias without collecting additional annotations. These results suggest that statistical techniques from survey research can improve model performance. We conclude with practical recommendations for improving the representativity of training data and model performance.

1 Introduction and Inspiration

When a hate speech detection model flags harmless expressions as toxic, or a content moderation system fails to identify genuinely harmful content, the root cause often lies not in the model architecture, but in who labeled the training data. While NLP models aim to serve broad populations, the human judgments used to train these systems typically come from a non-representative pool of annotators – crowdworkers and convenience samples whose demographics, cultural contexts, and worldviews may differ from the communities the models ultimately impact (Sorensen et al., 2024; Fleisig et al., 2024). These non-representative annotator pools can have real consequences, because annotator characteristics like age, education level, and cultural background impact how content is labeled (Sap et al., 2022; Fleisig et al., 2023; Kirk et al., 2024). Models trained on non-representative data can perpetuate the biases and blind spots of their training data (Smart et al., 2024; Berinsky et al., 2012; Ouyang et al., 2022; Mehrabi et al., 2021; Rolf et al., 2021; Favier et al., 2023; Hebert-Johnson et al., 2018; Hüllermeier and Waegeman, 2021).

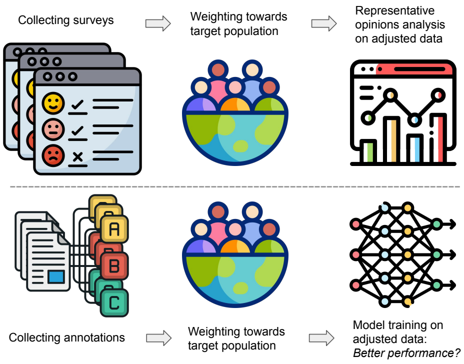

Figure 1: Top: Adjusting survey data to match population produces high quality results. Bottom: Can a similar adjustment in data annotations also improve model performance?

Fortunately, survey researchers have developed robust statistical techniques to produce population-level estimates from non-representative samples (Bethlehem et al., 2011). The top panel of Figure 1 shows a simple survey workflow: collecting survey data, creating statistical weights to match the sample to the population, and estimating population parameters. We propose adapting this approach to the machine learning context, enabling models to better align with target populations even when trained on non-representative annotator pools (bottom panel).

To test this approach, we conduct a rigorous simulation study following established practices in statistical research (Burton et al., 2006; Valliant, 2019; Morris et al., 2019). Using controlled experiments, simulated populations, and multiple training datasets varying in annotator composition, we investigate two questions:

- **RQ1:** How do non-representative annotator

pools impact model calibration and accuracy?
- **RQ2:** Can survey techniques effectively mitigate these annotator pool effects?

Our results demonstrate that models trained on nonrepresentative annotator pools perform poorly. However, simple adjustment methods can improve performance without collecting additional data. These findings suggest that insights from survey methodology can help AI systems better represent the populations they serve.

2 Related Work

Several strands of related work inform our approach to identifying and mitigating bias due to the use of non-representative annotators:

Annotator Impact on Data and Models. Annotator characteristics and attitudes significantly influence label quality, particularly for subjective tasks like toxicity detection (Giorgi et al., 2024; Prabhakaran et al., 2021; Fleisig et al., 2023; Sap et al., 2022). For example, annotators' political views and racial attitudes affect their toxicity judgments (Sap et al., 2022). Models trained on nonrepresentative annotator pools inherit these biases and generalize poorly (Smart et al., 2024; Berinsky et al., 2012; Ouyang et al., 2022; Mehrabi et al., 2021; Rolf et al., 2021; Favier et al., 2023).

Annotator Demographics. Several researchers advocate collecting annotator demographics to assess representation and identify biases (Bender and Friedman, 2018; Prabhakaran et al., 2021).[1] However, collecting and releasing these data can raise privacy concerns (Fleisig et al., 2023).

Debiasing & Data Augmentation Methods. Prior work has proposed various approaches to reduce bias in training data features and labels (Calders et al., 2009; Kamiran and Calders, 2012; Sharma et al., 2020). Most similar to our work is the resampling and reweighting approaches of Calders et al. (2009) and Kamiran and Calders (2012), and the oversampling of minority class cases of Ling and Li (1998). PAIR adapts these methods to balance *annotator* characteristics rather than class labels or sensitive observation-level features. PAIR retains the simplicity and interpretability of earlier resampling methods while extending them to a "Learning with Disagreement" (Uma et al., 2021; Leonardelli et al., 2023) setting with multiple annotations per observation, by replicating labels from underrepresented annotator groups.

3 Annotation Simulation and Model Training

To address our research questions, we imagine a population made up of two types of people: those more likely to perceive offensive language and hate speech and those less likely. We create three sets of simulated annotations which differ in the mix of the annotator types. We then create a fourth data set, using the PAIR algorithm, to attempt to fix the imbalance in the annotators. We fine-tune RoBERTa models on the four data sets and evaluate the effect of annotator characteristics on model performance (RQ1) and the ability of the PAIR algorithm to improve performance (RQ2).

3.1 Simulating Annotations

We use a dataset of 3,000 English-language tweets, each with 15 annotations of both offensive language (OL: yes/no) and hate speech (HS: yes/no) (Kern et al., 2023).[2] We chose this dataset because the high number of annotations of each tweet gives us a diverse set of labels to work with. We randomly select (without replacement) 12 labels (of both OL and HS) of each tweet in the original dataset.[3] Let $p_{i,OL}$ be the proportion of the 12 annotators who labeled tweet i as OL and $p_{i,HS}$ defined similarly. Table 1 shows the distribution of these proportions across the 3,000 tweets. The HS labels are unevenly distributed, whereas the OL labels are relatively balanced. Because this work is a preliminary investigation of the PAIR approach, and balanced labels provide a clearer view of PAIR's abilities, the main body of the paper focuses on results with the OL labels; the HS results in Appendix §C.

Variable	25^{th} pctile.	Median	Mean	75^{th} pctile.
$p_{i,OL}$	0.167	0.667	0.564	0.917
$p_{i,HS}$	0.083	0.167	0.301	0.50

Table 1: Distribution of $p_{i,OL}$ and $p_{i,HS}$ in original data

We imagine a population made up of equal shares of two types of people. **Type A** people

[1]In our context, these characteristics are used only to analyze bias; because they are not available for unlabeled text, they are not features that the model can use.

[2]https://huggingface.co/datasets/soda-lmu/tweet-annotation-sensitivity-2

[3]As shown in Table 2, we can more carefully control the construction of our data sets when the number of labels per tweet is divisible by four.

are *less likely* to say a tweet contains OL. **Type B** people are *more likely*:

$$p^A_{i,OL} = \max(p_{i,OL} - \beta, 0) \quad (1)$$

$$p^B_{i,OL} = \min(p_{i,OL} + \beta, 1) \quad (2)$$

Here β captures the magnitude of the bias. We vary β from $[0.05, 0.3]$ by 0.05, corresponding to an increase or decrease in the probability to judge a tweet as OL by five to 30 percentage points. This range is large on the probability scale and covers most reasonable situations. With these seven values of β, we create seven vectors of probabilities $(p^A_{i,OL}, p^B_{i,OL})$ for each tweet.

We then create four datasets, each with 3,000 tweets (Table 2), for each value of β. The **Representative** Dataset contains OL labels from six A annotators (drawn from Bernoulli($p^A_{i,OL}$)) and six B annotators (drawn from Bernoulli($p^B_{i,OL}$)). Because the proportion of A and B annotators in this dataset matches the population, the labels in this dataset are our gold standard.

Dataset	Labels per tweet	A labels	B labels
Representative	12	6	6
Non-representative 1	9	6	3
Non-representative 2	12	9	3
Adjusted	12	6	3 + 3*

* 3 B labels duplicated

Table 2: Four training datasets for each bias value (β)

We then create two unbalanced datasets. **Non-representative 1** randomly deletes three B labels for each tweet from the Representative Dataset. **Non-representative 2** adds three additional A labels, drawn from p^A_i, to the Non-representative 1 dataset. The Non-representative 2 Dataset is more unbalanced than Non-representative 1, but contains the same number of annotations as the Representative dataset.

Finally, we use the PAIR algorithm to create the **Adjusted** Dataset. It is the same as the Non-representative 1 dataset, but the three B annotations are duplicated. This duplication is an easy way to adjust the unbalanced training dataset to reflect the population. Appendix §A provides a general version of the PAIR algorithm which can handle imbalances across multiple annotator characteristics. Figure 2 shows the percentage of tweets labeled OL in the four datasets for each value of β.

3.2 Model Training and Evaluation

Training and Test Setup. We train models on each dataset. We divide each dataset, at the tweet

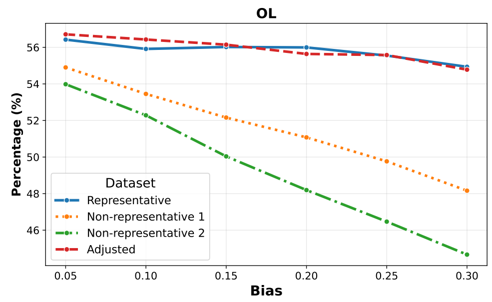

Figure 2: Percentage of tweets annotated as OL, by dataset and bias (β)

level, into training (2000 tweets), development (500), and test (500) sets. Each tweet appears 12 times in the Representative, Non-representative 2, and Adjusted data sets and nine times in the Non-representative 1 set.

Model Selection and Training. We used RoBERTa base (Liu et al., 2019) as our text classifier, training for five epochs on each dataset, with development set optimization. To ensure reliable results, we trained five versions with different random seeds and averaged their performance. Appendix §B contains details on model training.

Performance Metrics. We evaluate models using both **calibration and accuracy metrics** on the test set. While accuracy metrics directly measure classification performance, calibration metrics provide crucial insights into model reliability by assessing probability estimate quality – particularly important for high-stakes applications requiring trustworthy confidence measures.

For calibration, we report Absolute Calibration Bias (ACB), which measures how well model predicted probabilities match true frequencies in the Representative dataset (lower is better). $\text{ACB}_{OL} = \frac{1}{n}\sum_{i=1}^{n} \left|p_{i,OL} - \text{preds}_{i,OL}\right|$. For accuracy, we report the F1 score.

4 Results

Calibration. Figure 3(1) compares the ACB in the test set for models trained on the simulated datasets. The dark lines show average ACB across the five training runs and the shading shows the standard deviation.

The ACB for the models trained on the Adjusted dataset closely tracks that for the Representative (gold standard) data set and does not increase with β. ACB for the models trained on the two unbalanced datasets is greater and grows with β. These results demonstrate the effectiveness of our adjustment method. Duplicating the labels from the underrepresented annotator type to match population proportions improves model calibration.

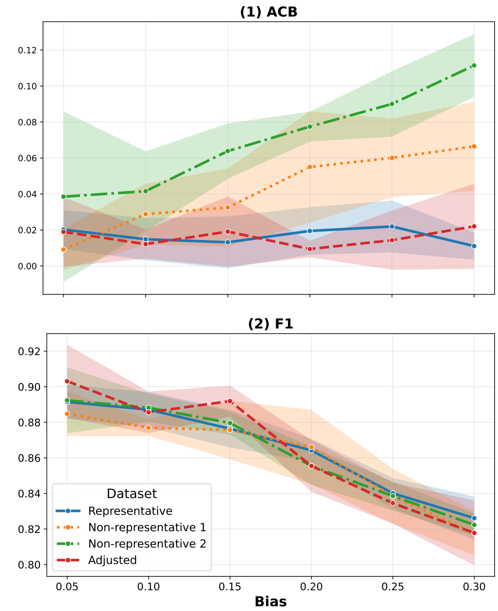

Figure 3: Metrics for OL models, by dataset and bias (β)

Accuracy. Figure 3(2) compares the models' F1 scores. In contrast to Figure 3(1), we do not see strong differences between the models trained on the different datasets. For all datasets, model performance declines with β: as the amount of bias in the labels increases, the models have a harder time predicting the binary OL label.

Because the F1 metric focuses on binary predictions, it is less sensitive to training biases than calibration metrics like ACB, which more explicitly capture biases through prediction scores. These findings suggest that calibration metrics provide a clearer view of the impact of annotators on models: binary classification metrics can obscure such effects. In decision-making, miscalibrated predictions can have harmful consequences when, for example, hateful content remains undetected (Van Calster et al., 2019).

Ablation Study 1: Hate Speech labels. Appendix §C gives the ACB and F1 results for the HS labels. The PAIR approach has limited effectiveness when label distributions are highly skewed. This failure is likely due to the rarity of hate speech labels in the dataset (only 16.7% positive labels) combined with our simulation setup, which further reduces positive labels in the unbalanced datasets. PAIR may need modifications, such as stratified sampling or adaptive replication, when working with highly imbalanced datasets.

Ablation Study 2: Difficult Tweets. A representative pool of annotators may be more important for instances that are difficult or about which people disagree. Filtering our tweets in this way also removes cases where floor and ceiling effects could mask the impact of annotator characteristics. To test PAIR performance on difficult tweets, we reran the models on tweets where $0.4 < p_{i,HS} < 0.6$, $0.4 < p_{i,OL} < 0.6$. Performance on OL is still good, and performance on HS is improved (Appendix §D).

5 Discussion & Recommendations

Our results show that (RQ1) OL prediction models perform less well when trained on data from non-representative annotator pools, and (RQ2) simple statistical adjustments can improve model calibration without collecting additional annotations. These findings establish a promising bridge between survey statistics and machine learning - offering a practical approach to make AI systems more representative of and responsive to the populations they serve, particularly for tasks involving subjective human judgments.

We recommend these four steps to reduce bias due to non-representative annotator pools:

1) **Use social science research** to identify the annotator characteristics that influence the propensity to engage in annotation and the annotations provided (Eckman et al., 2024).
2) **Collect these characteristics** from annotators and gathering corresponding population-level data from national censuses or high-quality surveys.[4]
3) **Calculate weights** that match the annotators to the population on those characteristics (Bethlehem et al., 2011; Valliant et al., 2013).
4) **Use these weights in model training**. Our simple duplication approach showed promise, future work should test more sophisticated weighting approaches.

[4] Collection and release of annotator characteristics or weights derived from them may raise confidentiality concerns. The survey literature contains relevant approaches (see Karr, 2016, for a review). Collecting annotator characteristics may also require involvement of Institutional Review Boards or other participant protection organizations (Kaushik et al., 2024).

Limitations

Stylized Biases and Simulated Data. Our simulation makes strong assumptions about annotator behavior, particularly in modeling consistent biases across annotator types. Real-world annotator biases may be more nuanced or context-dependent. Future work could incorporate more realistic biases and refine the proposed simulations and statistical techniques.

Sampling Variability. We have created only one version the four datasets for each label type and value of β, each of which contains random draws from the Bernoulli distribution. A more traditional statistical approach would create multiple versions of the datasets and train models on each one, to average over the sampling variability. We have not done that in this preliminary study because of the high cost and time needed to fine tune many RoBERTa models. As discussed, we used five seeds in model training.

Need for Population Benchmarks and Annotator Characteristics. PAIR requires high quality benchmark information about the relevant population. These benchmarks might come from national statistical offices or national surveys. Annotators must provide accurate data on the same characteristics available in the benchmark data. In addition, theory from survey science demonstrates that bias will be reduced only when the characteristics used in weighting correlate with the annotations (Eckman et al., 2024). In our simulation, differences in annotations were driven solely by group membership (A, B). In the real world, it is challenging to know what characteristics impact annotation behavior and to find good benchmarks for those characteristics.

Generalization Beyond Task Types. The study focuses only on binary classification tasks. Many real-world annotation tasks involve multiple classes or labels, which may show different bias patterns. Additional research is needed to extend these methods to more complex classification scenarios.

Evaluation Metrics. While we measured calibration and accuracy, we did not examine other important metrics such as fairness across subgroups or robustness to adversarial examples. Future work on training data adjustment should assess a broader range of performance measures.

Ethical Considerations

In this simulation study, we experiment on a publicly available dataset (Kern et al., 2023), which contains offensive and hateful tweets. We do not support the views expressed in the tweets. The simulation study itself does not collect any new data or raise any ethical considerations.

Acknowledgments

Authors acknowledge use of Claude models to edit the text of the paper and to assist with coding.

References

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Adam J. Berinsky, Gregory A. Huber, and Gabriel S. Lenz. 2012. Evaluating online labor markets for experimental research: Amazon.com's mechanical turk. *Political Analysis*, 20(3):351–368.

Jelke Bethlehem, Fannie Cobben, and Barry Schouten. 2011. *Handbook of Nonresponse in Household Surveys*. Wiley.

Andrea Burton, Douglas G. Altman, Patrick Royston, and Roger L. Holder. 2006. The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24):4279–4292.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *ICDMW '09: Proceedings of the 2009 IEEE International Conference on Data Mining Workshops*, pages 13–18.

Stephanie Eckman, Barbara Plank, and Frauke Kreuter. 2024. Position: Insights from survey methodology can improve training data. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 12268–12283. PMLR.

Marco Favier, Toon Calders, Sam Pinxteren, and Jonathan Meyer. 2023. How to be fair? a study of label and selection bias. *Machine Learning*, 112(12):5081–5104.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore. Association for Computational Linguistics.

Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024. The perspectivist paradigm shift: Assumptions and challenges of capturing human labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2279–2292, Mexico City, Mexico. Association for Computational Linguistics.

Tommaso Giorgi, Lorenzo Cima, Tiziano Fagni, Marco Avvenuti, and Stefano Cresci. 2024. Human and LLM biases in hate speech annotations: A socio-demographic analysis of annotators and targets. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing*.

Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. 2018. Multicalibration: Calibration for the (Computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1939–1948. PMLR.

Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506.

Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33.

Alan F. Karr. 2016. Data sharing and access. *Annual Review of Statistics and Its Application*, 3(Volume 3, 2016):113–132.

Divyansh Kaushik, Zachary C. Lipton, and Alex John London. 2024. Resolving the human-subjects status of ml's crowdworkers. *Commun. ACM*, 67(5):52–59.

Christoph Kern, Stephanie Eckman, Jacob Beck, Rob Chew, Bolei Ma, and Frauke Kreuter. 2023. Annotation sensitivity: Training data collection methods affect model performance. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14874–14886, Singapore. Association for Computational Linguistics.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Preprint*, arXiv:2404.16019.

Elisa Leonardelli, Alexandra Uma, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, and Massimo Poesio. 2023. Semeval-2023 task 11: Learning with disagreements (lewidi). *Preprint*, arXiv:2304.14803.

Charles X. Ling and Chenghui Li. 1998. Data mining for direct marketing: problems and solutions. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, KDD'98, page 73–79. AAAI Press.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6):1–36.

Tim P. Morris, Ian R. White, and Michael J. Crowther. 2019. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop,*.

Andreas Quatember. 2015. *Pseudo-Populations: A Basic Concept in Statistical Surveys*. Springer.

Esther Rolf, Theodora T Worledge, Benjamin Recht, and Michael Jordan. 2021. Representation matters: Assessing the importance of subgroup allocations in training data. In *International Conference on Machine Learning*, pages 9040–9051. PMLR.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Shubham Sharma, Yunfeng Zhang, Jesús M Ríos Aliaga, Djallel Bouneffouf, Vinod Muthusamy, and Kush R Varshney. 2020. Data augmentation for discrimination prevention and bias disambiguation. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society*, pages 358–364.

Andrew Smart, Ding Wang, Ellis Monk, Mark Díaz, Atoosa Kasirzadeh, Erin Van Liemt, and Sonja Schmer-Galunder. 2024. Discipline and label: A weird genealogy and social theory of data annotation. *Preprint*, arXiv:2402.06811.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: A roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 46280–46302. PMLR.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Richard Valliant. 2019. Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, 8(2):231–263.

Richard Valliant, Jill A Dever, and Frauke Kreuter. 2013. *Practical tools for designing and weighting survey samples*, volume 1. Springer.

Ben Van Calster, David J. McLernon, Maarten van Smeden, Laure Wynants, Ewout W. Steyerberg, Patrick Bossuyt, Gary S. Collins, Petra Macaskill, David J. McLernon, Karel G. M. Moons, Ewout W. Steyerberg, Ben Van Calster, Maarten van Smeden, and Andrew J. Vickers. 2019. Calibration: the achilles heel of predictive analytics. *BMC Medicine*, 17(1):230.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A PAIR Algorithm Implementation Details

The adjustment we use to make the annotator pool more representative of the target population is a form of **pseudo-population** generation (Quatember, 2015). We create the pseudo-population by first constructing post-stratification weights, performing weight normalization to ensure the sum of the weights equals the size of the target population, and then duplicating each observation proportionally to its weight via deterministic replication.

Post-stratification is a method of statistical adjustment that makes a selected sample more closely resemble a target population (Bethlehem et al., 2011; Valliant et al., 2013). We form strata (groups) of annotators. For example, we might form 6 strata from three categories of region and two categories of age. Post-stratification requires population-level totals or proportions for each stratum and corresponding case-level observations. The weight for each unit i in stratum s is:

$$w_i = \frac{P_s}{S_s} \tag{3}$$

where P_s is the true population proportion (or total) for stratum s and S_s is the sample proportion (or total) for stratum s. In our case, the strata of interest was a single variable (annotator Types A & B). However, post-stratification can involve multiple variables if their joint distribution is known at the population level.

Although the post-stratified weights will preserve the ratios of the strata in the target population, the weighted totals themselves may not match those in the target population. **Weight normalization** can be used to address this by updating the survey weights so that they sum to a desired total. The normalized weight for unit i can be calculated by:

$$w_i^{\text{normalized}} = w_i^{\text{initial}} \cdot \frac{T}{\sum_{i=1}^{n} w_i^{\text{initial}}} \tag{4}$$

where T is the target total. Since we want the Adjusted dataset to match the size of our gold standard Representative dataset, the target total for the simulation was 12 labels per tweet.

Lastly, to construct a pseudo-population from our weighted data, we perform **deterministic replication** by replicating each unit n_a times where:

$$n_i = \text{round}(w_i^{\text{normalized}}) \tag{5}$$

Rounding the normalized weights ensures that n_a are integers and prevents fractional replicates.

In this initial work, we prioritized this approach for its simple interpretation and reproducibility. However, researchers may prefer other approaches, such as replication via resampling, if they are interested in how the adjustment varies across samples, or incorporating weights directly into the loss function if they want to avoid the normalization and rounding steps.

In our simulation, after post-stratification and weight normalization, each Type A label in the Non-representative 1 dataset receives a weight of 1 and each Type B label receives a weight of 2. This resulted in an Adjusted dataset where each of the Type A labels stays the same and each of the Type B labels is duplicated once.

B Model Training Details

Our implementation of RoBERTa models was based on the libraries `pytorch` (Paszke et al., 2019) and `transformers` (Wolf et al., 2020). During training, we used the same hyperparameter settings of the respective models for our five training conditions to keep these variables consistent for comparison purposes. We report the hyperparameter settings of the models in Table 3. To avoid random effects on training, we trained each model variation with five random seeds $\{10, 42, 512, 1010, 3344\}$ and took the average across the models. All experiments were conducted on an NVIDIA® A100 80 GB RAM GPU.

Hyperparameter	Value
encoder	`roberta-base`
epochs_trained	5
learning_rate	$3e^{-5}$
batch_size	32
warmup_steps	500
optimizer	`AdamW`
max_length	128

Table 3: Hyperparameter settings of RoBERTa models

C Ablation Study 1: HS Results

We show the results on HS labels in this section. We construct the HS probabilities for the A and B annotators are defined in the same way as the OL probabilities: $p^A_{i,HS} = \max(p_{i,HS} - \beta, 0)$, $p^B_{i,HS} = \min(p_{i,HS} + \beta, 1)$. We also construct the four data sets (Representative, Non-representative 1, Non-representative 2, Adjusted) in the same way we did for the OL case.

Label Distribution during Simulations. Figure 4 shows the percentage of tweets labeled HS in the four datasets for each value of β. As expected, the percentage in the Adjusted dataset is similar to that in the Representative dataset for all values of β, and the two unbalanced datasets have lower rates of HS, because they overrepresent the A annotators, who are less likely to label HS.

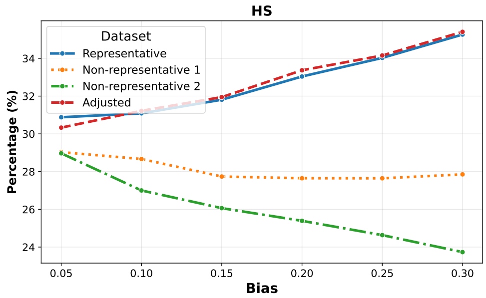

Figure 4: Percentage of tweets annotated as HS, by dataset and bias (β)

HS is rare in our dataset (16.7% of HS labels are yes, Table 1). And, our simulation strategy overrepresents A annotators in the two Non-representative data sets, who are less likely to perceive HS (Table 2). For these reasons, a higher proportion of $p^A_{i,HS}$ are 0 while the $p^B_{i,HS}$ probabilities continue to increase. This issue leads the proportion of “yes” labels in the Representative and Adjusted datasets to increase with β in the HS data set, which have more B labels than the unadjusted datasets.

Results. Figure 5 (1) contains the ACB results and Figure 5 (2) the F1 score results for the HS datasets. The PAIR approach does not improve calibration or accuracy: the adjusted model performs similarly to the Non-representative models. This effect is likely due to the combination of label rarity and our simulation design. With few positive labels to begin with, the impact of varying annotator characteristics through our β parameter may be overwhelmed by the baseline scarcity of hate speech annotations. Ablation Study 2 (Appendix §D) addresses this point.

D Ablation Study 2: Difficult Tweets

Our simulations assumed that all tweets are impacted the same way (Eq. (2)), which is an oversimplification. More realistically, **annotator characteristics likely have more impact for ambiguous tweets**. For this reason, we repeat model training and recompute metrics for those tweets where $0.4 \leq p_i \leq 0.6$. This approach not only focuses on those tweets where annotator characteristics likely play a larger role, it also eliminates the floor and ceiling effects in Eq. 2). The filtered datasets contain 267 (OL) and 360 (HS) tweets.

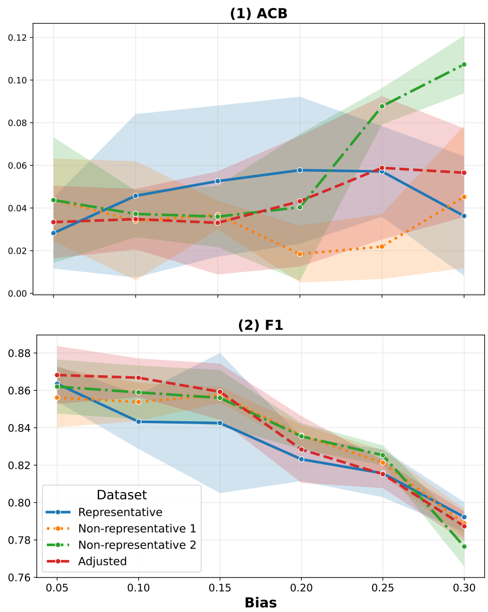

Figure 5: Metrics for HS models, by dataset and bias (β)

Figures 6 and 7 show the results for two metrics (ACB, F1) for filtered OL and HS labels. In Figure 6(1): the Representative and Adjusted models have similar ACB and are lower than the Non-representative models. The F1 scores do not show differences between the models. These results are similar to those on the full set of tweets (Figure 3). In the two HS figures (7), we see signs that the Representative and Adjusted models perform similarly, and better than the two Non-representative models, on both metrics. These results are more promising than those on the full set of tweets (Figure 5).

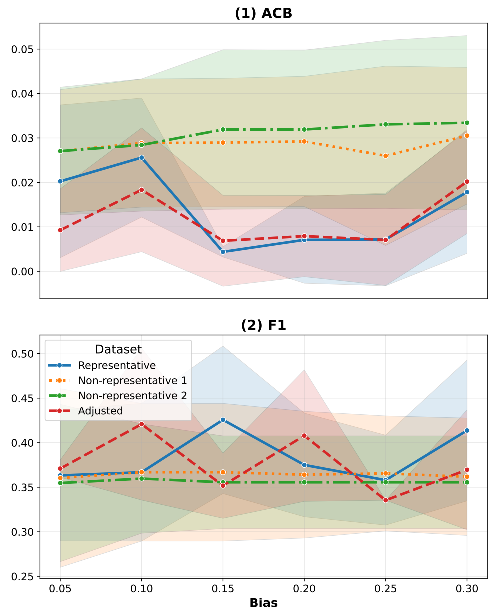

Figure 6: Metrics for OL models for filtered tweets ($0.4 \leq p_{i,OL} \leq 0.6$), by dataset and bias ($\beta$)

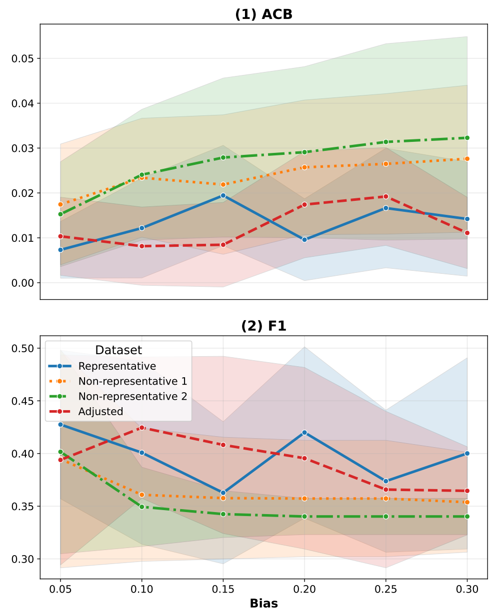

Figure 7: Metrics for HS models for filtered tweets ($0.4 \leq p_{i,HS} \leq 0.6$), by dataset and bias ($\beta$)