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 when annotator pools are not representative. Since collecting representative labels is challenging, we propose Population-Aligned Instance Replication (PAIR), a method to address this bias through statistical adjustment. Using a simulation study of hate speech and offensive language detection, 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. However, PAIR, which duplicates labels from underrepresented annotator groups to match population proportions, significantly reduces bias without requiring new data collection. These results suggest statistical techniques from survey research can help align model training with target populations even when representative annotator pools are unavailable. We conclude with three practical recommendations for improving training data quality.

1 Introduction and Inspiration

NLP models should align with the interests and judgments of the population they impact (Sorensen et al., 2024; Fleisig et al., 2024). However, the training and feedback data for these models often comes from crowdworkers or convenience samples of annotators (e.g. student assistants). These populations differ from the general population on important characteristics like age, education, and cultural context (Smart et al., 2024; Berinsky et al., 2012; Ouyang et al., 2022), and these characteristics impact the labels they assign (Sap et al., 2022; Fleisig et al., 2023; Kirk et al., 2024).

Fortunately, survey researchers have developed robust statistical techniques to estimate population-level parameters from non-representative samples (Eckman et al., 2024; Bethlehem et al., 2011a). The top panel of Figure 1 shows a simple workflow of collecting survey data and then creating statistical weights to match the data to the population. We propose that similar techniques could help align machine learning models with target populations, even when working with imperfect annotator pools (bottom panel).

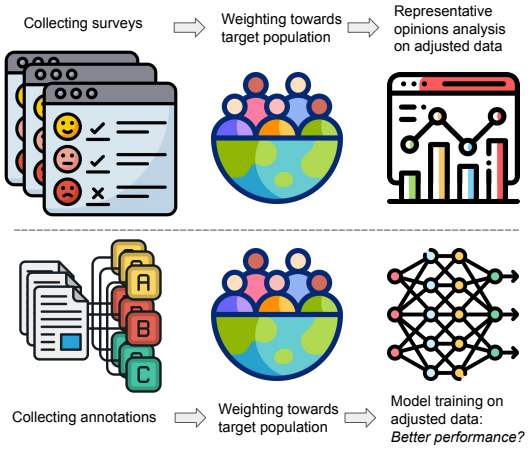

Figure 1: Top: Adjusting survey data to match population produces high quality results. Bottom: Can a similar adjustment in data annotations also improve model performance?

To test this approach, we use a simulation study, a common approach in the statistics literature (Burton et al., 2006; Valliant, 2019; Morris et al., 2019). We simulate seven populations, create several labeled datasets with varying mixes of annotators, and train model on each dataset. We investigate two research questions: **RQ1:** How does the composition of the annotator pool impact model calibration and performance? **RQ2:** Can techniques from survey methodology mitigate the effects of non-representative annotator pools?

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 help make AI systems more representative of the populations they serve.

2 Annotation Simulation

To address our research questions, we create populations with two types of people: those more likely to perceive offensive language and hate speech and those less likely. We then simulate labels, varying the mix of the two types of people, to approximate when happens when the pool of annotators does not represent the characteristics of the population.

Data. We use a dataset[1] of 3,000 tweets sampled from Davidson et al. (2017). Each tweet has 15 annotations of both offensive language (OL: yes/no) and hate speech (HS: yes/no) (Kern et al., 2023). 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) from the 15 labels of each tweet in the original dataset.[2] 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.

Variable	p25	Median	Mean	p75
$p_{i,OL}$	0.167	0.667	0.564	0.917
$p_{i,HS}$	0.083	0.167	0.301	0.50

p refers to percentile

Table 1: Distribution of $p_{i,OL}$ and $p_{i,HS}$ in Gold Dataset

Population Setup. We imagine populations made up of equal shares of two types of people. **Type A** people are *less likely* to say a tweet contains OL or HS. **Type B** people are *more likely*:

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

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

The HS probabilities are defined similarly. Here β captures the magnitude of the bias. We vary β from $[0.05, 0.3]$ by 0.05 to create seven populations and seven vectors of probabilities $(p^A_{i,OL}, p^B_{i,OL}, p^A_{i,HS}, p^B_{i,HS})$ for each tweet.

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

[2] For reasons that will become clear, it is helpful for the number of labels per tweet to be divisible by four.

Label Simulation. For each value of β, we create four datasets, each with 3,000 tweets (Table 2). The **Balanced** Dataset contains OL labels from six A annotators (drawn from Bernoulli($p^A_{i,OL}$)) and six B annotators (drawn from Bernoulli($p^B_{i,OL}$)). 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
Balanced	12	6	6
Unbalanced 1	9	6	3
Unbalanced 2	12	9	3
Adjusted	12	6	3 + 3*

NA = Not Applicable; * 3 B labels duplicated

Table 2: Four Training Datasets for each label (OL, HS) and each bias value (β)

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

Finally, we create the **Adjusted** Dataset. It is the same as the Unbalanced 1 dataset, but the B annotations are duplicated. This duplication is an easy way to adjust the unbalanced training dataset to reflect the population; a generalized version of this adjustment is provided in A.5.

Appendix Figure 5 shows the percentage of tweets labeled OL and HS in the four datasets for each value of β.

3 Model Training and Evaluation

Training and Test Setup. We train models on each dataset. We divide each dataset, at the tweet level, into training (2000 tweets), development (500), and test (500) sets.

Model Selection and Training. We used RoBERTa base (Liu et al., 2019) as our text classifier. The model trained 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 §A.2 contains additional details.

Performance Metrics. We measure model performance using two metrics.

- **Absolute Calibration Bias (ACB):** This metric compares the observed proportion of OL/HS labels in the Balanced dataset to each the model's predicted probabilities (preds_i):

$$\text{ACB}_{OL} = \frac{1}{n}\sum_{i=1}^{n}\left|p_{i,OL} - \text{preds}_{i,OL}\right| \quad (3)$$

  ACB simplifies the ECE metric (Naeini et al., 2015; Guo et al., 2017) by removing bins, and makes the single difference more straightforward by incorporating true frequencies directly. A low ACB score means the model's predicted probabilities match the true frequencies in the Gold dataset.
- **F1 Score:** This standard classification metric balances precision and recall.

The appendix also presents ECE (similar to ACB) and ROC-AUC (similar to F1) metrics (see §A.4). We report all metrics on the test set.

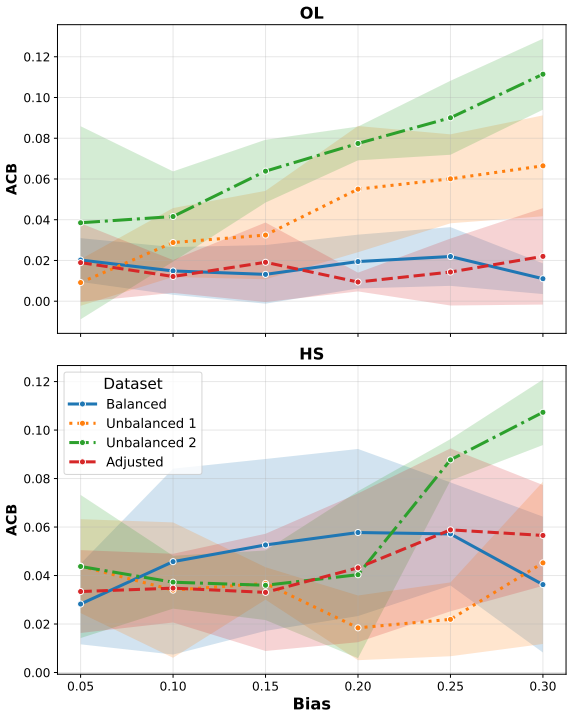

Figure 2: Model ACB scores, by dataset and bias (β)

4 Results

Model Calibrations. Figure 2 compares the ACB in the test set for models trained on the simulated datasets. Lower ACB indicates better model calibrations with the test labels. The dark lines show average ACB across the five training runs and the shading shows the standard deviation. See Appendix Figure 7 for the ECE results, which are similar.

In the OL graph, ACB for the models trained on the Balanced and Adjusted datasets do not increase with β and are close together. 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 calibration.

In the HS graph, the trends are less clear. Because HS is rarer in the dataset (Table 1), the A annotations are often 0, which complicates interpretation of the HS panel of Figure 2. We address this issue with additional analyses below.

Model Predictions. Figure 3 compares the models' F1 scores. In contrast to Figure 2, 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 noise in the labels increases, the models have a harder time predicting the labels. The ROC-AUC graphs show the same pattern (Appendix Figure 8).

Because the F1 metric focuses on binary predictions, it is less sensitive to training biases compared to calibration metrics, 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, and binary classification performance alone can obscure such effects. In decision-making, miscalibrated predictions can have harmful consequences when, e.g., hateful content remains undetected (Van Calster et al., 2019).

Results with 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. 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 4 shows the ACB for models trained on the filtered datasets. In the OL graph, the patterns in the ACB scores are similar to Figure 2. The Balanced and Adjusted models have similar ACB

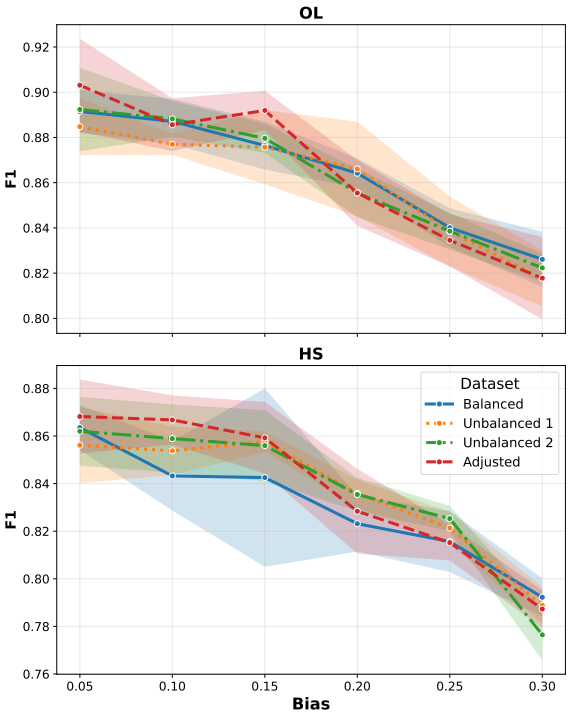

Figure 3: Model F1 scores, by dataset and bias (β)

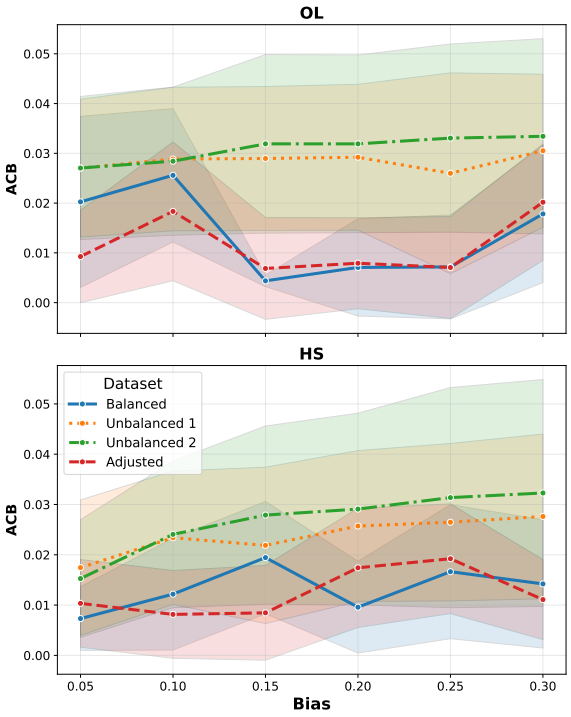

Figure 4: Model ACB scores for filtered tweets ($0.4 \leq p_i \leq 0.6$), by dataset and bias (β)

and are lower than the Unbalanced models. The HS graph shows the same pattern, which is in contrast to Figure 2 where HS performed differently.

The F1 results for this subset of tweets are in Appendix Figure 6. The OL graph shows higher F1 scores for the Balanced and Adjusted models than for the two unbalanced models. There is no clear pattern in the F1 scores for the F1 model, though we do see an unexpected decrease in agreement in the Adjusted model when $\beta = 0.3$.

5 Discussion & Recommendations

Our results demonstrate two key findings about annotator representation in training data. First, models perform less well when trained on data from non-representative annotator pools. Second, simple statistical adjustments can help correct for these biases without collecting additional data. These findings have important implications for dataset creation and model training.

Recommendations If these findings hold up in future work, we recommend the AI/ML field take three steps:

1) Identify which annotator characteristics influence labeling and feedback decisions for different tasks. Research should integrate social science methods to understand how demographics, attitudes, and behaviors shape annotation patterns (Eckman et al., 2024). This knowledge would help determine which characteristics need to be balanced across annotator pools for specific tasks.

2) Begin collecting relevant characteristics from annotators and gather corresponding population-level data. National censuses and large-scale surveys could provide useful population benchmarks.[3]

3) Implement data adjustment methods that account for differences between annotator and population characteristics. While our simple duplication approach showed promise, more sophisticated statistical techniques from survey research may yield better results.

Limitations

Stylized Biases. 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

[3]Collection and release of annotator characteristics or weights derived from them may raise concerns about annotator confidentiality. This topic is outside the scope of this paper, however, the survey literature contains useful approaches (see Karr, 2016, for a review). We also note that collecting annotator characteristics may require involvement of Institutional Review Boards or other participant protection organizations (Kaushik et al., 2024).

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 have used five seeds in model training.

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 classification accuracy, we did not examine other important metrics such as fairness across subgroups or robustness to adversarial examples. Future work should 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 the Claude model to edit the text of the paper and to assist in coding.

References

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. 2011a. *Handbook of Nonresponse in Household Surveys*. Wiley.

Jelke Bethlehem, Fannie Cobben, and Barry Schouten. 2011b. *Weighting Adjustment Techniques*, chapter Eight. John Wiley & Sons, Ltd.

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.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515.

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.

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.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1321–1330. JMLR.org.

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.

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*.

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.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1).

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.

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

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.

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.

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

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 Appendix

A.1 Data

Figure 5 shows the percentage of tweets labeled OL and HS in the four datasets for each value of β. The percentage in the Adjusted dataset is similar to that in the Balanced dataset for all values of β. The two unbalanced datasets have lower rates of OL and HS, because they overrepresent the A annotators, who are less likely to label OL and HS.

Because HS is rare in our dataset, as β increases, a higher proportion of $p^A_{i,HS}$ are 0 while the $p^B_{i,HS}$ values increase. This issue leads to a higher proportion of "yes" HS labels in the Balanced and Adjusted datasets, which have more B labels than the unadjusted datasets.[4]

A.2 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 $\{10, 42, 512, 1010, 3344\}$

[4] If we had picked the B labels to overrepresent, we would have seen a converse problem with the OL labels.

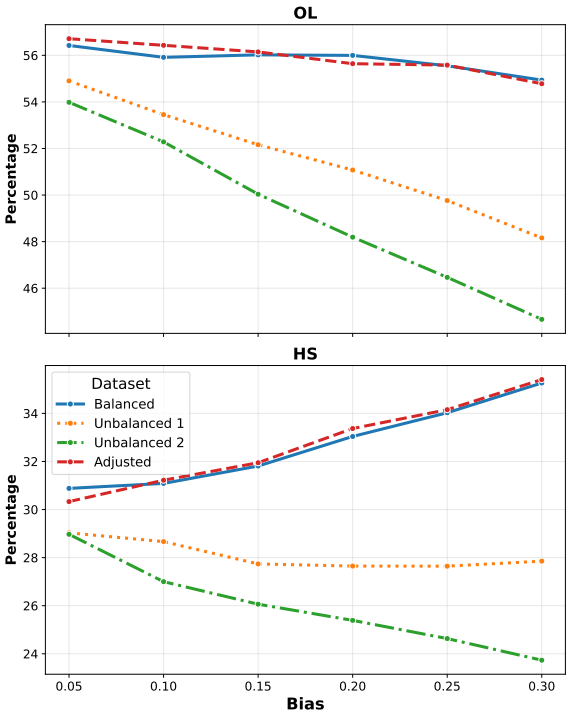

Figure 5: Percentage of OL, HS "yes" labels, by dataset and bias (β)

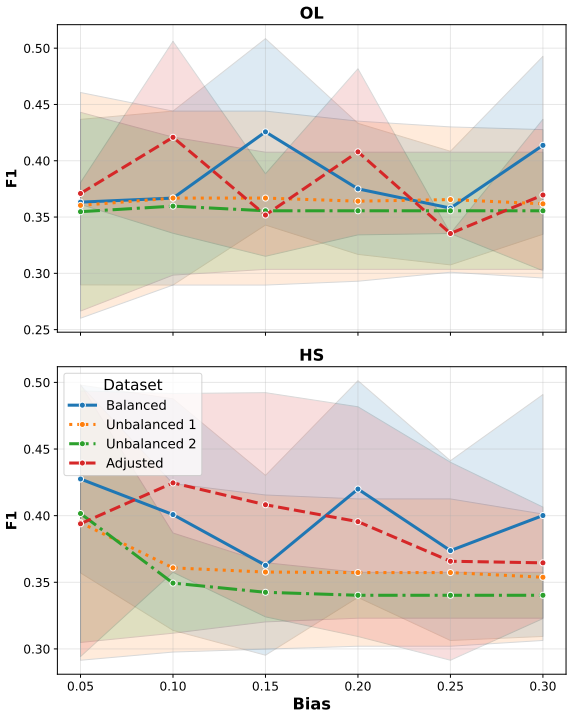

Figure 6: Model F1 scores for filtered tweets ($0.4 \leq p_i \leq 0.6$), by dataset and bias (β)

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

A.3 Additional Results

We present the results of F1 scores for filtered tweets ($0.4 \leq p_i \leq 0.6$) in Figure 6.

A.4 Additional Metrics

We also use the Expected Calibration Error (ECE) metric to evaluate model calibration. This metric quantifies calibration quality by measuring the weighted average absolute difference between accuracy and predicted confidence across M bins:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \cdot |\text{acc}(B_m) - \text{conf}(B_m)| \quad (4)$$

where B_m is the set of samples in the m-th bin, $|B_m|$ its size, n the total number of samples, $\text{acc}(B_m)$ the accuracy, and $\text{conf}(B_m)$ the mean predicted confidence in B_m.

Figure 7 shows the results. Consistent with our main findings in §4, we observe that for OL labels, the ECE values are relatively stable on the Balanced and Adjusted datasets ($t = -0.48, p = 0.64$), with divergence with $\beta = 0.15$. This contrasts with the significant divergence in ECE values for the Unbalanced subsets as the β value increases.

Additionally, for model performance, we use another metric ROC-AUC in Figure 8. The results are very similar to the F1 results in §4.

A.5 Adjustment Details

The adjustment we use to make the labeler 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 (Bethlehem et al., 2011b) is a method of statistical adjustment that makes a selected sample more closely resemble a target popu-

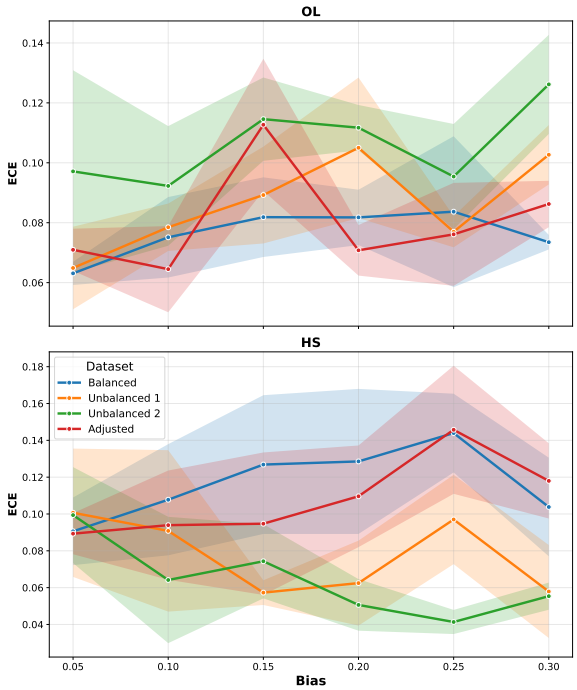

Figure 7: Model ECE metrics, by dataset and bias (β)

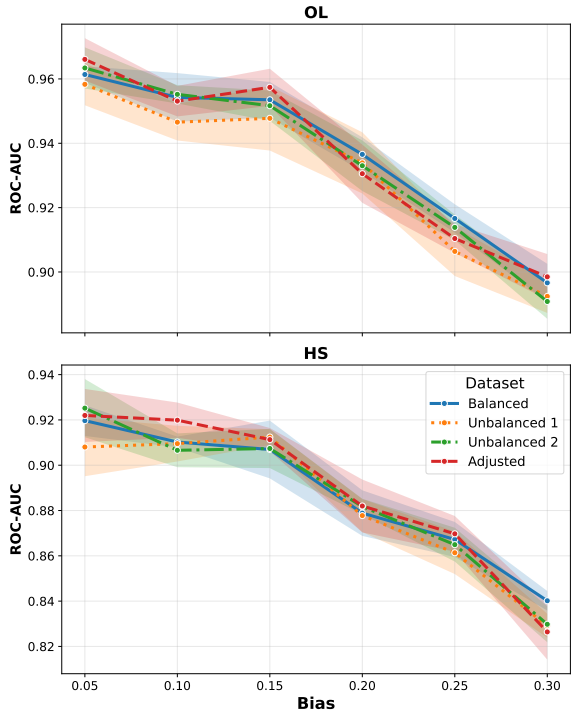

Figure 8: Model ROC-AUC metrics, by dataset and bias (β)

lation. Post-stratification requires population level totals or proportions and corresponding case-level observations of the same characteristic. to be available for each sample stratum that will be used in the weighting. The weight for unit i is determined for each stratum by:

$$w_i = \frac{P_s}{S_s} \tag{5}$$

where P_s is the true population proportion (or total) for stratum (or group) s and S_s is the sample proportion (or total) for stratum s. In our case, the strata of interest was a single variable (labeler 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:

$$w_i^{\text{normalized}} = w_i^{\text{initial}} \cdot \frac{T}{\sum_{i=1}^{n} w_i^{\text{initial}}} \tag{6}$$

where T is the target total. Since we want the Adjusted dataset to match the size of our gold standard Balanced 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_i times. In this initial work, we appreciated the simple interpretation and reproducibility of this approach. However, researcher may prefer other approaches, such as replication via resampling, if they are interested in how the adjustment varies across samples.

In our case, after post-stratification and weight normalization, each Type A label in the Unbalanced 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.