Training and Evaluating with Human Label Variation: An Empirical Study

Kemal Kurniawan[1*], Meladel Mistica[1], Timothy Baldwin[2,1], and Jey Han Lau[1]

[1]School of Computing and Information Systems, University of Melbourne
kurniawan.k@unimelb.edu.au

[2]Mohamed bin Zayed University of Artificial Intelligence

Human label variation (HLV) challenges the standard assumption that a labeled instance has a single ground truth, instead embracing the natural variation in human annotation to train and evaluate models. While various training methods and metrics for HLV have been proposed, it is still unclear which methods and metrics perform best in what settings. We propose new evaluation metrics for HLV leveraging fuzzy set theory. Because these new proposed metrics are differentiable, we then in turn experiment with using these metrics as training objectives. We conduct an extensive study over 6 HLV datasets testing 14 training methods and 6 evaluation metrics. We find that training on either disaggregated annotations or soft labels performs best across metrics, outperforming training using the proposed training objectives with differentiable metrics. We also show that our proposed soft micro F_1 score is one of the best metrics for HLV data.[1]

1. Introduction

Human label variation (HLV) challenges the standard assumption that an example has a single ground truth (Plank 2022). With HLV, different human annotations for the same instance are viewed as useful signal rather than undesirable noise. In this context, model training and evaluation are no longer straightforward. For example, standard metrics such as accuracy assume that a single ground truth label exists for each instance. To address this issue, recent work (Pavlick and Kwiatkowski 2019; Peterson et al. 2019; Nie, Zhou, and Bansal 2020; Uma et al. 2020, 2021; Cui 2023; Gajewska 2023; Maity et al. 2023; Wan and Badillo-Urquiola 2023, inter alia) has proposed various methods to train

* Corresponding author.

[1] Code and supplementary materials are available at: https://github.com/kmkurn/train-eval-hlv.

Action Editor: Anh Tuan Luu. Submission received: 23 March 2025; revised version received: 22 August 2025; accepted for publication: 8 October 2025.

https://doi.org/10.1162/COLI.a.578

© 2025 Association for Computational Linguistics
Published under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license

models and metrics to evaluate their performance in the HLV context. However, the lack of a systematic study means it is still unclear which methods and metrics perform best.

In natural language processing (NLP), most existing metrics for HLV represent human judgments as probability distributions over classes and use information-theoretic measures such as divergence. It has largely ignored the field of remote sensing that has dealt with issues in model evaluation against fuzzy ground truths (Foody 1996; Binaghi et al. 1999; Lewis and Brown 2001; Pontius and Cheuk 2006; Pontius and Connors 2006; Silván-Cárdenas and Wang 2008; Gómez, Biging, and Montero 2008). In remote sensing, the task is to classify patches of a satellite image of a region into categories such as land, water, and so forth. As an image patch often contains multiple categories, the field has developed evaluation metrics that can handle this variation in the ground truth labels.

Taking inspiration from remote sensing research, we represent human judgment distributions as degrees of membership over fuzzy sets and generalize standard metrics such as accuracy into their soft versions using fuzzy set operations. In contrast to information-theoretic measures, each of these soft metrics has an intuitive interpretation that corresponds to the standard counterpart. For example, soft accuracy is interpreted as the proportion of correctly predicted judgment distributions. To the best of our knowledge, we are the first to propose utilizing soft metrics for HLV in NLP.

As these soft metrics are differentiable, we next explore the use of each soft metric as the training objective. We perform an extensive study testing 14 training methods and 6 evaluation metrics on 6 HLV datasets across 2 pretrained models of different sizes to investigate which training methods are best for HLV data across the different metrics. This study is then followed by an empirical meta-evaluation of the evaluation metrics to understand which evaluation metrics are best for HLV data. To the best of our knowledge, we are the first to perform such an empirical meta-evaluation of HLV evaluation metrics.

To summarize, we: (a) propose soft metrics to measure model performance in the context of HLV, taking inspiration from remote sensing research; (b) propose new training methods based on the soft evaluation metrics; and (c) conduct an extensive study covering 6 HLV datasets, 14 training methods, 2 pretrained models, and 6 evaluation metrics to understand the best training methods and evaluation metrics.

We find that two of the simplest methods for HLV training perform surprisingly well, outperforming more complex methods including training with each soft metric as the objective and attaining the best performance in most cases. The first method considers each annotation of an instance as a separate instance–label pair. Training is then performed on this disaggregated data where the same instance may occur multiple times with a different label each. The second method trains the model to predict the full label distribution induced by the instance's annotations using the standard cross-entropy loss. This is in contrast to the standard model training where the prediction target is the mode of the label distribution.

For evaluation metrics, we find that an existing metric based on Jensen-Shannon divergence (JSD) and our proposed soft micro F_1 score are two of the best HLV metrics. Our analysis shows that in single-label classification, the two metrics are highly correlated but we prove that soft accuracy (because micro F_1 is equal to accuracy in this case) is upper-bounded by the JSD-based metric. As a result, we show that the JSD-based metric can give a score close to its maximum value to wrong predictions, which is potentially misleading, unlike our proposed soft accuracy. We thus recommend future work in HLV to report both metrics but focus on our proposed soft metrics when interpretability is important.

2. Related Work

Training Methods for Human Label Variation. First introduced by Plank (2022), the term **human label variation** (HLV) captures the fact that disagreements in annotations can be well-founded and thus signal for data-driven methods. It challenges the traditional notion in machine learning that an instance has a single ground truth. Training methods that accommodate such variation had been proposed before the term was coined (Sheng, Provost, and Ipeirotis 2008; Peterson et al. 2019; Uma et al. 2020, inter alia). Newer methods have also been proposed in the literature (Deng et al. 2023; Lee, An, and Thorne 2023; Chen et al. 2024; Rodríguez-Barroso et al. 2024) and in the Learning with Disagreement shared task (Leonardelli et al. 2023), which provides a benchmark for HLV. While a systematic investigation of some of these methods exists (Uma et al. 2020), it uses smaller pretrained language models and covers only binary or multiclass classification tasks. In contrast, our work additionally employs a large language model and covers multilabel tasks. Furthermore, we also propose new soft evaluation metrics for HLV.

Evaluation of Soft Classification in Remote Sensing. While evaluating in the HLV context is a relatively new concept in NLP, evaluating against a fuzzy[2] reference is a well-studied area in remote sensing research. Early work by Foody (1996) used a measure of distance that is equivalent to twice the JSD (Lin 1991). Binaghi et al. (1999) argued that entropy-based measures are sensible only if the reference is crisp, and proposed the use of fuzzy set theory to compute the soft version of standard evaluation metrics such as accuracy. A similar approach was also proposed by Harju and Mesaros (2023) for audio data. Key to this approach is the minimum function used to compute the class membership scores given a fuzzy reference and a fuzzy model output. Other proposed functions include product (Lewis and Brown 2001), sum (Pontius and Connors 2006), and a composite of such functions designed to ensure diagonality of the confusion matrix when the reference and the model output match perfectly (Pontius and Cheuk 2006; Silván-Cárdenas and Wang 2008). Approaches that take the cost of misclassifications into account have also been proposed (Gómez, Biging, and Montero 2008). In this work, we take the fuzzy set approach by Binaghi et al. (1999) and apply it to text classification tasks.

Meta-evaluation of HLV Evaluation Metrics. There is little existing work on the meta-evaluation of HLV metrics. Most studies used information theory-based measures such as cross-entropy (Uma et al. 2020; Leonardelli et al. 2023, inter alia), presumably because of the standard probabilistic approach in modern NLP. Rizzi et al. (2024) proposed some theoretical properties that an ideal soft metric should satisfy and performed a theoretical meta-evaluation of existing HLV metrics for single-label classification. In contrast, we propose new soft metrics for both single- and multilabel classification and perform an *empirical* meta-evaluation of these metrics.

3. Evaluation Metrics for HLV

We now discuss the evaluation metrics relevant for HLV data, first on multiclass metrics (Section 3.1), then multilabel metrics (Section 3.2), and finally the relationship between

2 The terms "fuzzy" and "crisp" are sometimes used in remote sensing literature to mean "soft" and "hard" in NLP literature.

Table 1
HLV metrics for classification tasks.

	Multiclass	Multilabel
Existing:	Accuracy Macro F_1 PO-JSD Entropy correlation	Micro F_1 Macro F_1
Proposed:	Soft accuracy Soft macro F_1	Soft micro F_1 Soft macro F_1 Multilabel PO-JSD Multilabel entropy correlation

these metrics (Section 3.3). In these discussions, we will also introduce our use of soft metrics and explain how they relate to their hard variant counterparts. Table 1 provides a summary of the existing and proposed/new evaluation metrics.

Notation. Let $P_{ik} \geq 0$ denote the human judgment for example i and class k, i.e., proportion of humans labeling example i with class k. Let $Q_{ik} \geq 0$ denote the value of P_{ik} predicted by a model. An evaluation metric m is a function such that the scalar $m(\mathbf{P}, \mathbf{Q})$ measures how well (i.e., a positive orientation) $\mathbf{Q}$ captures the judgment distribution in $\mathbf{P}$. To complete our notation, let N and K denote the number of examples and classes, respectively.

3.1 Multiclass Metrics

In multiclass classification tasks, an example can only be assigned to one class. This is reflected by the unity constraints $\sum_k P_{ik} = 1$ and $\sum_k Q_{ik} = 1$ for all i. We study a total of 6 multiclass evaluation metrics m. The first two are standard hard metrics, the next two are their soft versions that we propose inspired by prior work (Binaghi et al. 1999; Harju and Mesaros 2023), and the last two are existing HLV metrics.

Let I^P_{ik} and I^Q_{ik} denote $\mathbb{1}(\arg\max_l P_{il} = k)$ and $\mathbb{1}(\arg\max_l Q_{il} = k)$, respectively. The 6 metric definitions under our notation are as follows:

1. **(Hard) accuracy**, where:

$$m(\mathbf{P}, \mathbf{Q}) \stackrel{\text{def}}{=} \frac{1}{N} \sum_{ik} I^P_{ik} I^Q_{ik}.$$

This is the standard evaluation metric for classification tasks.

2. **(Hard) macro F_1**, where

$$m(\mathbf{P}, \mathbf{Q}) \stackrel{\text{def}}{=} \frac{1}{K} \sum_k \frac{2 \sum_i I^P_{ik} I^Q_{ik}}{\sum_i \left(I^P_{ik} + I^Q_{ik} \right)}.$$

This metric addresses the issue with accuracy that is biased toward the majority class when class proportions are imbalanced.

3. **Soft accuracy**, where:

$$m(\mathbf{P}, \mathbf{Q}) \stackrel{\text{def}}{=} \frac{1}{N} \sum_{ik} \min(P_{ik}, Q_{ik}).$$

This metric is novel to this work, for evaluation in the context of HLV. This is a special case of the soft micro F_1 score proposed later in Section 3.2.

4. **Soft macro F_1**, where:

$$m(\mathbf{P}, \mathbf{Q}) \stackrel{\text{def}}{=} \frac{1}{K} \sum_{k} \frac{2 \sum_i \min(P_{ik}, Q_{ik})}{\sum_i (P_{ik} + Q_{ik})}.$$

This metric is also novel to this work, to address the issue of class imbalance with soft accuracy and other existing HLV metrics.[3] Note that this metric implies the existence of soft *class-wise* metrics (see Appendix).

5. **Jensen-Shannon divergence** (Uma et al. 2021), which we modify into its positively oriented version (**PO-JSD** for short) where:

$$m(\mathbf{P}, \mathbf{Q}) \stackrel{\text{def}}{=} 1 - \frac{1}{N} \sum_{i} \text{JSD}(\mathbf{p}_i, \mathbf{q}_i).$$

Vectors $\mathbf{p}_i, \mathbf{q}_i$ denote the i-th row of $\mathbf{P}, \mathbf{Q}$, respectively. The scalar $\text{JSD}(\mathbf{a}, \mathbf{b}) = \frac{1}{2}\left(\text{KL}(\mathbf{a}, \frac{1}{2}(\mathbf{a} + \mathbf{b})) + \text{KL}(\mathbf{b}, \frac{1}{2}(\mathbf{a} + \mathbf{b}))\right)$ is the JSD (Lin 1991), where $\text{KL}(\mathbf{a}, \mathbf{b}) = \sum_k a_k \log_2 \frac{a_k}{b_k}$ is the Kullback-Leibler divergence.[4]

6. **Entropy correlation** (Uma et al. 2020), where:

$$m(\mathbf{P}, \mathbf{Q}) \stackrel{\text{def}}{=} \frac{\sum_i \zeta_i^P \zeta_i^Q}{\sqrt{\sum_i {\zeta_i^P}^2} \sqrt{\sum_i {\zeta_i^Q}^2}},$$
$$\zeta_i^P = \eta_i^P - \bar{\eta}^P, \text{ and}$$
$$\zeta_i^Q = \eta_i^Q - \bar{\eta}^Q.$$

Scalar $\eta_i^P = -\sum_k \frac{P_{ik} \log P_{ik}}{\log K}$ is the normalized entropy of row i of $\mathbf{P}$, and $\bar{\eta}^P = \frac{1}{N} \sum_i \eta_i^P$ is the mean of $\{\eta_i^P\}_{i=1}^N$. Scalars η_i^Q and $\bar{\eta}^Q$ are defined analogously.

3 In the context of HLV, we understand the concept of class imbalance analogously to that in the context where class assignments are hard: There exists a class k such that $\sum_i P_{ik} >> \sum_i P_{il}$ for all $l \neq k$.

4 Jensen-Shannon divergence of two distributions has an upper bound of $\log_b 2$ if the logarithms used in $\text{KL}(\cdot)$ are of base b. Normalizing this bound to 1 results in logarithms of base 2 as $\log_b x / \log_b 2 = \log_2 x$.

Interpretation of Soft Accuracy. By noting that $N = \sum_{ik} P_{ik}$, soft accuracy can be interpreted as the proportion of the judgment distribution that is correctly predicted. For example, if soft accuracy is 70%, then the model predicts 70% of judgment distribution correctly. This interpretation is similar to the hard counterpart (i.e., proportion of examples that are correctly predicted) and arguably more intuitive than that of existing metrics because it is stated directly in terms of the problem of predicting human judgment distribution.

Besides interpretation, soft accuracy's similarities to the hard counterpart include symmetry (due to the symmetry of min) and boundedness (between 0 and 1), where the lower and the upper bounds are achieved if and only if $P_{ik}Q_{ik} = 0$ and $P_{ik} = Q_{ik}$, respectively, for all i, k. Note also that *soft accuracy is reduced to the standard hard accuracy* when all rows of both $\mathbf{P}$ and $\mathbf{Q}$ are one-hot vectors (i.e., no label variation).

While functions other than min have been used to compute soft accuracy (see Section 2), they lead to the soft accuracy being <1 in the case where $\mathbf{P} = \mathbf{Q}$. This is problematic because intuitively, good HLV evaluation metrics should produce the highest score when the predicted judgment distribution perfectly matches the true counterpart. Therefore, we use the min function in this work.

Other evaluation metrics we considered included entropy similarity (Uma et al. 2021), (negative) cross-entropy (Peterson et al. 2019; Pavlick and Kwiatkowski 2019), (negative) Kullback-Leibler divergence (Nie, Zhou, and Bansal 2020), (the positively oriented version of) information closeness (Foody 1996), and (the positively oriented version of) Jensen-Shannon distance (Nie, Zhou, and Bansal 2020). However, we ultimately excluded them from this study because of their expected high correlation with either PO-JSD or entropy correlation.

3.2 Multilabel Metrics

Next, we consider the case where the unity constraints do not hold, i.e., an example can be (soft) assigned to multiple classes, forming a *multilabel* classification task.

By considering this case as multiple independent binary classification tasks, some existing metrics can be extended straightforwardly by taking the average of the metric values over classes. For example, PO-JSD can be modified into the multilabel version:

$$m(\mathbf{P}, \mathbf{Q}) \stackrel{\text{def}}{=} 1 - \frac{1}{NK} \sum\nolimits_{ik} \text{JSD}(P^{\text{bin}}_{ik}, Q^{\text{bin}}_{ik})$$

where $P^{\text{bin}}_{ik} = [P_{ik}, 1 - P_{ik}]$ is a judgment distribution over two classes, and Q^{bin}_{ik} is defined analogously. Entropy correlation can also be modified similarly into the multilabel version:

$$m(\mathbf{P}, \mathbf{Q}) \stackrel{\text{def}}{=} \frac{1}{K} \sum_{k} \frac{\sum_i \zeta^P_{ik} \zeta^Q_{ik}}{\sqrt{\sum_i {\zeta^P_{ik}}^2} \sqrt{\sum_i {\zeta^Q_{ik}}^2}}$$

where $\zeta^P_{ik} = \eta^P_{ik} - \bar{\eta}^P_k$, $\eta^P_{ik} = -\frac{1}{\log 2}\big(P_{ik} \log P_{ik} + (1 - P_{ik}) \log(1 - P_{ik})\big)$, $\bar{\eta}^P_k = \frac{1}{N} \sum_i \eta^P_{ik}$, and scalars ζ^Q_{ik}, η^Q_{ik}, and $\bar{\eta}^Q_k$ are defined analogously.

While it is possible to define (hard) accuracy for this multilabel case, a more common evaluation metric is the micro F_1 score, which in our notation can be expressed as:

$$m(\mathbf{P}, \mathbf{Q}) \stackrel{\text{def}}{=} 2\frac{\sum_{ik} J_{ik}^P J_{ik}^Q}{\sum_{ik}\left(J_{ik}^P + J_{ik}^Q\right)}$$

where $J_{ik}^P = \mathbb{1}(P_{ik} > 0.5)$, and J_{ik}^Q is defined analogously. Similarly, we define the soft version of the micro F_1 score as follows (see derivations in the Appendix):

$$m(\mathbf{P}, \mathbf{Q}) \stackrel{\text{def}}{=} 2\frac{\sum_{ik} \min(P_{ik}, Q_{ik})}{\sum_{ik}(P_{ik} + Q_{ik})}$$

The (hard) micro F_1 score is a special case of this soft counterpart when $P_{ik} \in \{0, 1\}$ and $Q_{ik} \in \{0, 1\}$ (i.e., no label variation). Furthermore, the soft accuracy in Section 3.1 is a special case of this soft micro F_1 score when the unity constraints hold. Therefore, *the soft micro F_1 score is a general evaluation metric for classification tasks*.

3.3 Relationship Between Metrics

To analyze the relationship between metrics, we compute Pearson correlation between each pair of multiclass metrics (excluding the macro ones), similarly to prior work (Chicco and Jurman 2020). For a given number of classes, we sample B matrices $\mathbf{P}$ and $\mathbf{Q}$ and compute the value of each metric on these B pairs of $\mathbf{P}$ and $\mathbf{Q}$. The Pearson correlations between pairs of metrics are then computed based on these B data points. We set $B = 500$ for computational reasons. We report the correlation coefficients in Table 2.

Table 2 shows that the metrics are generally weakly correlated with each other, with the exception of soft accuracy and PO-JSD where the two are consistently strongly correlated ($r > 0.9$). This strong correlation with an established HLV metric gives an empirical assurance that our proposed soft accuracy is a sensible metric. The table also shows that soft accuracy is only moderately correlated with (hard) accuracy in most cases, where this correlation gets weaker when the number of classes is large ($K = 100$), or either $\mathbf{P}$ or $\mathbf{Q}$ is dense ($\alpha = 10$ or $\beta = 10$). This finding suggests that hard and soft accuracy capture different aspects when humans disagree. Furthermore, the table shows that entropy correlation is poorly correlated with all metrics, which suggests that it is an outlier HLV metric.

Soft Accuracy and PO-JSD. Despite their strong positive correlation, soft accuracy and PO-JSD have a notable difference: Soft accuracy is upper-bounded by PO-JSD.

Theorem 1
Consider the same definitions of P_{ik} and Q_{ik} used in Section 3. The soft accuracy and the PO-JSD between $\mathbf{P}$ and $\mathbf{Q}$ satisfy the following inequality:

$$\frac{1}{N}\sum_{ik}\min(P_{ik}, Q_{ik}) \leq 1 - \frac{1}{N}\sum_{i}\text{JSD}(\mathbf{p}_i, \mathbf{q}_i)$$

Table 2
Pearson correlations between a pair of evaluation metrics $m(\mathbf{P}, \mathbf{Q})$ for 1K examples and various number of classes (K) where the rows of $\mathbf{P}$ and $\mathbf{Q}$ are drawn from a symmetric Dirichlet with parameters α and β, respectively. A, J, E, and S denote the accuracy, PO-JSD, entropy correlation, and soft accuracy, respectively.

K	A-J	A-E	A-S	J-E	J-S	E-S
			$\alpha = \beta = 10$			
10	0.272347	0.050255	0.289984	0.035927	0.942263	0.100303
100	0.047588	0.040964	0.051956	0.091122	0.924129	0.087560
			$\alpha = \beta = 0.1$			
10	0.627773	0.063820	0.733743	0.126164	0.963214	0.162816
100	0.194069	0.093113	0.251355	0.014869	0.961086	0.032225
			$\alpha = 10, \beta = 0.1$			
10	0.226175	−0.015882	0.206464	−0.018527	0.975144	−0.002286
100	−0.004257	−0.062204	−0.026900	0.050080	0.957816	0.061640
			$\alpha = 0.1, \beta = 10$			
10	0.211874	0.016467	0.181706	−0.016632	0.975830	−0.019553
100	0.103478	−0.021530	0.057268	−0.005990	0.967420	0.005447

where $\text{JSD}(\mathbf{p}_i, \mathbf{q}_i)$ denote the Jensen-Shannon divergence between the i-th rows of $\mathbf{P}$ and $\mathbf{Q}$.

Proof. Given a scalar $0 \leq \pi \leq 1$ and discrete distributions $\mathbf{a}, \mathbf{b}$, it has been shown that the following bound holds (Lin 1991, Theorem 4):

$$\sum_k \min(\pi a_k, (1-\pi) b_k) \leq \frac{1}{2} \left(H([\pi, 1-\pi]) - \text{JSD}_\pi(\mathbf{a}, \mathbf{b}) \right)$$

where $H(\mathbf{u})$ is the entropy of $\mathbf{u}$, and $\text{JSD}_\pi(\mathbf{a}, \mathbf{b}) = H(\pi\mathbf{a} + (1-\pi)\mathbf{b}) - \pi H(\mathbf{a}) - (1-\pi)H(\mathbf{b})$ is the general form of JSD with two distributions. It can be easily shown that JSD_π is reduced to the standard Jensen-Shannon divergence when $\pi = \frac{1}{2}$. Performing this substitution and noting that $H([\frac{1}{2}, \frac{1}{2}]) = 1$,[5] we have

$$\sum_k \min(\frac{1}{2} P_{ik}, \frac{1}{2} Q_{ik}) \leq \frac{1}{2} \left(1 - \text{JSD}(\mathbf{p}_i, \mathbf{q}_i)\right)$$

for all i. Doubling both sides and taking the average of this inequality over i gives the desired result. □

This fact implies that soft accuracy penalizes a model more heavily than PO-JSD, as illustrated in Figure 1. It shows that a skewed predicted judgment distribution of $(0.2, 0.8)$ has a soft accuracy of 0.7, much less than the PO-JSD (>0.9). The high PO-JSD is potentially misleading as it is very close to the maximum possible value of 1.

5 All logarithms are of base 2.

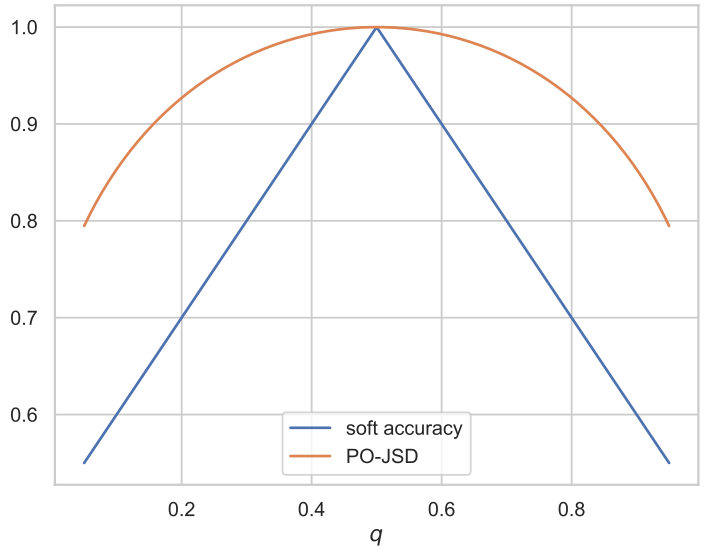

Figure 1
Soft accuracy and PO-JSD graphs for a binary classification problem on a single example where the true and the predicted judgment for the positive class is 0.5 and q, respectively.

4. Training Methods for HLV

We now discuss the training methods. In total, we experiment with 14 HLV training methods, as summarized in Table 3. The first method (ReL) is the simplest and uses standard cross-entropy loss as the training objective based on prior work (Sheng, Provost,

Table 3
Overview of HLV training methods used, along with details of whether annotator identity is required (I), multilabel tasks are supported (M), and training data is enlarged (D), which makes training run much longer.

Name	I	M	D
Repeated labeling (**ReL**)	✗	✓	✓
Majority voting (**MV**)	✗	✓	✗
Annotator ranking (**AR**)	✓	✓	✗
Annotator ranking (hard) (**ARh**)	✓	✓	✗
Ambiguous labeling (**AL**)	✗	✗	✓
Soft labeling (**SL**)	✗	✓	✗
Multitask of **SL** and **MV** (**SLMV**)	✗	✓	✗
Annotator ensemble (**AE**)	✓	✓	✗
Annotator ensemble (hard) (**AEh**)	✓	✓	✗
Below are proposed in this work			
Jensen-Shannon divergence (**JSD**)	✗	✓	✗
Soft micro F_1 (**SmF1**)	✗	✓	✗
Soft macro F_1 (**SMF1**)	✗	✓	✗
Loss aggregation with min (**LA-min**)	✗	✓	✗
Loss aggregation with max (**LA-max**)	✗	✓	✗

and Ipeirotis 2008). The next 8 methods (*MV*, *AR*, *ARh*, *AL*, *SL*, *SLMV*, *AE*, *AEh*) are from the Learning with Disagreements shared task (Leonardelli et al. 2023) which also use standard cross-entropy loss. The next 3 methods (*JSD*, *SmF1*, *SMF1*) use soft metrics that are differentiable as the training objective. The last 2 methods (*LA-min*, *LA-max*) aggregate the cross-entropy losses of all annotations for a given input. In detail:

1. *ReL* (repeated labeling) trains on the disaggregated labels directly (Sheng, Provost, and Ipeirotis 2008). For example, if the training data only has a single instance x with annotations y_1 and y_2, then *ReL* constructs new training data $\{(x, y_1), (x, y_2)\}$.
2. *MV* trains on the majority-voted labels. This is the baseline model given by the shared task organizers.
3. *AR* (annotator ranking) and *ARh* (annotator ranking hard) weight each training instance equal to the sum of ranking scores of its annotators (Cui 2023). Suppose that training instance x_1 is labeled as A, A, B, and A by 4 annotators; and x_2 is labeled as A, B, and B by only the first 3 annotators. Then, to compute the ranking score of an annotator:

   (a) *AR* computes the average judgment of the majority-voted label given by the annotator. The annotator ranking scores are (in order) $\frac{3}{4}$, $\frac{1}{2}\left(\frac{3}{4}+\frac{2}{3}\right)$, $\frac{2}{3}$, and $\frac{3}{4}$.

   (b) *ARh* computes the average number of times the annotator agrees with the majority. The scores are $\frac{1}{2}$, $\frac{2}{2}$, $\frac{1}{2}$, and $\frac{1}{1}$.

   For both methods, the ranking score is zero if the annotator never agrees with the majority. Given an instance, the training objective is maximizing the probability of its majority-voted label(s), scaled by its weight as defined above.
4. *AL* (ambiguous labeling) labels *ambiguous* instances, i.e., instances where annotators are not unanimous in their judgment, with all selected classes exactly once (Gajewska 2023). Each instance-label pair is included in the training set. For example, if training instance x_1 is labeled as y_1, y_1, and y_2 by 3 annotators, and x_2 is labeled as y_2, y_2, and y_2, then *AL* produces $\{(x_1, y_1), (x_1, y_2), (x_2, y_2)\}$ as the training data. Because it was developed for single-label tasks, we exclude this method from multilabel tasks.[6]
5. *SL* (soft labeling) trains on soft labels as targets (Maity et al. 2023; Wan and Badillo-Urquiola 2023). For example, if an instance is labeled as positive by 80% of annotators then the log-likelihood is $0.8 \log p(1) + 0.2 \log p(0)$, where $p(1)$ and $p(0)$ denote the predicted probability of the positive and the negative class, respectively.
6. *SLMV* performs multi-task learning using both the soft (*SL*) and the majority-voted (*MV*) labels as targets (Grötzinger, Heuschkel, and Drews 2023). Two models with a shared encoder are trained to predict the two

6 Extending this method to multilabel tasks is beyond the scope of this work.

types of labels respectively, which is similar to the method of Fornaciari et al. (2021). For example, if an instance is labeled as positive by 80% of annotators then the multitask log-likelihood is $0.8 \log p_{\theta_1}(1) + 0.2 \log p_{\theta_1}(0) + \log p_{\theta_2}(1)$, where θ_1 and θ_2 denote the parameters of the first and the second model, respectively. We predict only the soft labels for evaluation.

7. *AE* (annotator ensembling) and *AEh* (annotator ensembling hard) model each annotator separately followed by ensembling (Sullivan, Yasin, and Jacobs 2023; Vitsakis et al. 2023). In this approach, we have as many models as there are annotators. Each model is trained on the labels given by a distinct annotator. In practice, we share the encoder parameters across models. Suppose that there are 3 annotators, and the predicted judgments of the positive class for an instance are 60%, 30%, and 90%. To ensemble these predictions at test time:

    (a) *AE* computes a simple mean (Sullivan, Yasin, and Jacobs 2023). The final predicted judgment for the positive class is $\frac{1}{3}(0.6 + 0.3 + 0.9)$.

    (b) *AEh* computes the distribution of most likely labels (Vitsakis et al. 2023). The final predicted judgment for the positive class is $\frac{2}{3}$ because 2 out of 3 judgments are over 50%.

8. *JSD*, *SmF1* (soft micro F_1), and *SMF1* (soft macro F_1) train using JSD, $1-$ soft accuracy for multiclass or soft micro F_1 for multilabel tasks, and $1-$ soft macro F_1 as the objective respectively (see Section 3 for definitions).[7] These approaches leverage the differentiability of soft metrics, resulting in a single objective for both training and inference. By directly optimizing the evaluation metric at training time, we expect the models to exhibit superior performance, especially when evaluated with the corresponding metric.

9. *LA-min* (loss aggregation min) and *LA-max* (loss aggregation max) aggregate the losses of all annotations for a given training instance using the min and the max functions, respectively. For example, given a training instance x with annotations y_1 and y_2, we compute the losses $l_1 = \mathcal{L}(x, y_1)$ and $l_2 = \mathcal{L}(x, y_2)$ where $\mathcal{L}$ is the cross-entropy loss function. Then, we update model parameters based on the gradient of $g(l_1, l_2)$ where g is either the min or the max function. Using the min (resp., max) function means selecting the least (resp., most) "surprising" annotation for the model. Using the mean aggregation is mathematically equivalent to *SL* and thus excluded.

7 We exclude entropy correlation because it doesn't have a unique maximizer.

Table 4
Datasets used in the evaluation of HLV training methods, along with the language (L), type of annotators (O) where E and C mean Experts and Crowds, respectively, number of examples to the nearest thousand (N), average number of annotations per example (J), number of classes (K), whether annotator identity is present (I), and whether the task is multilabel classification (M).

Name	Task	L	O	N	J	K	I	M
HS-Brexit (Akhtar, Basile, and Patti 2021)	Hate speech	EN	E	1	6.0	2	✓	✗
MD-Agreement (Leonardelli et al. 2021)	Offensiveness	EN	C	10	5.0	2	✓	✗
ArMIS (Almanea and Poesio 2022)	Misogyny	AR	E	1	3.0	2	✓	✗
ChaosNLI (Nie, Zhou, and Bansal 2020)	NLI	EN	C	3	100.0	3	✗	✗
MFRC (Trager et al. 2022)	Moral sentiment	EN	E	18	3.4	8	✓	✓
TAG (private dataset)	Legal area	EN	E	11	5.5	33	✓	✓

5. Evaluation of HLV Training Methods

We now conduct an extensive study covering 6 datasets, 14 training methods, and 6 evaluation metrics across 2 pretrained models to understand the performance landscape of HLV data. Table 4 shows an overview of the 6 datasets. These datasets are selected to represent diverse factors such as language, type of annotators, number of annotations for each example, number of classes, presence of annotator identity, and type of classification tasks (i.e., single- vs multilabel). By using these diverse datasets, consistent patterns that emerge across datasets are more likely to hold true in general rather than just in a specific type of dataset.

5.1 Experimental Set-up

Public Datasets. We use 5 public datasets with disaggregated annotations that have been used in previous HLV studies, four of which are in English and one in Arabic. For HS-Brexit, MD-Agreement, and ArMIS, we use the same train–test splits as Leonardelli et al. (2023). For both ChaosNLI and MFRC, we use 10-fold cross-validation with 10% of the training portion used as a development set. For ChaosNLI, we include only the SNLI and MNLI subsets (each has 1.5K examples) because of their standard NLI setup of one premise and one hypothesis.

"Text Annotation Game" (TAG) Dataset. We also include a private dataset called TAG, which consists of English texts describing legal problems written by non-expert legal help seekers. Because it is based on real-world confidential legal requests from help-seekers, we are unable to distribute the dataset. However, we still include the TAG dataset due to its importance for the meta-evaluation later in Section 6. We have access to this dataset because of our collaboration with Justice Connect,[8] an Australian public benevolent institution[9] providing free legal assistance to laypeople facing legal

8 https://justiceconnect.org.au.
9 As defined by the Australian government: https://www.acnc.gov.au/charity/charities/4a24f21a-38af-e811-a95e-000d3ad24c60/profile.

problems. Each text in the dataset was annotated by an average of 5.5 practicing lawyers, who each selected one or more out of the 33 areas of law[10] that applied to the problem. Thus, it is a multilabel classification task. Example areas of law include *Neighborhood disputes* and *Housing and residential tenancies*. Average inter-annotator agreement (α) over the areas of law is 0.454, which is modest.[11] This is because lawyers often have different interpretations of the same problem due to their different legal specializations and years of experience. However, these different interpretations are all valid as human label variation because these registered lawyers are subject matter experts who play a crucial part in interpreting the law. The dataset has a total of 11K texts collected between July 2020 and early December 2023. We randomly split the dataset 8:1:1 for the training, development, and test sets, respectively.

Models. We experiment with both small and large language models. Because HS-Brexit, MD-Agreement, and ArMIS use Twitter as source data, we use TwHIN-BERT (Zhang et al. 2023), which was pretrained on Twitter data and supports both English and Arabic. For other datasets, we use RoBERTa (Liu et al. 2019). In addition, we experiment with 8B LLaMA 3 (Grattafiori et al. 2024) for the English datasets.[12] We replace the output layer of these pretrained models with a linear layer with K output units. We use the base versions of both RoBERTa and TwHIN-BERT (with roughly 100M parameters each).

Training. We implement all methods using FlairNLP (Akbik et al. 2019). For RoBERTa and TwHIN-BERT models, we tune the learning rate and the batch size using random search. Because there are multiple evaluation metrics that aren't necessarily comparable, we select the best hyperparameters based on the geometric mean of their values[13] to avoid biasing towards one metric. We exclude both hard and soft macro F_1 scores from this mean as most datasets are balanced. For LLaMA, we use FlairNLP's default hyperparameters[14] for computational reasons. We use LoRA (Hu et al. 2022) to finetune LLaMA efficiently. All models are finetuned for 10 epochs. We truncate inputs longer than 512 tokens in the TAG dataset, affecting only 4% of instances. For ChaosNLI and MFRC, we train the final model used for evaluation of each fold on the concatenation of the training and the development sets.

Evaluation. When cross-validation is used, we evaluate the concatenated test predictions. Otherwise, we evaluate on the test set. We perform the evaluation across 3 training runs.

5.2 Results

We compute the difference between the performance of a single run of an HLV training method and the mean performance of *MV*. We then report the mean differences across runs. We present the results on ArMIS in Figure 2 (the only non-English dataset),

10 Including a special label "Not a legal issue".
11 We compute this agreement on random 10% selection of the data due to the high computational cost.
12 While LLaMA 3 was trained on multilingual data, its intended use is English only.
13 We transform entropy correlation to a value between 0 and 1 before taking the mean.
14 Learning rate and batch size are 5×10^{-5} and 32, respectively.

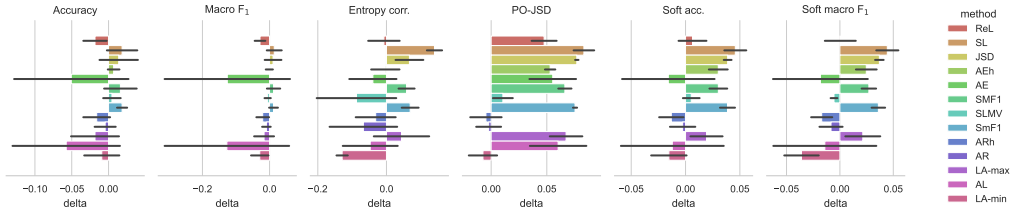

Figure 2
Performance difference (delta) with mean *MV* performance of TwHIN-BERT on ArMIS. The methods are sorted by their mean ranking across datasets, models, and metrics.

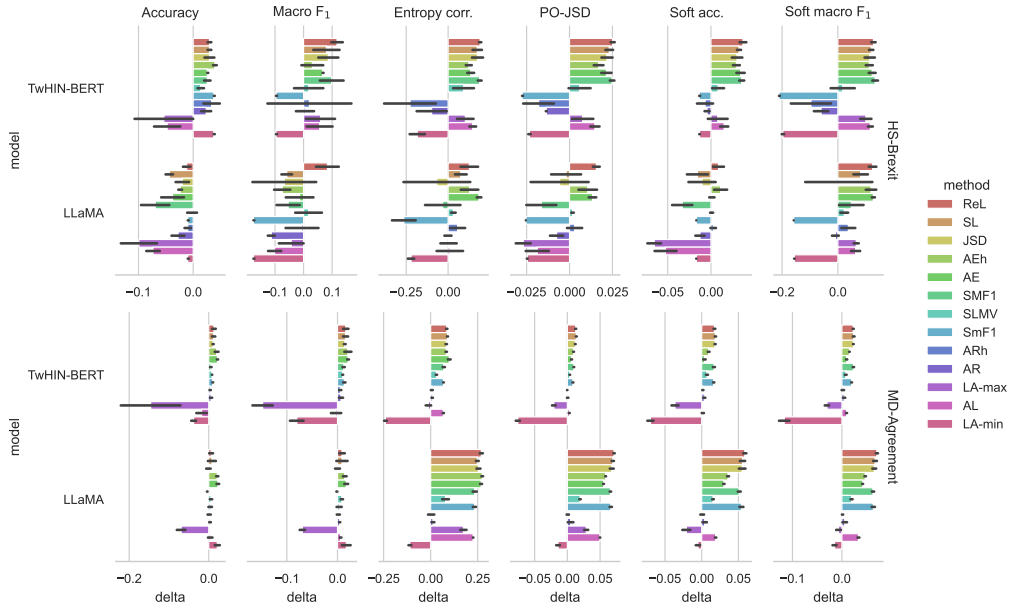

Figure 3
Performance difference (delta) with mean *MV* performance on HS-Brexit and MD-Agreement. *SmF1* with TwHIN-BERT predicts zero for all test instances on HS-Brexit so its entropy correlation is undefined. The methods are sorted by their mean ranking across datasets, models, and metrics.

HS-Brexit and MD-Agreement in Figure 3 (Learning with Disagreements shared task datasets), and the remaining datasets in Figure 4 (multiclass and multilabel datasets). We report performance of *MV* in Table B.1 in the Appendix. We make the following observations.

First, HLV training methods have different effectiveness compared with *MV*, with some methods performing very poorly. For example, on ArMIS and both portions of ChaosNLI, *LA-min* consistently performs worse than *MV*. This finding suggests the importance of choosing the right training method when embracing HLV.

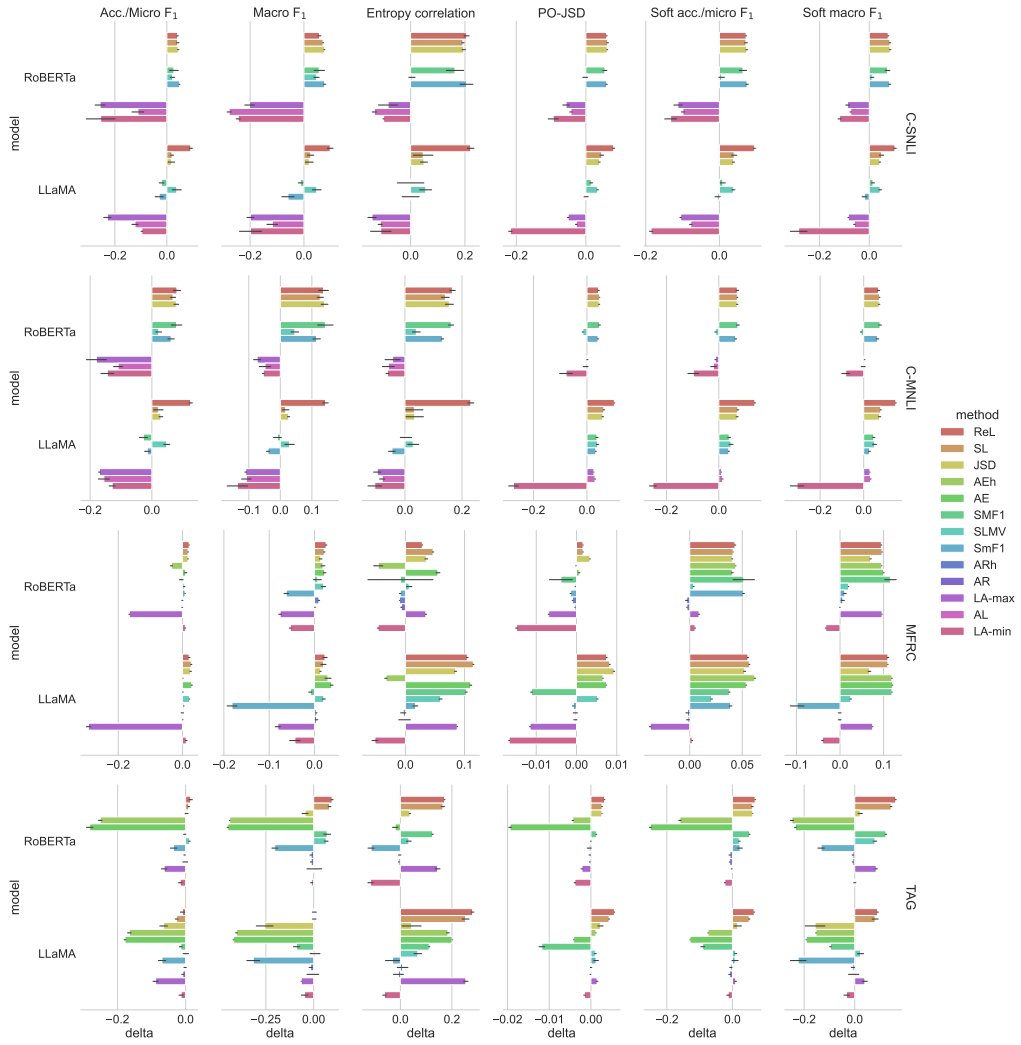

Figure 4
Performance difference (delta) with mean *MV* performance on ChaosNLI (both the SNLI and the MNLI portions), MFRC, and TAG datasets. ChaosNLI doesn't have annotator identity information, so *AE*, *AEh*, *AR*, *ARh* are inapplicable. *AL* is omitted from MFRC and TAG because it is developed for single-label rather than multilabel tasks. *AEh* with RoBERTa predicts zero for all test instances on several classes in TAG, so its entropy correlation is undefined. The methods are sorted by their mean ranking across datasets, models, and metrics.

Second, both *ReL* and *SL* substantially outperform *MV* in 68 out of 78 settings (87.2% of all settings),[15] where a setting is a 3-tuple of a dataset, a model, and an evaluation metric.[16] Furthermore, *ReL* or *SL* is among the top-performing methods in 89.7% of all

15 Negative cases for *ReL*: all metrics except PO-JSD on ArMIS (5 settings), LLaMA on HS-Brexit and MD-Agreement with accuracy (2), LLaMA on MD-Agreement with macro F_1 score (1), LLaMA on TAG with hard metrics (2). Negative cases for *SL*: TwHIN-BERT on ArMIS with hard metrics (2), LLaMA on HS-Brexit with all metrics except entropy correlation and soft macro F_1 score (4), LLaMA on MD-Agreement and TAG with hard metrics (4).

16 We consider the SNLI and the MNLI portions of ChaosNLI as separate datasets for this purpose.

settings,[17] including when compared to *JSD* and *SmF1* evaluated using PO-JSD and soft accuracy/micro F_1, respectively. Concretely, *ReL* and *SL* are on par with (or outperform in some cases with LLaMA) the training method whose learning objective is the same as the evaluation metric in 23 out of 26 settings (88.5%),[18] demonstrating the strength of *ReL* and *SL*.

Third, *ReL* and *SL* are generally competitive with each other. However, there are two exceptions to this trend. First, on ArMIS, *SL* outperforms *ReL*. Second, on both HS-Brexit and ChaosNLI, *ReL* outperforms *SL* with LLaMA. For the former finding, however, ArMIS is an outlier dataset: It is the only Arabic dataset, has the smallest number of examples and annotations per example, and all methods generally have higher variance on this dataset. The latter finding suggests that *ReL* is better suited to large language models than *SL*.

Fourth, *SMF1* and *JSD* generally also perform well, although they lag behind *ReL* and *SL*. Because both *SMF1* and *JSD* use a soft evaluation metric as the training objective, this finding underlines the value in using differentiable metrics for HLV model training.

We also observe that *SmF1* with TwHIN-BERT results in undefined entropy correlation on HS-Brexit. Looking closely, we find that the method degenerates to predicting $Q_{ik} = 0$ for all i. We observe the same degenerate results for *AEh* with RoBERTa on the TAG dataset. Moreover, *SmF1* with RoBERTa attains much lower soft micro F_1 score than the top-performing *ReL* on TAG, even though *SmF1* optimizes the evaluation metric directly during training. Our investigation suggests that these unexpected observations are due to the class imbalance present in both HS-Brexit and TAG. When the training objective is fair (e.g., *SMF1*), the problem disappears.

5.3 Discussion

ReL and *SL* appear to perform best as a training method, and that is somewhat surprising because they are the most straightforward methods to incorporate HLV in model training. That said, our finding that *SL* is a strong performer is in line with what Uma et al. (2021) discovered. The results of *ReL*, however, stand in contrast with prior work (Uma et al. 2021; Kurniawan et al. 2024). This difference may be due to the structured prediction tasks that the prior work considered which have an exponentially large output space. We note that Uma et al. (2021) also found that *ReL* performs really well in simple, multiclass classification tasks, in line with our findings.

Moreover, the results suggest that *ReL* is likely to outperform *SL* when large language models are used. A possible explanation is that with *SL*, the model observes the complete judgment distribution of an instance as target. Thus, there is not much flexibility as the model has to predict that distribution to minimize the training loss. In contrast, with *ReL*, the model only observes one annotation of the instance as target. Therefore, it has more flexibility in how it distributes the judgment distribution mass across all the labels of that instance over training steps. We hypothesize that large language models have sufficient learning capacity to benefit from this flexibility while smaller models do not. This explanation is consistent with our findings on ChaosNLI

17 Negative cases: TwHIN-BERT on HS-Brexit with accuracy (1 setting); RoBERTa on MFRC with entropy correlation, PO-JSD, and soft macro F_1 score (3); LLaMA on MFRC with macro F_1 score, PO-JSD, and both micro and macro soft F_1 scores (4).

18 We consider only PO-JSD and soft accuracy/micro F_1 evaluation metrics to get the total of 26 settings. Negative cases: RoBERTa on MFRC (2 settings), LLaMA on MFRC with PO-JSD (1).

that *ReL* has much larger gains compared to *SL*. ChaosNLI has 100 annotations per instance, the largest number out of all datasets. We leave further investigation of this hypothesis to future work.

Despite its strengths, *ReL* has substantial computational cost: because it keeps the annotations disaggregated, the size of the training data can be enormous. For example, in the ChaosNLI dataset, *ReL* makes the training data 100 times larger because there are exactly 100 annotations for each training instance. That *ReL* performs better with LLaMA than RoBERTa on ChaosNLI exacerbates the computational inefficiency of *ReL*: both the data and the model must be large for it to perform optimally. Mitigating this inefficiency can be an interesting avenue for future work.

Because the training data is much larger, one may think that the superiority of *ReL* over *SL* is due to its much longer training. However, this is not the case. We find that increasing the number of training iterations of *SL* to match that of *ReL* does not improve performance.

6. Meta-Evaluation of HLV Evaluation Metrics

In the previous section, we saw the best training methods that perform consistently well across metrics. In this section, we address the next question: What is the best evaluation metric for HLV data? We answer this by conducting a meta-evaluation on the evaluation metrics. We focus on the TAG dataset specifically for this meta-evaluation experiment as we have access to practicing lawyers as annotators who can provide high quality annotations. It does mean we are unable to distribute the dataset for ethical reasons, but we believe the insights derived from the meta-evaluation is worth the trade-off.

A good evaluation metric should produce a ranking of training methods (henceforth "metric ranking") that correlates to the ranking produced by human judgments (henceforth "human ranking"). To create the latter (human ranking), we frame it as a pairwise comparison task where we pit two methods against each other and ask lawyers to select one of them. Given a large number of human-judged pairwise comparisons, we then use an algorithm to create a ranking for the training methods. Intuitively, the algorithm will rank a method that "wins" consistently higher than another method that "loses" most of the time. Once we have the human ranking, we can then compute its correlation with each metric ranking, and the best metrics are the ones with the highest correlations.

6.1 Experimental Set-up

To create the ranking of training methods produced by human judgments, given a language model (RoBERTa or LLaMA) we randomly sample a pair of methods (e.g., *ReL* and *SL*) and present the instance text (i.e., a legal problem) and two results (i.e., distribution of the areas of law)[19] produced by the two methods, and then ask the lawyers which result is more accurate. The lawyers can choose either one of them, both, or neither of them. See Appendix Figure C.2 for an illustration of the annotation task. On average, for a given model (RoBERTa or LLaMA) each pair of methods has 8 judgments made by 4.4 lawyers across 3–4 randomly sampled instance texts. This gives a total of

19 Note that because it is multilabel task, the distribution doesn't sum to one. We only show areas of law whose probability exceeds a threshold of 0.1. We choose this threshold because it results in similar numbers of areas of law whose probability exceeds the value across all methods.

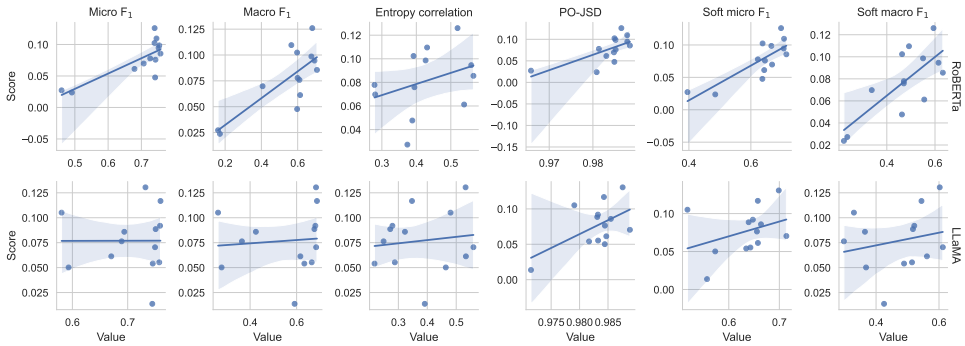

Figure 5
Relationships between method performance as given by the evaluation metrics (Value) and method scores produced by human judgments (Score) on the TAG dataset. The shaded area denotes a 95% confidence interval. Each data point corresponds to a specific training method (e.g., *ReL*). *AEh* is excluded from RoBERTa with entropy correlation because of its degenerate results.

1.2K pairwise judgments[20] across 445 instances, 18 lawyers, and 2 models (RoBERTa and LLaMA). Each pair of instance text and two results is annotated by an average of 2.3 lawyers.

To make sure the lawyers form their own expectation on the areas of law relevant to an instance before making this pairwise selection, we first show the instance text and ask them to select the most relevant areas of law (Figure C.1 in the Appendix). After completing this task for an instance text, we then ask them to make the pairwise selection. The two tasks must be completed for an instance text before the lawyers can move to the next instance text. We conduct several pilot studies with the lawyers to refine the interface design and ensure the task is clear to them.

At the end of the annotation task, we have a collection of human-judged pairwise comparisons for the training methods. We next use the rank centrality algorithm (Negahban, Oh, and Shah 2012) to compute the "score" for each method to ultimately produce a ranking of the methods. This algorithm casts the problem as a random walk on a weighted directed graph where each node corresponds to a method and the weight of the edge from method i to method j is the probability that method j wins over model i.[21] The score of a method is then its stationary probability for this random walk. Intuitively, the score is high if it wins against other high-scoring methods or against many low-scoring methods. To assess the evaluation metrics and understand which metrics are the best, for each metric we compute the Pearson correlation between: (1) method performance as given by the evaluation metric; and (2) the method scores produced by human judgments.

6.2 Results

We draw the regression plots between method performance as given by the evaluation metrics and method scores produced by human judgments in Figure 5. The figure

20 8 judgments $\times \left(\frac{1}{2} \times 13 \times (13-1)\right)$ pairs of methods $\times$ 2 models $=$ 1,248 judgments.
21 We determine that method j wins over method i if the annotator select method j but not method i.

Table 5
Correlation coefficients between method performance as given by the evaluation metrics and method scores produced by human judgments on the TAG dataset along with their p-values computed with a permutation test. Asterisks denote statistical significance ($p < .05$).

Model	Type	Metric	Pearson r	p-value
RoBERTa	Hard	Micro F_1	0.787*	0.001
		Macro F_1	0.774*	0.005
	Soft	Entropy correlation	0.339	0.281
		PO-JSD	0.674*	0.005
		Soft micro F_1	0.804*	0.001
		Soft macro F_1	0.753*	0.006
LLaMA	Hard	Micro F_1	0.002	0.984
		Macro F_1	0.087	0.776
	Soft	Entropy correlation	0.125	0.689
		PO-JSD	0.539	0.070
		Soft micro F_1	0.343	0.249
		Soft macro F_1	0.213	0.488

shows that not all metrics exhibit a strong linear relationship with human judgments. For example, both hard F_1 scores and entropy correlation show almost no relationship for LLaMA. All metrics except entropy correlation show a strong linear relationship for RoBERTa. This suggests that entropy correlation is a poor HLV metric.

In Table 5, we show the Pearson correlation coefficients for all metrics along with the p-values. The table confirms that all metrics are positively correlated with human judgments but with different magnitudes. As before, both hard F_1 scores and entropy correlation are very weakly correlated with human judgments for LLaMA, and entropy correlation is the weakest for RoBERTa. Furthermore, it is the only metric whose correlation is not statistically significant. Across both pretrained models, both PO-JSD and soft micro F_1 consistently have the strongest correlation, but each is best for different models: PO-JSD is best for LLaMA while soft micro F_1 score is best for RoBERTa. This finding suggests that both metrics are good for evaluation in the HLV context.

6.3 Method Ranking Based on Human Judgments

Given the scores computed from human judgments, we can rank the methods based on their scores (human ranking). Table 6 shows that the human rankings have some similarities to TAG results in Figure 4. For instance, *ReL*, *JSD*, and *SL* are generally strong performers (exception: *ReL* with LLaMA). That said, we also see some discrepancies. For example, *SLMV* is highly ranked here but not reflected in Figure 4. We contend, however, that the meta-evaluation study is ultimately based on one dataset and so we should interpret these results cautiously.

6.4 Discussion

Putting all the results together, PO-JSD and soft micro F_1 are arguably some of the best metrics for HLV data. That said, Uma et al. (2021) argued that the value of

Table 6
Rankings of HLV training methods based on their scores produced by human judgments on the TAG dataset.

RoBERTa				LLaMA			
Rank 1–7		Rank 8–13		Rank 1–7		Rank 8–13	
Method	Score	Method	Score	Method	Score	Method	Score
SMF1	0.126	**MV**	0.076	**SL**	0.130	**ReL**	0.070
JSD	0.109	**SmF1**	0.070	**SLMV**	0.117	**LA-max**	0.061
ARh	0.102	**LA-max**	0.061	**AE**	0.105	**ARh**	0.055
SLMV	0.098	**AR**	0.048	**MV**	0.092	**LA-min**	0.054
SL	0.095	**AE**	0.027	**AR**	0.088	**AEh**	0.050
ReL	0.085	**AEh**	0.024	**JSD**	0.086	**SMF1**	0.014
LA-min	0.078			**SmF1**	0.076		

PO-JSD is typically confined within a small range and lacks an intuitive interpretation. We believe that this is its biggest weakness, which is shared by other information-theoretic measures such as cross-entropy. In contrast, soft micro F_1 is more interpretable as it is analogous to the standard F_1 score. Therefore, our general recommendation is to report both metrics, but focus on the soft micro F_1 score when an accessible interpretation is important (e.g., communicating with non-technical people) or one is restricted to a single metric (e.g., in hyperparameter tuning).

7. Conclusions

We propose new evaluation metrics for evaluating model predictions with human label variation (HLV). Taking inspiration from remote sensing research, we represent human judgment distributions as degrees of membership over fuzzy sets and generalize standard metrics such as accuracy into their soft versions using fuzzy set operations. Therefore, our proposed metrics have intuitive interpretations and reduce to standard hard metrics if there is no label variation. While our analysis shows that our proposed soft accuracy metric is strongly correlated with an existing metric based on Jensen-Shannon divergence, we mathematically prove that the former is upper-bounded by the latter. We further show that the JSD-based metric can produce a misleadingly high score as a result.

Because the soft metrics are differentiable, we propose 3 new training methods for HLV using the metrics as the training objective. Additionally, we propose 2 new training methods that aggregate losses over annotations of the same instance, bringing the total of our proposed training methods to 5 methods. We test our proposed methods on 6 datasets spanning binary, multiclass, and multilabel classification tasks, as well as crowd and expert annotators. We evaluate against 9 existing HLV training methods across 2 pretrained models and use a total of 6 HLV evaluation metrics including both existing and proposed ones to find the best methods for HLV data. Then, we perform an empirical meta-evaluation of the evaluation metrics to understand which metrics are best for HLV data.

We find that simple methods such as training on disaggregated annotations (*ReL*) or soft labels (*SL*) perform best in most cases. They often outperform not only training on the majority-voted labels but also more complex HLV training methods including our proposed training methods with the differentiable metrics. Our meta-evaluation shows that our proposed soft micro F_1 score is one of the best metrics for HLV data. This metric reduces to our proposed soft accuracy in single-label classification. Given its intuitive interpretation and positive meta-evaluation result, we recommend further research to include soft micro F_1 when reporting model performance in the HLV context.

Limitations

The TAG dataset is private and cannot be released publicly, which is a limitation of our work in terms of reproducibility. This is because the data is based on real-world confidential legal requests from help-seekers. More importantly, their safety and privacy is of high concern: Some cases are so unique that even if you were to anonymize the cases, re-identfication may still be possible. Nevertheless, we believe our work still offers valuable scientific knowledge on the investigation of HLV training methods and evaluation metrics.

Our empirical meta-evaluation is limited to a single multilabel classification dataset. Thus, it is unclear if the findings extend to other datasets and binary or single-label tasks. Future work could expand on this limitation by experimenting with a different or diverse selection of datasets.

Appendix A: Derivation of Soft F_1 Scores

A.1 Fuzzy Sets

Below we provide a brief overview of fuzzy sets drawn from the work of Kruse et al. (2022). A fuzzy set A is defined by its membership function $\mu_A : X \to [0, 1]$, where $\mu_A(x)$ represents the degree of membership of $x \in X$ in A. The cardinality of A is defined as the sum of the degrees of membership of all elements of X:

$$\sum_{x \in X} \mu_A(x).$$

If A and B are fuzzy sets over the same universe X then their intersection is a fuzzy set whose membership function is defined as

$$\mu_{A \cap B}(x) = t(\mu_A(x), \mu_B(x))$$

where t is a function that satisfies 3 properties: commutativity, associativity, and monotonicity.[22] The function t is called a **t-norm (triangular norm)**. A special t-norm is the min function because it is also idempotent, i.e., $t(\alpha, \alpha) = \alpha$.

22 $\beta \leq \gamma \Rightarrow t(\alpha, \beta) \leq t(\alpha, \gamma)$ where $0 \leq \alpha, \beta, \gamma \leq 1$.

A.2 Soft F_1 Scores

Let H_k and R_k denote the predicted and the true (crisp) sets of examples in class k, respectively. Standard precision and recall scores for class k are defined as:

$$\text{Prec}_k = \frac{|H_k \cap R_k|}{|H_k|}$$
$$\text{Rec}_k = \frac{|H_k \cap R_k|}{|R_k|}$$

Embracing HLV, we view judgment distributions as degrees of memberships of fuzzy sets, each corresponding to a class, over a universe of examples. Specifically, we consider P_{ik} and Q_{ik} as the true and the predicted degrees of membership of example i in class k. Therefore, we can define the soft precision and recall scores using fuzzy set operations as follows (Harju and Mesaros 2023):

$$\text{SoftPrec}_k = \frac{\sum_i \min(P_{ik}, Q_{ik})}{\sum_i Q_{ik}} \tag{A.1}$$

and

$$\text{SoftRec}_k = \frac{\sum_i \min(P_{ik}, Q_{ik})}{\sum_i P_{ik}} \tag{A.2}$$

The soft F_1 score is the harmonic mean between the soft precision and recall scores as usual:

$$\text{SoftF}_{1_k} = 2\frac{\sum_i \min(P_{ik}, Q_{ik})}{\sum_i (P_{ik} + Q_{ik})} \tag{A.3}$$

Taking the mean of Equation (A.3) over classes results in the soft macro F_1 score. To obtain the micro variant, we simply modify the sums in both the numerator and the denominator of Equation (A.1) and (A.2) to also iterate over classes to get the soft micro precision and recall scores, and then take the harmonic mean of the two scores as normal.

Appendix B: Performance of MV

Table B.1 reports the performance of *MV* across datasets and evaluation metrics.

Table B.1
Mean (± std) performance of *MV* measured by entropy correlation, PO-JSD, and various versions of F_1 scores. MD-Agr refers to MD-Agreement. C-SNLI and C-MNLI refer to the SNLI and the MNLI portions of ChaosNLI, respectively. Hard (resp., soft) accuracy scores (for multiclass tasks) are reported in the hard (resp., soft) micro F_1 score column.

Dataset	Model	Micro F_1	Macro F_1	Entropy corr.	PO-JSD	Soft micro F_1	Soft macro F_1
HS-Brexit	TwHIN-BERT	.903 ± .009	.581 ± .093	.470 ± .119	.941 ± .010	.883 ± .013	.676 ± .071
	LLaMA	.950 ± .012	.660 ± .104	.330 ± .065	.939 ± .002	.886 ± .003	.623 ± .012
MD-Agr	TwHIN-BERT	.809 ± .003	.779 ± .002	.379 ± .002	.921 ± .001	.811 ± .002	.790 ± .003
	LLaMA	.810 ± .001	.784 ± .001	.215 ± .019	.863 ± .003	.770 ± .002	.743 ± .003
ArMIS	TwHIN-BERT	.713 ± .014	.705 ± .014	.062 ± .015	.769 ± .011	.702 ± .009	.695 ± .008
C-SNLI	RoBERTa	.632 ± .005	.551 ± .007	.123 ± .014	.830 ± .005	.672 ± .004	.629 ± .007
	LLaMA	.601 ± .010	.551 ± .008	.118 ± .022	.830 ± .005	.668 ± .005	.633 ± .007
C-MNLI	RoBERTa	.496 ± .014	.369 ± .014	.044 ± .007	.854 ± .009	.668 ± .010	.642 ± .009
	LLaMA	.514 ± .017	.439 ± .010	.100 ± .018	.827 ± .005	.643 ± .006	.615 ± .006
MFRC	RoBERTa	.657 ± .001	.456 ± .002	.404 ± .002	.954 ± .000	.591 ± .001	.418 ± .000
	LLaMA	.649 ± .001	.454 ± .005	.353 ± .005	.948 ± .000	.576 ± .001	.395 ± .001
TAG	RoBERTa	.742 ± .002	.610 ± .014	.394 ± .002	.985 ± .000	.645 ± .002	.471 ± .003
	LLaMA	.759 ± .006	.681 ± .005	.279 ± .010	.983 ± .000	.648 ± .002	.521 ± .003

Appendix C: Annotation Interface Examples

Figures C.1 and C.2 show the interface of the area of law and the preference annotation tasks, respectively.

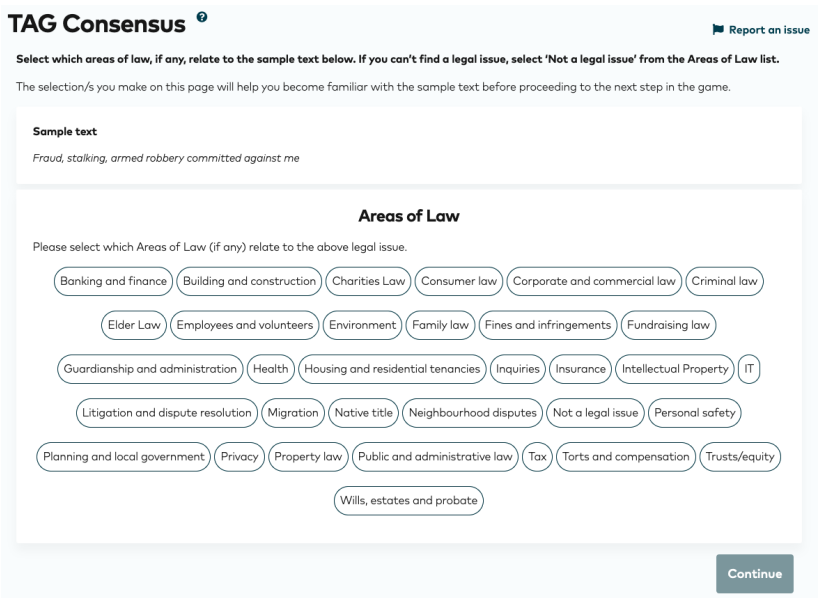

Figure C.1
Areas of law annotation interface.

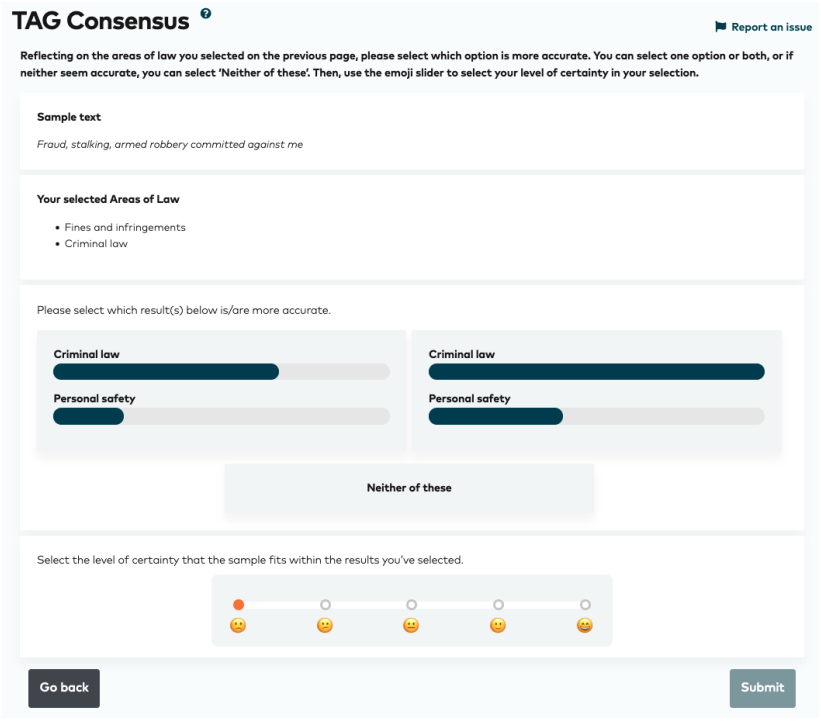

Figure C.2
Preference annotation interface.

Acknowledgments

This research is supported by the Australian Research Council Linkage Project LP210200917[23] and funded by the Australian Government. This research is done in collaboration with Justice Connect, an Australian public benevolent institution.[24] We thank Kate Fazio, Tom O'Doherty, and Rose Hyland from Justice Connect for their support throughout the project. We thank David McNamara, Yashik Anand, and Mark Pendergast also from Justice Connect for the development of the meta-evaluation annotation interface. This research is supported by The University of Melbourne's Research Computing Services and the Petascale Campus Initiative.

References

Akbik, Alan, 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.

Akhtar, Sohail, Valerio Basile, and Viviana Patti. 2021. Whose opinions matter? Perspective-aware models to identify opinions of hate speech victims in abusive language detection. *CoRR*, cs.CL/2106.15896.

Almanea, Dina and Massimo Poesio. 2022. ArMIS - The Arabic Misogyny and Sexism corpus with annotator subjective disagreements. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2282–2291.

Binaghi, Elisabetta, Pietro A. Brivio, Paolo Ghezzi, and Anna Rampini. 1999. A fuzzy set-based accuracy assessment of soft classification. *Pattern Recognition Letters*, 20(9):935–948. https://doi.org/10.1016/S0167-8655(99)00061-6

Chen, Beiduo, Xinpeng Wang, Siyao Peng, Robert Litschko, Anna Korhonen, and Barbara Plank. 2024. "Seeing the big through the small": Can LLMs approximate human judgment distributions on NLI from a few explanations? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14396–14419. https://doi.org/10.18653/v1/2024.findings-emnlp.842

Chicco, Davide and Giuseppe Jurman. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1):6. https://doi.org/10.1186/s12864-019-6413-7, PubMed: 31898477

Cui, Xia. 2023. Xiacui at SemEval-2023 Task 11: Learning a model in mixed-annotator datasets using annotator ranking scores as training weights. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1076–1084. https://doi.org/10.18653/v1/2023.semeval-1.148

Deng, Naihao, Xinliang Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You are what you annotate: Towards better models through annotator representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12475–12498. https://doi.org/10.18653/v1/2023.findings-emnlp.832

Foody, G. M. 1996. Approaches for the production and evaluation of fuzzy land cover classifications from remotely-sensed data. *International Journal of Remote Sensing*, 17(7):1317–1340. https://doi.org/10.1080/01431169608948706

Fornaciari, Tommaso, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597. https://doi.org/10.18653/v1/2021.naacl-main.204

Gajewska, Ewelina. 2023. Eevvgg at SemEval-2023 Task 11: Offensive language classification with rater-based information. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 171–176. https://doi.org/10.18653/v1/2023.semeval-1.24

Gómez, D., G. Biging, and J. Montero. 2008. Accuracy statistics for judging soft classification. *International Journal of Remote*

23 https://dataportal.arc.gov.au/NCGP/Web/Grant/Grant/LP210200917.

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

Sensing, 29(3):693–709. https://doi.org/10.1080/01431160701311325

Grattafiori, Aaron, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The Llama 3 herd of models. *CoRR*, cs.AI/2407.21783.

Grötzinger, Dennis, Simon Heuschkel, and Matthias Drews. 2023. CICL_DMS at SemEval-2023 Task 11: Learning with disagreements (Le-Wi-Di). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1030–1036. https://doi.org/10.18653/v1/2023.semeval-1.141

Harju, Manu and Annamaria Mesaros. 2023. Evaluating classification systems against soft labels with fuzzy precision and recall. In *Proceedings of the 8th Detection and Classification of Acoustic Scenes and Events 2023 Workshop (DCASE2023)*, pages 46–50.

Hu, Edward J., 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*.

Kruse, Rudolf, Sanaz Mostaghim, Christian Borgelt, Christian Braune, and Matthias Steinbrecher. 2022. Introduction to fuzzy sets and fuzzy logics. In *Computational Intelligence: A Methodological Introduction*. Springer, pages 373–405. https://doi.org/10.1007/978-3-030-42227-1_15

Kurniawan, Kemal, Meladel Mistica, Timothy Baldwin, and Jey Han Lau. 2024. To aggregate or not to aggregate. That is the question: A case study on annotation subjectivity in span prediction. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 362–368. https://doi.org/10.18653/v1/2024.wassa-1.29

Lee, Noah, Na Min An, and James Thorne. 2023. Can large language models capture dissenting human voices? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4585. https://doi.org/10.18653/v1/2023.emnlp-main.278

Leonardelli, Elisa, 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. https://doi.org/10.18653/v1/2023.semeval-1.314

Leonardelli, Elisa, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539. https://doi.org/10.18653/v1/2021.emnlp-main.822

Lewis, H. G. and M. Brown. 2001. A generalized confusion matrix for assessing area estimates from remotely sensed data. *International Journal of Remote Sensing*, 22(16):3223–3235. https://doi.org/10.1080/01431160152558332

Lin, Jianhua. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151. https://doi.org/10.1109/18.61115

Liu, Yinhan, 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. *CoRR*, cs.CL/1907.11692.

Maity, Ankita, Pavan Kandru, Bhavyajeet Singh, Kancharla Aditya Hari, and Vasudeva Varma. 2023. IREL at SemEval-2023 Task 11: User conditioned modeling for toxicity detection in subjective tasks. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2133–2136. https://doi.org/10.18653/v1/2023.semeval-1.294

Negahban, Sahand, Sewoong Oh, and Devavrat Shah. 2012. Iterative ranking from pair-wise comparisons. In *Advances in Neural Information Processing Systems*, volume 25.

Nie, Yixin, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143. https://doi.org/10.18653/v1/2020.emnlp-main.734

Pavlick, Ellie and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694. https://doi.org/10.1162/tacl_a_00293

Peterson, Joshua, 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. https://doi.org/10.1109/ICCV.2019.00971

Plank, Barbara. 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. https://doi.org/10.18653/v1/2022.emnlp-main.731

Pontius, R. G. and M. L. Cheuk. 2006. A generalized cross-tabulation matrix to compare soft-classified maps at multiple resolutions. *International Journal of Geographical Information Science*, 20(1):1–30. https://doi.org/10.1080/13658810500391024

Pontius, R. G. and John Connors. 2006. Expanding the conceptual, mathematical, and practical methods for map comparison. In *Proceedings of the Meeting of Spatial Accuracy*, pages 64–79.

Rizzi, Giulia, 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.

Rodríguez-Barroso, Nuria, Eugenio Martínez Cámara, Jose Camacho Collados, M. Victoria Luzón, and Francisco Herrera. 2024. Federated learning for exploiting annotators' disagreements in natural language processing. *Transactions of the Association for Computational Linguistics*, 12:630–648. https://doi.org/10.1162/tacl_a_00664

Sheng, Victor S., Foster Provost, and Panagiotis G. Ipeirotis. 2008. Get another label? Improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 614–622. https://doi.org/10.1145/1401890.1401965

Silván-Cárdenas, J. L. and L. Wang. 2008. Sub-pixel confusion–uncertainty matrix for assessing soft classifications. *Remote Sensing of Environment*, 112(3):1081–1095. https://doi.org/10.1016/j.rse.2007.07.017

Sullivan, Michael, Mohammed Yasin, and Cassandra L. Jacobs. 2023. University at Buffalo at SemEval-2023 Task 11: MASDA–Modeling annotator sensibilities through disaggregation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 978–985. https://doi.org/10.18653/v1/2023.semeval-1.135

Trager, Jackson, Alireza S. Ziabari, Aida Mostafazadeh Davani, Preni Golazizian, Farzan Karimi-Malekabadi, Ali Omrani, Zhihe Li, Brendan Kennedy, Nils Karl Reimer, Melissa Reyes, Kelsey Cheng, Mellow Wei, Christina Merrifield, Arta Khosravi, Evans Alvarez, and Morteza Dehghani. 2022. The moral foundations Reddit corpus. *CoRR*, cs.CL/2208.05545.

Uma, Alexandra, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2020. A case for soft loss functions. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8(1):173–177. https://doi.org/10.1609/hcomp.v8i1.7478

Uma, Alexandra N., 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. https://doi.org/10.1613/jair.1.12752

Vitsakis, Nikolas, Amit Parekh, Tanvi Dinkar, Gavin Abercrombie, Ioannis Konstas, and Verena Rieser. 2023. iLab at SemEval-2023 Task 11 Le-Wi-Di: Modeling disagreement or modeling perspectives? In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1660–1669. https://doi.org/10.18653/v1/2023.semeval-1.231

Wan, Ruyuan and Karla Badillo-Urquiola. 2023. Dragonfly_captain at SemEval-2023 Task 11: Unpacking disagreement with investigation of annotator demographics and task difficulty. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1978–1982. https://doi.org/10.18653/v1/2023.semeval-1.272

Zhang, Xinyang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2023. TwHIN-BERT: A socially-enriched pre-trained language model for multilingual tweet representations at Twitter. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5597–5607. https://doi.org/10.1145/3580305.3599921