Annotator-Centric Active Learning for Subjective NLP Tasks

Anonymous ACL submission

Abstract

Active Learning (AL) addresses the high costs of collecting human annotations by strategically annotating the most informative samples. However, for subjective NLP tasks, incorporating a wide range of perspectives in the annotation process is crucial to capture the variability in human judgments. 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). In a subjective task, a data sample is typically labeled by a set of annotators, and differences in annotation are reconciled via majority voting, resulting in a single (supposedly, true) "gold label" (Uma et al., 2021). However, this approach has been criticized for treating label variation exclusively as noise, which is especially problematic in sensitive subjective tasks (Aroyo and Welty, 2015) since it can lead to exclusion of minority voices (Leonardelli et al., 2021).

Subjectivity can be addressed by modeling the full distribution of annotations for each data sample instead of employing gold labels (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 costly annotations from a diverse set of annotators (Bakker et al., 2022).

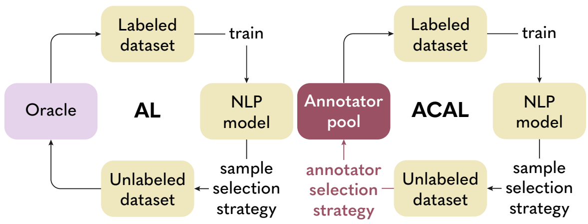

Figure 1: Active Learning (AL) approaches (left) use a sample selection strategy to pick samples to be annotated by an oracle. The Annotator-Centric Active Learning (ACAL) approach (right) extends AL by introducing an annotator selection strategy to choose the annotators who annotate the selected samples.

One way to handle 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 obtain maximally informative samples, 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. Demanding all available annotators to annotate all 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 is as critical as deciding which samples to annotate.

In most practical applications, annotators are randomly selected. This results in an annotation distribution insensitive to outlier annotators —most annotations reflect the majority voices and fewer reflect the minority voices. This may not be desirable in applications such as hate speech, where the opinion of majority and minority should be valued equally. In such cases, a more deliberate annotator selection is required. To ensure a balanced representation of majority and minority voices, we can leverage strategies inspired by Rawls' principle of fairness (Rawls, 1973), which advocates that a fair society is achieved when the well-being of the worst-off members of society (the minority annotators, in this case) is maximized.

We introduce Annotator-Centric Active Learning (ACAL) to emphasize and control who annotates which sample. In ACAL (Figure 1), the sample selection strategy of traditional AL is followed by an *annotator selection strategy*, indicating which of the available annotators should annotate each selected data sample.

Contributions (1) We present ACAL as an extension of the AL approach and introduce three annotator selection strategies aimed at collecting a balanced distribution of minority and majority annotations. (2) We introduce a suite of annotator–centric evaluation metrics to measure how individual and minority annotators are modeled. (3) We demonstrate ACAL's effectiveness in three datasets with subjective tasks—hate speech detection, moral value classification, and safety judgments.

Our experiments show that the proposed ACAL methods can approximate the distribution of human judgments similar to AL while requiring a lower annotation budget and modeling individual and minority voices more accurately. However, our evaluation shows how the task's annotator agreement and the number of available annotations impact ACAL's effectiveness—ACAL is most effective when a large pool of diverse annotators is available. Importantly, our experiments show how the ACAL framework controls how models learn to represent majority and minority annotations, which is crucial for subjective and sensitive applications.

2 Related work

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). Changing annotation 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 (Peterson et al., 2019; Müller et al., 2019; 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 behavior (Mokhberian et al., 2023), or encoding the annotator's socio-demographic information (Beck et al., 2023). 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 capture opinion diversity (Orlikowski et al., 2023).

We represent diversity based on *which* annotators annotated *what* and *how*. We experiment with 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 by judiciously choosing a few training examples (Settles, 2012). In a typical AL scenario, a large collection of unlabeled data is available, and an oracle (e.g., a human expert) is asked to annotate this unlabeled data. A *sampling strategy* is used to iteratively select the next batch of unlabeled data for annotation (Ren et al., 2021).

AL has found widespread application in 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. Existing works propose to align model and annotator uncertainties (Baumler et al., 2023), 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). These methods ignore a crucial part of learning with human variation: the diversity among annotators. We focus on selecting annotators such that they best inform 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 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):

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

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.

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

In the sample selection strategies, a batch of data of a given size B is queried at each iteration. Our experiments compare the following strategies:

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 $x_i \in U$, and selects B samples with the highest prediction entropy (the samples the model is most uncertain about).

3.3 Annotator-Centric Active Learning

ACAL builds on AL. 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 three annotator selection strategies to gather a distribution that uniformly contains all possible (majority and minority) labels, inspired by Rawls' principle of fairness (Rawls, 1973). The strategies vary in the type of information used to represent differences between annotators, including *what* or *how* the annotators have annotated thus far. Our experiments compare the following strategies:

Random ($\mathcal{T}_R$) randomly selects an annotator a_j.

Label Minority ($\mathcal{T}_L$) considers only the labels that annotators have assigned. The minority label is selected as the class with the smallest annotation count in the available dataset D_n thus far. 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.

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 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

We use three datasets which vary 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). Samples have 73 annotations on average. 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* (average count), and *betrayal* (highest count). Samples have 4 annotations on average. We create three binary classifications to predict the presence of the respective elements. As most tweets were labeled as 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. Samples have 3 annotations on average. We perform a multi-class classification with the annotated Likert scores for each task.

The datasets and tasks differ in levels of annotator agreement, measured via entropy of the annotation distribution. DICES and MHS generally have medium entropy scores, whereas the MFTC entropy is highly polarized (divided between samples with very high and very low agreement). Appendix A.5 provides details of the entropy scores.

4.2 Evaluation metrics

The ACAL strategies aim to guide the model to learn a representative distribution of the annotator's perspectives while reducing 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 model can represent the annotation distribution.

Further, since ACAL shifts the focus to annotators, we introduce novel annotator-centric evaluation metrics. First, we report the average among annotators. Second, in line with Rawls' principle of fairness, the result for the worst-off annotators:

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

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

These metrics allow us to measure the trade-offs between modeling the majority agreement, a representative distribution of annotations, and accounting for minority voices. In the next section, we describe how we obtained the results.

4.3 Training procedure

We test the annotator selection strategies proposed in Section 3.3 by comparing all combinations of the two sample selection strategies ($\mathcal{S}_R$ and $\mathcal{S}_U$) and the four annotator selection strategies ($\mathcal{T}_R$, $\mathcal{T}_L$, $\mathcal{T}_S$, and $\mathcal{T}_D$). At each iteration, we use $\mathcal{S}$ to select B unique samples from the unlabeled data pool U. We select B as 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}.

We split each dataset into 80% train, 10% validation, and 10% test. We start the training procedure with a warmup iteration of B randomly selected annotations (Zhang et al., 2022). We proceed with the ACAL iterations by combining $\mathcal{S}$ and $\mathcal{T}$. We select the model iteration that led to the best JS performance on the validation set and evaluate it on the test set. We repeat this process across three data splits and model initializations. We report the average scores on the test set. Appendix A contains additional details on training.

We compare ACAL 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 as in Alg. 1. Further, 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. Similar to ACAL, the AL and PL baselines are averaged over three seeds.

5 Results

We start by highlighting the benefits of ACAL over AL and PL (Section 5.1). Next, we closely examine ACAL on efficiency and fairness (Section 5.2). Then, we select a few cases of interest and dive deeper into the strategies' behavior during training (Section 5.3). Finally, we investigate ACAL across varying levels of subjectivity (Section 5.4).

5.1 Highlights

Our experiments show that ACAL can have a beneficial impact over using PL and AL. Figure 2 highlights two main findings: (1) ACAL strategies can more quickly learn to represent the annotation distribution with a large pool of annotators, and (2) when agreement between annotators is polarized, ACAL leads to improved results compared to learning from aggregated labels. In the next sections, we provide a deeper understanding of the conditions in which ACAL works well.

5.2 Efficiency and Fairness

Table 1 presents the results of evaluating the best models on the test set. We analyze the results along two dimensions: (a) *efficiency*: what is the impact of the different strategies on the trade-off between

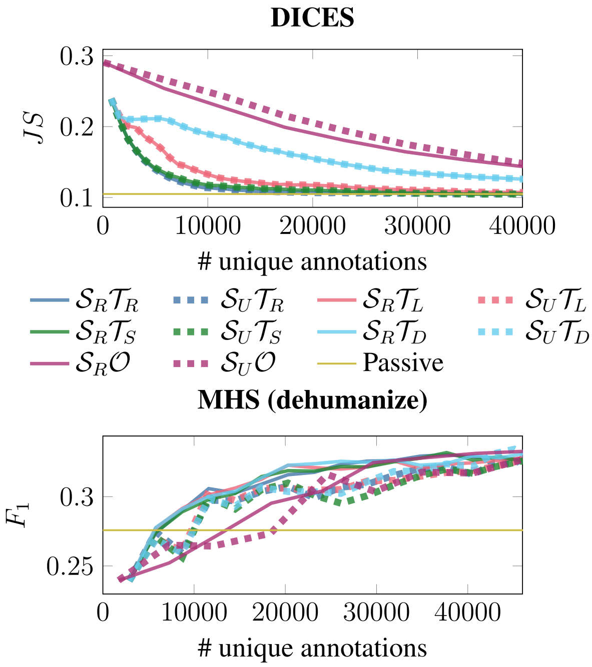

Figure 2: Learning curves showing model performance on the validation set. On DICES, ACAL approaches are quicker than AL in obtaining similar performance to passive learning. On MHS, ACAL surpasses passive learning in F_1 when data has high disagreement.

annotation budget and performance? (b) *fairness*: do the selection strategies that aim for a balanced consideration of minority and majority views lead to better performance in the human-centric evaluation metrics? For MFTC we focus on *care* because it has an average number of samples available, and for MHS we focus on *dehumanize* because it has high levels of disagreement. Appendix B presents additional results.

Efficiency We discuss the performance on F_1 and JS to measure how well the proposed strategies model label distributions and examine the used annotator budget. Across all tasks and datasets, ACAL and AL consistently yield comparable or superior F_1 and JS with a lower annotation budget than PL. When comparing ACAL with AL, the results vary depending on the task and dataset. For DICES, there is a significant benefit to using ACAL, as it can save up to ∼40% of the annotation budget while yielding better scores across all metrics than AL. With AL, we observe only a small reduction in annotation cost. For MFTC, AL with $\mathcal{S}_U$ leads to the largest cost benefits (∼12% less annotation budget), but at a cost in terms of absolute JS and F_1. ACAL slightly outperforms AL but does not lead to a decrease in annotation budget.

	App.	F_1	JS	Average F_1^a	JS^a	Worst-off F_1^w	JS^w	$\Delta\%$
DICES	$\mathcal{S}_R\mathcal{T}_R$	53.2	.100	43.2	.186	16.7	.453	-36.8
	$\mathcal{S}_R\mathcal{T}_L$	55.5	.101	42.4	.187	15.5	.450	-32.7
	$\mathcal{S}_R\mathcal{T}_S$	61.0	.103	44.2	.186	16.4	.447	-35.5
	$\mathcal{S}_R\mathcal{T}_D$	58.9	.142	43.1	.203	16.9	**.370**	-30.0
	$\mathcal{S}_U\mathcal{T}_R$	53.2	.100	43.2	.186	16.7	.453	-36.8
	$\mathcal{S}_U\mathcal{T}_L$	55.5	.101	42.4	.187	15.5	.450	-32.7
	$\mathcal{S}_U\mathcal{T}_S$	**63.1**	**.098**	**43.9**	**.187**	**18.4**	.447	**-38.2**
	$\mathcal{S}_U\mathcal{T}_D$	58.9	.142	43.1	.203	16.9	**.370**	-30.0
	$\mathcal{S}_R\mathcal{O}$	59.1	.112	41.4	.191	13.3	.425	-0.1
	$\mathcal{S}_U\mathcal{O}$	46.2	.110	38.4	.192	11.7	.427	-0.1
	PL	59.0	.105	37.1	.211	12.3	.479	–
MFTC (*care*)	$\mathcal{S}_R\mathcal{T}_R$	78.9	.038	61.1	**.141**	37.7	.247	-1.6
	$\mathcal{S}_R\mathcal{T}_L$	78.5	.037	**61.6**	.142	39.2	.249	-0.4
	$\mathcal{S}_R\mathcal{T}_S$	78.1	.039	60.0	.145	35.1	.248	-1.7
	$\mathcal{S}_R\mathcal{T}_D$	76.6	.040	60.4	.144	35.7	**.243**	-1.7
	$\mathcal{S}_U\mathcal{T}_R$	79.4	.038	61.2	.143	37.7	.252	-5.6
	$\mathcal{S}_U\mathcal{T}_L$	80.7	.037	58.9	.142	**42.3**	.248	-2.5
	$\mathcal{S}_U\mathcal{T}_S$	79.1	.037	60.8	.143	39.9	.258	-1.1
	$\mathcal{S}_U\mathcal{T}_D$	78.1	.040	58.6	.145	35.7	.253	-2.5
	$\mathcal{S}_R\mathcal{O}$	79.0	.037	58.6	**.141**	39.2	.255	-0.2
	$\mathcal{S}_U\mathcal{O}$	79.4	.037	58.3	.144	35.7	.253	**-12.7**
	PL	**81.1**	**.032**	51.2	.179	37.7	.251	–
MHS (*dehumanize*)	$\mathcal{S}_R\mathcal{T}_R$	**33.6**	.081	31.5	.394	0.0	.489	-50.0
	$\mathcal{S}_R\mathcal{T}_L$	33.1	.081	32.2	.397	0.0	**.478**	**-62.5**
	$\mathcal{S}_R\mathcal{T}_S$	30.5	.079	31.3	.397	0.0	.480	**-62.5**
	$\mathcal{S}_R\mathcal{T}_D$	32.4	.081	31.8	.398	0.0	.479	**-62.5**
	$\mathcal{S}_U\mathcal{T}_R$	32.4	.080	32.2	.389	0.0	.508	-7.8
	$\mathcal{S}_U\mathcal{T}_L$	33.1	.080	32.8	.388	0.0	.507	-7.8
	$\mathcal{S}_U\mathcal{T}_S$	**33.6**	.080	32.6	.388	0.0	.506	-7.8
	$\mathcal{S}_U\mathcal{T}_D$	33.0	.079	32.6	**.384**	0.0	.513	-3.0
	$\mathcal{S}_R\mathcal{O}$	32.8	.077	**33.9**	.387	0.0	.496	-60.1
	$\mathcal{S}_U\mathcal{O}$	33.3	.080	33.1	.390	0.0	.497	-24.7
	PL	28.0	**.075**	20.2	.424	0.0	.547	–

Table 1: Test set results on the DICES, MFTC (*care*), and MHS (*dehumanize*) datasets. $\Delta\%$ denotes the reduction in the annotation budget with respect to passive learning. In bold, the best performance per column and per dataset (higher F_1 are better, lower JS are better).

For MHS, both AL and ACAL significantly reduce the annotation cost ($\sim$60%) while yielding better scores than PL—however, AL and ACAL do not show substantial performance differences.

Overall, we conclude that ACAL is most efficient when the pool of available annotators for one sample is large (as with the DICES dataset), whereas the difference between ACAL and AL is negligible with a small pool of annotators per data sample (as with MFTC and MHS).

Fairness We investigate the extent to which the models represent individual annotators fairly and capture minority opinions via the annotator-centric evaluation metrics (F_1^a, JS^a, F_1^w, and JS_w). We observe a substantial improvement when using AL or ACAL over PL. Further, we observe no single winner-takes-all approach: high F_1 and JS scores do not consistently cooccur with high scores for the annotator-centric metrics. This highlights the need for a more comprehensive evaluation to assess models for subjective tasks. We observe that ACAL slightly outperforms AL in modeling individual annotators (JS^a and F_1^a). This trend is particularly evident with DICES, again likely due to the large pool of annotators available per data sample. Lastly, ACAL is best in the worst-off metrics (JS^w and F_1^w), showing the ability to better represent minority opinions as a direct consequence of the proposed annotator selection strategies on DICES and MFTC. However, all approaches score 0 for F_1^w on MHS. This is due to the high disagreement in this dataset: the 10% worst-off annotators always disagree with a hard label derived from the predicted label distribution.

In conclusion, our experiments show that, when a large pool of annotators is available, a targeted sampling of annotators requires fewer annotations and is fairer. That is, minority opinions are better represented without large sacrifices in performance compared to the overall label distribution.

5.3 Convergence

The evaluation on the test set paints a general picture of the advantage of using ACAL over AL or PL. In this section, we assess how different ACAL strategies converge over iterations. We describe the major patterns across our experiments by analyzing six examples of interest with F_1^a and JS^w (Figure 3). We select F_1^a because it reveals how well individual annotators are modeled on average, and JS^w to measure how strategies deviate from modeling the majority perspective. Appendix B.2 provides an overview of all metrics.

First, we notice that the trends for F_1^a and JS^w are both increasing—the first is expected, but the second requires an explanation. As the model is exposed to more annotations over the training iterations, the predicted label distribution starts to fit the true label distribution. However, here we consider each annotator individually: JS^w reports the average of the 10% lowest JS scores per annotator. The presence of disagreement implies the existence of annotators that annotate differently from the majority. Since our models predict the full distribution, they assign a proportional probability to dissenting annotators. Thus, learning to model the full distribution of annotations leads to

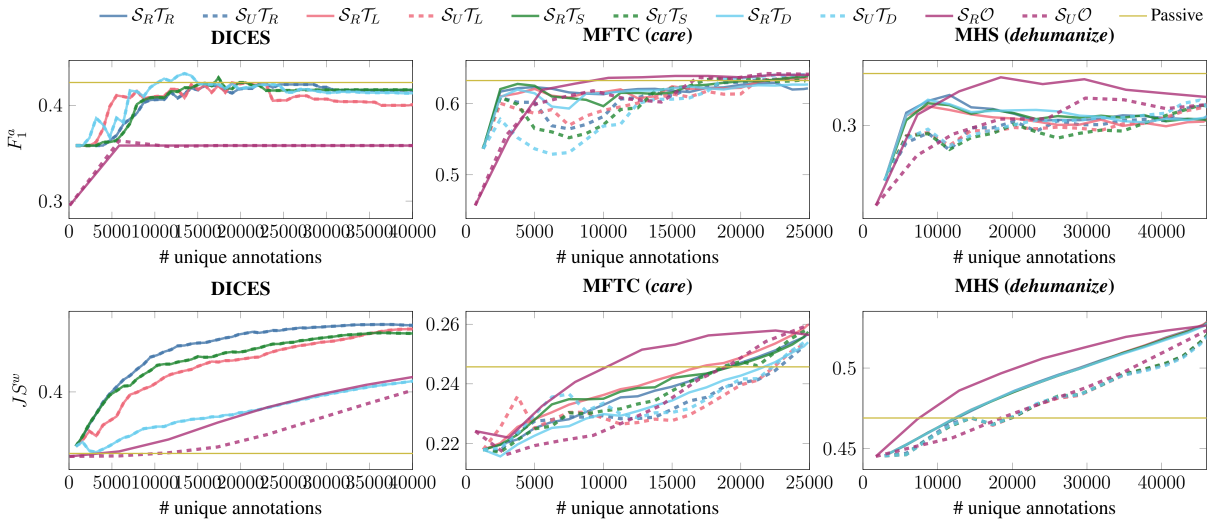

Figure 3: Selected plots showing the F_1^a and JS^w performance on the validation set through the ACAL and AL iterations for DICES, MFTC (*care*), and MHS (*dehumanize*). Higher F_1^a is better, lower JS^w is better.

an increase in JS^w.

Second, we notice a difference between ACAL and AL. On MFTC and MHS, ACAL, compared to AL, yields overall smaller JS^w at the cost of a slower convergence in F_1^a, showing the trade-off between modeling all annotators and representing minorities. However, with DICES the trend is the opposite. This is due to AL having access to the complete label distribution: it can model a balanced distribution, leading to lower worst-off performance. With a large number of annotations, ACAL requires more iterations to get the same balanced predicted distribution.

Third, we observe differences among the annotator selection strategies ($\mathcal{T}$). $\mathcal{T}_D$ shows the most differences—both JS^w and F_1^a increase slower than for the other strategies. This suggests that selecting annotators based on the average embedding of the annotated content strongest emphasizes diverging label behavior.

Finally, we analyze the impact of the sample selection strategies ($\mathcal{S}$, dotted vs. solid lines in Figure 3). For DICES, $\mathcal{S}_R$ and $\mathcal{S}_U$ lead to comparable results, likely due to the low number of samples. Using $\mathcal{S}_U$ in MFTC leads to F_1^a performance decreasing at the start of training. The strategy prioritizes obtaining annotations for already added samples to lower their entropy, while the variation in labels is irreconcilable (since there are limited labels available, and they are in disagreement). We see a similar pattern for MHS.

These results further underline our main finding that ACAL is effective in representing diverse annotation perspectives when there is a (1) heterogeneous pool of annotators, and (2) a task that facilitates human label variation.

5.4 Impact of subjectivity

We further investigate ACAL strategies on (1) label entropy, and (2) cross-task performance.

Alignment of ACAL strategies during training We want to investigate how well the ACAL strategies align with the overall subjective annotations: do they drive the model entropy in the right direction? We measure the entropy of the samples in the labeled training set at each iteration and compare it to the actual entropy of those samples. Higher entropy suggests that the selection strategy overestimates uncertainty. Lower entropy indicates that the model may not sufficiently account for disagreement. When the entropy matches the true entropy, the selection strategy is well-calibrated. We focus on DICES as a case study due to the wide range of entropy scores. We group each sample based on the true label entropy into low, medium, and high[1]. We apply the same categorization at each training iteration for samples labeled thus far. Subsequently, we plot the proportion of data points for which the selection strategy results in excessively high or excessively low entropy.

Figure 4 visualizes the proportions. At the beginning of training, entropy is generally low because samples have few annotations. Over time, the selected annotations better align with the true entropy.

[1]Entropy bins: low (< 0.43), medium ($0.43 - 0.72$) high (> 0.72).

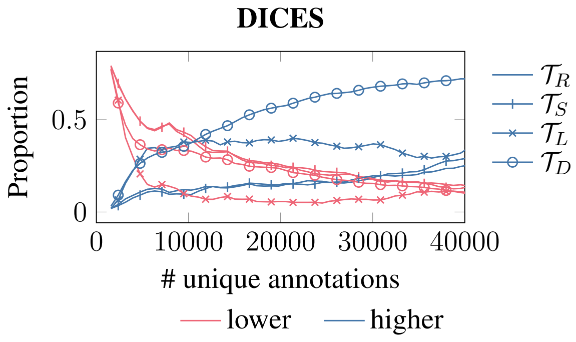

Figure 4: Proportion of data samples that result in higher or lower entropy than the target label distribution per ACAL strategy.

However, when and how much strategies succeeded in representing the true label distribution differs: $\mathcal{T}_S$ and $\mathcal{T}_R$ take longer to increase label entropy than the other two strategies. They are conservative in adding diverse labels. $\mathcal{T}_L$ and $\mathcal{T}_D$ increase the proportion of well-aligned data points earlier in the training process, achieving a balanced entropy alignment sooner. However, both strategies start to overshoot the target entropy, whereas the others show a more gradual alignment with the true entropy. This effect is strongest for $\mathcal{T}_D$. This finding suggests that minority-aware annotator-selection strategies achieve the best results in the early stages of training. They are effective for quickly raising entropy but can lead to overrepresentation.

Cross-task performance Figure 5 compares the two annotator-centric metrics on the three tasks of MFTC and MHS—the datasets for which we have seen the least impact of ACAL over AL and PL. We select a data sampling ($\mathcal{S}_R$) and annotator sampling strategy ($\mathcal{T}_S$), based on its strong performance on DICES for comprehensive comparison.

When evaluating MFTC *loyalty*, which has the highest disagreement, JS^w is more accurately approximated with PL. Similarly, ACAL is outperformed by AL on F_1^a for the *dehumanize* (high disagreement) task. However, for the less subjective task *genocide*, ACAL leads to higher F_1^a. This suggests that the effectiveness of annotation strategies varies depending on the task's degree of subjectivity *and* the 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 sample selection. We also observe that there is a trade-off between modeling the majority of annotators equally (F_1^a) and prioritizing the minority (JS^w).

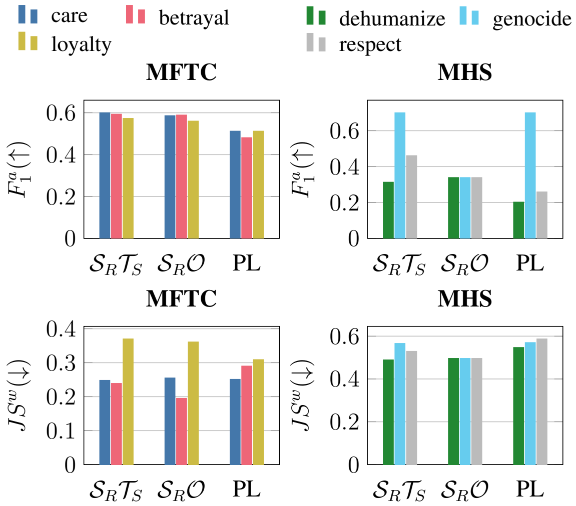

Figure 5: Comparison of ACAL, AL, and PL across different MFTC and MHS tasks. Higher F_1^a is better, and lower JS^w is better.

6 Conclusion

We present ACAL as an extension of AL to emphasize the selection of diverse annotators. We introduce three novel annotator selection strategies and four annotator-centric metrics and experiment with tasks across three different datasets. We find that the ACAL approach is especially effective in reducing the annotation budget when the pool of available annotators is large. However, its effectiveness is contingent on data characteristics such as the number of annotations per sample, the number of annotations per annotator, and the nature of disagreement in the task annotations. Furthermore, our novel evaluation metrics display the trade-off between modeling overall distributions of annotations and adequately accounting for minority voices, showing that different strategies can be tailored to meet different goals. Especially early in the training process, strategies that are aggressive in obtaining diverse labels have a beneficial impact Furthermore, we recognize that gathering a distribution that uniformly contains all possible (minority and majority) labels can be overly sensitive to small minorities or noise. Future work can integrate methods that account for noisy annotations (Weber-Genzel et al., 2024) or that strike a balance between egalitarian and utilitarian approaches (Lera-Leri et al., 2024).

Limitations

The main limitation of this work is that the experiments are based on simulated AL 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 other methods for measuring uncertainty (such as ensemble methods ()), different classification models, the extensive turning of hyperparameters, or even different training paradigms (such as low-rank adaptation ()). 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*.

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.

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.

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.

Elisa Leonardelli, 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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Roger X. Lera-Leri, Enrico Liscio, Filippo Bistaffa, Catholijn M. Jonker, Maite Lopez-Sanchez, Pradeep K. Murukannaiah, Juan A. Rodriguez-Aguilar, and Francisco Salas-Molina. 2024. Aggregating value systems for decision support. *Knowledge-Based Systems*, 287:111453.

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.

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.

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.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine de Marneffe, and Barbara Plank. 2024. Varierr nli: Separating annotation error from human label variation.

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 Dataset details

We provide an overview of the datasets used in our work in Table A1. 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 A2, A3, and A4, 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.

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 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`[2]. 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.

A.5 Disagreement rates

We report the average disagreement rates per dataset and task in Figure A1, for each of the dataset and task combinations.

B Detailed results overview

B.1 Annotator-Centric evaluation for other MFTC and MHS tasks

We show the full annotator-centric metrics results for MFTC *betrayal*, MFTC *loyalty*, MHS *genocide*, and MHS *respect* in Table B1. This follows the same format at Table 1. The results in this table also form the basis for Figure 5.

B.2 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 B1 through B7.

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

Dataset	Task (*dimension*)	# Samples	# Annotators	# Annotations	# Annotations per item
DICES	Safety Judgment	990	172	72,103	72.83
MFTC	Morality (*care*)	8,434	23	31,310	3.71
MFTC	Morality (*loyalty*)	3,288	23	12,803	3.89
MFTC	Morality (*betrayal*)	12,546	23	47,002	3.75
MHS	Hate Speech (*dehumanize*, *genocide*, *respect*)	17,282	7,807	57,980	3.35

Table A1: Overview of the datasets and tasks employed in our work.

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

Table A2: Hyperparameters for the passive learning.

Parameter	Dataset (task)	Value
learning rate	all	1e-05
batch size	all	128
epochs per round	all	20
num iterations	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 A3: Hyperparameters for the oracle-based active learning approaches.

Parameter	Dataset	Value
learning rate	all	1e-05
num iterations	DICES	50
num iterations	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 A4: Hyperparameters for the annotator-centric active learning approaches.

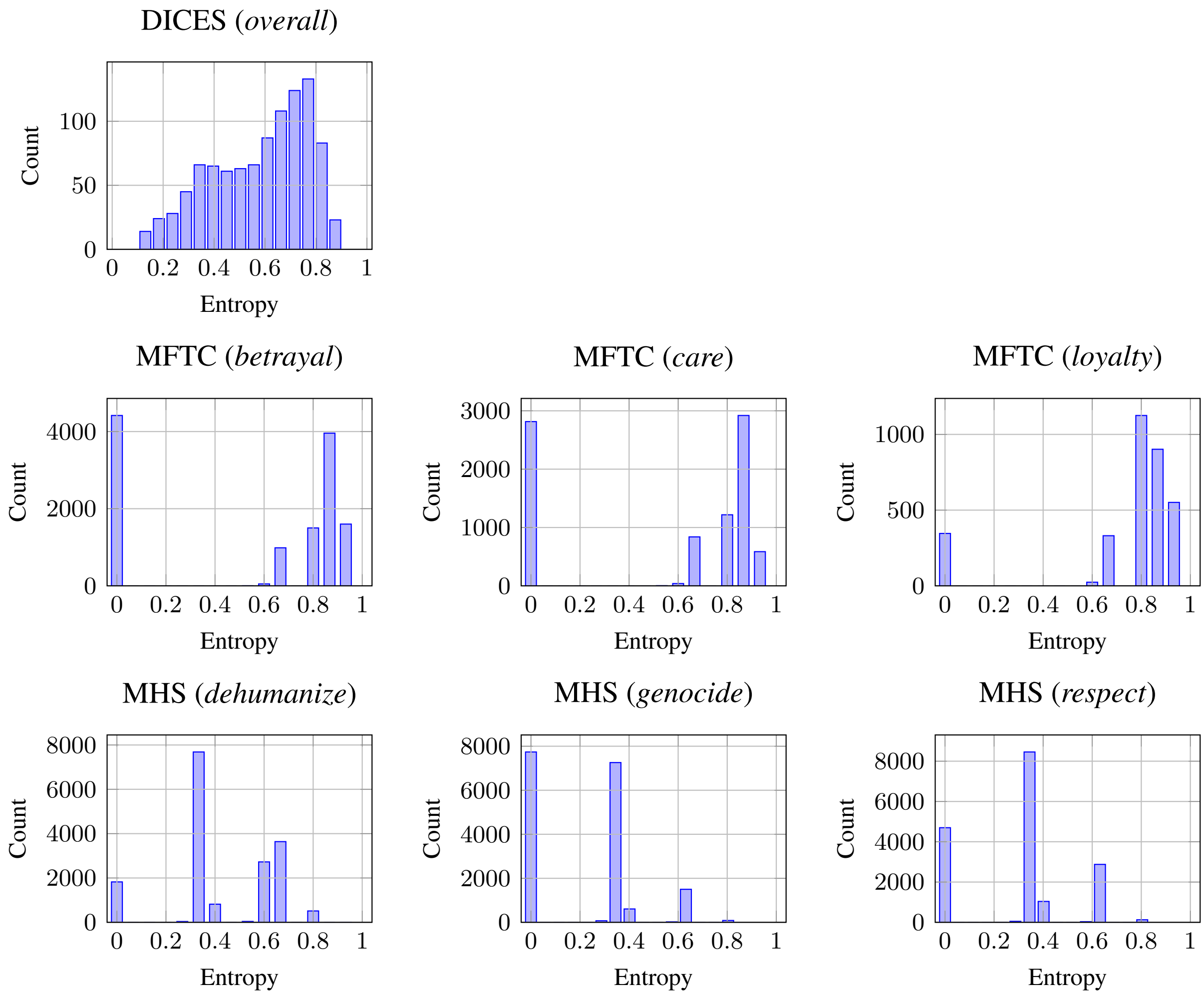

Figure A1: Histogram of entropy score over all annotations per sample for each dataset and task combination.

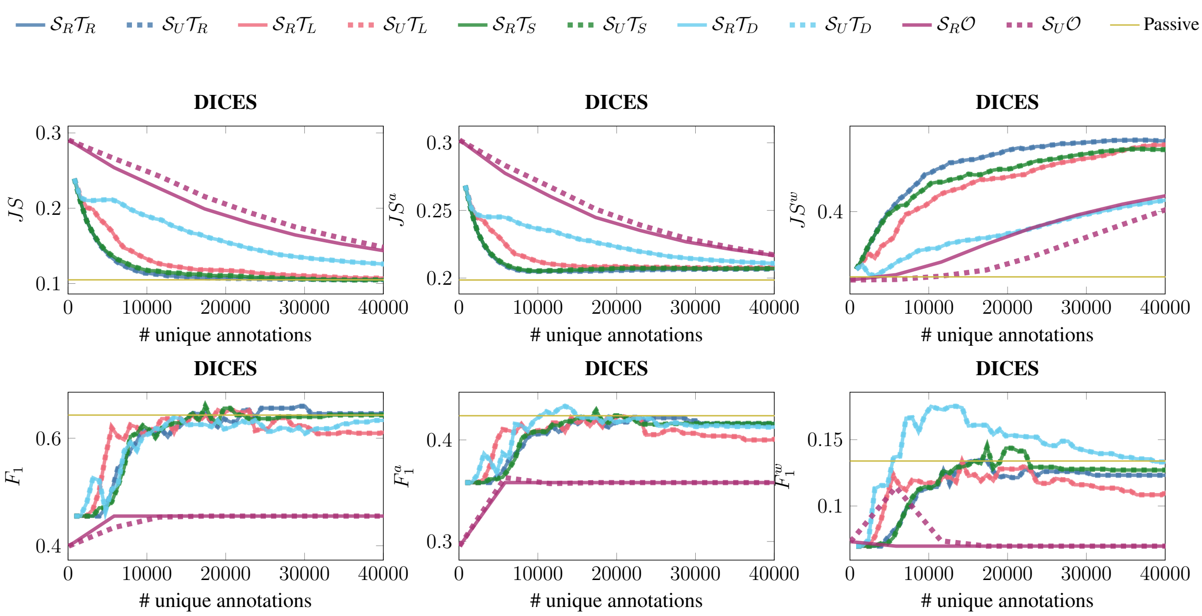

Figure B1: Validation set performance across all metrics for DICES during training.

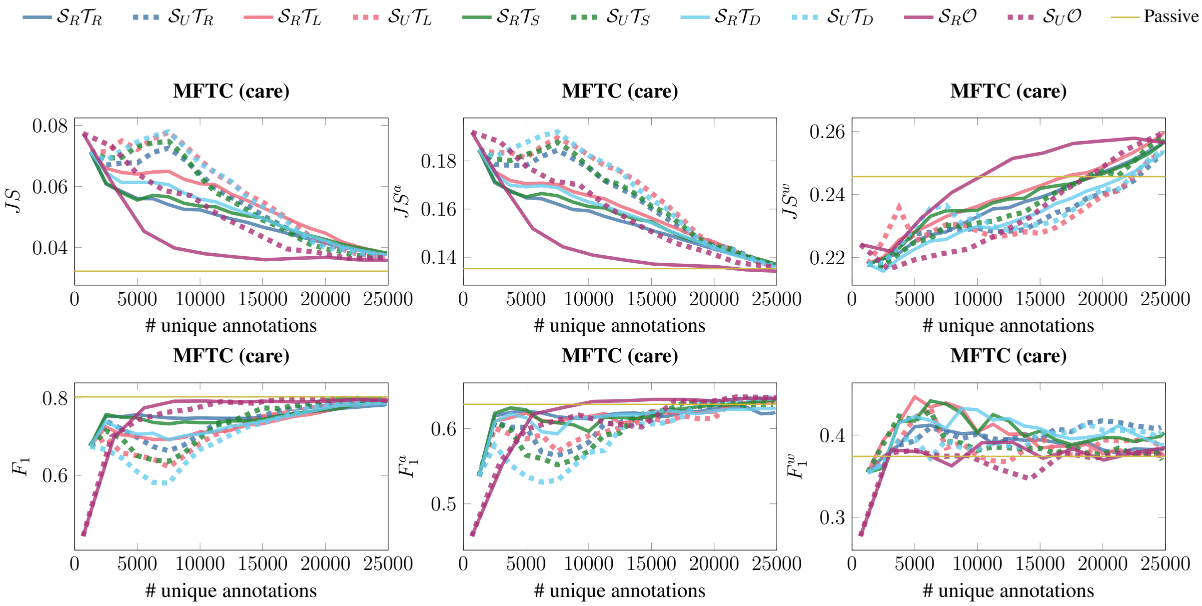

Figure B2: Validation set performance across all metrics for MFTC (care) during training

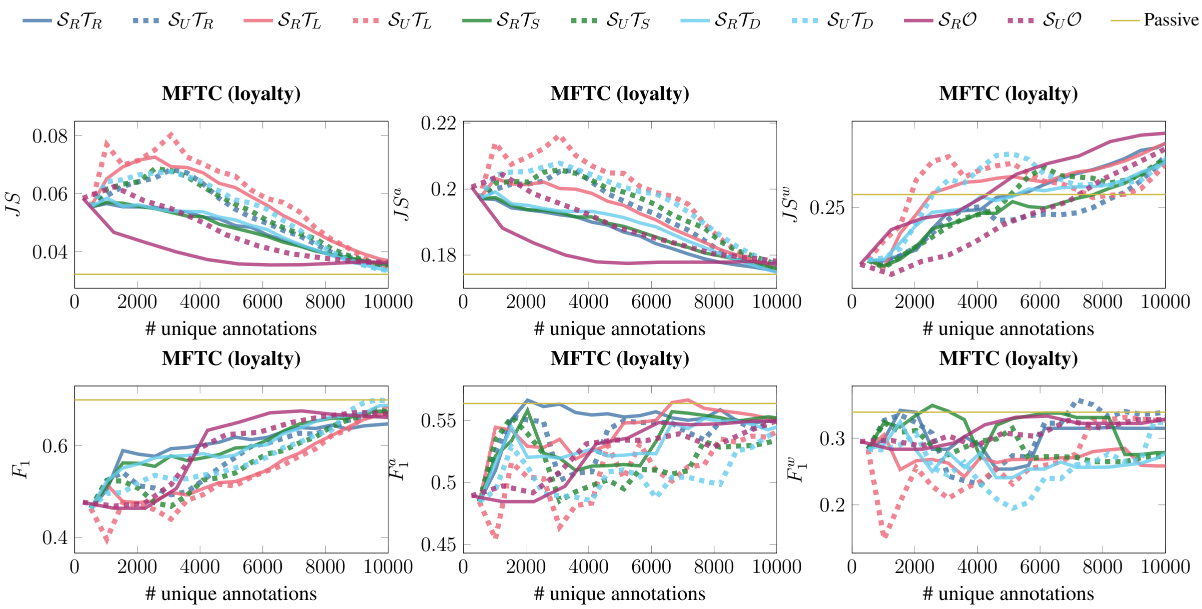

Figure B3: Validation set performance across all metrics for MFTC (loyalty) during training

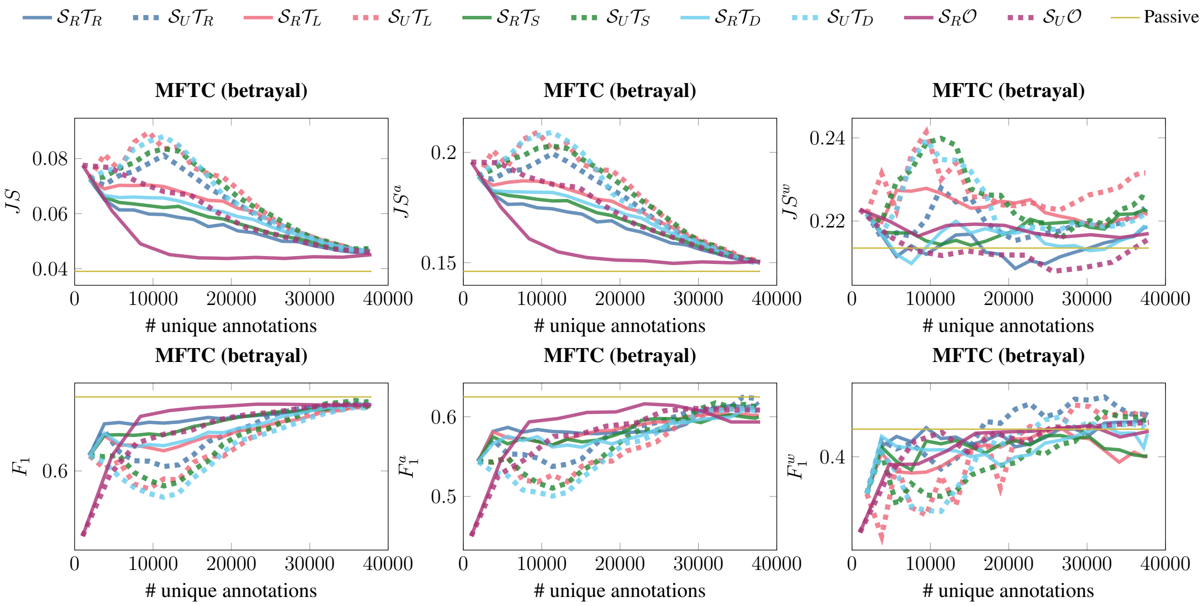

Figure B4: Validation set performance across all metrics for MFTC (betrayal) during training

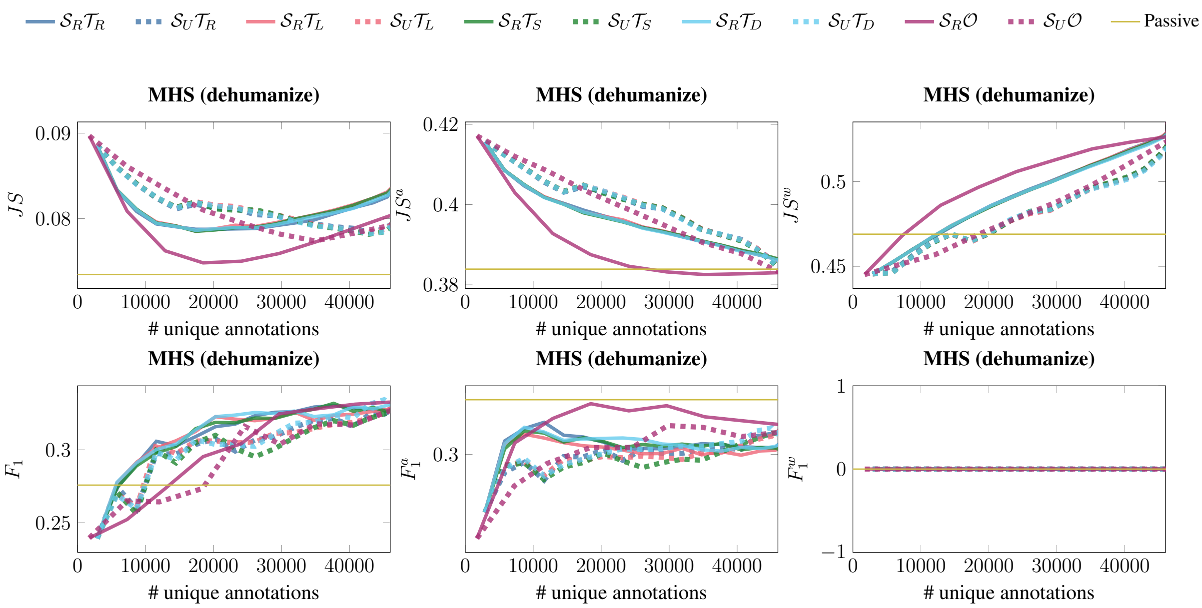

Figure B5: Validation set performance across all metrics for MHS (dehumanize) during training

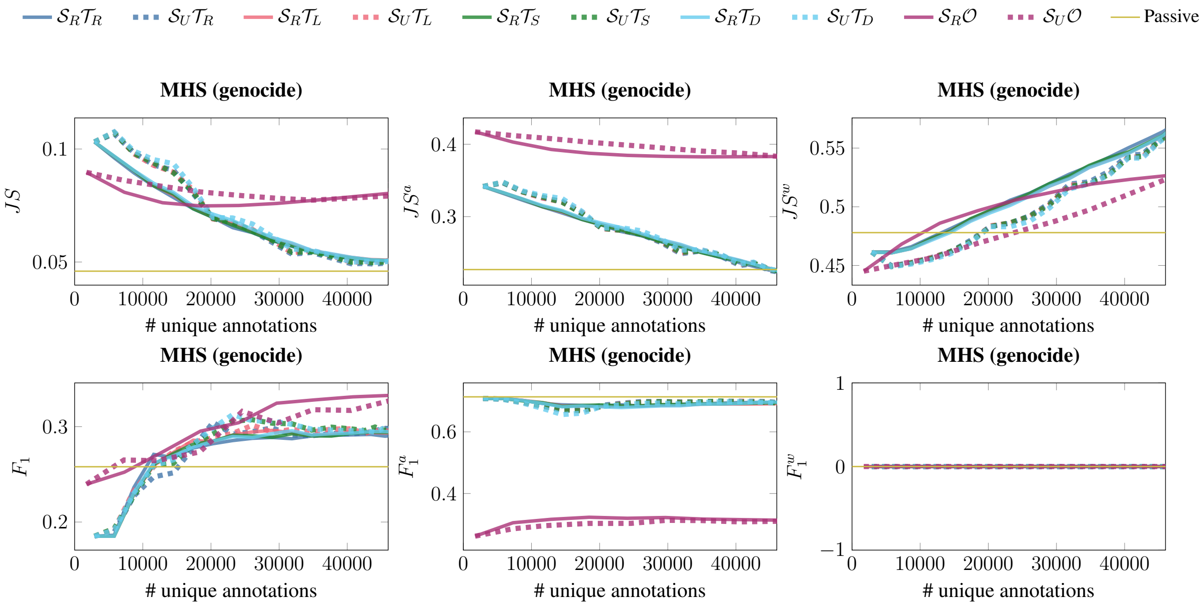

Figure B6: Validation set performance across all metrics for MHS (genocide) during training

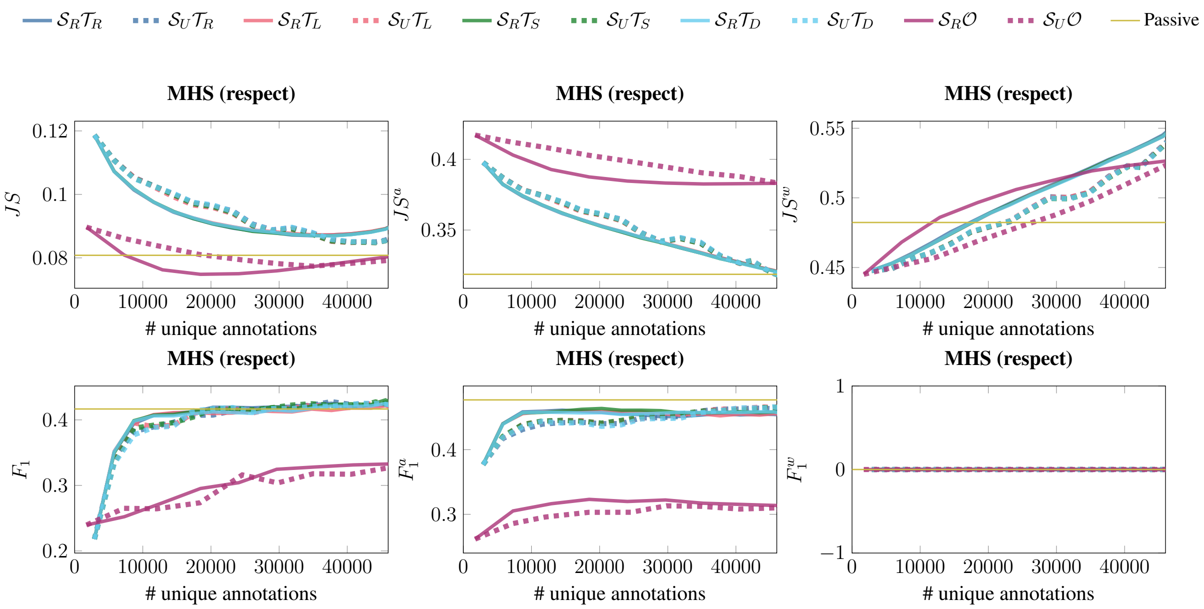

Figure B7: Validation set performance across all metrics for MHS (respect) during training

				Average		Worst-off		
	App.	F_1	JS	F_1^a	JS^a	F_1^w	JS^w	$\Delta\%$
MFTC (*betrayal*)	$\mathcal{S}_R\mathcal{T}_R$	71.5	.047	57.8	**.147**	42.0	.199	-1.6
	$\mathcal{S}_R\mathcal{T}_L$	71.2	.046	58.1	.149	43.3	.212	-1.6
	$\mathcal{S}_R\mathcal{T}_S$	71.2	.051	59.3	.161	43.0	.239	-5.0
	$\mathcal{S}_R\mathcal{T}_D$	71.0	.046	58.3	.148	42.9	.199	-1.6
	$\mathcal{S}_U\mathcal{T}_R$	72.6	.042	**59.4**	.150	41.9	.203	-2.5
	$\mathcal{S}_U\mathcal{T}_L$	73.6	.045	58.4	.148	43.4	.200	-1.3
	$\mathcal{S}_U\mathcal{T}_S$	74.0	.045	58.8	.149	**43.5**	.204	-1.0
	$\mathcal{S}_U\mathcal{T}_D$	73.2	.044	59.1	.149	42.8	**.194**	-2.6
	$\mathcal{S}_R\mathcal{O}$	72.1	.046	58.9	**.147**	43.1	.195	**-48.6**
	$\mathcal{S}_U\mathcal{O}$	71.8	.047	58.9	.149	43.0	.200	-0.0
	PL	**75.2**	**.037**	48.1	.199	36.0	.290	0.0
MFTC (*betrayal*)	$\mathcal{S}_R\mathcal{T}_R$	66.9	.034	56.4	.177	22.2	.372	-0.4
	$\mathcal{S}_R\mathcal{T}_L$	68.9	.032	56.3	.176	22.2	.374	-0.3
	$\mathcal{S}_R\mathcal{T}_S$	67.1	.031	**57.3**	.176	22.2	.370	-0.3
	$\mathcal{S}_R\mathcal{T}_D$	68.4	.031	55.1	**.175**	22.2	.373	-0.3
	$\mathcal{S}_U\mathcal{T}_R$	61.3	.032	55.7	.177	21.7	.357	-1.1
	$\mathcal{S}_U\mathcal{T}_L$	66.5	.032	54.1	.177	22.2	.355	-0.8
	$\mathcal{S}_U\mathcal{T}_S$	62.4	.033	55.6	.177	22.2	.358	-0.9
	$\mathcal{S}_U\mathcal{T}_D$	64.4	.031	55.8	.177	22.2	.358	-1.3
	$\mathcal{S}_R\mathcal{O}$	**71.5**	.030	56.0	.176	22.2	.361	**-29.1**
	$\mathcal{S}_U\mathcal{O}$	66.5	.033	55.9	.177	22.2	.366	-0.1
	PL	62.5	**.029**	51.2	.183	**26.1**	**.309**	0.0
MHS (*genocide*)	$\mathcal{S}_R\mathcal{T}_R$	26.5	.050	70.0	.227	0.0	.560	-6.3
	$\mathcal{S}_R\mathcal{T}_L$	28.2	.051	69.8	.225	0.0	.565	-1.7
	$\mathcal{S}_R\mathcal{T}_S$	28.1	.051	70.0	**.224**	0.0	.566	-1.7
	$\mathcal{S}_R\mathcal{T}_D$	28.3	.050	70.2	**.224**	0.0	.565	-1.7
	$\mathcal{S}_U\mathcal{T}_R$	32.8	.077	71.1	.229	0.0	.549	-12.6
	$\mathcal{S}_U\mathcal{T}_L$	27.7	.048	70.7	.231	0.0	.548	-7.9
	$\mathcal{S}_U\mathcal{T}_S$	26.7	.048	70.9	.231	0.0	.548	-7.9
	$\mathcal{S}_U\mathcal{T}_D$	27.3	.048	**71.2**	.229	0.0	.547	-12.6
	$\mathcal{S}_R\mathcal{O}$	28.0	.048	33.9	.387	0.0	**.496**	**-60.1**
	$\mathcal{S}_U\mathcal{O}$	**33.3**	.080	33.1	.390	0.0	.497	-24.7
	PL	21.6	**.044**	70.0	.245	0.0	.570	–
MHS (*respect*)	$\mathcal{S}_R\mathcal{T}_R$	41.4	.086	46.0	.331	0.0	.528	-18.8
	$\mathcal{S}_R\mathcal{T}_L$	40.8	.087	45.6	.331	0.0	.530	-18.8
	$\mathcal{S}_R\mathcal{T}_S$	41.2	.086	46.1	.331	0.0	.529	-18.8
	$\mathcal{S}_R\mathcal{T}_D$	40.6	.086	46.0	.331	0.0	.528	-18.8
	$\mathcal{S}_U\mathcal{T}_R$	32.8	**.077**	**46.6**	**.323**	0.0	.533	-4.9
	$\mathcal{S}_U\mathcal{T}_L$	41.0	.085	46.3	**.323**	0.0	.532	-4.9
	$\mathcal{S}_U\mathcal{T}_S$	**41.8**	.084	45.9	.324	0.0	.531	-4.9
	$\mathcal{S}_U\mathcal{T}_D$	40.6	.085	46.2	.324	0.0	.532	-4.9
	$\mathcal{S}_R\mathcal{O}$	41.7	.085	33.9	.387	0.0	**.496**	**-60.1**
	$\mathcal{S}_U\mathcal{O}$	33.3	.080	33.1	.390	0.0	.497	-24.7
	PL	41.0	.080	25.9	.405	0.0	.587	–

Table B1: Test set results on the MFTC (*betrayal*), MFTC (*loyalty*), MHS (*genocide*), and MHS (*respect*) datasets. $\Delta\%$ denotes the reduction in the annotation budget with respect to passive learning.