Annotator-Centric Active Learning for Subjective NLP Tasks

Michiel van der Meer
LIACS
Leiden University
m.t.van.der.meer@liacs.leidenuniv.nl

Neele Falk
Institute for Natural Language Processing
University of Stuttgart
neele.falk@ims.uni-stuttgart.de

Pradeep K. Murukannaiah
Interative Intelligence
TU Delft
p.k.murukannaiah@tudelft.nl

Enrico Liscio
Interative Intelligence
TU Delft
e.liscio@tudelft.nl

Abstract

To accurately capture the variability in human judgments for subjective NLP tasks, incorporating a wide range of perspectives in the annotation process is crucial. Active Learning (AL) addresses the high costs of collecting human annotations by strategically annotating the most informative samples. We introduce Annotator-Centric Active Learning (ACAL), which incorporates an annotator selection strategy following data sampling. Our objective is two-fold: (1) to efficiently approximate the full diversity of human judgments, and (2) to assess model performance using annotator-centric metrics, which emphasize minority perspectives over a majority. We experiment with multiple annotator selection strategies across seven subjective NLP tasks, employing both traditional and novel, human-centered evaluation metrics. Our findings indicate that ACAL improves data efficiency and excels in annotator-centric performance evaluations. However, its success depends on the availability of a sufficiently large and diverse pool of annotators to sample from.

1 Introduction

A challenging aspect of natural language understanding (NLU) is the variability of human judgment and interpretation in subjective tasks (e.g., hate speech detection) (Plank, 2022). While humans can navigate subjectivity naturally, most machine learning methods are insensitive to individual differences (Sandri et al., 2023) and underrepresented perspectives (van der Meer et al., 2024).

Modern NLU approaches are commonly trained and tested on annotated datasets. In a subjective task, each data sample is typically labeled by a set of annotators, and differences in annotation are reconciled through aggregation techniques (e.g., majority voting), resulting in a single "gold label" (Uma et al., 2021). This approach, though effective for training ML algorithms, neglects the labels of minorities, which becomes problematic, especially,

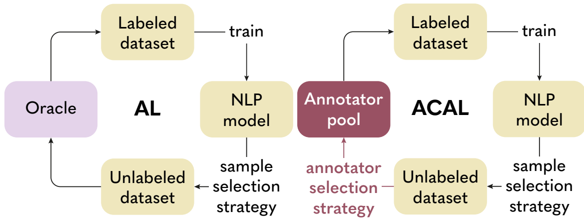

Figure 1: Active Learning (AL) approaches (left) use a sample selection strategy to pick samples to be annotated by an oracle. The Annotator-Centric Active Learning (ACAL) approach (right) extends AL by introducing an annotator selection strategy to choose the annotators who annotate the selected samples.

in the case of sensitive subjective tasks.

Subjectivity has been addressed by modeling the full distribution of annotations for each data sample as opposed to aggregating them (Plank, 2022). However, resources for such approaches are scarce, as most datasets do not (yet) make fine-grained annotation details available (Cabitza et al., 2023), and representing a full range of perspectives is contingent on obtaining annotations from a diverse crowd (Bakker et al., 2022).

One way of accounting for a limited annotation budget is to use Active Learning (Settles, 2012, AL). Given a pool of unannotated data samples, AL employs a sample selection strategy to select maximally informative samples for training, retrieving the corresponding annotations from a ground truth oracle (e.g., a single human expert). However, in subjective tasks there is no such oracle, instead we rely on a set of available annotators. Given this practical constraint, we argue that informativeness for AL manifests in both samples and annotations, as the model should also be guided to reflect the distribution of annotations. Demanding all available annotators to annotate all selected samples would provide a truthful representation of the annotation distribution, but is often unfeasible, especially if the pool of annotators is large. Thus, deciding *which annotator(s)* should annotate the selected samples is as critical as selecting which samples to annotate.

We introduce Annotator-Centric Active Learning (ACAL) to account for annotation diversity in subjective tasks. In ACAL, the sample selection strategy of traditional AL is followed by an *annotator selection strategy* as Figure 1 shows. For each data sample selected through the sample selection strategy, the annotator selection strategy selects an annotator from the available annotators. We make the following contributions: (1) We create ACAL, extending the AL approach to optimize for diversity among annotators when learning soft labels in subjective tasks. (2) We introduce a suite of annotator-centric evaluation metrics to measure both representativeness and diversity. (3) We demonstrate our approach's effectiveness on three diverse datasets with subjective tasks—hate speech detection, moral value classification, and safety judgments.

Our experiments show that ACAL works to better approximate the distribution of human judgments with a lower annotation budget. However, this effectiveness requires a large pool of diverse annotators, as is the case for one of our datasets. In other cases, the differences between ACAL and traditional AL become smaller. Through our annotator-centric evaluation, we show that task agreement and the number of available annotations both influence the effectiveness of ACAL, hinting at a direct trade-off between learning from a majority versus being sensitive to minority annotations.

2 Related work

We review related works on annotator disagreement and active learning. Our work is novel in combining these fields to (1) represent annotation distributions through soft labels, (2) incorporate annotator selection strategies in the active learning loop, and (3) evaluate with annotator-centric metrics next to traditional evaluation.

2.1 Learning with annotator disagreement

Modeling annotator disagreement is garnering increasing attention (Aroyo and Welty, 2015; Uma et al., 2021; Plank, 2022; Cabitza et al., 2023). For instance, some aggregation methods can lead to a fairer representation than simple majority (Hovy et al., 2013; Tao et al., 2018). Alternatively, the full annotation distribution can be modeled using soft labels (e.g., Peterson et al., 2019; Müller et al., 2019; Fornaciari et al., 2021; Collins et al., 2022). Other approaches leverage annotator-specific information, e.g., by including individual classification heads per annotator (Davani et al., 2022), embedding annotator-specific behavior (Mokhberian et al., 2023), or encoding the annotator's sociodemographic information (Beck et al., 2023).

Yet, representing annotator diversity remains challenging. Standard calibration metrics under human label variation may be unsuitable, especially when the variation is high (Baan et al., 2022). Trade-offs ought to be made between collecting more samples or more annotations (Gruber et al., 2024). Further, solely measuring differences among sociodemographic traits is not sufficient to fully capture opinion diversity (Orlikowski et al., 2023). To this end, we represent diversity based on *which* annotators have annotated, *what* they annotated, and *how* they have annotated. We experiment with different annotator selection strategies to reveal what aspects impact task performance and annotation budget.

2.2 Active Learning

AL enables a supervised learning model to achieve high performance with a few training examples if chosen judiciously (Settles, 2012). In a typical AL scenario, a vast collection of unlabeled data is available, and an oracle (e.g., a human expert) can be asked to annotate this unlabeled data. A *sampling strategy* is employed to iteratively (and smartly) select the next batch of unlabeled data for annotation by the oracle (Ren et al., 2021).

AL has found widespread application in the field of NLP (Zhang et al., 2022). Two main strategies are employed, either by selecting the unlabeled samples on which the model prediction is most uncertain (Zhang et al., 2017), or by selecting samples that are most representative of the unlabeled dataset (Erdmann et al., 2019; Zhao et al., 2020).

The combination of AL and annotator diversity is a novel direction that has not garnered much attention yet. Existing work proposes to align model uncertainty with annotator uncertainty (Baumler et al., 2023), whereas others adapt annotator-specific classification heads in AL settings (Wang and Plank, 2023), or select texts to annotate based on annotator preferences (Kanclerz et al., 2023).

Existing methods ignore a crucial part of learning with human variation: the diversity among annotators. We focus on which annotators should annotate, such that it best informs us about the underlying label diversity.

3 Method

First, we define the soft-label prediction task we use to train a supervised model. Then, we introduce the traditional AL and the novel ACAL approaches.

3.1 Soft-Label prediction

Consider a dataset composed of triples (x_i, a_j, y_{ij}), where x_i is a data sample (i.e., a piece of text) and $y_{ij} \in C$ is the class label assigned by annotator a_j. The multiple labels assigned to a sample x_i by the different annotators are usually combined into an aggregated label $\hat{y}_i$. For training with soft labels, the aggregation typically takes the form of maximum likelihood estimation (Uma et al., 2021):

$$\hat{y}_i(x) = \frac{\sum_{i=1}^{N}[x_i = x][y_{ij} = c]}{\sum_{i=1}^{N}[x_i = x]} \qquad (1)$$

In our experiments, We use a passive learning approach that uses all available $\{x_i, \hat{y}_i\}$ to train a model f_θ with cross-entropy loss as a baseline.

3.2 Active Learning

AL imposes a sampling technique for inputs x_i, such that the most *informative* sample(s) are picked for learning. In a typical AL approach, a set of unlabelled data points U is available. At every iteration, a sample selection strategy $\mathcal{S}$ selects samples $x_i \in U$ to be annotated by an oracle $\mathcal{O}$ that provides the ground truth label distribution $\hat{y}_i$. The selected samples and annotations are added to the labeled data D, with which the model f_θ is trained. Alg. 1 provides an overview of the procedure. In our sample selection strategies, a batch of data of a given size B is queried at each iteration. In our experiments, we compare the following strategies:

Algorithm 1: AL approach.

input : Unlabeled data U, Data sampling strategy $\mathcal{S}$, Oracle $\mathcal{O}$
$D_0 \leftarrow \{\}$
for $n = 1..N$ **do**
  sample data points x_i from U using $\mathcal{S}$
  obtain annotation $\hat{y}_i$ for x_i from $\mathcal{O}$
  $D_{n+1} = D_n + \{x_i, \hat{y}_i\}$
  train f_θ on D_{n+1}
end

Random ($\mathcal{S}_R$) selects a B samples uniformly at random from U.
Uncertainty ($\mathcal{S}_U$) predicts a distribution over class labels with $f_\theta(x_i)$ for each $x_i \in U$. Select the B samples with the highest prediction entropy (i.e., the samples on which the model is most uncertain).

3.3 Annotator-Centric Active Learning

The ACAL approach builds on the AL approach. In contrast to AL, which retrieves an aggregated annotation $\hat{y}_i$, ACAL employs an annotator selection strategy $\mathcal{T}$ to select one annotator and their annotation for each selected data point x_i. Alg. 2 describes the ACAL approach.

Algorithm 2: ACAL approach.

input : Unlabeled data U, Data sampling strategy $\mathcal{S}$, Annotator sampling strategy $\mathcal{T}$
$D_0 \leftarrow \{\}$
for $n = 1..N$ **do**
  sample data points x_i from U using $\mathcal{S}$
  sample annotators a_j for x_i using $\mathcal{T}$
  obtain annotation y_{ij} from a_j for x_i
  $D_{n+1} = D_n + \{x_i, y_{ij}\}$
  train f_θ on D_{n+1}
end

We propose annotator selection strategies that include annotations from diverse annotators. The strategies vary in the type of information used to represent differences between annotators, and include *what* or *how* the annotators have annotated thus far. We test the following strategies:
Random ($\mathcal{T}_R$) selects one random annotator a_j.
Label Minority ($\mathcal{T}_L$) considers only the labels that annotators have assigned. Given a new sample x_i, $\mathcal{T}_L$ selects the available annotator that has the largest bias toward the minority label compared to the other available annotators, i.e., who has annotated other samples with the minority label the most. The minority label is selected as the class with the smallest annotation count in the available dataset D_n thus far.
Semantic Diversity ($\mathcal{T}_S$) considers only information on *what* each annotator has annotated so far (i.e., the samples that they have annotated). Given a new sample x_i selected through $\mathcal{S}$, $\mathcal{T}_S$ selects the available annotator for whom x_i is semantically the most different from what the annotator has labeled so far. To measure this difference for an annotator

Dataset	Task (dimension)	Num. Samples	Num. Annotators	Num. Annotations	Avg. Annotations per item
DICES	Safety Judgment	990	172	72,103	72.83
MFTC	Morality (care)	8434	23	31310	3.71
MFTC	Morality (loyalty)	3288	23	12803	3.89
MFTC	Morality (betrayal)	12546	23	47002	3.75
MHS	Hate Speech (dehumanize, genocide, respect)	17282	7807	57980	3.35

Table 1: Overview of the datasets and tasks employed in our work.

a_j, we employ a sentence embedding model to measure the cosine distance between the embeddings of x_i and embeddings of all the samples annotated by a_j. We then take the average of all semantic similarities. The annotator with the lowest average similarity score is selected.

Representation Diversity ($\mathcal{T}_D$) selects the annotator that has the lowest similarity with the other annotators available for that item. We create a simple representation for each annotator based on the items together with the respective label that they have annotated, followed by computing the pairwise cosine similarity between all annotators.

4 Experimental Setup

We describe the experimental setup for the comparisons between ACAL strategies. In all our experiments, we employ a TinyBERT model (Jiao et al., 2019) to reduce the number of trainable parameters. Appendix A includes a detailed overview of the computational setup and hyperparameters. We will provide our codebase upon publication.

4.1 Datasets

Table 1 introduces the three datasets that we use, with variation in domain, annotation task (in *italics*), annotator count, and annotations per instance.

The **DICES Corpus** (Aroyo et al., 2023) is composed of 990 conversations with an LLM where 172 annotators provided judgments on whether a generated response can be deemed safe (3-way judgments: yes, no, unsure). We perform a multi-class classification with the scores.

The **MFTC Corpus** (Hoover et al., 2020) is composed of 35K tweets that 23 annotators annotated with any of the 10 moral elements from the Moral Foundation Theory (Graham et al., 2013). We select the elements of *loyalty* (lowest annotation count), *care* (medium count), and *betrayal* (highest count) and perform three binary classifications to predict the presence of the respective elements. As most tweets were labeled non-moral (i.e., with no moral element), we balanced the datasets by subsampling the non-moral class.

The **MHS Corpus** (Sachdeva et al., 2022) consists of 50K social media comments on which 8K annotators judged three hate speech aspects—*dehumanize* (low inter-rater agreement), *respect* (medium agreement), and *genocide* (high agreement)—on a 5-point Likert scale. We perform a multi-class classification with the annotated Likert scores for each task.

The datasets and tasks differ in the entropy scores over annotations (Appendix A.5). DICES and MHS generally have medium normalized entropy scores (most lie between $0.15 < H < 0.85$), whereas the MFTC entropy scores are highly polarized.

4.2 Training procedure

We test the annotator selection strategies proposed in Section 3.3 by comparing all possible combinations of the two different sample selection strategies ($\mathcal{S}_R$ and $\mathcal{S}_U$) with the annotator selection strategies ($\mathcal{T}_R$, $\mathcal{T}_L$, $\mathcal{T}_S$, and $\mathcal{T}_D$). At each round, we use $\mathcal{S}$ to select B unique samples from the unlabeled data pool U. We empirically select B to be the smallest between 5% of the number of available annotations and the number of unique samples in the training set. For each selected sample x_i, we use $\mathcal{T}$ to select one annotator and retrieve their annotation y_{ij}.

To populate the annotation history for the annotation selection strategies, we perform a single warmup round with B randomly selected annotations before starting the ACAL iterations (Zhang et al., 2022). We report our training progress results on a validation set with 3-fold cross-validation, showing the average to account for stability across random data splits (into 80% train, 10% validation, and 10% test) and initialization. Then, we select the model iteration that led to the best performance (according to JS) on the validation set and evaluate it using a separate test set.

We compare our work with traditional Oracle-based AL approaches ($\mathcal{S}_R\mathcal{O}$ and $\mathcal{S}_U\mathcal{O}$), which use the data sampling strategies but obtain all possible annotations for each sample (following Alg. 1). Moreover, we employ a passive learning (PL) approach as an upper bound by training the model on the full dataset, thus observing all available samples and annotations. Our baselines follow the analogous cross-validation setup.

4.3 Evaluation metrics

The ACAL strategies aim to guide the algorithm to model a representative distribution of the annotator's perspectives while reducing human annotation effort. To this end, we evaluate the model both with a traditional evaluation metric and a metric aimed at comparing predicted and annotated distributions:

Macro F_1-score (F_1) For each sample in the test set, we select the label predicted by the model with the highest confidence, determine the golden label through a majority agreement aggregation, and compute the resulting macro F_1-score.

Jensen-Shannon Divergence (JS) The JS measures the divergence between the distribution of label annotation and prediction (Nie et al., 2020). We report the average JS for the samples in the test set to measure how well the algorithm can model the annotation distribution.

Next, since our proposed annotator selection strategies aim to promote diversity, we introduce novel annotator-centric evaluation metrics. First, we report the average among annotators. Second, in line with Rawls' principle of maximum fairness (Rawls, 1973), the result for the worst-off annotators:

Per-annotator F_1 (F_1^a) We compute the F_1 for each annotator in the test set using their annotations as golden labels, and average it.

Per-annotator JS (JS^a) We compute the JS for each annotator in the test set using their annotations as target distribution, and average it.

Worst per-annotator F_1 (F_1^w) We compute the F_1 for each annotator in the test set using their annotations as golden labels, and report the average of the lowest 10% (to mitigate noise).

Worst per-annotator JS (JS^w) We compute the JS for each annotator in the test set using their annotations as target distribution, and report the average of the lowest 10% (to mitigate noise).

These evaluation metrics allow us to measure the trade-offs between modeling the majority agreement, a representative distribution of annotations, and accounting for minority voices. We report these metrics on the validation set (as progress over the AL iterations) and test set (by using the best-performing model on the validation set), as described in Section 4.2.

5 Results

5.1 Test sets results

See Figure 2 for the performance of the DICES, MFTC, and MHS, respectively. For MFTC, we initially focus on *care*, since it is the task with neither the most nor least amount of data. For MHS, we start with *dehumanize*, since it saw the most medium-level disagreement. The rest of the results can be observed in Appendix B.

Combining our results across datasets, we see that data characteristics influence whether ACAL can learn performant models efficiently. In particular, we see that for DICES and MHS, ACAL may learn models that perform well using less data (38% and 62% reduction at best, respectively). Conversely, for MFTC, there is little impact of using ACAL over PL (5.6% less data used). A similar pattern holds when comparing ACAL to AL, though AL seems to be a strong baseline for MHS, where random sample selection leads to more efficient data usage (60%). AL with uncertainty sampling is more efficient for MFTC (13%).

When we compare the performance metrics, we see that the distributions obtained through ACAL are consistently closer to the ground truth distribution in DICES, as measured by JS than PL and AL. However, this pattern is not visible for MFTC and MHS. In terms of majority-voted F_1, ACAL again leads to better scores in both DICES and MHS. Since DICES and MHS are datasets with moderate disagreement, we may benefit from using ACAL in such scenarios. Further, if the dataset contains a large number of annotators per sample, annotator selection strategies are shown to pick a more informative set of annotators to learn from.

We highlight some further dataset-specific findings that shed light on the differences between the annotator selection strategies in ACAL. First, in **DICES**, we see that for three out of four annotator sampling strategies ($\mathcal{T}_R$, $\mathcal{T}_D$, $\mathcal{T}_L$), the choice of

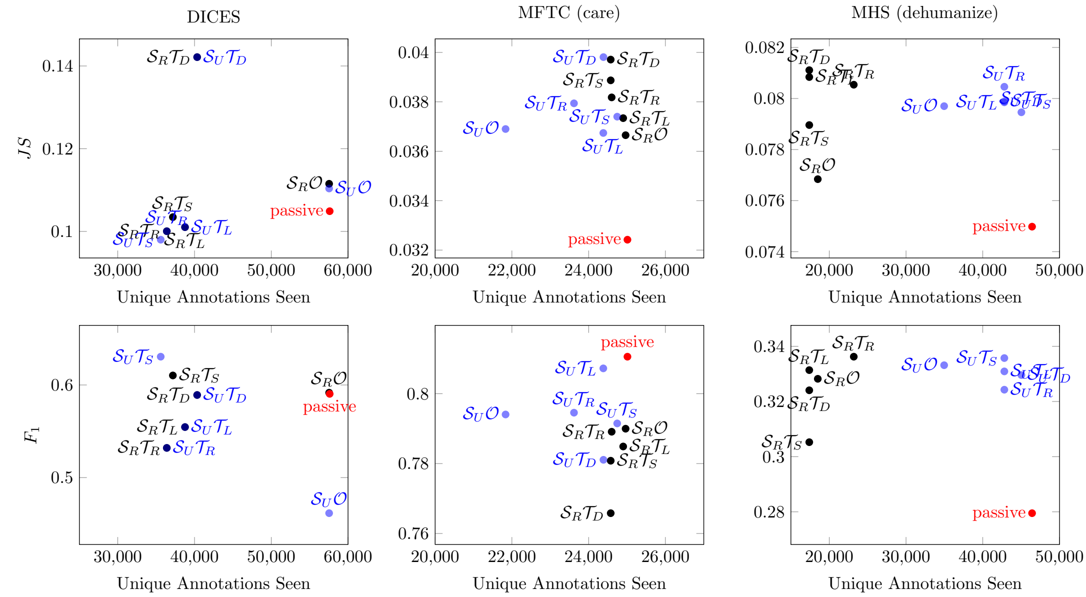

Figure 2: Test set evaluation of the ACAL, AL, and passive approaches across the three dataset/task combinations. For JS, strategies further to the bottom left are more data efficient (x-axis) and perform better (y-axis). For F_1, the top left contains well-performing, data-efficient approaches.

data sampling strategy has no impact on the performance of the model due to the low number of samples to choose from. Furthermore, only $\mathcal{T}_D$ performs worse in terms of JS, overrepresenting outlier annotators. This hints that selecting annotators based on the average embedding of the annotated content strongly emphasizes diverging label behavior. Second, **MFTC** was annotated by a limited fixed set of annotators for whom we can construct a rich annotation history. However, since there are few annotators per sample to pick from, ACAL cannot leverage this information effectively. Again, we see that strategies perform relatively similarly to one another, except for the F_1 scores. Third, in **MHS** we observe that all strategies using random sample selection require less data. Since the task has low inter-rater agreement scores, uncertainty-based sampling wrongly attempts to sample annotations for correct high-entropy predictions, while this is an accurate distribution.

Our findings highlight that with many labels per sample, ACAL is more data-efficient than traditional AL and passive learning in terms of overall evaluation. However, when data characteristics differ and few annotations are available per sample, ACAL has less of an impact. Polarized agreement scores (either high agreement or no agreement) make the use of ACAL and AL cause little to no improvements over passive learning. This corroborates that (AC)AL leads to improvements in specific cases (Dor et al., 2020). Furthermore, we found conflicting results depending on the metric used (JS and F_1). We closely examine the relationship between the evaluation metrics by turning to annotator-centric evaluation, observing how ACAL impacted predictions for individual annotators.

5.2 Annotator-centric evaluation

We show the annotator-centric evaluation metrics in Tables 2, 3, and 4 for DICES, MFTC (care) and MHS (dehumanize), respectively. We again describe per-dataset results. Again, for DICES and MHS, we observe a positive effect of using ACAL over PL, both in terms of data efficiency and final annotator-centric behavior. For these datasets, ACAL leads to a better representation of annotators on average (JS^a,F_1^a), as well as a better representation of the 10% most different annotators (JS^w, F_1^w). Compared ACAL to AL, we mainly observe improvements in the DICES dataset, showing less data used and a better annotator-centric F_1 score. We observe a strong JS^w for the $\mathcal{T}_D$ strategy and worse JS_a, corroborating our earlier finding that emphasizing diverging label behavior trades off with the averaged evaluation scores. Interestingly, this is not apparent in the F_1 scores. For MHS, all approaches using random data sampling ($\mathcal{S}_R$) require considerably less data than passive learning. Further, since the pool of annotators for MHS is large (7K+), there will always be some annota-

App.	Average F_1^a	JS^a	Worst-off F_1^w	JS^w	$\Delta\%$
$\mathcal{S}_R\mathcal{T}_R$	0.432	0.186	0.167	0.453	-36.8
$\mathcal{S}_R\mathcal{T}_L$	0.424	0.187	0.155	0.450	-32.7
$\mathcal{S}_R\mathcal{T}_S$	0.442	0.186	0.164	0.447	-35.5
$\mathcal{S}_R\mathcal{T}_D$	0.431	0.203	0.169	0.370	-30.0
$\mathcal{S}_U\mathcal{T}_R$	0.432	0.186	0.167	0.453	-36.8
$\mathcal{S}_U\mathcal{T}_L$	0.424	0.187	0.155	0.450	-32.7
$\mathcal{S}_U\mathcal{T}_S$	0.439	0.187	0.184	0.447	-38.2
$\mathcal{S}_U\mathcal{T}_D$	0.431	0.203	0.169	0.370	-30.0
$\mathcal{S}_R\mathcal{O}$	0.414	0.191	0.133	0.425	-0.1
$\mathcal{S}_U\mathcal{O}$	0.384	0.192	0.117	0.427	-0.1
Passive	0.371	0.211	0.123	0.479	–

Table 2: DICES annotator-centric evaluation scores. $\Delta\%$ denotes the relative change in the annotation budget with respect to passive learning.

App.	Average F_1^a	JS^a	Worst-off F_1^w	JS^w	$\Delta\%$
$\mathcal{S}_R\mathcal{T}_R$	0.611	0.141	0.377	0.247	-1.6
$\mathcal{S}_R\mathcal{T}_L$	0.616	0.142	0.392	0.249	-0.4
$\mathcal{S}_R\mathcal{T}_S$	0.600	0.145	0.351	0.248	-1.7
$\mathcal{S}_R\mathcal{T}_D$	0.604	0.144	0.357	0.243	-1.7
$\mathcal{S}_U\mathcal{T}_R$	0.612	0.143	0.377	0.252	-5.6
$\mathcal{S}_U\mathcal{T}_L$	0.589	0.142	0.423	0.248	-2.5
$\mathcal{S}_U\mathcal{T}_S$	0.608	0.143	0.399	0.258	-1.1
$\mathcal{S}_U\mathcal{T}_D$	0.586	0.145	0.357	0.253	-2.5
$\mathcal{S}_R\mathcal{O}$	0.586	0.141	0.392	0.255	-0.2
$\mathcal{S}_U\mathcal{O}$	0.583	0.144	0.357	0.253	-12.7
Passive	0.512	0.179	0.377	0.251	–

Table 3: MFTC (care) annotator-centric evaluation scores. $\Delta\%$ denotes the relative change in the annotation budget with respect to passive learning.

tors in disagreement with the output of our models, leading to a zero score on F_1^w.

5.3 Training plots

While the evaluation shows a pattern of efficient data use with ACAL under certain data conditions, it reveals little about how the metrics behave during training or how individual annotator strategies behave. To this end, we provide a complete overview of all metrics (as computed on the validation set) during training in App. B.3. Here we describe the major patterns reoccurring across our experiments using examples and show six of particular interest (Figure 3). Since the strategies only differ in what annotations are included during training, we only show plots related to the annotator-centric metrics.

We can see that there is an influence of both the data sampling and annotator strategies on the performance of the models. Only on DICES is the choice of $\mathcal{S}$ irrelevant, probably due to the low number of samples. Specifically $\mathcal{T}_D$ deteriorates slower for the worst-off annotators than the other strategies but does so without being able to uphold a competitive F_1^a score. In MFTC, we see that when using $\mathcal{S}_U$, performance on F_1^a dips at the start of training. Selecting annotators for samples with high predicted entropy initially leads to a decrease in average performance. The strategy seeks to first lower the entropy for the labels already encountered, though some of the variation in labels is irreconcilable. A similar reasoning holds for MHS, where the differences between strategies are even less impacted by the choice of $\mathcal{T}$. These two plots further underline our main finding that for ACAL to be impactful in representing diverse annotation perspectives, we need to ensure a (1) heterogeneous pool of annotators, and (2) a task that facilitates human label variation.

App.	Average F_1^a	JS^a	Worst-off F_1^w	JS^w	$\Delta\%$
$\mathcal{S}_R\mathcal{T}_R$	0.315	0.394	0.000	0.489	-50.0
$\mathcal{S}_R\mathcal{T}_L$	0.322	0.397	0.000	0.478	-62.5
$\mathcal{S}_R\mathcal{T}_S$	0.313	0.397	0.000	0.480	-62.5
$\mathcal{S}_R\mathcal{T}_D$	0.318	0.398	0.000	0.479	-62.5
$\mathcal{S}_U\mathcal{T}_R$	0.322	0.389	0.000	0.508	-7.8
$\mathcal{S}_U\mathcal{T}_L$	0.328	0.388	0.000	0.507	-7.8
$\mathcal{S}_U\mathcal{T}_S$	0.326	0.388	0.000	0.506	-7.8
$\mathcal{S}_U\mathcal{T}_D$	0.326	0.384	0.000	0.513	-3.0
$\mathcal{S}_R\mathcal{O}$	0.339	0.387	0.000	0.496	-60.1
$\mathcal{S}_U\mathcal{O}$	0.331	0.390	0.000	0.497	-24.7
Passive	0.202	0.424	0.000	0.547	–

Table 4: MHS (dehumanize) annotator-centric evaluation scores. $\Delta\%$ denotes the relative change in the annotation budget with respect to passive learning.

5.4 Change in task

In Fig. 4, we present a comparative analysis of two annotator-centric metrics across the three distinct tasks of MFTC and MHS, for which we have seen the least impact of ACAL over AL and PL. We cannot conclude that the chosen ACAL approach ($\mathcal{S}_R\mathcal{T}_S$) offers a consistent improvement over sampling all annotations ($\mathcal{S}_R\mathcal{O}$), particularly given that the models using ACAL occasionally require more data to converge (Tables 8 to 11).

Initially, we hypothesized that tasks with a high degree of subjectivity would benefit from ACAL strategies, especially on metrics focused on the most marginalized (worst-off) annotators. These strategies typically involve selecting an annotator whose patterns of annotation diverge from the majority, either in terms of their annotation behavior or in the semantic content of their past annotations.

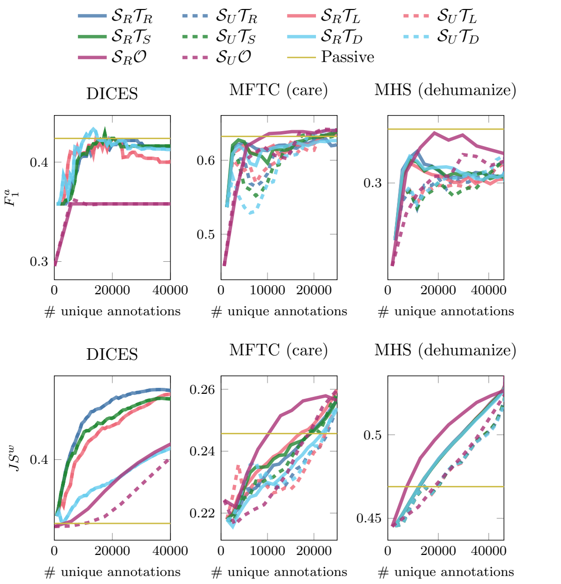

Figure 3: Selected validation set performance plots. We show progress for DICES, MFTC (care), and MHS (dehumanize) for F_1^a and JS^w.

However, as depicted in Figure 4, when examining the task of *dehumanize* (high entropy), it becomes apparent that ACAL does not consistently outperform AL. ACAL demonstrates a lower F_1^a-score than AL for this task, and on the other hand, a higher F_1^a-score for a task that is less subjective, such as *genocide*. Similarly, when evaluating *loyalty*, which involves the moral dimension with the highest disagreement among annotators, the lower 10% of annotators are more accurately approximated with PL. This suggests that the effectiveness of annotation strategies varies depending on the task's degree of subjectivity *and* available pool of annotators. The more heterogeneous the annotation behavior, indicative of a highly subjective task, the larger the pool of annotators required for each item selection. However, due to the limited annotations available per item in both datasets MFTC and MHS, even carefully selecting specific annotators may not adequately represent divergent annotation behavior in general, which challenges the generalization to unseen data. Finally, we can observe that there is a trade-off between modeling the majority of annotators equally, as reflected in the F_1^a-score and prioritizing the minority viewpoint (JS^w). A better performance in one aspect does not necessarily guarantee superiority in the other.

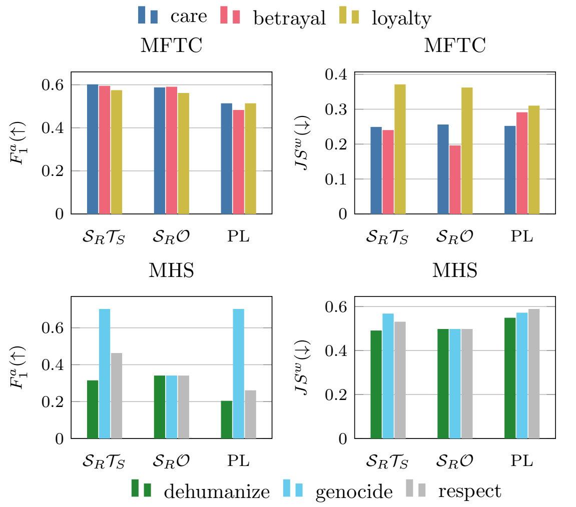

Figure 4: Relative performance across MFTC and MHS tasks, comparing one ACAL and AL approach to PL.

6 Conclusion

We introduce Annotator-Centric Active Learning (ACAL), an active learning approach that incorporates annotator selection strategies aimed at capturing label variation among annotators. We experiment with tasks across three different datasets, each leading to different ACAL behaviors. One of these datasets, DICES, is the most realistic application of ACAL since the pool of possible annotators is the largest. Here, ACAL leads to more diverse label distributions using fewer annotations. However, we find that the effectiveness of the ACAL paradigm is contingent on data characteristics. These characteristics include the number of annotations per sample, the number of annotations per annotator, and the nature of disagreement in the task annotations. Our analysis shows that we can use these conditions to help explain the often disappointing results for AL in NLP applications.

Including annotator-centric evaluation reveals how methods with similar averaged performance deal with different levels of disagreement among annotators. We show that evaluation can be enhanced by focusing on individual annotators, as there is a large gap between conventional, averaged, and worst-off performance. Furthermore, many aspects of our ACAL approach can be experimented with, e.g. by swapping the order in which samples are selected (in our case first) and annotators (second), or investigating the impact of including annotator-specific demographic information, as it is inconsistently predictive of annotation behavior (Orlikowski et al., 2023; Beck et al., 2024).

Limitations

The main limitation of this work is that the experiments are based on simulated active learning which is known to bear potential issues (Margatina and Aletras, 2023). In our study, a primary challenge arises with two of the datasets (MFTC, MHS), which, despite having a large pool of annotators, lack annotations from every annotator for each item. Consequently, in real-world scenarios, the annotator selection strategies for these datasets would benefit from access to a more extensive pool of annotators. This limitation likely contributes to the underperformance of ACAL on these datasets compared to DICES. We emphasize the need for more datasets that feature a greater number of annotations per item, as this would significantly enhance research efforts aimed at modeling human disagreement.

Since we evaluate four different annotator selection strategies and two sample selection strategies across three datasets and seven tasks, the amount of experiments is high. This did not allow for further investigation of the difference using different classification models, the extensive turning of hyperparameters, or even different training paradigms. Lastly, a limitation of our annotator selection strategies is that they rely on a small annotation history. This is why we require a warmup phase for some of the strategies, for which we decided to take a random sample of annotations. Incorporating more informed warmup strategies or incorporating ACAL strategies that do not rely on annotator history may positively impact its performance and data efficiency.

Ethical Considerations

Our goal is to approximate a good representation of human judgments over subjective tasks. We want to highlight the fact that the *performance* of the models differs a lot depending on which metric is used. We tried to account for a less majority-focussed view when evaluating the models which is very important, especially for more human-centered applications, such as hate-speech detection. However, the evaluation metrics we use do not fully capture the diversity of human judgments. The selection of metrics should align with the specific goals and motivations of the application, and there is a pressing need to develop more metrics to accurately reflect human variability in these tasks.

Our experiments are conducted on English datasets due to the scarcity of unaggregated datasets in other languages. In principle, ACAL can be applied to other languages (given the availability of multilingual models to semantically embed textual items for some particular strategies used in this work). We encourage the community to enrich the dataset landscape by incorporating more perspective-oriented datasets in various languages, ACAL potentially offers a more efficient method for creating such datasets in real-world scenarios.

References

Lora Aroyo, Alex S Taylor, Mark Diaz, Christopher M Homan, Alicia Parrish, Greg Serapio-Garcia, Vinodkumar Prabhakaran, and Ding Wang. 2023. Dices dataset: Diversity in conversational ai evaluation for safety. *arXiv preprint arXiv:2306.11247*.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

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

Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, et al. 2022. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35:38176–38189.

Connor Baumler, Anna Sotnikova, and Hal Daumé III. 2023. Which Examples Should be Multiply Annotated? Active Learning When Annotators May Disagree. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10352–10371. ACL.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2023. How (Not) to Use Sociodemographic Information for Subjective NLP Tasks. In *ArXiv*.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615, St. Julian's, Malta. Association for Computational Linguistics.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6860–6868.

Katherine M Collins, Umang Bhatt, and Adrian Weller. 2022. Eliciting and learning with soft labels from every annotator. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, pages 40–52.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Liat Ein Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active learning for bert: an empirical study. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 7949–7962.

Alexander Erdmann, David Joseph Wrisley, Benjamin Allen, Christopher Brown, Sophie Cohen-Bodénès, Micha Elsner, Yukun Feng, Brian Joseph, Béatrice Joyeux-Prunel, and Marie Catherine de Marneffe. 2019. Practical, Efficient, and Customizable Active Learning for Named Entity Recognition in the Digital Humanities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '19, pages 2223–2234, Minneapolis, Minnesota, USA. ACL.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, Massimo Poesio, et al. 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*. Association for Computational Linguistics.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. In *Advances in Experimental Social Psychology*, volume 47, pages 55–130. Elsevier, Amsterdam, the Netherlands.

Cornelia Gruber, Katharina Hechinger, Matthias Assenmacher, Göran Kauermann, and Barbara Plank. 2024. More labels or cases? assessing label variation in natural language inference. In *Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language*, pages 22–32, Malta. Association for Computational Linguistics.

Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, Gabriela Moreno, Christina Park, Tingyee E. Chang, Jenna Chin, Christian Leong, Jun Yen Leung, Arineh Mirinjian, and Morteza Dehghani. 2020. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11:1057–1071.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with mace. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*.

Kamil Kanclerz, Konrad Karanowski, Julita Bielaniewicz, Marcin Gruza, Piotr Miłkowski, Jan Kocoń, and Przemyslaw Kazienko. 2023. Pals: Personalized active learning for subjective tasks in nlp. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13326–13341.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Katerina Margatina and Nikolaos Aletras. 2023. On the limitations of simulating active learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4402–4419, Toronto, Canada. Association for Computational Linguistics.

Negar Mokhberian, Myrl G Marmarelis, Frederic R Hopp, Valerio Basile, Fred Morstatter, and Kristina Lerman. 2023. Capturing perspectives of crowdsourced annotators in subjective learning tasks. *arXiv preprint arXiv:2311.09743*.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? *Advances in neural information processing systems*, 32.

Yixin Nie, 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, Online. Association for Computational Linguistics.

Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. 2023. The Ecological Fallacy in Annotation: Modeling Human Label Variation goes beyond Sociodemographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics Volume 2: Short Papers*, pages 1017–1029. ACL.

Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. 2019. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9617–9626.

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. 1973. *A Theory of Justice*. Oxford University Press, Oxford.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. 2021. A Survey of Deep Active Learning. *ACM Computing Surveys*, 54(9):1–40.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia. Association for Computational Linguistics.

Burr Settles. 2012. *Active Learning*. Morgan & Claypool.

Dapeng Tao, Jun Cheng, Zhengtao Yu, Kun Yue, and Lizhen Wang. 2018. Domain-weighted majority voting for crowdsourcing. *IEEE transactions on neural networks and learning systems*, 30(1):163–174.

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

Michiel van der Meer, Piek Vossen, Catholijn M Jonker, and Pradeep K Murukannaiah. 2024. An empirical analysis of diversity in argument summarization. In *(To appear) Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*.

Xinpeng Wang and Barbara Plank. 2023. Actor: Active learning with annotator-specific classification heads to embrace human label variation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2046–2052.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Ye Zhang, Matthew Lease, and Byron C. Wallace. 2017. Active Discriminative Text Representation Learning. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 3386–3392, San Francisco, California, USA.

Zhisong Zhang, Emma Strubell, and Eduard Hovy. 2022. A Survey of Active Learning for Natural Language Processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, EMNLP '22, pages 6166–6190. ACL.

Yuekai Zhao, Haoran Zhang, Shuchang Zhou, and Zhihua Zhang. 2020. Active Learning Approaches to Enhancing Neural Machine Translation. In *Findings of the Association for Computational Linguistics*, EMNLP 2020, pages 1796–1806, Online. ACL.

A Detailed Experimental Setup

A.1 Cross validation details

We split the data on samples, meaning that all annotations for any given sample are completely contained in each separate split.

A.2 Hyperparameters

We report the hyperparameters for training passive, AL, and ACAL in Tables 5, 6, and 7, respectively. For turning the learning rate for passive learning, on each dataset, we started with a learning rate of 1e-06 and increased it by a factor of 3 in steps until the model showed a tendency to overfit quickly (within a single epoch). All other parameters are kept on their default setting.

Parameter	**Value**
learning rate	1e-04 (constant)
max epochs	50
early stopping	3
batch size	128
weight decay	0.01

Table 5: Hyperparameters for the passive learning.

A.3 Training details

Experiments were largely run between January and April 2024. Obtaining the ACAL results for a single run takes up to an hour on a Nvidia RTX4070. For large-scale computation, our experiments were run on a cluster with heterogeneous computing infrastructure, including RTX2080 Ti, A100, and Tesla T4 GPUs. Obtaining the results of all experiments required a total of 231 training runs, combining: (1) two data sampling strategies, (2) four annotator sampling strategies, plus an additional Oracle-based AL approach, (3) a passive learning approach. Each of the above were run for (1) three folds, each with a different seed, and (2) the seven

Parameter	Dataset (task)	Value
learning rate	all	1e-05
batch size	all	128
epochs per round	all	20
num rounds	all	10
sample size	DICES	79
sample size	MFTC (care)	674
sample size	MFTC (betrayal)	1011
sample size	MFTC (loyalty)	263
sample size	MHS (dehumanize), MHS (genocide), MHS (respect)	1728

Table 6: Hyperparameters for the oracle-based active learning approaches.

Parameter	Dataset	Value
learning rate	all	1e-05
num rounds	DICES	50
num rounds	MFTC (all), MHS (all)	20
epochs per round	DICES, MHS (all)	20
epochs per round	MFTC (all)	30
sample size	DICES	792
sample size	MFTC (care)	1250
sample size	MFTC (betrayal)	1894
sample size	MFTC (loyalty)	512
sample size	MHS (dehumanize), MHS (genocide), MHS (respect)	2899

Table 7: Hyperparameters for the annotator-centric active learning approaches.

tasks across three datasets. For training all our models, we employ the AdamW optimizer (Loshchilov and Hutter, 2018). Our code is based on the Huggingface library (Wolf et al., 2019), unmodified values are taken from their defaults.

A.4 ACAL Annotator Strategy details

Some of the strategies used for selecting annotators to provide a label to a sample

$\mathcal{T}_S$ uses a sentence embedding model to represent the content that an annotator has annotated. We use `all-MiniLM-L6-v2`[1]. We select annotators that have not annotated yet (empty history) before picking from those with a history to prioritize filling the annotation history for each annotator.

$\mathcal{T}_L$ creates an average embedding for the content annotated by each annotator and selects the most different annotator. We use the same sentence embedding model as $\mathcal{T}_S$. To avoid overfitting, we perform PCA and retain only the top 10 most informative principal components for representing each annotator.

[1] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

App.	Average F_1^a	Average JS^a	Worst-off F_1^w	Worst-off JS^w	$\Delta\%$
$\mathcal{S}_R\mathcal{T}_R$	0.578	0.147	0.420	0.199	-1.6
$\mathcal{S}_R\mathcal{T}_L$	0.581	0.149	0.433	0.212	-1.6
$\mathcal{S}_R\mathcal{T}_S$	0.593	0.161	0.430	0.239	-5.0
$\mathcal{S}_R\mathcal{T}_D$	0.583	0.148	0.429	0.199	-1.6
$\mathcal{S}_U\mathcal{T}_R$	0.594	0.150	0.419	0.203	-2.5
$\mathcal{S}_U\mathcal{T}_L$	0.584	0.148	0.434	0.200	-1.3
$\mathcal{S}_U\mathcal{T}_S$	0.588	0.149	0.435	0.204	-1.0
$\mathcal{S}_U\mathcal{T}_D$	0.591	0.149	0.428	0.194	-2.6
$\mathcal{S}_R\mathcal{O}$	0.589	0.147	0.431	0.195	-48.6
$\mathcal{S}_U\mathcal{O}$	0.589	0.149	0.430	0.200	-0.0
passive	0.481	0.199	0.360	0.290	0.0

Table 8: MFTC (betrayal) annotator-centric evaluation scores. $\Delta\%$ denotes the relative change in the annotation budget with respect to passive learning.

A.5 Disagreement rates

We report the average disagreement rates per dataset and task in Figure 5, for each of the dataset and task combinations.

B Detailed Results Overview

B.1 Test set evaluation other MFTC and MHS tasks

See Table 6 for the trade-off between data efficiency and test-set performance for the two conventional metrics (JS and F_1). We include copy the earlier mentioned results for MFTC (care) and MHS (dehumanize) for convenience.

B.2 Annotator-Centric evaluation for other MFTC and MHS tasks

We show the full annotator-centric metrics results for MFTC betrayal (Table 8), MFTC loyalty (Table 9), MHS genocide (Table 10), and MHS respect (Table 11).

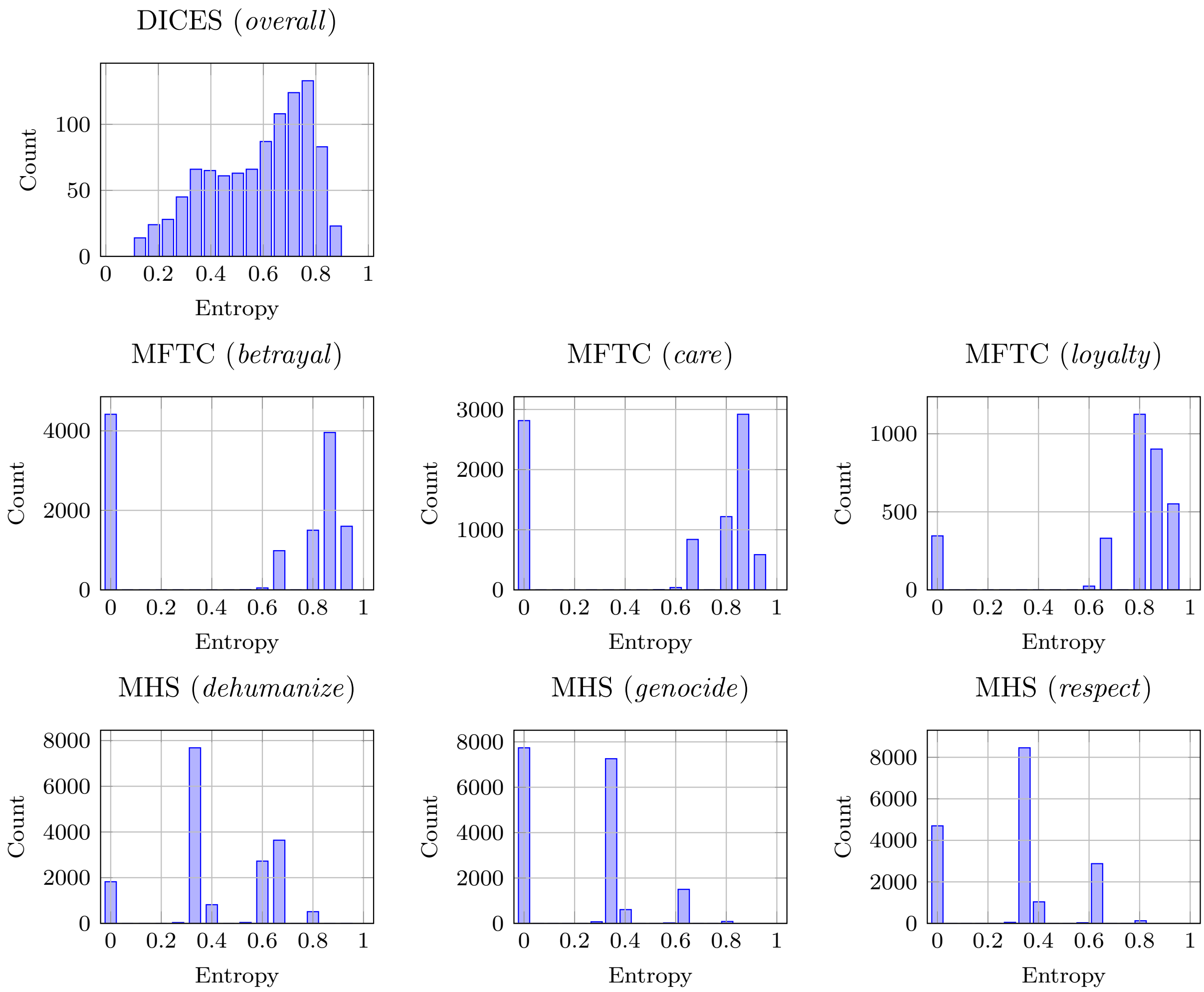

Figure 5: Histogram of entropy score over all annotations per sample for each dataset and task combination.

B.3 Training process

In our main paper, we report a condensed version of all metrics during the training phase of the active learning approaches. Below, we provide a complete overview of all approaches over all metrics. The results can be seen in Figures 7 through 13.

App.	Average F_1^a	Average JS^a	Worst-off F_1^w	Worst-off JS^w	$\Delta\%$
$\mathcal{S}_R\mathcal{T}_R$	0.564	0.177	0.222	0.372	-0.4
$\mathcal{S}_R\mathcal{T}_L$	0.563	0.176	0.222	0.374	-0.3
$\mathcal{S}_R\mathcal{T}_S$	0.573	0.176	0.222	0.370	-0.3
$\mathcal{S}_R\mathcal{T}_D$	0.551	0.175	0.222	0.373	-0.3
$\mathcal{S}_U\mathcal{T}_R$	0.557	0.177	0.217	0.357	-1.1
$\mathcal{S}_U\mathcal{T}_L$	0.541	0.177	0.222	0.355	-0.8
$\mathcal{S}_U\mathcal{T}_S$	0.556	0.177	0.222	0.358	-0.9
$\mathcal{S}_U\mathcal{T}_D$	0.558	0.177	0.222	0.358	-1.3
$\mathcal{S}_R\mathcal{O}$	0.560	0.176	0.222	0.361	-29.1
$\mathcal{S}_U\mathcal{O}$	0.559	0.177	0.222	0.366	-0.1
passive	0.512	0.183	0.261	0.309	0.0

Table 9: MFTC (loyalty) annotator-centric evaluation scores. $\Delta\%$ denotes the relative change in the annotation budget with respect to passive learning.

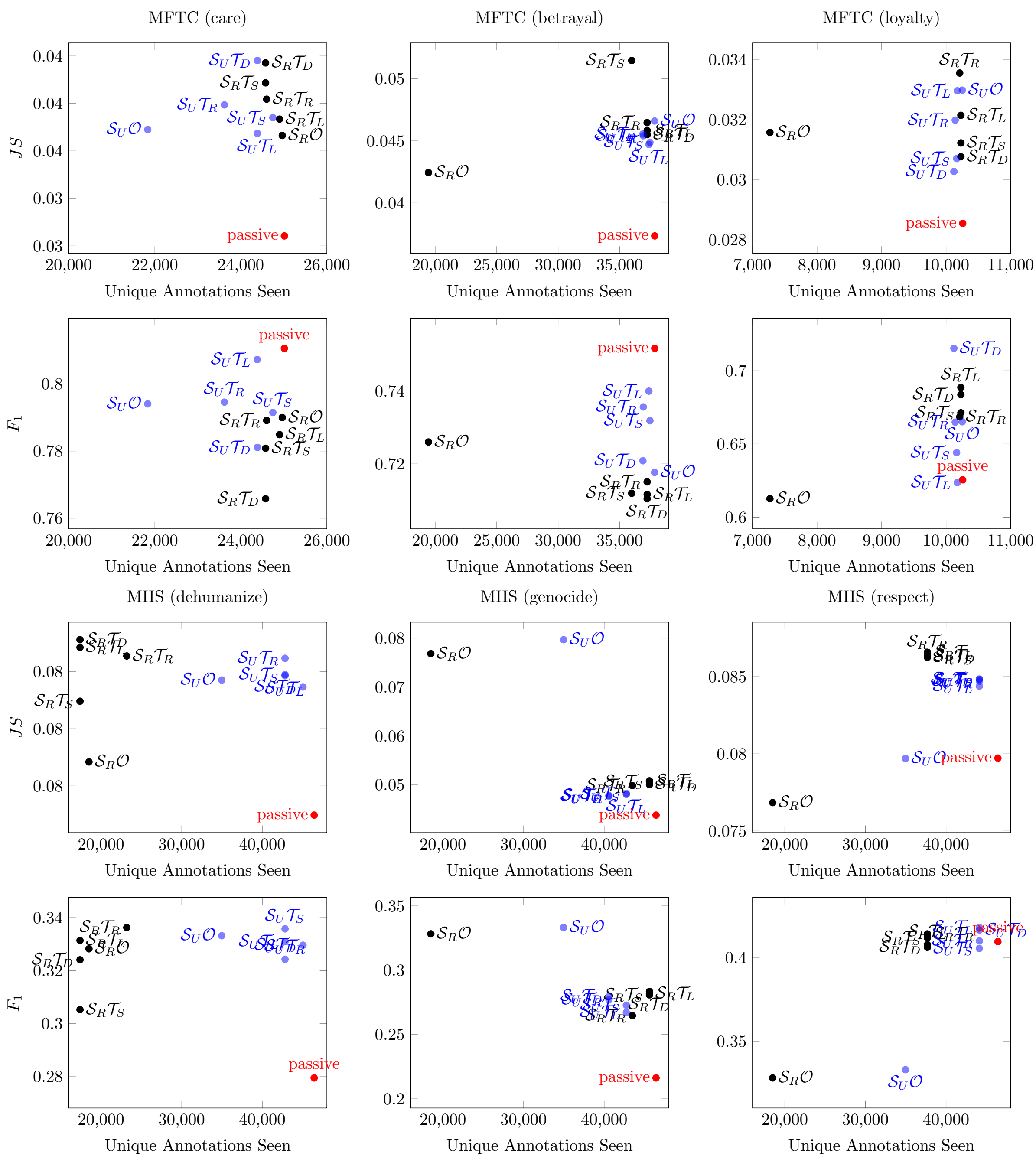

Figure 6: Test set evaluation of the ACAL, AL, and passive approaches across the extra two MFTC and MHS tasks. The leftmost column is repeated from Figure 2. For JS, strategies further to the bottom left are more data efficient (x-axis) and perform better (y-axis). For F_1, the top left contains well-performing, data-efficient approaches.

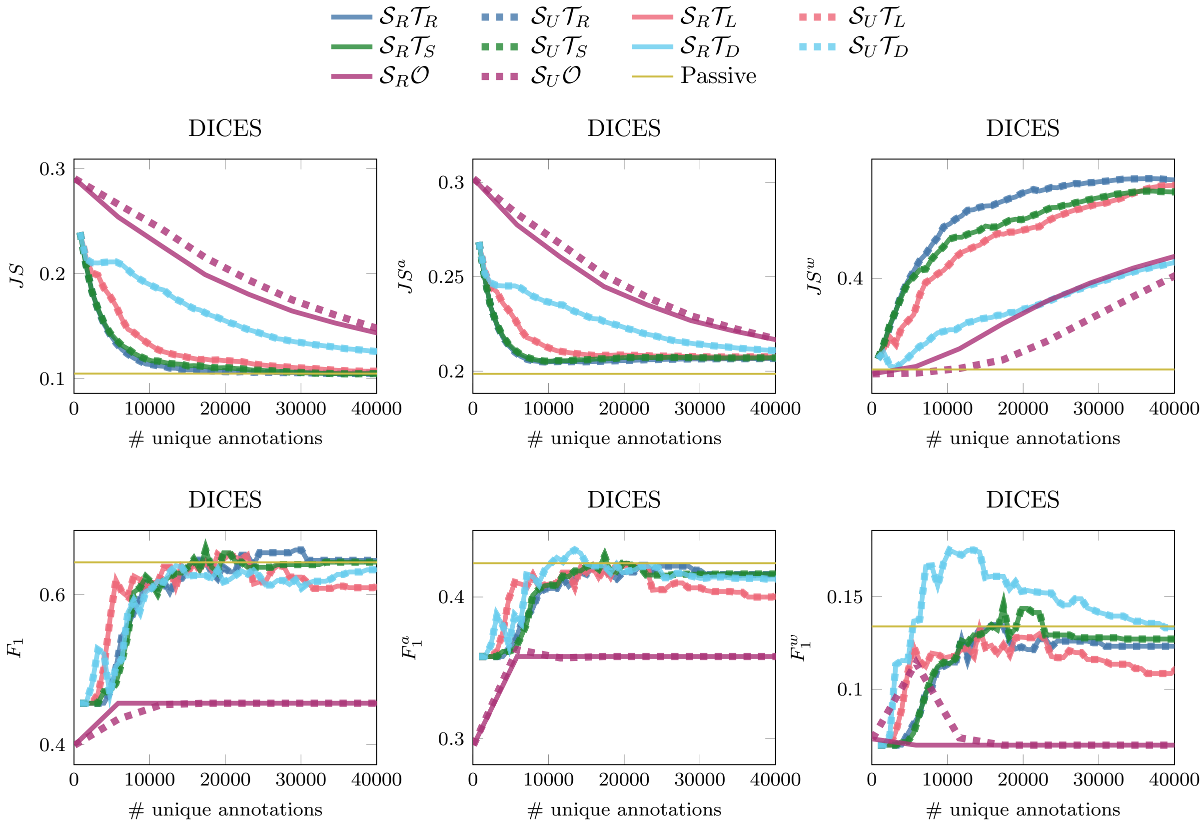

Figure 7: Validation set performance across all metrics for DICES during training.

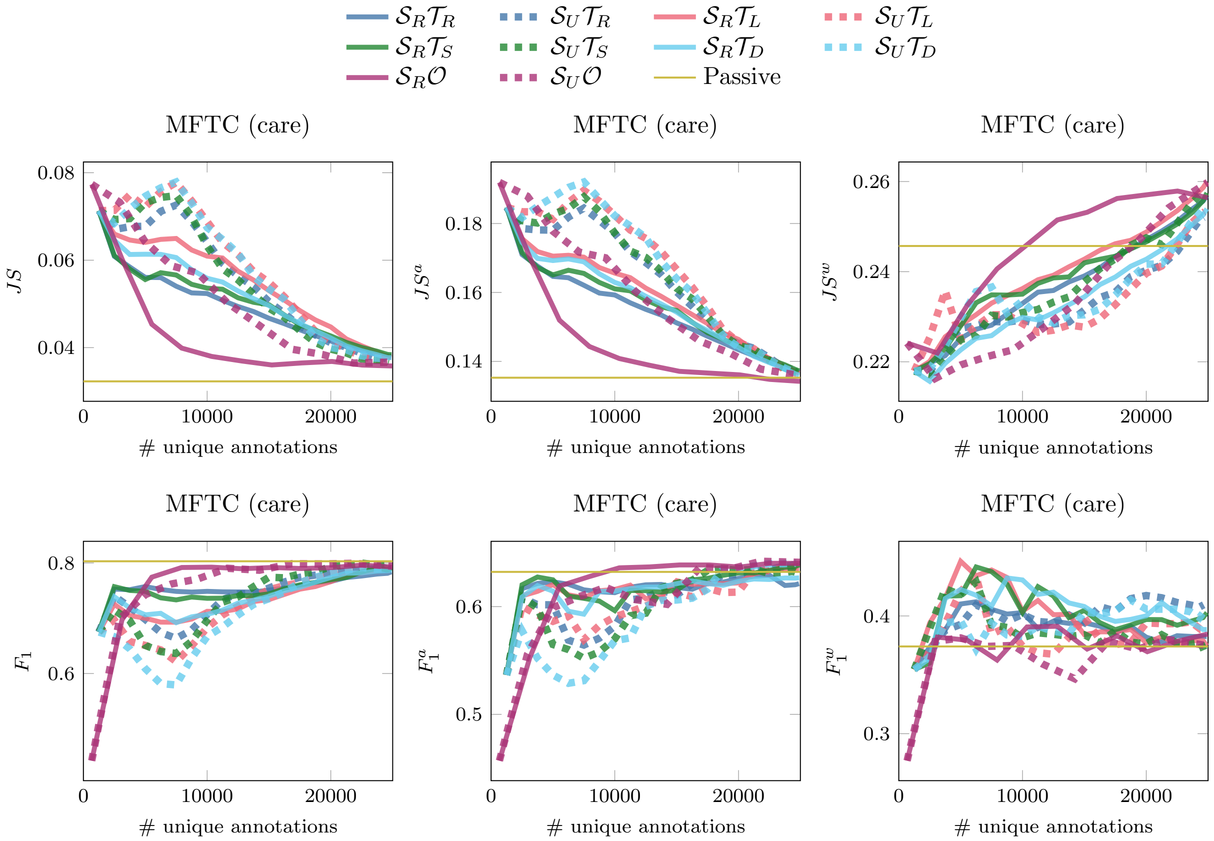

Figure 8: Validation set performance across all metrics for MFTC (care) during training

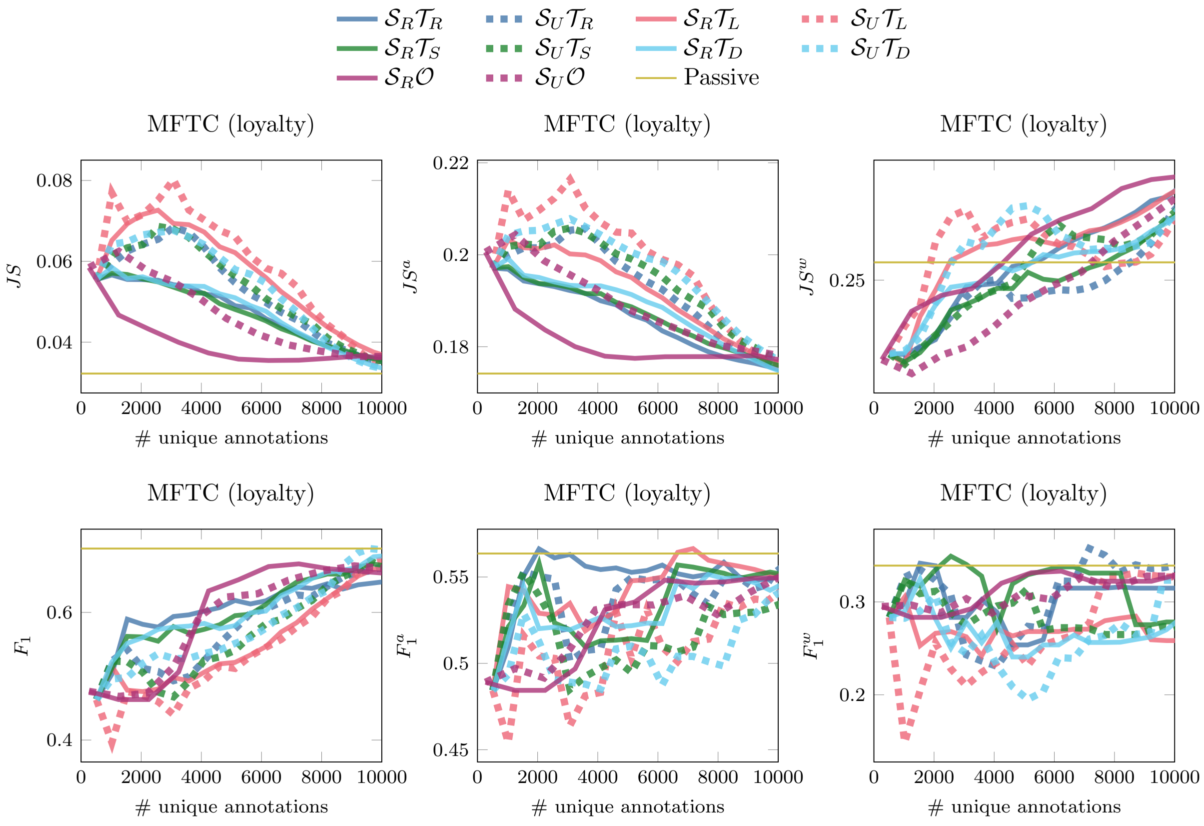

Figure 9: Validation set performance across all metrics for MFTC (loyalty) during training

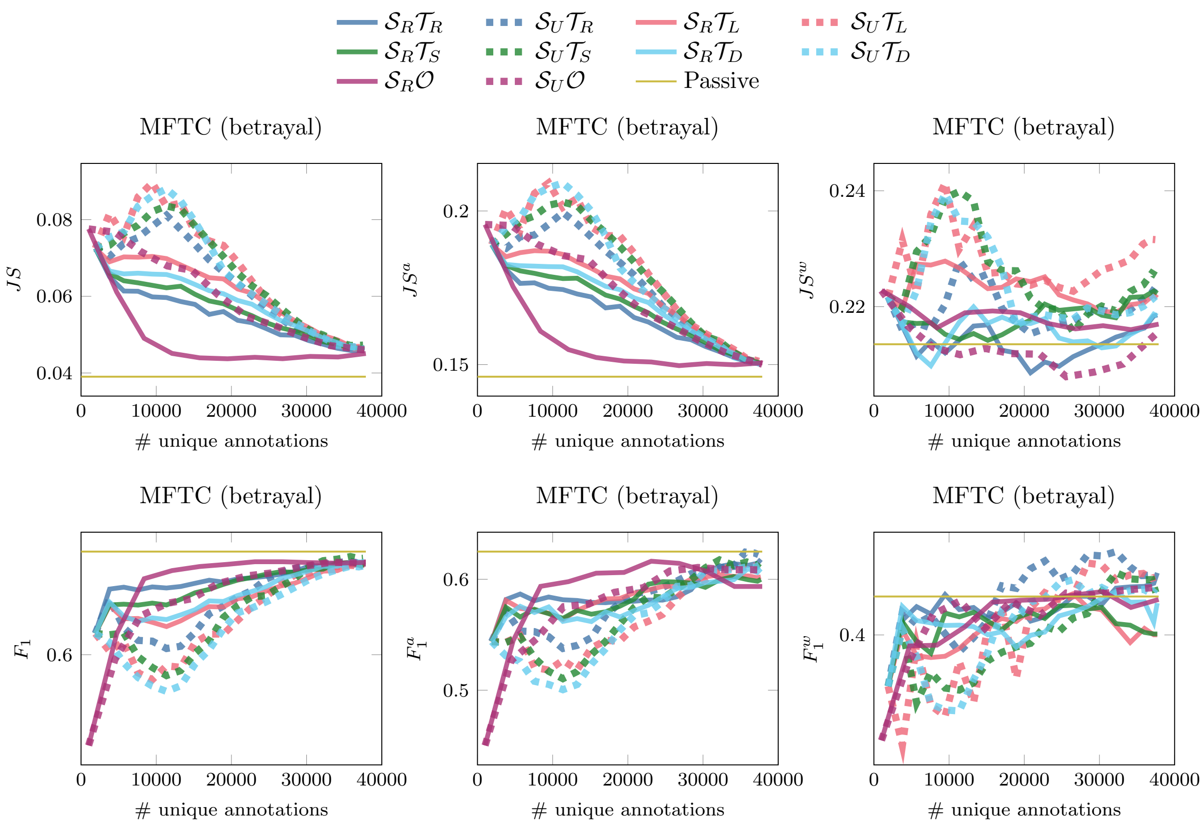

Figure 10: Validation set performance across all metrics for MFTC (betrayal) during training

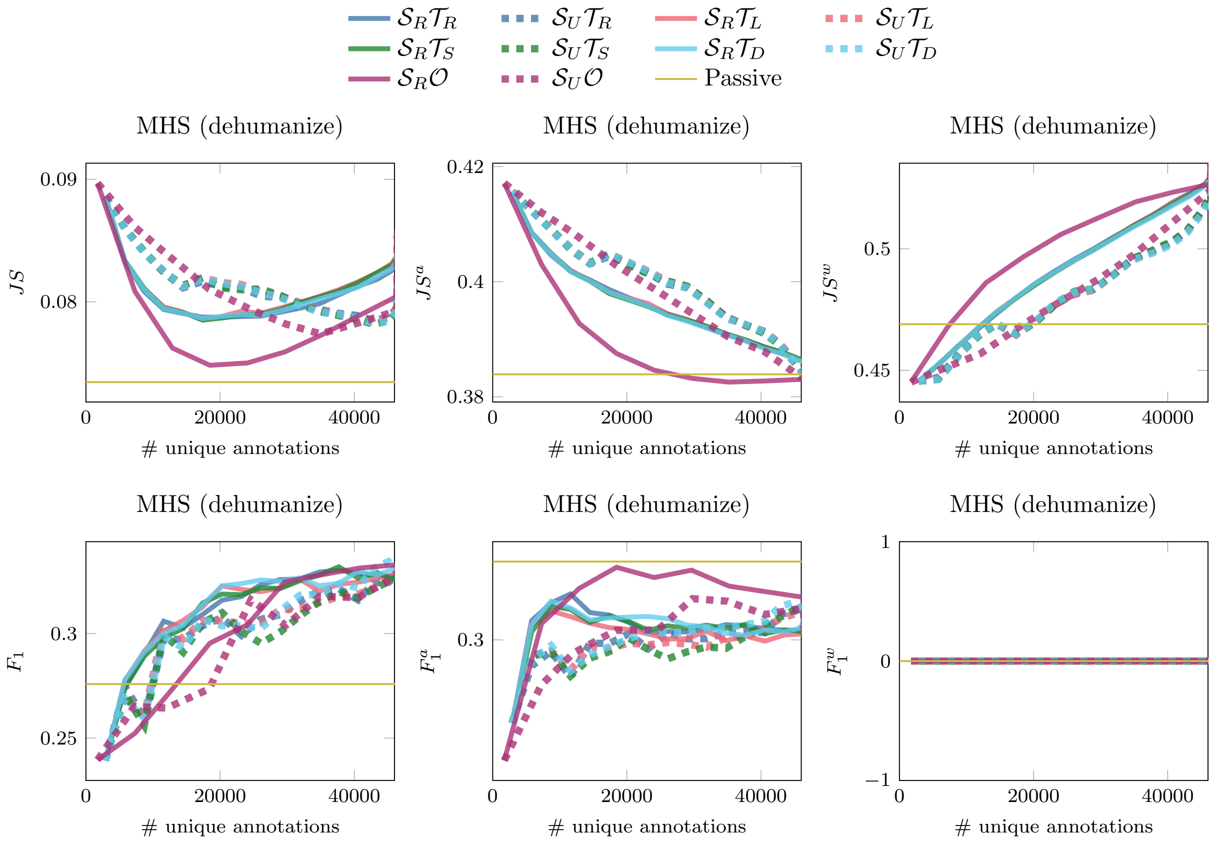

Figure 11: Validation set performance across all metrics for MHS (dehumanize) during training

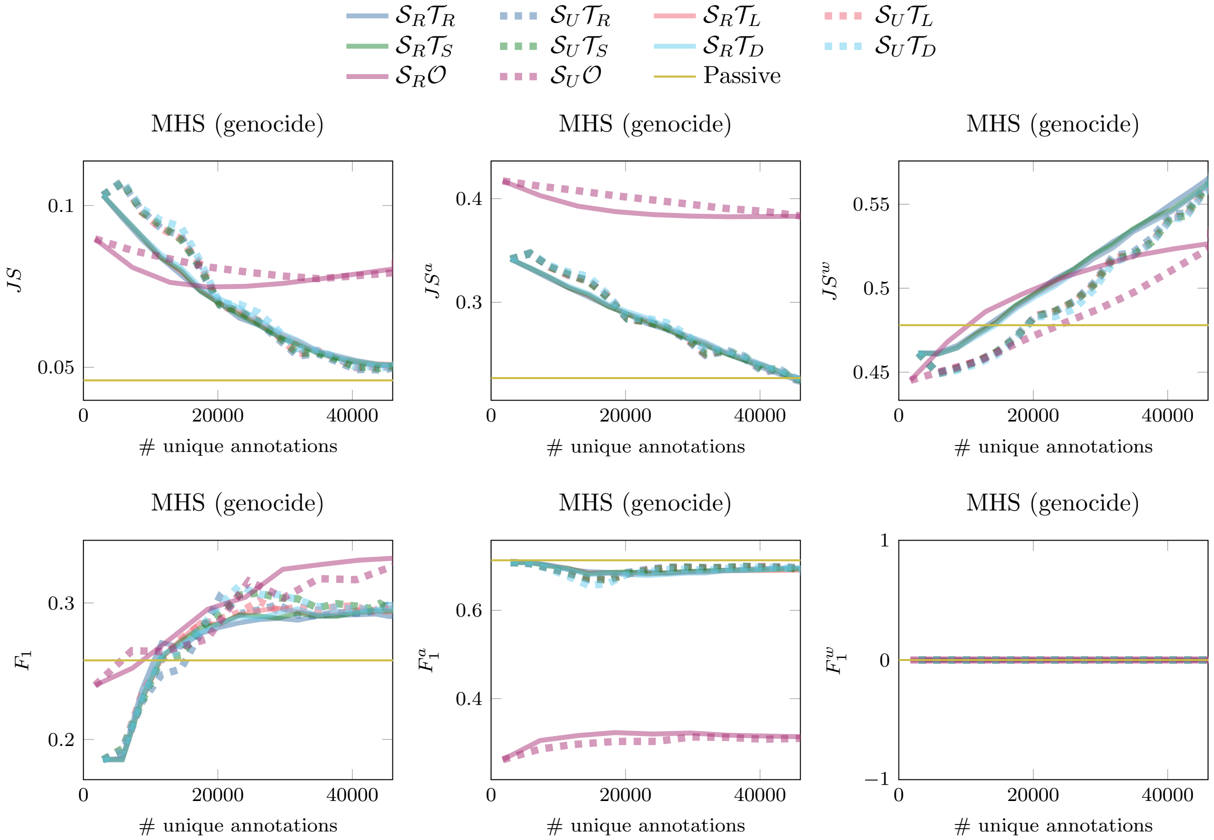

Figure 12: Validation set performance across all metrics for MHS (genocide) during training

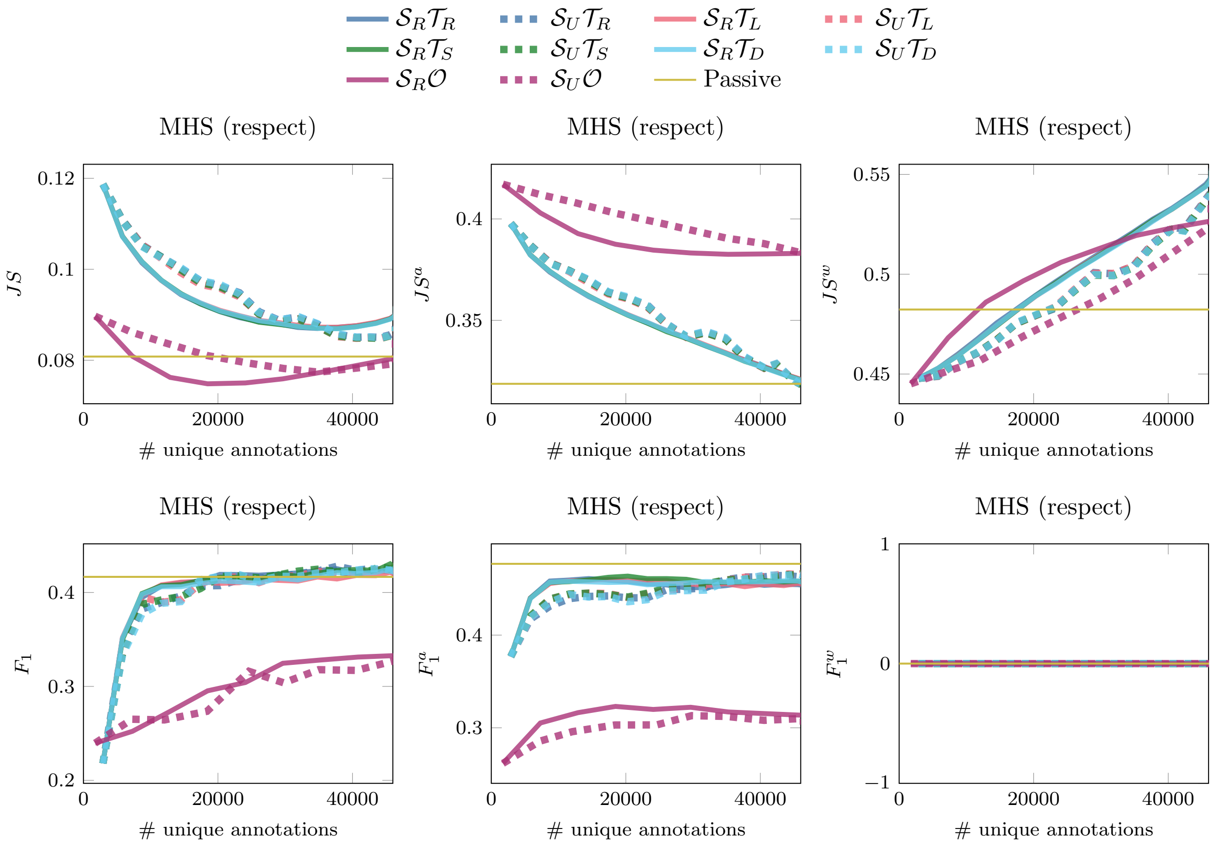

Figure 13: Validation set performance across all metrics for MHS (respect) during training

App.	Average F_1^a	Average JS^a	Worst-off F_1^w	Worst-off JS^w	Δ%
$\mathcal{S}_R\mathcal{T}_R$	0.700	0.227	0.000	0.560	-6.3
$\mathcal{S}_R\mathcal{T}_L$	0.698	0.225	0.000	0.565	-1.7
$\mathcal{S}_R\mathcal{T}_S$	0.700	0.224	0.000	0.566	-1.7
$\mathcal{S}_R\mathcal{T}_D$	0.702	0.224	0.000	0.565	-1.7
$\mathcal{S}_U\mathcal{T}_R$	0.711	0.229	0.000	0.549	-12.6
$\mathcal{S}_U\mathcal{T}_L$	0.707	0.231	0.000	0.548	-7.9
$\mathcal{S}_U\mathcal{T}_S$	0.709	0.231	0.000	0.548	-7.9
$\mathcal{S}_U\mathcal{T}_D$	0.712	0.229	0.000	0.547	-12.6
$\mathcal{S}_R\mathcal{O}$	0.339	0.387	0.000	0.496	-60.1
$\mathcal{S}_U\mathcal{O}$	0.331	0.390	0.000	0.497	-24.7
passive	0.700	0.245	0.000	0.570	–

Table 10: MHS (genocide) annotator-centric evaluation scores. Δ% denotes the relative change in the annotation budget with respect to passive learning.

App.	Average F_1^a	Average JS^a	Worst-off F_1^w	Worst-off JS^w	Δ%
$\mathcal{S}_R\mathcal{T}_R$	0.460	0.331	0.000	0.528	-18.8
$\mathcal{S}_R\mathcal{T}_L$	0.456	0.331	0.000	0.530	-18.8
$\mathcal{S}_R\mathcal{T}_S$	0.461	0.331	0.000	0.529	-18.8
$\mathcal{S}_R\mathcal{T}_D$	0.460	0.331	0.000	0.528	-18.8
$\mathcal{S}_U\mathcal{T}_R$	0.466	0.323	0.000	0.533	-4.9
$\mathcal{S}_U\mathcal{T}_L$	0.463	0.323	0.000	0.532	-4.9
$\mathcal{S}_U\mathcal{T}_S$	0.459	0.324	0.000	0.531	-4.9
$\mathcal{S}_U\mathcal{T}_D$	0.462	0.324	0.000	0.532	-4.9
$\mathcal{S}_R\mathcal{O}$	0.339	0.387	0.000	0.496	-60.1
$\mathcal{S}_U\mathcal{O}$	0.331	0.390	0.000	0.497	-24.7
passive	0.259	0.405	0.000	0.587	–

Table 11: MHS (respect) annotator-centric evaluation scores. Δ% denotes the relative change in the annotation budget with respect to passive learning.