On the Interplay between Human Label Variation and Model Fairness

Kemal Kurniawan[1] **Meladel Mistica**[1] **Timothy Baldwin**[2,1] **Jey Han Lau**[1]
[1]School of Computing and Information Systems, University of Melbourne
[2]Mohamed bin Zayed University of Artificial Intelligence
{kurniawan.k,misticam,laujh}@unimelb.edu.au tb@ldwin.net

Abstract

The impact of human label variation (HLV) on model fairness is an unexplored topic. This paper examines the interplay by comparing training on majority-vote labels with a range of HLV methods. Our experiments show that without explicit debiasing, HLV training methods have a positive impact on fairness under certain configurations.[1]

1 Introduction

Human label variation (HLV; Plank, 2022) has been shown to improve performance through better generalisation (Peterson et al., 2019; Uma et al., 2021b; Kurniawan et al., 2025). However, the impact of HLV on model fairness remains unexplored. One might conjecture that HLV would improve fairness because it preserves minority views and avoids collapsing them into a single point like majority voting. However, it is well established that there is a trade-off between performance and fairness (Han et al., 2022; Shen et al., 2022), so one might also conjecture that HLV would similarly negatively impact fairness (since HLV improves performance). We explore this interplay between HLV and model fairness in this paper, in what we believe to be the first systematic analysis of this interaction. We compare training on majority-vote labels with four HLV training methods from prior work (Kurniawan et al., 2025), and evaluate on two tasks: offensiveness classification with SBIC (Sap et al., 2020) and legal area classification with TAG, an in-house legal dataset, developed with practising lawyers.

We find that across SBIC and TAG, training with HLV methods improves overall performance without harming fairness. Over TAG, training with HLV even *improves* fairness. Since these findings are only from a single set of class and group weights (*configuration* hereafter), we examine many possible configurations to draw more robust conclusions. For both datasets, HLV does not reduce model fairness for most configurations, and this trend is clearer when the configuration prioritises all groups/classes equally. Interestingly, HLV can improve model fairness for a substantial number of configurations.

We then investigate why HLV improves fairness by introducing temperature scaling to an HLV training method to control the weighting of minority annotations. Our TAG experiments suggest fairness gains do indeed come from minority annotations.

Our findings suggest that training with HLV improves performance without compromising fairness. Additionally, our findings underscore the effect of class and group weightings in measuring fairness in the HLV context. Therefore, it is important to understand what fairness means for each application before incorporating HLV.

2 Related Work

HLV (Plank, 2022) challenges the standard assumption that an instance has one ground truth, in line with data perspectivism (Cabitza et al., 2023) which advocates for the consideration of multiple perspectives in data annotation. Previous work on this topic includes model generalisation (Uma et al., 2021b,a; Leonardelli et al., 2023), evaluation metrics (Uma et al., 2021b; Rizzi et al., 2024; Kurniawan et al., 2025), model calibration (Baan et al., 2022), and separating annotation signal from noise (Weber-Genzel et al., 2024; Ivey et al., 2025). HLV has also been shown to improve model generalisation for computer vision tasks (Peterson et al., 2019). Our work contributes to this growing discipline by examining the interplay between HLV and fairness.

Previous work covers the interplay between model fairness and other aspects of machine learning models. Hessenthaler et al. (2022) studied how

[1]Code and supplementary materials: https://github.com/kmkurn/hlv-fairness-interplay

Findings of the Association for Computational Linguistics: EACL 2026, pages 967–976
March 24-29, 2026

model fairness and environmental costs affect each other. Zhao et al. (2023); Brandl et al. (2024) explored the interaction between model fairness and explainability, as did Ferry et al. (2025) but with privacy considerations. In contrast, we study how model fairness is impacted by HLV, which to our knowledge, is the first examination of this interplay.

3 Method

3.1 Datasets

SBIC Each instance in the Social Bias Inference Corpus (Sap et al., 2020) is an English social media post annotated with an offensiveness class and a list of target minority groups.[2] Target groups are divided into 7 broad categories[3] with each post annotated by 3 crowd annotators. We compute its offensiveness class distribution and union of the broad categories over these 3 annotations.

TAG A private dataset developed together with Justice Connect,[4] a legal non-profit that helps connect legal help-seekers to pro-bono lawyers. The dataset is drawn from an intake triage process, where each instance is an English description of a legal problem by a help-seeker and annotated with one or more areas of law (e.g., *Criminal law*, *Not a legal issue*) by practising lawyers, rendering the task multi-label. In addition, each instance is associated with one or more self-reported help-seeker demographic identities/cohorts (e.g., seniors, low income earners).[5] Justice Connect provides a weighting for the classes (i.e. legal areas) and groups (i.e. cohorts) reflecting their priorities in the intake process. We use this weighting for the fairness evaluation in Section 3.2. The average inter-annotator agreement[6] over the legal areas is Krippendorff's α = 0.454, which is modest. Manual analyses and discussions with the organisation confirm that the annotation disagreements are often valid, suggesting they are genuine variations rather than noise.

3.2 Evaluation

We focus on a *group-level parity* definition of fairness in this study due to its prevalence in prior work (Shen et al., 2022; Han et al., 2023). To evaluate this, we first compute a fairness score s_{kg} (defined later) for each class k and group g. Next, following Han et al. (2023), we aggregate s_{kg} over groups then classes using weighted generalised mean to obtain the overall fairness score. The group-wise aggregation is defined as

$$\bar{s}_k = \left(\sum_g w_g s_{kg}^p \right)^{\frac{1}{p}} \tag{1}$$

where $p \neq 0$ is a scalar exponent controlling the contribution of smaller-valued s_{kg} and w_g is a non-negative importance weights for group g satisfying $\sum_g w_g = 1$. The class-wise aggregation is performed analogously over $\bar{s}_k$ using separate class-specific weights and exponent.

To define the groups, we use the broad target categories for SBIC, and the defined cohorts for TAG. Therefore, for SBIC, the group information is associated with the *target* of the social media post, while the group information for TAG is associated with the *help-seeker* who *produced* the instance.

In both datasets, each instance can belong to zero or more groups. Therefore, we partition the instances into two subsets for each group based on in- or out-of-group membership, respectively.[7] We compute s_{kg} by first evaluating performance for class k on both subsets then taking the ratio between the worse- and the better-performing subsets, following Yeh et al. (2024).[8]

Computing s_{kg} requires a performance metric for each class that works in the HLV context (i.e., soft ground truths). Thus, we use soft F_1, the only metric satisfying the requirement (Kurniawan et al., 2025). For a given class k, it measures the amount of overlap between the ground-truth and predicted probabilities of class k relative to their total.[9]

Aggregation configurations For TAG results in Section 4.1, the class and group aggregation weights w_g are provided by the legal organisation who co-developed the dataset (Section 3.1). For SBIC, as we have no a priori preference we weight each class and group equally. For both datasets, we set $p = 1$ for both aggregations, resulting in standard weighted average. In Section 4.2, we experiment with multiple randomly-sampled configurations for more robust conclusions.[10]

[2] Offensiveness classes = Yes/No/Maybe; example target groups = black folks, asian folks, etc.

[3] Body, culture, disabled, gender, race, social, victim.

[4] https://justiceconnect.org.au

[5] More details (incl. splits) are given in Appendix A.

[6] Computed on 10% of the data due to computational costs.

[7] See Appendix B for subset size statistics.

[8] See Appendix C.1 for a more formal definition.

[9] See Appendix C.2 for details.

[10] Details are given in Appendix E.

Overall performance To assess performance in the HLV context, we use soft *micro* F_1, a natural extension of class-wise soft F_1 which correlates well with human judgement (Kurniawan et al., 2025).[11]

3.3 Approaches

We experiment with two pretrained models: base RoBERTa (Liu et al., 2019, ≈100M parameters) and Nov'24 7B OLMo 2 (Walsh et al., 2025). We replace the output layer of each model with the appropriate classification layer[12] and finetune on the train set.

For HLV training methods, we use the four most successful methods identified by Kurniawan et al. (2025): repeated labelling (REL), soft labelling (SL), minimising Jensen-Shannon divergence (JSD), and maximising soft micro F_1 score (SMF1).[13] For non-HLV baselines, we finetune them on majority-vote classes (MV).[14]

4 Results

4.1 Single Fairness Configuration

We report the overall performance and fairness of each method in Table 1. For SBIC, the table shows that HLV training methods consistently outperform MV for overall performance across the two models. These improvements are statistically[15] significant (with bootstrap tests) for 6 out of 8 model–method pairs.[16] As for fairness, the HLV training methods can improve or reduce fairness, but the differences are not statistically significant for 7 out of 8 model–method pairs.[17]

Looking at the TAG results, for both overall performance and fairness, the HLV training methods consistently outperform MV across both models, with statistically significant improvements for 6 out of 8 model–method pairs.[18] Taken together, the results suggest that HLV training improves overall performance without sacrificing fairness, and in some cases (e.g., TAG), can even improve fairness.

Discussion Unsuprisingly, the different HLV methods attain different levels of fairness. However, these variations are consistent with their overall performance. For example, REL, a strong method performance-wise, is also among the strongest in terms of fairness, while SMF1 is among the weakest. That SMF1 attains poor performance is not surprising because Kurniawan et al. (2025) showed that the method can result in degenerate predictions where all probability mass is concentrated on the majority class. Looking into class- and group-wise performance for each method, we find that all noticeable differences seem to align well with overall performance trends.

Model	Method	SBIC		TAG	
		Perf	Fair	Perf	Fair
OLMo	MV	75.3	45.7	64.4	62.3
	REL	**78.9***	46.0	**70.4***	**72.8***
	SL	**78.3***	46.3	**69.0***	**72.2***
	JSD	**78.6***	44.2	**66.3***	**68.9***
	SMF1	**79.6***	42.6	64.8	64.7
RoBERTa	MV	77.8	42.9	64.5	67.3
	REL	**78.7***	41.9	**71.5***	**72.8***
	SL	78.0	48.5	**70.6***	**72.3***
	JSD	78.1	43.1	**70.7***	**71.1***
	SMF1	**79.1***	39.2*	66.9	70.4

Table 1: Mean performance (Perf) and fairness (Fair) of the MV baseline and HLV training methods over 3 runs. Asterisks (*) indicate statistical significance ($p < 0.05$) against MV with a two-tailed bootstrap test (MacKinnon, 2009). Significant gains over MV are boldfaced.

4.2 Multiple Fairness Configurations

The results in Section 4.1 were obtained with one configuration, i.e. fixed values for w_g and p in Equation (1) for both group- and class-wise aggregation. Here, we test multiple randomly-sampled configurations to draw more robust conclusions.[19] This experimental setup is especially important for SBIC: the single configuration used in Section 4.1 is divorced from any intended application of the model.

For each randomly-sampled configuration,[20] we perform a two-tailed bootstrap test across 3 runs to check if the mean fairness score difference between each HLV training method and MV is statistically significant ($p < 0.05$). If so, we record if the mean fairness score of MV is higher (i.e. fairer). Otherwise, we deem the HLV training method as fair as MV. Note that overall performance is unaffected by fairness configurations. Therefore, we compare

[11] See Appendix C.2 for details.
[12] A linear layer then either a softmax (SBIC) or sigmoid (TAG) activation functions.
[13] See Appendix F for more details.
[14] Further training details are reported in Appendix G.
[15] We report effect sizes as a measure of *practical* significance in Appendix H.
[16] Exceptions: SL and JSD with RoBERTa
[17] Exception: SMF1 with RoBERTa
[18] Exception: SMF1

[19] See Appendix E for more details.
[20] Note that performance is constant regardless of configurations as they configure only the fairness score computation.

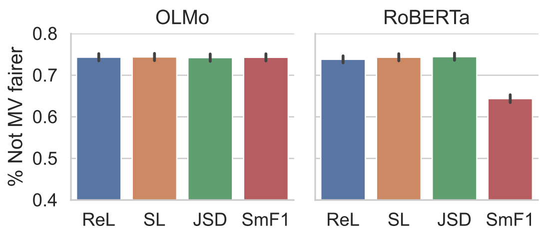


Figure 1: Fraction of randomly sampled configurations where MV is not significantly fairer than each HLV training method on SBIC.

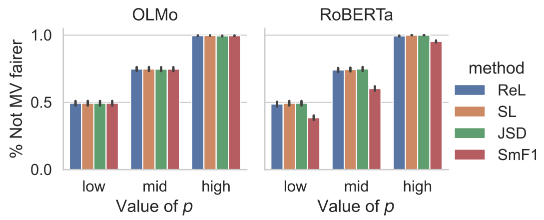


Figure 2: Fraction of randomly sampled configurations where MV is not significantly fairer than each HLV training method on SBIC for various levels of exponent p in the group-wise aggregation.

only the fairness scores.

We report the fraction of configurations where MV is found *not* to be fairer than each HLV method on SBIC in Figure 1. The result on TAG follows the same trend, which we show in Appendix I.

Figure 1 shows for most model–method pairs, training with HLV does not hurt model fairness in over 70% of configurations. An exception is SMF1 with RoBERTa, but the proportion is still over 60%. Remarkably, we find that training with HLV improves fairness for up to 25% of configurations (see Appendix I). Overall, the results suggest that training with HLV generally does not hurt fairness, further supporting our findings in Section 4.1.

That said, Figure 1 also shows a substantial number of configurations where MV is fairer than training with HLV. This finding highlights the need to determine fairness configurations specific to the application before incorporating HLV.

5 Analysis

5.1 Fairness Configurations

This section investigates the role p in Equation (1) for both group- and class-wise aggregation. We categorise the value of p into 3 levels: low (< -5), mid (–5 to 5), and high (> 5). In Figure 2, we plot the fraction of configurations where training with HLV does not decrease fairness for each level of p in the group-wise aggregation on SBIC. Figure 2 shows that when p is low, the fraction is about 50%, lower than when p is mid-ranging. In contrast, the fraction increases to almost 100% when p is high. A nearly identical trend is observed for the class-wise aggregation (see Appendix J). TAG results are similar to SBIC (see Appendix J).

Scalar p controls the contribution of smaller-valued s_{kg} in group-wise (Equation (1)) and class-wise aggregation. Smaller/larger p means prioritising lower-scoring groups or classes more/less. Therefore, the results suggest that when one cares about all classes or groups similarly regardless of their scores (mid), training with HLV does not reduce fairness. This setting corresponds to most existing fairness evaluation (Han et al., 2023). However, when one prioritises the lowest-scoring classes or groups in fairness (low, similar to Rawls (2001)), training with HLV is comparable to MV.[21]

5.2 Minority Annotations

Here, we investigate factors that make HLV beneficial for model fairness. We focus on the TAG dataset on which HLV generally outperforms MV in both performance and fairness as reported in Section 4.1. We hypothesise that the fairness improvements come from the minority annotations.

To test this, we introduce temperature scaling to SL. Each class probability is proportional to the number of annotators that select the class raised to the power of $\frac{1}{\tau}$ where $\tau > 0$ is the temperature. As $\tau \to 0$, the most-voted class will have probability close to 1, resembling MV. In contrast, the class distribution will reduce to uniform as $\tau \to \infty$, thereby weighting up the minority annotations.[22] Standard SL corresponds to $\tau = 1$.

We report how fairness changes as τ grows in Figure 3. The figure shows that when τ is small, fairness is low. As τ increases, fairness improves and stabilises after peaking at $\tau = 1$. These findings suggest minority annotations contain useful signals for fairness. We offer a possible explanation for this in Appendix K and leave the full investigation on this direction for future work.

Unsurprisingly, the figure also shows that performance peaks at $\tau = 1$. This is because the scaling

[21] The results also suggest that training with HLV generally never harms fairness when one focuses more on the highest-scoring classes or groups (high). However, this overestimates model fairness and seems unused in practice.

[22] In practice, distributions will be uniform over all most-voted classes as $\tau \to 0$ and over selected classes as $\tau \to \infty$.

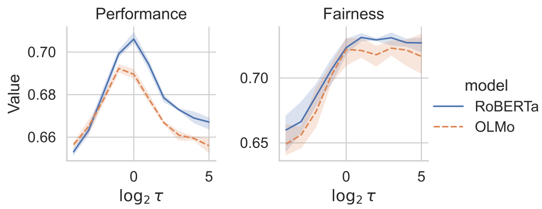


Figure 3: Impact of minority annotations (τ) on fairness (right) on TAG. Larger τ means minority annotations are weighted more. Impact on performance (left) is shown for completeness.

is only applied to the train set, i.e. test ground truths are computed normally ($\tau = 1$).

6 Conclusions

While HLV has been shown to improve generalisation, this is the first work to study the interplay between HLV and fairness. We first showed empirically that training with HLV does not harm model fairness, and indeed, it can improve fairness. We then showed that this observation is robust to various fairness configurations. However, when the lowest-scoring groups or classes in fairness were prioritised, training with HLV provided no fairness benefits compared to training on majority-vote classes. Lastly, we confirmed that the fairness improvements with HLV training come from the minority annotations.

Our findings highlight the performance and fairness benefits of training with HLV under certain configurations. Therefore, we recommend NLP practitioners to: (1) consider what fairness means in their application, (2) define it in terms of the fairness evaluation configuration described in Section 3.2, (3) incorporate HLV if the configuration prioritises all groups/classes roughly equally, and (4) evaluate fairness with the configuration to confirm that HLV does not harm fairness compared to majority voting.

Limitations

Because the TAG dataset contains real-world confidential legal help requests, we cannot distribute it as it used in the study. However, we still include it because of its importance to our analysis in Section 5.2, and we believe that the insight obtained from the dataset in the study is still useful. Although we cannot distribute the dataset, we can distribute ground-truth and predicted judgement distributions of the dev and the test sets.[23] This will allow other researchers to replicate our reported numbers and build on our work to some extent.

Prior work has evaluated fairness using synthetic, targeted check lists (Manerba and Tonelli, 2021). Therefore, a reasonable solution may be to follow a similar approach to address the data privacy issue. We agree in theory that it would allow us to not only measure fairness at a high level but also understand in what cases the models break down. However, our goal is to understand fairness and its relationship with HLV rather than address specific impacts of model bias. Furthermore, we foresee a few complications: (1) it is not always immediately clear what the biases might be as we need to know them in order to create the synthetic check lists); (2) in our case, the language of the TAG data is non-standard English, and so it is difficult to mimic that genre; and (3) we will need to distribute our TAG-trained models, which is restricted because of IP constraints and possible risks of leaking the sensitive TAG training data through the model. Therefore, this work does not take this approach.

We evaluate using soft evaluation metrics in all experiments. One could argue that this setup favours HLV training, so the comparisons with MV are unfair. However, we note that previous work has found that MV outperforms many HLV training methods in this setup (Kurniawan et al., 2025), suggesting that the argument does not hold in practice.

Another limitation of our study is the exclusive use of English datasets. This is because of the scarcity of datasets that have both disaggregated annotations for HLV and instance attributes (e.g., cohort information) which are appropriate for fairness evaluation. However, we expect that our findings would generalise to non-English languages, as our experimental design is language-independent. We encourage future work to develop such datasets for non-English languages to test this empirically.

Acknowledgments

This research is supported by the Australian Research Council Linkage Project LP210200917[24] and funded by the Australian Government. This research is done in collaboration with Justice Con-

[23] https://github.com/kmkurn/hlv-fairness-interplay

[24] https://dataportal.arc.gov.au/NCGP/Web/Grant/Grant/LP210200917

nect, an Australian public benevolent institution.[25] We thank Kate Fazio, Tom O'Doherty, and Rose Hyland from Justice Connect for their support throughout the project. This research is supported by The University of Melbourne's Research Computing Services and the Petascale Campus Initiative.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernandez. 2022. Stop Measuring Calibration When Humans Disagree. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1892–1915.

Stephanie Brandl, Emanuele Bugliarello, and Ilias Chalkidis. 2024. On the Interplay between Fairness and Explainability. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pages 94–108.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a Perspectivist Turn in Ground Truthing for Predictive Computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Julien Ferry, Ulrich Aïvodji, Sébastien Gambs, Marie-José Huguet, and Mohamed Siala. 2025. Taming the Triangle: On the Interplays between Fairness, Interpretability, and Privacy in Machine Learning. *Computational Intelligence*, 41(4):e70113.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2023. Fair Enough: Standardizing Evaluation and Model Selection for Fairness Research in NLP. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 297–312.

Xudong Han, Aili Shen, Yitong Li, Lea Frermann, Timothy Baldwin, and Trevor Cohn. 2022. FairLib: A Unified Framework for Assessing and Improving Fairness. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 60–71.

Marius Hessenthaler, Emma Strubell, Dirk Hovy, and Anne Lauscher. 2022. Bridging Fairness and Environmental Sustainability in Natural Language Processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7817–7836.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *The Tenth International Conference on Learning Representations*.

Jonathan Ivey, Susan Gauch, and David Jurgens. 2025. NUTMEG: Separating Signal From Noise in Annotator Disagreement. *Preprint*, arXiv:2507.18890.

Kemal Kurniawan, Meladel Mistica, Timothy Baldwin, and Jey Han Lau. 2025. Training and Evaluating with Human Label Variation: An Empirical Study. *Computational Linguistics*, pages 1–27.

Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. SemEval-2023 Task 11: Learning with Disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318.

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. *Preprint*, arXiv:1907.11692.

James G. MacKinnon. 2009. Bootstrap hypothesis testing. In *Handbook of Computational Econometrics*, chapter 6, pages 183–213. Wiley.

Marta Marchiori Manerba and Sara Tonelli. 2021. Fine-Grained Fairness Analysis of Abusive Language Detection Systems with CheckList. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 81–91.

Joshua Peterson, Ruairidh Battleday, Thomas Griffiths, and Olga Russakovsky. 2019. Human Uncertainty Makes Classification More Robust. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9616–9625.

Barbara Plank. 2022. The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682.

John Rawls. 2001. *Justice as Fairness: A Restatement*. Harvard University Press.

Giulia Rizzi, Elisa Leonardelli, Massimo Poesio, Alexandra Uma, Maja Pavlovic, Silviu Paun, Paolo Rosso, and Elisabetta Fersini. 2024. Soft metrics for evaluation with disagreements: An assessment. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 84–94.

[25]As defined by the Australian government: `https://www.acnc.gov.au/charity/charities/4a24f21a-38af-e811-a95e-000d3ad24c60/profile`

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 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.

Shlomo Sawilowsky. 2009. New Effect Size Rules of Thumb. *Journal of Modern Applied Statistical Methods*, 8(2).

Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2022. Optimising Equal Opportunity Fairness in Model Training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4073–4084.

Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021a. SemEval-2021 Task 12: Learning with Disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347.

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

Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, and 20 others. 2025. 2 OLMo 2 Furious. *Preprint*, arXiv:2501.00656.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating Annotation Error from Human Label Variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269.

Min-Hsuan Yeh, Blossom Metevier, Austin Hoag, and Philip Thomas. 2024. Analyzing the Relationship Between Difference and Ratio-Based Fairness Metrics. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 518–528.

Yuying Zhao, Yu Wang, and Tyler Derr. 2023. Fairness and Explainability: Bridging the Gap towards Fair Model Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):11363–11371.

A TAG Dataset Details

The TAG dataset has a total of 11K instances split randomly 8:1:1 for train, dev, and test sets. Each instance is a legal problem description, e.g., *my landlord evicted me w/o notice*, annotated by an average of 5.5 lawyers with one or more legal areas out of 33 options, e.g., *Criminal law*, *Elder law*, *Not a legal issue*. Each instance is also associated with one or more cohorts out of 6 choices, e.g., seniors, low income earners. Sensitive identifying information had been anonymised by the owner organisation prior to the dataset use in our work.

Group	Subset	
	In-group	Out-of-group
Body	58	4640
Culture	495	4203
Disabled	112	4586
Gender	503	4195
Race	819	3879
Social	104	4594
Victim	215	4483

(a) SBIC

Group	Subset	
	In-group	Out-of-group
ATS	35	317
HOM	154	64
LGB	29	8
LOW	135	148
PUB	19	109
SEN	164	90

(b) TAG (ATS = indigenous, HOM = experiencing or at risk of homelessness, LGB = LGBTQ+ individuals, LOW = low-income earners, PUB = public housing dwellers, SEN = seniors)

Table 2: Number of instances in the in- and out-of-group subsets defined by each group in SBIC (top) and TAG (bottom).

B Subset Size Statistics

Table 2 reports the number of instances in the in- and out-of-group subsets for each group in the datasets. For TAG, the help-seekers were allowed to select "Prefer not to say" when self-reporting. For fairness evaluation, we include only those that did not select that option, i.e. their identities were given explicitly.

C Evaluation Details

C.1 Definition of s_{kg}

If F^1_{kg} and F^0_{kg} denote the performance for class k on the in- and out-of-group subsets defined by group g respectively, then the fairness score s_{kg} is defined as

$$s_{kg} = \begin{cases} 1 & F_{kg}^0 = F_{kg}^1 = 0 \\ \frac{\min\left(F_{kg}^0, F_{kg}^1\right)}{\max\left(F_{kg}^0, F_{kg}^1\right)} & \text{otherwise.} \end{cases} \tag{2}$$

C.2 Performance

To compute F_{kg}^1 and F_{kg}^0 in Equation (2), we use the soft F_1 score (Kurniawan et al., 2025). The score for class k is defined as

$$2\frac{\sum_i \min\left(P_{ik}, Q_{ik}\right)}{\sum_i \left(P_{ik} + Q_{ik}\right)} \tag{3}$$

where P_{ik} and Q_{ik} denote the ground-truth and the predicted probability of class k for instance i respectively.

To evaluate overall performance with soft ground truths, we use the soft micro F_1 score (Kurniawan et al., 2025). It is defined formally as

$$2\frac{\sum_{ik} \min\left(P_{ik}, Q_{ik}\right)}{\sum_{ik} \left(P_{ik} + Q_{ik}\right)}.$$

This equation is similar to Equation (3), but the summations are performed over both instances and classes.

D Difference in Fairness Interpretations

Due to the different group definitions for both datasets, the interpretations of fairness are also different. Consider an arbitrary group in SBIC. Here, a fair model is one that performs similarly in identifying offensiveness on social media posts that target and do not target the minority category corresponding to the group. This is sensible because we do not want a model to identify offensive posts less accurately when the posts target minorities. In contrast, for an arbitrary group in TAG, a fair model is one that achieves similar levels of performance for help-seekers that belong and do not belong to the group. This is also sensible because we do not want a model to more frequently fail at recognising relevant legal areas for a particular cohort (e.g., seniors). Thus, while there is a difference, the fairness definitions are still reasonable for both cases.

E Configuration Sampling

To get the results in Section 4.2, we randomly sample ten thousand configurations for each of group- and class-wise aggregation. Each configuration consists of a set of item weights and a scalar exponent p that parameterises the weighted generalised mean used for aggregation. Specifically, the weights and exponent p are drawn from a flat Dirichlet distribution and $U(-15, 15)$ respectively. Smaller and larger values of p make the generalised mean focus more on lower and higher fairness scores respectively. In particular, as $p \to \infty$ and $p \to -\infty$, generalised mean reduces to $\max$ and $\min$ respectively. In other words, the exponent p controls how much the lowest-scoring groups or classes are prioritised.

F HLV Training Methods

We experiment with four HLV training methods:

1. REL (repeated labelling) which treats each annotation of an instance as a separate instance–class pair. In other words, an instance may appear multiple times in the training data, each with a different class. The loss function is standard cross-entropy.
2. SL (soft labelling) which computes a class distribution for each instance and uses the distributions as soft ground truths for training. Each class probability is proportional to the number of annotators that select the class. Standard cross-entropy is used as the loss function.
3. JSD which computes class distributions similarly to SL but uses Jensen-Shannon divergence as the loss function.
4. SMF1 which also computes class distributions as SL does but maximises the soft micro F_1 score during training (see Appendix C.2).

We refer the readers to the work by Kurniawan et al. (2025) for more details on these methods.

G Training Details

We implement all methods using FlairNLP (Akbik et al., 2019).[26] For RoBERTa, we tune the learning rate and the batch size using random search, optimising only for overall performance (without fairness). Table 3 reports the best values found. For OLMo, we use FlairNLP's default hyperparameters[27] for computational reasons. We use LoRA (Hu et al., 2022) to finetune OLMo efficiently. All models are finetuned for 10 epochs. On a single NVIDIA A100 GPU with 80GB of memory, the finetuning runs for:

[26] https://flairnlp.github.io

[27] Learning rate and batch size are 5×10^{-5} and 32 respectively.

Dataset	Method	Learning rate	Batch size
SBIC	MV	5.4×10^{-7}	4
	REL	5.4×10^{-7}	4
	SL	1.7×10^{-5}	4
	JSD	4.6×10^{-5}	512
	SMF1	4.6×10^{-5}	512
TAG	MV	5.2×10^{-6}	16
	REL	3.1×10^{-5}	8
	SL	3.1×10^{-5}	8
	JSD	6.1×10^{-5}	16
	SMF1	2.6×10^{-5}	64

Table 3: Best hyperparameter values for RoBERTa.

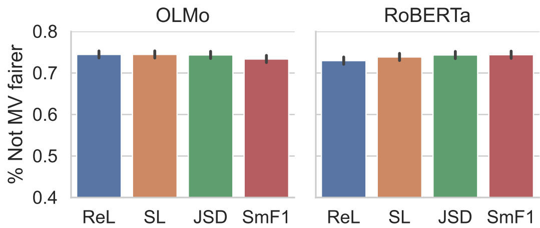


Figure 4: Fraction of randomly sampled configurations where MV is not significantly fairer than each HLV training method on TAG. Whiskers indicate 95% bootstrap confidence intervals.

- less than an hour with RoBERTa for all methods except REL on SBIC and TAG,
- slightly over 2 hours with RoBERTa for REL on SBIC and TAG,
- slightly over 10 hours with OLMo for all methods except REL on SBIC,
- about 1.3 days with OLMo for REL on SBIC,
- about 9 hours with OLMo for all methods except REL on TAG, and
- about 2 days with OLMo for REL on TAG.

We truncate inputs longer than 512 tokens in TAG to reduce memory consumption, affecting only 4% of instances.

H Effect Sizes

To complement our main result in Table 1, we also report the effect sizes between each HLV training method and MV as a measure of *practical* significance (as opposed to statistical). We report the 95% bootstrap confidence intervals of Cohen's d for both performance and fairness evaluations in Table 4. The table shows that all statistically significant results in Table 1 have an effect size greater than 2.0, which is considered "huge" (Sawilowsky, 2009). This observation demonstrates the practical significance of our findings.

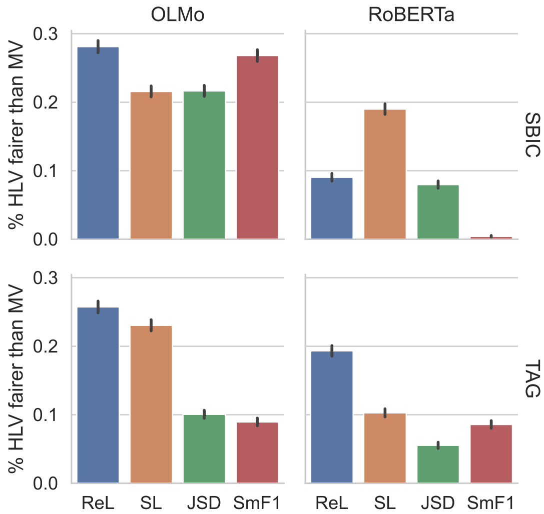


Figure 5: Fraction of randomly sampled configurations where each HLV training method is significantly fairer than MV. Whiskers indicate 95% bootstrap confidence intervals.

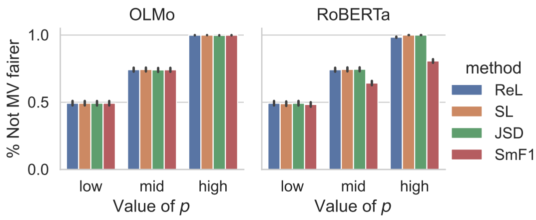


Figure 6: Fraction of randomly sampled configurations where MV is not significantly fairer than each HLV training method on SBIC for various levels of exponent p in the class-wise aggregation. Whiskers indicate 95% bootstrap confidence intervals.

I More Multiple Fairness Configurations Results

Figure 4 shows the fraction of configurations where MV is not fairer than each HLV training method on TAG. The figure shows a similar trend to that of SBIC reported in Section 4.2.

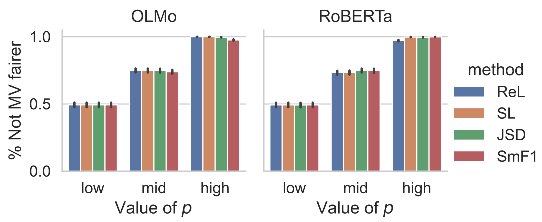
Figure 5 shows the fraction of configurations where each HLV training method is fairer than MV. We observe that except for SMF1 with RoBERTa on SBIC, this fraction is substantial for all dataset–model–method tuples. In particular, the fraction is over 25% for REL with OLMo across the two datasets.

J More Fairness Configurations Analysis Results

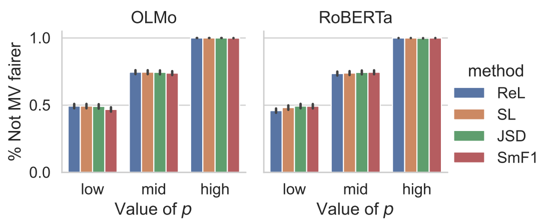
This section shows the fraction of configurations where MV is not fairer than each HLV training

Model	Method	SBIC				TAG			
		Perf		Fair		Perf		Fair	
		low	high	low	high	low	high	low	high
OLMo	ReL	23.7	193.8	−7.1	6.5	24.0	104.1	6.8	87.6
	SL	29.4	239.9	−8.1	3.1	16.7	80.5	5.9	61.8
	JSD	26.1	180.6	−7.9	0.5	5.4	44.3	3.6	23.7
	SmF1	22.4	234.2	−107.2	−1.7	0.0	6.5	0.7	9.8
RoBERTa	ReL	4.9	27.1	−7.0	0.0	40.3	1957.4	18.8	486.1
	SL	−0.1	359.1	2.0	53.3	25.4	3154.8	9.5	58.6
	JSD	1.1	12.0	−2.5	3.1	38.4	939.3	10.7	46.7
	SmF1	6.0	208.5	−12.6	−2.7	3.4	36.2	2.3	35.5

Table 4: 95% bootstrap confidence intervals of Cohen's d between each HLV training method and MV. Values greater than 2.0 are considered "huge" effect size (Sawilowsky, 2009).



(a) Group-wise



(b) Class-wise

Figure 7: Fraction of randomly sampled configurations where MV is not significantly fairer than each HLV training method on TAG for various levels of exponent p in the group- (top) or class-wise (bottom) aggregation. Whiskers indicate 95% bootstrap confidence intervals.

method for various levels of exponent p. Figure 6 shows the results on SBIC for the exponent p in the class-wise aggregation (the group-wise counterpart is shown in Figure 2). In contrast, Figure 7 shows the same results on TAG for the exponent p in both the group- and class-wise aggregation. The figures show that the trend is similar to that reported in Section 5.1.

K Possible Explanation on Minority Annotations Improving Fairness on TAG

We offer a possible explanation for the finding in Section 5.2 that the minority annotations on TAG contain useful signals that improve fairness. TAG was annotated by lawyers, each with a specific set of legal areas of specialisation. Intuitively, some legal specialisations are more common than others. We hypothesise that minority annotations are correlated with rare specialisations, which also correlate with certain cohorts. This would imply that incorporating the minority annotations would result in fairer models for the defined cohorts.