Cost-Efficient Subjective Task Annotation and Modeling through Few-Shot Annotator Adaptation

Anonymous ACL submission

Abstract

In subjective NLP tasks, where a single ground truth does not exist, the inclusion of diverse annotators becomes crucial as their unique perspectives significantly influence the annotations. In realistic scenarios, the annotation budget often becomes the main determinant of the number of perspectives (i.e., annotators) included in the data and subsequent modeling. We introduce a novel framework for annotation collection and modeling in subjective tasks that aims to minimize the annotation budget while maximizing the predictive performance for each annotator. Our framework has a two-stage design: first, we rely on a small set of annotators to build a multitask model, and second, we augment the model for a new perspective by strategically annotating a few samples per annotator. To test our framework at scale, we introduce and release a unique dataset, Moral Foundations Subjective Corpus, of 2000 Reddit posts annotated by 24 annotators for moral sentiment. We demonstrate that our framework surpasses the previous SOTA in capturing the annotators' individual perspectives with as little as 25% of the original annotation budget on two datasets. Furthermore, our framework results in more equitable models, reducing the performance disparity among annotators.

1 Introduction

The common pipeline for supervised learning in Natural Language Processing (NLP) starts by collecting annotations from multiple annotators. These annotations are often aggregated through majority voting (Talat and Hovy, 2016) to construct a *ground truth* or *gold standard* on which the subsequent modeling is performed. In recent years, researchers have advocated for a transition from single ground-truth labels to annotator-level modeling, aiming to capture diverse perspectives, enhance contextual understanding, and incorporate cultural nuances (Uma et al., 2021), and have proposed different frameworks that take into account unique perspectives of the annotators by modeling them as separate subtasks (Davani et al., 2022; Kanclerz et al., 2022).

The impact of individual annotators' backgrounds and life experiences on annotations in subjective tasks signifies the importance of incorporating a diverse set of annotators. Nevertheless, the primary constraint on achieving this diversity is often the annotation budget, limiting the number and, consequently, the diversity of perspectives considered. In this paper, we introduce a novel framework for annotation collection and modeling in subjective tasks. Our framework is designed to minimize the annotation budget required to model a fixed number of annotators, while maximizing the predictive performance for each annotator.

Our framework operates in two stages. In the first stage, data is collected from a small pool of annotators. This data serves as a foundation for building a multitask model that captures the general patterns for the task and provides a signal of differences among individual annotators. Informed by the first stage annotations, the second stage involves collecting a few samples from each new annotator that best capture their differences from the general patterns. We use this data to augment the model from the first stage to learn the new annotators' perspective from a few examples (Figure 1).

We introduce a unique dataset that enables the study of detecting moral content, an understudied subjective task, at a scale that was not possible before[1]. The Moral Foundations Subjective Corpus (MFSC) is a collection of 2000 Reddit posts, each annotated by 24 annotators for moral content along with annotators' responses to a range of psychological questionnaires (§4.1).

We use MFSC in conjunction with the Brexit Hate Dataset (Akhtar et al., 2021) to extensively study each component of our proposed framework. First, we empirically investigate the effect of pool-

[1]The dataset will be released as part of the accepted paper

ing data from varying numbers of annotators in a multitask model (§5.1) and demonstrate that increasing the number of annotators does not improve annotator-level modeling in a multitask model. Second, we showcase the efficacy of our framework in capturing diverse annotator perspectives under budget constraints (§5.2). Third, in §5.3, we study the impact of various sample selection strategies. Our framework achieves a 4.3% increase in F_1 score with access to just 25% of the annotation budget in moral sentiment prediction. Furthermore, our results demonstrate a 6.7% improvement over the previous state-of-the-art in hate speech detection when using only 50% of the original annotation budget. Next, we show that our proposed framework consistently yields a more equitable model by minimizing the performance disparity across various annotators (§5.4). Specifically, on the lowest budget scenarios, our approach reduces the standard deviation in performance across annotators by 7.5% and 1.1% on hate speech detection and moral foundation classification, respectively. Finally, we extend our analysis to investigate whether the selection of the initial set of annotators in the first stage of our framework is related to the model's performance (§5.5).

Our experiments on two subjective datasets revealed that our framework consistently surpasses previous state-of-the-art models with access to as little as 25% of the original annotation budget. In addition, our framework produced more equitable models with reduced performance disparities among the annotators. By minimizing data requirements, our cost-efficient framework for subjective tasks enables us to scale the number of included annotators and, hence, improve the diversity of captured perspectives. Furthermore, the two-stage design of our framework facilitates the integration of new annotators into pre-existing datasets.

2 Related Work

Subjective Tasks in NLP: In recent years, the variety of tasks for which NLP is used has significantly expanded. In many of these tasks, a single ground truth does not exist, making them inherently *subjective* in nature. In subjective tasks, researchers have argued that disagreements in particular labels should not be treated as statistical noise (Larimore et al., 2021; Pavlick and Kwiatkowski, 2019; Plank, 2022), as they are often indicative of individual differences which are driven by different backgrounds and lived experiences of the annotators (Akhtar et al., 2019; Plank et al., 2014; Prabhakaran et al., 2021; Díaz et al., 2018; Garten et al., 2019; Ferracane et al., 2021). For example, Davani et al. (2023) revealed how the stereotypes of annotators influence their behavior when annotating hate speech. In a similar context, Sap et al. (2021) demonstrate that annotators' identity and beliefs impact their ratings of toxicity. Sang and Stanton (2022) conducted a study showing that differences in age and personality among annotators result in variations in their annotations. Larimore et al. (2021) explored how annotators' perceptions of racism differ based on their own racial identity. Basile (2020) calls for a paradigm shift away from majority aggregated ground truths, and towards representative frameworks preserving unique perspectives of the annotators. In their later work, Basile et al. (2021) define the phenomena of *Data Perspectivism*, and share recommendations and outlines to advance the perspectivist stance in machine learning.

Capturing the Perspectives: To capture annotator-level labels, Akhtar et al. (2020) proposed dividing annotators into groups based on similar personal characteristics and creating different sets of gold standards for each group. Kanclerz et al. (2022) and Deng et al. (2023) incorporated knowledge about annotators into their models to make them personalized. Davani et al. (2022) propose a multitask approach, modeling each annotators' perspective as a subtask, while having a shared encoder across the subtasks. Baumler et al. (2023) and Wang and Plank (2023) propose active learning methods for reducing the budget of data collection by proposing methods for collecting samples based on model confidence and annotators' disagreement. Casola et al. (2023) also proposes ensembling perspective-aware models based on their confidence.

3 Method

Problem Formulation: To formalize the task, suppose we have a set of annotators $\mathcal{A} = \{a_1, ..., a_n\}$ and input texts $X = \{x_1, x_2, ..., x_m\}$ and their corresponding annotations $Y = \{y_1, y_2, ..., y_m\}$. Let $D = \{D_{a_i} | a_i \in \mathcal{A}\}$ be the entire annotations and $D_{a_i} = \{X_{a_i}, Y_{a_i}\}$ denote data collected from annotator a_i. Then the budget $B = |D|$ is defined

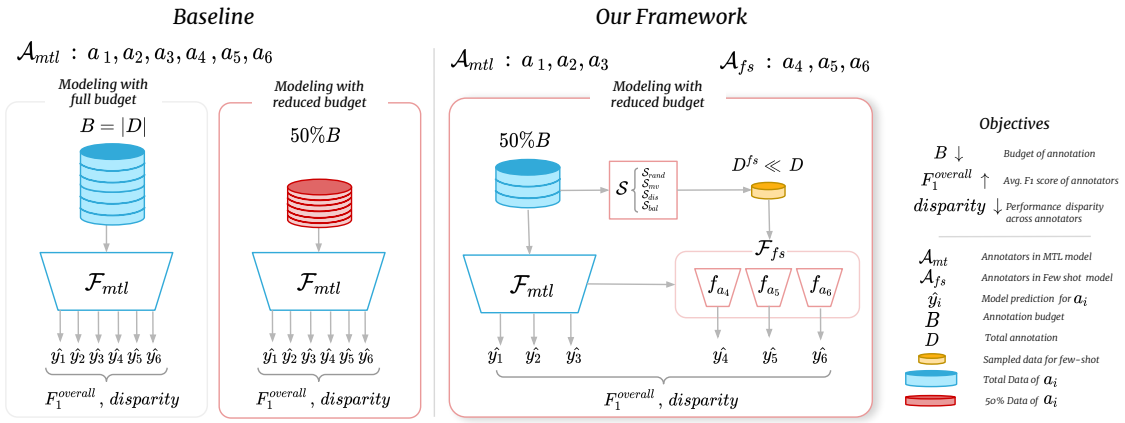


Figure 1: **Left:** The baseline approach for annotator-level modeling, in full and reduced budget scenarios. **Right:** Our two-stage proposed framework, designed to achieve the outlined objectives

as the total number of annotations collected. Let $\mathcal{F} = \{f_{a_i} | a_i \in \mathcal{A}\}$ and f_{a_i} denote the model capturing labels assigned by annotator a_i.

Proposed Framework: We design our framework with two objectives: first, maximizing the average performance over all annotators. Second, minimizing the budget (B) required to achieve the first goal. The second objective allows us to increase the number of annotators' perspectives ($|\mathcal{A}|$) captured with a given budget. Our framework design is based on two key intuitions. Firstly, as we show in Section 5.1, multitask learning, which has often been treated as the upper bound by previous work, does not improve in performance as the number of annotators grows. Secondly, even in subjective tasks, there exists a substantial number of texts on which annotators mostly agree, particularly when these texts are randomly drawn from a source. Therefore, obtaining many annotations on such instances is not beneficial in learning a new perspective. In line with these intuitions, our framework consists of two stages (Figure 1). In the first stage, we learn the commonalities between annotators through a multitask model $\mathcal{F}_{mtl}$. A crucial difference of our approach in comparison to previous multitask methods is that we only collect annotations from a small subset of annotators $\mathcal{A}_{mtl} \subset \mathcal{A}$. In the second stage, we learn the perspectives of new annotators $\mathcal{A}_{fs} = \mathcal{A} - \mathcal{A}_{mtl}$ with only a few shots. Specifically, we collect annotations for k input texts $\mathcal{S}(X) \subset X$, where $\mathcal{S}$ is a sampling function that ideally helps in capturing patterns specific to individual annotators' perspectives. Let $D^{fs}_{a_i} = \{(x, y_{ai}) | x \in \mathcal{S}(X)\}$ and $|D^{fs}_{a_i}| = k << |D_{a_i}|$. We initialize $\mathcal{F}_{\mathcal{A}_{fs}}$ with $\mathcal{F}_{mtl}$ and train it on $D^{fs}_{a_i}$.

Sampling Function ($\mathcal{S}$): We explore four different sampling functions: 1) $\mathcal{S}_{rand}$: selects a random sample for each annotator 2) $\mathcal{S}_{mv}$: selects a balanced sample determined by the majority vote of the annotators. For a set of annotators $\mathcal{A}_{mtl}$, we calculate the majority vote among these annotators and select k samples that have an equal number of each label based on that majority vote. 3) $\mathcal{S}_{dis}$ selects the samples from $\mathcal{A}_{mtl}$ with highest disagreement score, and 4) $\mathcal{S}_{bal}$ acts as an oracle, selecting a balanced sample based on a specific annotator's label, not the majority vote. Therefore, if we have a new annotator, $\mathcal{S}_{bal}$ would select a balanced sample based on the annotations of that specific annotator. One frequent challenge in some subjective tasks is the heavy imbalance in class frequencies. Hence, we chose $\mathcal{S}_{mv}$ and $\mathcal{S}_{bal}$ to provide a more balanced sample to the few-shot model for each annotator. We added $\mathcal{S}_{dis}$ with the goal of providing samples that differentiate the individual annotator perspectives to the model. We use the "item disagreement" and "annotator disagreement" measure from Davani et al. (2023) to select samples in $\mathcal{S}_{dis}$.

4 Experiments

4.1 Datasets

We run experiments on two datasets annotated for subjective tasks: Brexit Hate dataset (Akhtar et al., 2021) and the Moral Foundations Subjective Corpus (MFSC), which we created as part of this work to explore this less-studied subjective task. Both datasets contain per-annotator labels for instances, with every instance being annotated by all annotators. This ensures that any observed performance

gains are attributed to the used methods and are not driven by the specific samples annotated by each annotator. Subsequently, we also evaluate our framework on the Gab Hate Corpus (GHC) (Kennedy et al., 2018), where the number of samples annotated by different annotators varies. Detailed experiments on this dataset are presented in Appendix D.

Moral Foundations Subjective Corpus (MFSC): We introduce a new dataset, Moral Foundations Subjectivity Corpus (MFSC), consisting of 2000 Reddit posts annotated by 24 annotators for moral sentiment based on the Moral Foundations Theory (MFT; Graham et al., 2013; Atari et al., 2022). Morality is widely acknowledged to be a subjective concept, strongly influenced by cultural backgrounds (Graham et al., 2016), one that has not been explored much in the NLP community. We asked annotators to label each text for the specific moral concern (i.e., Purity, Harm, Loyalty, Authority, Proportionality or Equality), and if one of the concerns is chosen, we set the label as *moral*, else as *non-moral*; This dataset was collected following the same procedure as Trager et al. (2022). This dataset also contains additional metadata information, such as confidence for each instance using a 3-level measure (i.e., *confident*, *somewhat confident*, and *not confident*). We also collected annotator responses for the "Big Five Inventory-2-Short" questionnaire (Soto and John, 2017). MFSC provides an opportunity to explore the subjective nature of morality. The substantial number of annotators in this dataset along with questionnaire responses enables future researchers to investigate the modeling of subjective tasks on a larger scale. The demographics of the annotators is shown in Appendix A.1. In our experiments, we use the annotations for the *moral* label. In addition, we evaluate our framework on *Care* label which is discussed in Appendix C.

Brexit Hate dataset: Hate speech detection has become one of the primary subjective tasks studied in the NLP community (Akhtar et al., 2019; Sang and Stanton, 2022; Sap et al., 2021). The Brexit Hate dataset (Brexit) proposed by Akhtar et al. (2021), consists of 1,120 English tweets collected with keywords related to immigration and Brexit. The dataset was annotated with hate speech (in particular xenophobia and islamophobia), aggressiveness, offensiveness, and stereotype, by six annotators belonging to two distinct groups: a target group of three Muslim immigrants in the UK, and a control group who were researchers with Western background. For our experiments, we use the overall hate label. Additional dataset statistics can be found in Appendix A.

4.2 Experiment Setup

We designed our experiments to study the impact of each component of the framework towards our two objectives: maximizing average performance and minimizing annotation budget.

We use multitask learning (MTL) on all the annotators as our baseline and assess the efficacy of our framework compared to this baseline in capturing individual annotators' perspectives under a range of budget constraints. Specifically, for our approach, we vary the budget B by changing the size of $|\mathcal{A}_{mtl}|$. Recall that $B = |D| = \sum |D_{a_i}|$ and $|D^{fs}_{a_i}| = k << |D_{a_i}|$. Also, recall that under our proposed framework the annotators $\mathcal{A}$ are divided into two sets $\mathcal{A}_{mtl}$ and $\mathcal{A}_{fs}$. Since the cost of annotating a few samples per new annotator is negligible ($\frac{|D^{fs}_{a_i}|}{|D_{a_i}|}$ is close to 0) the budget under our proposed framework can be reduced to

$$
\begin{aligned}
B_{ours} &\approx \sum_{a_i \in \mathcal{A}_{\text{mtl}}} |D_{a_i}| \\
&= \frac{\sum_{a_i \in \mathcal{A}_{\text{mtl}}} |D_{a_i}|}{\sum_{a_i \in \mathcal{A}} |D_{a_i}|} \times B \\
&= \frac{|\mathcal{A}_{mtl}|}{|\mathcal{A}|} \times B
\end{aligned}
$$

For example, the MFSC dataset has $|\mathcal{A}| = 24$ annotators. Hence, $25\%B$ shows the scenarios where $|\mathcal{A}_{mtl}| = 6$. Whereas, for the baseline, we vary the budget B by changing the size of D_{a_i} for all annotators. In the given example, a $25\%B$ for the baseline means using only 25% of D_{a_i} for each a_i.

To ensure that our results are not driven by the specific choices of $\mathcal{A}_{mtl}$, we run our experiments for each budget on multiple samples of $\mathcal{A}_{mtl} \subset \mathcal{A}$. Specifically, we run our models with all possible choices of $\mathcal{A}_{mtl}$ for Brexit dataset and 20 different samples of $\mathcal{A}_{mtl}$ for the MFSC dataset.

For each annotator a_i, $F_1^{a_i}$ denotes the performance on predicting a_i's labels. We use F_1^{fs} and F_1^{mtl} to denote the average of $F_1^{a_i}$ scores when $a_i \in \mathcal{A}_{fs}$ and $a_i \in \mathcal{A}_{mtl}$ respectively. For our framework, we also calculate the overall performance for all annotators F_1^{overall} as the weighted average of F_1^{fs} and F_1^{mtl}.

4.3 Implementation Details

We use RoBERTa-base as our base model (Liu et al., 2019). All multitask models undergo hyperparameter tuning, for learning rate and weight decay, and are trained for 5 epochs. The best model is selected based on the validation F_1 score, and its best hyperparameters are applied in subsequent fine-tuning. Few-shot models are trained for 50 epochs with all parameters updated. Experiments are repeated with three random seeds.

For the Brexit dataset, we utilize predefined train, validation, and test splits provided within the dataset[2]. In the case of the MFSC dataset, we allocate 80% for training, 10% for validation, and the remaining 10% for testing. Further implementation details are available in Appendix E.1.

5 Results and Analysis

5.1 Few Annotators are Enough

Previous state-of-the-art methods for subjective tasks achieve performance gains by jointly learning annotations from multiple annotators, as opposed to independently learning each annotator's perspectives (Davani et al., 2022). In this analysis, we further explore whether adding more annotators to the multitask model results in a better model for all annotators. To answer this question, we look at the multitask performance F_1^{mtl} as the number of annotators in the multitask model, $|\mathcal{A}_{mtl}|$, grows. There is no statistically significant correlation between the number of annotators and F_1^{mtl} ($r = -0.28, p > 0.05$). Figure 2 shows that average F_1^{mtl} for all annotators does not increase as the number of annotators increases. Note that this pattern is consistent in both datasets and as Figure 2 shows, the F_1^{mtl} scores are reliable with low standard deviations across models with different numbers of annotators. This observation motivates our first design choice; in the first stage of our framework, we can only rely on a small number of annotators to get a reliable multitask model.

5.2 Towards Better Performance with Less Annotation Budget

A successful framework for modeling subjective tasks should, above all, demonstrate the ability to accurately predict labels from all of the annotators. Therefore, we assess the overall performance (F_1^{overall}) of our proposed framework in comparison to the baseline of multitask learning. Recall the second component of our framework is to augment the multitask learning model of the first stage $\mathcal{F}_{mtl}$ to predict new annotators' labels using only a few samples. To isolate the impact of the few-shot learning model from the specific choice of samples, we sample the k shots randomly.

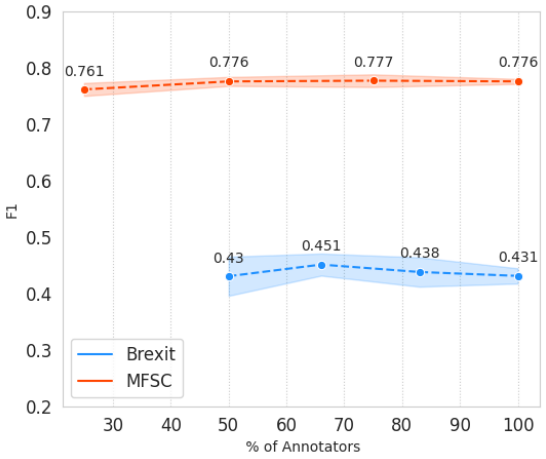


Figure 2: Multitask models F_1 performance (F_1^{mtl}) as the number of annotators increase. Mean and standard deviation are reported across three repeated runs

As shown in Table 1, our framework outperforms the multitask baseline on both datasets when it comes to predicting labels of all annotators. Specifically, on MFSC, our approach outperforms multitask learning using 100% of data with access to as little as 25% of the original annotation budget. On Brexit, our framework matches the performance of multitask learning on 100% budget with access to only 50% of the original budget. The findings demonstrate the success of our framework in achieving its dual objectives: enhancing performance across all annotators while reducing annotation budget requirements. Consequently, our framework facilitates increased diversity by incorporating additional annotator perspectives while adhering to a specific budget constraint. We also conduct an ablation study by omitting the first MTL stage and employing random few-shot sampling for each annotator (Appendix B).

5.3 Few-shot Sampling Strategies

Recall that in the second stage of our framework, we only select a small subset of the input texts $\mathcal{S}(X) \subset X$ to be annotated by the new annotators (Figure 1). Intuitively, we are relying on the first stage of our framework (i.e., multitask learning) to learn the commonalities between annotators.

[2] https://le-wi-di.github.io/

$metric = F_1^{overall} \uparrow$	Brexit				MFSC			
	50%	66%	83%	100%	25%	50%	75%	100%
$X\% \times \lvert D_{a_i} \rvert$	$50\%\lvert D_{a_i} \rvert$	$66\%\lvert D_{a_i} \rvert$	$83\%\lvert D_{a_i} \rvert$	$\lvert D_{a_i} \rvert$	$25\%\lvert D_{a_i} \rvert$	$50\%\lvert D_{a_i} \rvert$	$75\%\lvert D_{a_i} \rvert$	$\lvert D_{a_i} \rvert$
MTL	$0.417_{(0.049)}$	$0.449_{(0.027)}$	$0.418_{(0.018)}$	$0.431_{(0.014)}$	$0.763_{(0.016)}$	$0.773_{(0.011)}$	$0.772_{(0.015)}$	$0.776_{(0.004)}$
$X\% \times \lvert \mathcal{A} \rvert$	$50\%\lvert \mathcal{A} \rvert$	$66\%\lvert \mathcal{A} \rvert$	$83\%\lvert \mathcal{A} \rvert$		$25\%\lvert \mathcal{A} \rvert$	$50\%\lvert \mathcal{A} \rvert$	$75\%\lvert \mathcal{A} \rvert$	
Ours $k = 16$	$0.422_{(0.012)}$	$0.44_{(0.025)}$	$0.455_{(0.007)}$		$0.795_{(0.009)}$	$0.79_{(0.006)}$	$0.785_{(0.005)}$	
Ours $k = 32$	$0.428_{(0.006)}$	$0.447_{(0.019)}$	$0.452_{(0.016)}$		$0.795_{(0.01)}$	$0.791_{(0.006)}$	$0.785_{(0.004)}$	
Ours $k = 64$	$0.433_{(0.012)}$	$0.451_{(0.015)}$	$\mathbf{0.456}_{(0.013)}$		$0.797_{(0.009)}$	$0.791_{(0.006)}$	$0.785_{(0.004)}$	
Ours $k = 128$	$\mathbf{0.439}_{(0.015)}$	$\mathbf{0.457}_{(0.012)}$	$0.455_{(0.011)}$		$\mathbf{0.798}_{(0.008)}$	$\mathbf{0.791}_{(0.005)}$	$\mathbf{0.786}_{(0.004)}$	

Table 1: Mean and standard deviation of F_1 scores on Brexit and MFSC dataset across three runs, with varying annotation budgets ($\%B$). The budget allocation differs in the baseline and our framework; For example, $50\%|D_{a_i}|$ indicates selecting 50% of each annotator's data, while $50\%|\mathcal{A}|$ denotes selecting 50% of annotators and using all their data (i.e., 3 out of 6 annotators for Brexit, and 12 out of 24 annotators for MFSC).

Hence, the goal of the selection function $\mathcal{S}(X)$ is to choose input texts that best capture individual annotators' differences. To study the impact of data selection strategies, we compare the fewshot performance (F_1^{fs}) of each strategy controlling for the number of shots k (see §3 for definitions of $\mathcal{S}(.)$).

On Brexit dataset, the $\mathcal{S}_{bal}$ strategy, which selects a balanced sample for each annotator outperforms other approaches consistently with 50% budget. With 66% budget, $\mathcal{S}_{bal}$ strategy performs the best for all values of k. Notably, with this budget, the performance gap is closed by the $\mathcal{S}_{dis}$ for higher number of shots. Although the differences in performance are less pronounced for higher budget scenarios (i.e., the $83\%B$), $\mathcal{S}_{bal}$ and $\mathcal{S}_{dis}$ still achieve better results especially when more shots are available. We acknowledge that the $\mathcal{S}_{bal}$ strategy relies on access to a balanced sample for each annotator which is often not readily available. However, this study showcases the potential of this approach and future work can explore methods that attempt to select instances that approximate a balanced sample per annotator.

On the MFSC dataset, sampling strategies persistently outperform the multitask learning baseline in all scenarios. While different strategies seem to perform comparably, a crucial differentiating factor is the impact of various strategies on the stability of the results. Specifically, compared to $\mathcal{S}_{rand}$, all sampling functions result in more stable performances (i.e., lower standard deviation) across all budgets and values of k.

Overall, our results on two datasets show that the fewshot stage of our framework, results in models that consistently outperform the multitask learning baseline. Our experiments clearly demonstrate the importance of sample selection strategies both in terms of performance improvement and stability. Furthermore, our results on both datasets motivate the exploration of sampling strategies that can approximate a balanced sample per annotator.

5.4 Reduced Performance Disparities across Annotators

Ensuring a comprehensive representation of annotators' viewpoints is crucial in modeling subjective tasks. To achieve this goal, a critical criterion is to create models that not only improve the aggregated performance but also demonstrate fair and equitable performance across all annotators. For example, if the F_1 scores of one model for two annotators are 0.6 and 0.8, respectively, while the second model scores 0.7 for both annotators, the latter is considered a better model. Although the average performance is the same for both models, the first model has a disparate negative impact on the first annotator. This is important because performance disparities among social groups (in our case annotators) can lead to biased models, limiting the system's ability to accurately reflect diverse perspectives and potentially perpetuating inequalities in the outputs of subjective tasks (Buolamwini and Gebru, 2018). Merely relying on aggregated performance measures, like the average across all annotators, falls short of providing a complete understanding of how well the model is capturing annotators' varying perspectives. For example, it is not clear whether the average performance is improving because an approach improves on capturing only a subset of annotators or for everyone. Hence, we look into the standard deviation of performance across all annotators

$metric = F_1^{fs} \uparrow$		Brexit				MFSC			
		50%	66%	83%	100%	25%	50%	75%	100%
$X\% \times \lvert D_{a_i} \rvert$		$50\%\lvert D_{a_i} \rvert$	$66\%\lvert D_{a_i} \rvert$	$83\%\lvert D_{a_i} \rvert$	$\lvert D_{a_i} \rvert$	$25\%\lvert D_{a_i} \rvert$	$50\%\lvert D_{a_i} \rvert$	$75\%\lvert D_{a_i} \rvert$	$\lvert D_{a_i} \rvert$
MTL		$0.417_{(0.049)}$	$0.449_{(0.027)}$	$0.418_{(0.018)}$	$0.431_{(0.014)}$	$0.763_{(0.016)}$	$0.773_{(0.011)}$	$0.772_{(0.015)}$	$0.776_{(0.004)}$
$X\% \times \lvert \mathcal{A} \rvert$		$50\%\lvert \mathcal{A} \rvert$	$66\%\lvert \mathcal{A} \rvert$	$83\%\lvert \mathcal{A} \rvert$		$25\%\lvert \mathcal{A} \rvert$	$50\%\lvert \mathcal{A} \rvert$	$75\%\lvert \mathcal{A} \rvert$	
$k = 16$	$\mathcal{S}_{bal}$	$0.442_{(0.003)}$	$0.457_{(0.016)}$	$0.458_{(0.044)}$		$0.785_{(0.006)}$	$0.781_{(0.004)}$	$0.778_{(0.01)}$	
	$\mathcal{S}_{dis}$	$0.399_{(0.013)}$	$0.412_{(0.006)}$	$0.457_{(0.036)}$		$0.793_{(0.01)}$	$0.795_{(0.004)}$	$0.785_{(0.01)}$	
	$\mathcal{S}_{mv}$	$0.409_{(0.018)}$	$0.416_{(0.026)}$	$0.442_{(0.055)}$		$0.799_{(0.003)}$	$0.794_{(0.004)}$	$0.786_{(0.015)}$	
	$\mathcal{S}_{rand}$	$0.4_{(0.017)}$	$0.414_{(0.045)}$	$0.44_{(0.037)}$		$0.815_{(0.026)}$	$0.814_{(0.033)}$	$0.805_{(0.033)}$	
$k = 32$	$\mathcal{S}_{bal}$	$0.454_{(0.013)}$	$0.469_{(0.004)}$	$0.461_{(0.026)}$		$0.787_{(0.006)}$	$0.782_{(0.003)}$	$0.779_{(0.012)}$	
	$\mathcal{S}_{dis}$	$0.402_{(0.008)}$	$0.413_{(0.02)}$	$0.455_{(0.03)}$		$0.796_{(0.01)}$	$0.797_{(0.004)}$	$0.787_{(0.011)}$	
	$\mathcal{S}_{mv}$	$0.404_{(0.029)}$	$0.426_{(0.029)}$	$0.461_{(0.024)}$		$0.801_{(0.003)}$	$0.795_{(0.003)}$	$0.787_{(0.017)}$	
	$\mathcal{S}_{rand}$	$0.412_{(0.006)}$	$0.436_{(0.026)}$	$0.422_{(0.027)}$		$0.816_{(0.028)}$	$0.815_{(0.035)}$	$0.807_{(0.031)}$	
$k = 64$	$\mathcal{S}_{bal}$	$0.462_{(0.01)}$	$0.47_{(0.018)}$	$0.467_{(0.011)}$		$0.788_{(0.005)}$	$0.784_{(0.001)}$	$0.782_{(0.011)}$	
	$\mathcal{S}_{dis}$	$0.429_{(0.018)}$	$0.458_{(0.018)}$	$0.518_{(0.017)}$		$0.797_{(0.011)}$	$0.799_{(0.005)}$	$0.788_{(0.012)}$	
	$\mathcal{S}_{mv}$	$0.409_{(0.017)}$	$0.41_{(0.05)}$	$0.467_{(0.011)}$		$0.802_{(0.004)}$	$0.797_{(0.002)}$	$0.79_{(0.018)}$	
	$\mathcal{S}_{rand}$	$0.423_{(0.02)}$	$0.447_{(0.016)}$	$0.447_{(0.019)}$		$0.818_{(0.027)}$	$0.815_{(0.033)}$	$0.809_{(0.033)}$	
$k = 128$	$\mathcal{S}_{bal}$	$\mathbf{0.498}_{(0.009)}$	$\mathbf{0.517}_{(0.025)}$	$\mathbf{0.524}_{(0.021)}$		$0.79_{(0.005)}$	$0.785_{(0.001)}$	$0.785_{(0.01)}$	
	$\mathcal{S}_{dis}$	$0.456_{(0.011)}$	$0.476_{(0.026)}$	$0.511_{(0.023)}$		$0.798_{(0.009)}$	$0.799_{(0.003)}$	$0.786_{(0.012)}$	
	$\mathcal{S}_{mv}$	$0.425_{(0.028)}$	$0.428_{(0.035)}$	$0.46_{(0.025)}$		$0.805_{(0.004)}$	$0.798_{(0.003)}$	$0.79_{(0.019)}$	
	$\mathcal{S}_{rand}$	$0.433_{(0.023)}$	$0.466_{(0.006)}$	$0.443_{(0.027)}$		$\mathbf{0.819}_{(0.026)}$	$\mathbf{0.816}_{(0.032)}$	$\mathbf{0.81}_{(0.031)}$	

Table 2: Few-shot F_1 results (F_1^{fs}) on BREXIT and MFSC datasets across varying percentages of the full budget ($\%B$). Mean and standard deviation calculated over three repeated runs.

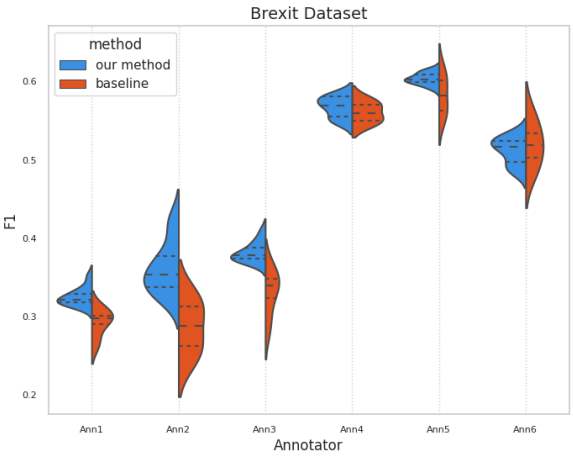


Figure 3: Comparison of Annotator level F1 scores ($F_1^{a_i}$) on the Brexit dataset between MTL model and our framework, leveraging the $\mathcal{S}_{bal}$ sampling method for all budgets and shots

$d = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(F_1^{a_i} - \overline{F_1^{overall}})^2}$. Lower standard deviations are indicative of more equitable models. As shown in Table 3, regardless of the sampling strategy, our approach results in lower performance variances compared to the MTL baseline on lower budgets (50% and 66%). Comparing the different sampling strategies on Brexit dataset, we observe that S_{bal} achieves the most equitable models across all values of k. On MFSC, using a balanced sampling strategy consistently reduces the disparities among annotators compared to the MTL baseline for all budgets. Figure 3 visualizes the performance of our framework in comparison to the multitask learning baseline for each annotator. Notably, our framework improves the performance for the annotators in the non-Western control group (i.e., the first three annotators) while maintaining the performance of the rest of the annotators.

Overall, these results suggest that our proposed framework not only improves the overall performance of all annotators but also yields models that are more fair and equitable. Interestingly, we observe similar patterns in fairness and overall performance improvements. In other words, the configurations with the best overall performance (see section 5.3) are also the ones that have the least performance disparities among all annotators.

5.5 Annotator-level Analysis

Here we delve into the relationship between annotator-level variables. Recall that our framework is trained on $\mathcal{A}_{mtl}$ in the initial stage, followed by fine-tuning for each $a \in \mathcal{A}_{fs}$. Hence, a practical question arises: does the choice of the set $\mathcal{A}_{mtl}$ matter? In other words, would the similarity or divergence in perspectives among annotators in this set impact the performance on $\mathcal{A}_{fs}$? Investigating this is crucial, as identifying such an effect

$metric = d \downarrow$		Brexit			MFSC		
		50%	66%	83%	25%	50%	75%
MTL		.168	.139	.131	.128	.136	.127
$k = 16$	$\mathcal{S}_{bal}$	.119	.127	.132	.119	.124	.124
	$\mathcal{S}_{dis}$	.141	.134	.127	.129	.130	.126
	$\mathcal{S}_{mv}$	.138	.137	.131	.131	.130	.127
	$\mathcal{S}_{rand}$	.137	.139	.135	.137	.135	.130
$k = 32$	$\mathcal{S}_{bal}$	.113	.125	.133	.120	.125	.124
	$\mathcal{S}_{dis}$	.145	.138	.131	.131	.130	.126
	$\mathcal{S}_{mv}$	.140	.137	.129	.130	.130	.127
	$\mathcal{S}_{rand}$	.139	.135	.138	.136	.134	.129
$k = 64$	$\mathcal{S}_{bal}$	.109	.119	.127	.119	.124	.123
	$\mathcal{S}_{dis}$	.137	.126	.123	.131	.130	.126
	$\mathcal{S}_{mv}$	.145	.146	.126	.130	.129	.127
	$\mathcal{S}_{rand}$	.131	.128	.132	.136	.135	.129
$k = 128$	$\mathcal{S}_{bal}$	**.093**	**.108**	**.117**	**.117**	**.122**	**.121**
	$\mathcal{S}_{dis}$	.111	.120	.124	.130	.129	.127
	$\mathcal{S}_{mv}$	.137	.142	.132	.126	.128	.127
	$\mathcal{S}_{rand}$	.131	.127	.136	.134	.133	.128

Table 3: $d \downarrow$ measure across annotators, for full budget $d = .13$ for both datasets; $\mathcal{S}_{bal}$, $\mathcal{S}_{dis}$, $\mathcal{S}_{mv}$, and $\mathcal{S}_{rand}$ refer to the sampling functions used in the second stage of our framework (§3)

would necessitate a thoughtful selection of $\mathcal{A}_{mtl}$. To examine this, we conduct the following analysis: Disagreement within $\mathcal{A}_{mtl}$ and performance on $\mathcal{A}_{fs}$: The aim of this analysis is to investigate whether there is a relationship between the disagreement within annotators in $\mathcal{A}_{mtl}$ and the performance of the newly adopted annotators in A_{fs}. To test this relationship, we employ a mixed-effects model to predict the performance of $a \in \mathcal{A}_{fs}$ by the *agreement within* $\mathcal{A}_{mtl}$ denoted as d^1 (Fleiss, 1971). The model controls for k, budget B, *and agreement between* $\mathcal{A}_{fs}$ *and* $\mathcal{A}_{mtl}$, denoted using d^2, incorporating random effects for $\mathcal{A}_{mtl}$ and $\mathcal{A}_{fs}$. The formula for this model is as follows:

$$\begin{aligned} f_{ij} =& \beta_0 + \beta_1 d^1_j + \beta_2 k_{ij} + \beta_3 B_j \\ &+ \beta_4 d^2_{ij} + u_{0i} + v_{1j} + e_{ij} \end{aligned} \quad (1)$$

where f_{ij} denotes the performance of i^{th} annotator in A_{fs} on the model trained on a j^{th} sample of A_{mtl}. The fixed effects coefficients are represented by β_0 to β_4, and the random effects for i and j are represented by u_{0i}, v_{1j} respectively. e_{ij} denotes the residual error term. To see the impact of sampling strategies, we run a total of four models, each corresponding to the performance results obtained from one of the strategies ($\mathcal{S}_{bal}, \mathcal{S}_{dis}, \mathcal{S}_{mv}, \mathcal{S}_{rand}$).

The findings regarding Brexit indicate that there is no statistically significant effect of agreement within $\mathcal{A}_{mtl}$ (d^1) on the performance. For the MFSC dataset, a significant effect was observed only for results obtained from $\mathcal{S}_{bal}$ ($\beta_1 = -0.052, SE = 0.012, p < 0.001$). This implies that a unit decrease in d^1, corresponding to moving from full agreement to full disagreement, is associated with a 0.052 increase in the F_1 score. This finding suggests that selecting a diverse $\mathcal{A}_{mtl}$ with high disagreement can potentially be advantageous.

6 Conclusion

We introduced a framework for annotation collection and annotator modeling in subjective tasks. Our framework aims to minimize the annotation budget required to model a fixed number of annotators while maximizing the predictive performance for each annotator. Our approach involves collecting annotations from an initial set of annotators and building a multitask model that captures general task patterns while signaling differences among individual annotators. Subsequently, we utilize the annotations from the first stage to select a small set of samples from new annotators that best highlight their deviations from the general patterns. Finally, we leverage this collected data to augment the initial model, enabling it to learn the new annotator's perspective from a limited number of examples. We explored four distinct methods for few-shot sample selection and found that the most effective approach involves balanced sample selection. We introduced a new subjective task dataset Moral Foundations Subjective Corpus (MFSC), of 2000 Reddit posts annotated by 24 annotators for moral sentiment which enabled us to test our framework in scale. Our experiments on MFSC and a hate speech dataset revealed that our framework consistently surpasses previous SOTA with access to as little as 25% of the original annotation budget. In addition, we showed that our framework yields more equitable models that reduce performance disparities among annotators. Our cost-effective framework for subjective tasks allows increasing the number of annotators, enhancing the diversity of perspectives captured, and facilitates the integration of new annotators into pre-existing datasets.

7 Limitations and Ethical Statement

We acknowledge that the datasets employed in our experiments are not representative of all annotator populations. While in MFSC we recruited a substantial number of annotators and efforts were made to diversify this pool, it is important to note that our sample is limited to undergraduate students at a private university in the US. Consequently, we advocate for the replication and extension of our work with non-student, non-US-based samples. Furthermore, we exclusively operate with English data and focus on datasets related to moral sentiment prediction and hate speech detection tasks. This may restrict the generalizability of our findings to a broader linguistic and thematic landscape. Despite these constraints, our research lays the groundwork for future research to extend and validate our approach across diverse languages and subjective NLP tasks. In our experiments, we do not consider the cost of collecting few-shot samples, as discussed in Section 4.2. We recognize that in certain cases, depending on the budget and the nature of the task, this assumption can be challenged. Even with the additional expense of annotating a few samples per new annotator, it is crucial to highlight that our proposed framework substantially reduces annotation cost, especially as the number of included perspectives grows.

In the MFSC dataset the annotators underwent four sessions of training, including guidance on avoiding potential adverse consequences of annotations, and were compensated at a rate of $17 per hour. The study protocol received approval from the Institutional Review Board (IRB), and all annotators consented to both the terms outlined in an information sheet provided by the IRB about the study and the sharing of their responses to the psychological questionnaires along with their annotations. We emphasize that MFSC is created with the intention of exploring subjectivity and different perspectives in this context and it should not be used for any other purposes.

References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2019. A new measure of polarization in the annotation of hate speech. In *International Conference of the Italian Association for Artificial Intelligence*, pages 588–603. Springer.

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2020. Modeling annotator perspective and polarized opinions to improve hate speech detection. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 151–154.

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2021. Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection. *arXiv preprint arXiv:2106.15896*.

Mohammad Atari, Jonathan Haidt, Jesse Graham, Sena Koleva, Sean T Stevens, and Morteza Dehghani. 2022. Morality beyond the weird: How the nomological network of morality varies across cultures.

Valerio Basile. 2020. It's the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks. In *2020 AIxIA Discussion Papers Workshop, AIxIA 2020 DP*, volume 2776, pages 31–40. CEUR-WS.

Valerio Basile, Federico Cabitza, Andrea Campagner, and Michael Fell. 2021. Toward a perspectivist turn in ground truthing for predictive computing. *arXiv preprint arXiv:2109.04270*.

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.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Silvia Casola, Soda Lo, Valerio Basile, Simona Frenda, Alessandra Cignarella, Viviana Patti, and Cristina Bosco. 2023. Confidence-based ensembling of perspective-aware models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3496–3507.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319.

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.

Naihao Deng, Xinliang Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You are what you annotate: Towards better models through annotator representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12475–12498, Singapore. Association for Computational Linguistics.

Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–14.

Elisa Ferracane, Greg Durrett, Junyi Jessy Li, and Katrin Erk. 2021. Did they answer? subjective acts and intents in conversational discourse. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1626–1644, Online. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Justin Garten, Brendan Kennedy, Joe Hoover, Kenji Sagae, and Morteza Dehghani. 2019. Incorporating demographic embeddings into language understanding. *Cognitive science*, 43(1):e12701.

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.

Jesse Graham, Peter Meindl, Erica Beall, Kate M Johnson, and Li Zhang. 2016. Cultural differences in moral judgment and behavior, across and within societies. *Current Opinion in Psychology*, 8:125–130.

Kamil Kanclerz, Marcin Gruza, Konrad Karanowski, Julita Bielaniewicz, Piotr Miłkowski, Jan Kocoń, and Przemyslaw Kazienko. 2022. What if ground truth is subjective? personalized deep neural hate speech detection. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, pages 37–45.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, et al. 2018. The gab hate corpus: A collection of 27k posts annotated for hate speech. *PsyArXiv. July*, 18.

Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. 2021. Reconsidering annotator disagreement about racist language: Noise or signal? In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 81–90.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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

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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. *arXiv preprint arXiv:2110.05699*.

Yisi Sang and Jeffrey Stanton. 2022. The origin and value of disagreement among data labelers: A case study of individual differences in hate speech annotation. In *International Conference on Information*, pages 425–444. Springer.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*.

Christopher J Soto and Oliver P John. 2017. Short and extra-short forms of the big five inventory–2: The bfi-2-s and bfi-2-xs. *Journal of Research in Personality*, 68:69–81.

Zeerak Talat and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Jackson Trager, Alireza S Ziabari, Aida Mostafazadeh Davani, Preni Golazazian, Farzan Karimi-Malekabadi, Ali Omrani, Zhihe Li, Brendan Kennedy, Nils Karl Reimer, Melissa Reyes, et al. 2022. The moral foundations reddit corpus. *arXiv preprint arXiv:2208.05545*.

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. *arXiv preprint arXiv:2310.14979*.

A Dataset Details

Table 4 displays dataset details, including Fleiss's kappa (Fleiss, 1971) measuring the inter-annotator agreement. The low agreement for these tasks highlights the subjective nature of them. Furthermore, the '%Pos.' column in Table 4 shows the class imbalance in these datasets and the scarcity of positive class annotations. For example, in the Brexit dataset, only 12% of samples, on average, were labeled as "Hate". See Table 5 for sample annotations.

Dataset	Size	$\lvert\mathcal{A}\rvert$	Kappa	%Pos.
Brexit	1120	6	0.34	12.86
MFSC (Moral)	2000	24	0.26	63.69
MFSC (Care)	2000	24	0.31	13.34

Table 4: Statistics of the datasets used in our experiments. $\lvert\mathcal{A}\rvert$ denotes the number of annotators, Kappa represents Fleiss's kappa inter-annotator agreement, and %Pos. indicates the average percentage of positive class annotations across annotators

A.1 Demographics of MFSC Annotators

We aimed to diversify the annotators for MFSC dataset across gender, sexual orientation, religion, and race. Even though our dataset is not balanced across these dimensions, we strived to include representative annotators from a cross-section of the aforementioned demographics. The distribution of the annotators across the mentioned demographics is presented in Figure 4.

B Additional Baseline and Ablation Study

First, we conduct an ablation study where we omit the first stage MTL. Essentially, this model is equivalent to few-shot adaptation for each annotator initializing from a pre-trained model. The resulting F_1 scores are presented in Table 6. Comparing these scores with our framework using random few-shot sampling for $k = 16$ from Table 2, we observe that our framework has a 21% gain for Brexit and a 15.5% gain for MFSC. This study demonstrates that the first stage MTL in our framework is crucial for its success.

Dataset	$k = 16$	$k = 32$	$k = 64$	$k = 128$
Brexit	0.23	0.22	0.20	0.28
MFSC	0.66	0.73	0.71	0.77

Table 6: F_1 scores for few-shot baseline with random sampling $\mathcal{S}_{rand}$ when the first stage MTL is omitted.

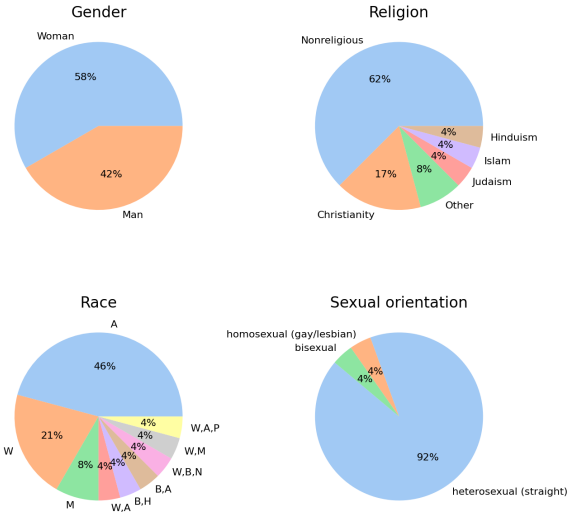


Figure 4: The abbreviations in the pie chart for race *W* stands for White or European American, *B* stands for Black or African American, *H* stands for Hispanic or Latino/Latinx, *P* stands for Native Hawaiian or Pacific Islander, *A* stands for Asian or Asian American, *M* stands for Middle Eastern or North African.

In the second ablation study, we omit the second stage few-shot sample selection from our framework. In other words, in the second stage, we use the entire annotated samples for each annotator instead of selecting only a few samples. Note that this is equivalent to using 100% of the budget and serves as an upper bound to the performance achieved with an ideal sampling function.

Additionally, we present *Ensemble*, a new baseline in which we train a separate model for each annotator using 100% of data from that annotator. The Ensemble baseline involves fine-tuning RoBERTa-basse directly for each annotator, calculating individual annotator F_1 scores, and reporting the average F_1 score across annotators. Hyperparameters and epoch numbers for training are consistent with those mentioned for the MTL model in Section 4.3. In line with the naming convention used in Davani et al. (2022), we name this baseline "Ensemble" to maintain consistency with the previous work in this domain. Figure 5 presents a comparison of 3 different strategies, using 100% of the budget (MTL, Ensemble, and ours). On the Brexit Dataset (top) our framework has as much as 7.4% performance gain compared to Ensemble baseline (when using $\frac{4}{6}$ annotators in MTL), and for MFSC dataset our framework has as large as 5% gain compared to Ensemble baseline (when using $\frac{12}{24}$ of annotators in MTL). These results show that even considering the 100% budget our frame-

Dataset	Examples	Annotations
Brexit	"THE MAJORITY WILL NEVER allow the Mentally Ill Globalists to turn the world into a SJW and Radical Islam SAFE SPACE #brexit #Trump2016"	"0,0,1,1,0,1"
	"A muslim Mayor of London? What!? This PC Sickness has become a pandemic. England turning into Little Asia. <url>"	"0,0,0,1,1,1"
	Not all foreign people who wants to go to the uk have bad intentions. Improve your law. The #Brexit isn't gonna help your economy.	"0,0,0,0,0,0"
MFSC	'm using the term *ethnicity* in all the meanings. A word doesn't have an inner, determined and true meaning; the meaning of a word is steadily constructing by the speakers of a language. If people in the US use *ethnicity* as a synonym of *race*, the term *ethnicity* acquires also that meaning, moreover the anthropological meaning.I guess if they have French flags celebrating Macron's victory, they'd likely feel French.",	16 annotators : 0 8 annotators : 1
	"More or less. Clinton was regarded as a ""female candidate"", but Le Pen was a ""far-right candidate"". The left calls for diversity of ethnicity/gender/etc, but conservatives don't seem to count. ",	4 annotators : 0 20 annotators:1
	"**ing bootlickers, all of them. They WANT a Trump monarchy. All of them are traitors."	24 annotators: 1

Table 5: Examples from each dataset alongside their annotations.

work outperforms both baselines demonstrating the benefit of our two-stage design.

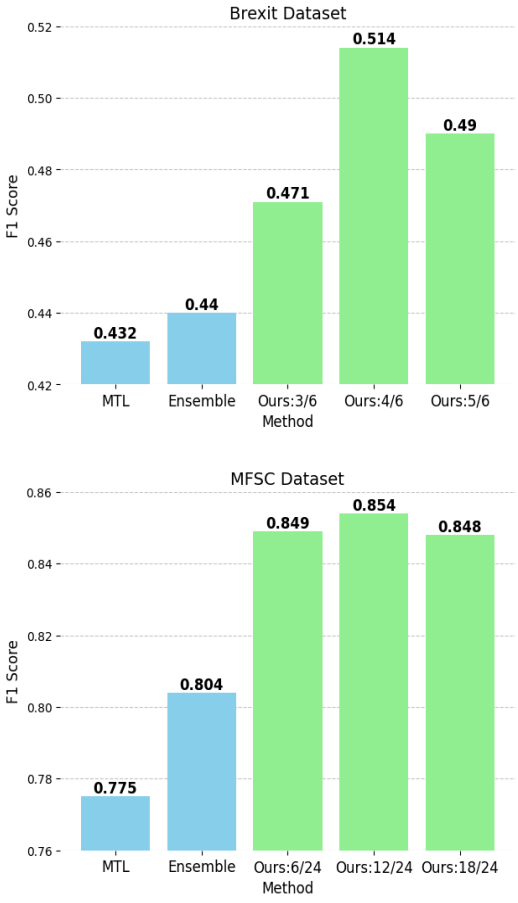


Figure 5: Comparison of our framework with baselines for the two datasets

Interestingly, Ensemble model outperforms MTL for these datasets, contrary to the findings of previous research comparing these two methods.

C Individual Moral Concern Results

We evaluate our framework on an additional binary label of *Care* moral concern from our MFSC dataset. This moral concern is defined as *"Care/Harm: Intuitions about avoiding emotional and physical damage or harm to another individual. It underlies virtues of kindness, gentleness, and nurturing, and vices of meanness, violence, and abuse."* (Trager et al., 2022). Table 7 presents the results for this task. Our framework outperforms the baseline MTL approach with 25% and 50% of the annotation budget. Notably, with only 25% of the budget, our framework has a 1.4% gain in F_1 score compared to MTL with 100% budget. The experiments were conducted with the same hyper-parameter tuning described in Section 4.3.

D Experiments on GHC

To ensure the generalizability of our framework, we evaluate it on a larger dataset with an imbalanced number of annotations among annotators.

Gab Hate Corpus (GHC) consists of 27,665 posts from the social network service gab.ai, each annotated by a minimum of three trained annotators, and 18 total annotators. It is coded for hate-based rhetoric and has labels of "assaults on human dignity" or "calls for violence". The annotators with

$metric = F_1^{Overall} \uparrow$		MFSC (Care) 25%	50%	75%	100%
$X\% \times \lvert D_{a_i} \rvert$		$25\%\lvert D_{a_i} \rvert$	$50\%\lvert D_{a_i} \rvert$	$75\%\lvert D_{a_i} \rvert$	$\lvert D_{a_i} \rvert$
MTL		0.474	0.476	0.49	0.469
$X\% \times \lvert \mathcal{A} \rvert$		$50\%\lvert \mathcal{A} \rvert$	$66\%\lvert \mathcal{A} \rvert$	$83\%\lvert \mathcal{A} \rvert$	
$k = 16$	$\mathcal{S}_{bal}$	0.462	0.471	0.485	
	$\mathcal{S}_{dis}$	0.46	0.467	0.485	
	$\mathcal{S}_{mv}$	0.476	0.473	**0.49**	
	$\mathcal{S}_{rand}$	0.469	0.468	0.482	
$k = 32$	$\mathcal{S}_{bal}$	0.467	0.477	0.487	
	$\mathcal{S}_{dis}$	0.463	0.463	0.483	
	$\mathcal{S}_{mv}$	0.475	0.475	0.488	
	$\mathcal{S}_{rand}$	0.47	0.468	0.484	
$k = 64$	$\mathcal{S}_{bal}$	0.47	0.475	0.486	
	$\mathcal{S}_{dis}$	0.467	0.471	0.478	
	$\mathcal{S}_{mv}$	0.479	0.48	0.487	
	$\mathcal{S}_{rand}$	0.472	0.477	**0.49**	
$k = 128$	$\mathcal{S}_{bal}$	0.473	0.477	0.488	
	$\mathcal{S}_{dis}$	0.474	0.474	0.481	
	$\mathcal{S}_{mv}$	0.477	**0.482**	0.488	
	$\mathcal{S}_{rand}$	**0.483**	0.481	0.487	

Table 7: Overall F_1 scores on MFSC dataset, *Care* label, with varying annotation budgets $(\%B)$.

less than 1000 annotations were filtered out resulting in 16 annotators. Figure 6 shows the number of annotated instances by each annotator.

Experiments: We replicate the experiment described in Section 4.2 with the same implementation details outlined in Section 4.3. We employ varying budgets of 25%, 50%, and 75%, using the two best-performing sampling methods identified in our experiments ($\mathcal{S}_{bal}$ and $\mathcal{S}_{rand}$), and compare the results to the MTL baseline. The overall results are presented in Table 8. It is evident that our framework consistently outperforms MTL across all numbers of shots, sampling methods, and budget variations. Specifically, with 25% of the budget, our model achieves a gain of 1.6% with $k = 64$ and $\mathcal{S}_{rand}$, and with 75% of the budget, our model performs the best, achieving a gain of 2%.

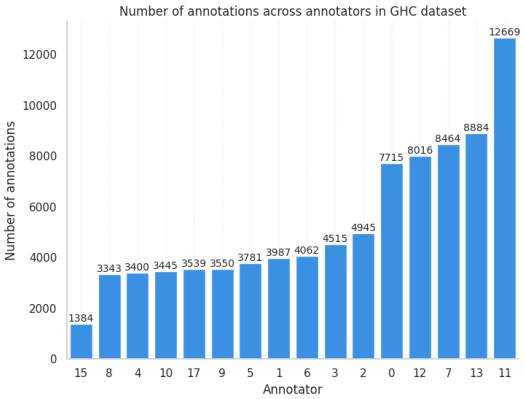


Figure 6: The number of annotated instances by each annotator in GHC dataset

$F_1^{Overall} \uparrow$		GHC 25%	50%	75%	100%
$X\% \times \lvert D_{a_i} \rvert$		$25\%\lvert D_{a_i} \rvert$	$50\%\lvert D_{a_i} \rvert$	$75\%\lvert D_{a_i} \rvert$	$\lvert D_{a_i} \rvert$
MTL		$.417_{(.004)}$	$.433_{(.007)}$	$.442_{(.013)}$	$.451_{(.006)}$
$X\% \times \lvert \mathcal{A} \rvert$		$50\%\lvert \mathcal{A} \rvert$	$66\%\lvert \mathcal{A} \rvert$	$83\%\lvert \mathcal{A} \rvert$	
$k = 16$	$\mathcal{S}_{bal}$	$.45_{(.004)}$	$.46_{(.002)}$	$.464_{(.003)}$	
	$\mathcal{S}_{rand}$	$.455_{(.008)}$	$.469_{(.005)}$	$.468_{(.003)}$	
$k = 32$	$\mathcal{S}_{bal}$	$.456_{(.002)}$	$.459_{(.001)}$	$.464_{(.003)}$	
	$\mathcal{S}_{rand}$	$.461_{(.003)}$	$.472_{(.001)}$	$.468_{(.001)}$	
$k = 64$	$\mathcal{S}_{bal}$	$.458_{(.004)}$	$.461_{(.002)}$	$.466_{(.003)}$	
	$\mathcal{S}_{rand}$	$\mathbf{.467}_{(.003)}$	$.474_{(.002)}$	$.468_{(.002)}$	
$k = 128$	$\mathcal{S}_{bal}$	$.466_{(.001)}$	$.466_{(0)}$	$.466_{(.003)}$	
	$\mathcal{S}_{rand}$	$.463_{(.007)}$	$\mathbf{.475}_{(.003)}$	$\mathbf{.47}_{(.001)}$	

Table 8: Overall F_1 scores on GHC dataset, *Hate* label, with varying annotation budgets $(\%B)$.

$F_1^{fs} \uparrow$		GHC 25%	50%	75%	100%
$X\% \times \lvert D_{a_i} \rvert$		$25\%\lvert D_{a_i} \rvert$	$50\%\lvert D_{a_i} \rvert$	$75\%\lvert D_{a_i} \rvert$	$\lvert D_{a_i} \rvert$
MTL		$.417_{(.004)}$	$.433_{(.007)}$	$.442_{(.013)}$	$.451_{(.006)}$
$X\% \times \lvert \mathcal{A} \rvert$		$50\%\lvert \mathcal{A} \rvert$	$66\%\lvert \mathcal{A} \rvert$	$83\%\lvert \mathcal{A} \rvert$	
$k = 16$	$\mathcal{S}_{bal}$	$.443_{(.004)}$	$.454_{(.006)}$	$.476_{(.005)}$	
	$\mathcal{S}_{rand}$	$.45_{(.009)}$	$.473_{(.008)}$	$.491_{(.006)}$	
$k = 32$	$\mathcal{S}_{bal}$	$.45_{(.002)}$	$.454_{(.004)}$	$.475_{(.005)}$	
	$\mathcal{S}_{rand}$	$.458_{(.003)}$	$.48_{(.003)}$	$.493_{(.008)}$	
$k = 64$	$\mathcal{S}_{bal}$	$.454_{(.003)}$	$.457_{(.003)}$	$.482_{(.005)}$	
	$\mathcal{S}_{rand}$	$\mathbf{.466}_{(.005)}$	$.484_{(.003)}$	$.493_{(.009)}$	
$k = 128$	$\mathcal{S}_{bal}$	$.465_{(.001)}$	$.468_{(.002)}$	$.483_{(.006)}$	
	$\mathcal{S}_{rand}$	$.461_{(.01)}$	$\mathbf{.485}_{(.007)}$	$\mathbf{.498}_{(.003)}$	

Table 9: Few-shot F_1 scores on GHC dataset, *Hate* label, with varying annotation budgets $(\%B)$.

Table 9 displays the results for the few-shot stage of our framework (F_1^{fs}). These results are in line with the overall findings, demonstrating that our framework outperforms MTL by 4.5% with 75% of the original budget.

D.1 Impact of the Imbalanced Number of Annotations on Performance

Results on the GHC dataset indicate a consistent and significant advantage of our framework, even when applied to larger datasets with imbalanced numbers of annotations across annotators. To further investigate the impact of varying numbers of annotations across annotators on the performance of our framework, we conducted a correlation analysis between each annotator's performance and their number of annotations. The results revealed no statistically significant correlation between the number of annotations and the overall F_1 score of an annotator, as indicated by the correlation coefficients for $\mathcal{S}_{rand}$ $(r = -0.17, p = 0.25)$ and $\mathcal{S}_{bal}$ $(r = -0.14, p = 0.32)$. The plots in Figure 7 illustrate the annotator-level F_1 scores as the number of annotations of the annotators increases.

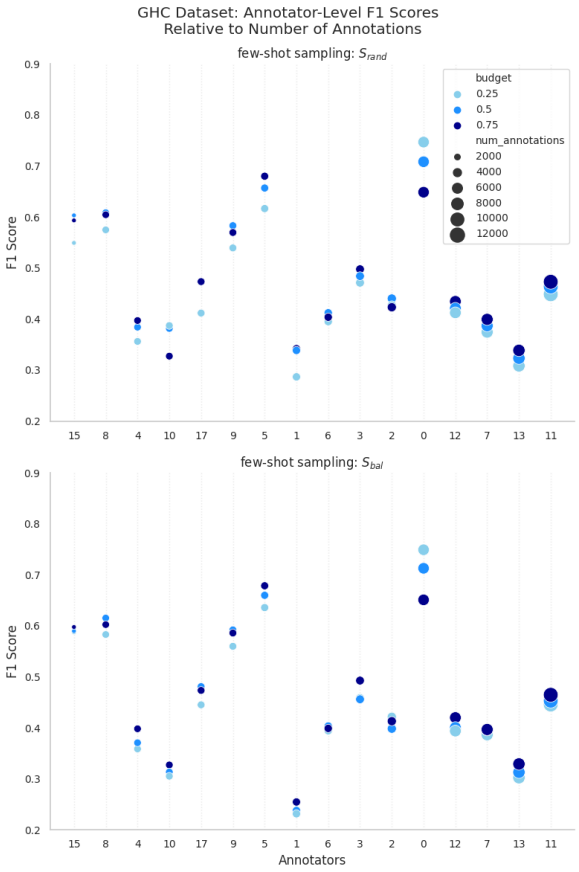


Figure 7: F_1 scores of annotators as the number of annotations increases

E Additional Details and Results

E.1 Implementation Details

Choice of model architecture: We chose to incorporate RoBERTa-base in our framework as a simple and robust model architecture. It is important to note that our focus is not on identifying the best model architecture. Rather, our primary goal is to demonstrate that even with a straightforward model, we can observe the advantages of our framework. This is a common practice in recent papers in the subjective modeling domain (Wang and Plank, 2023; Davani et al., 2022; Baumler et al., 2023).

We employ a weighted random sampler for the Brexit dataset to account for the imbalance in the labels of each annotator. Our training batch size is 128, and we use *AdamW* optimizer. In the training, we prevent overfitting by selecting the best model according to the validation F_1 score. All models converged within 5 epochs for MTL and 50 epochs for few-shot learning. Hyperparameter tuning was conducted for learning rates of $[3e-06, 5e-05, 1e-06, 2e-05]$ and weight decays of $[0, 0.01]$.

E.2 Hardware Configuration

The experiments were conducted on 4 NVIDIA RTX A6000 GPUs with 48GB RAM and overall they were completed in about 330 GPU hours.

E.3 Overall Performance for all Sampling Methods

For completeness of our analysis in Section 5.2, Table 10 presents the overall performance (F_1^{overall}) for all sampling strategies described in Section 3.

E.4 Impact of the Annotators' Disagreement on Performance

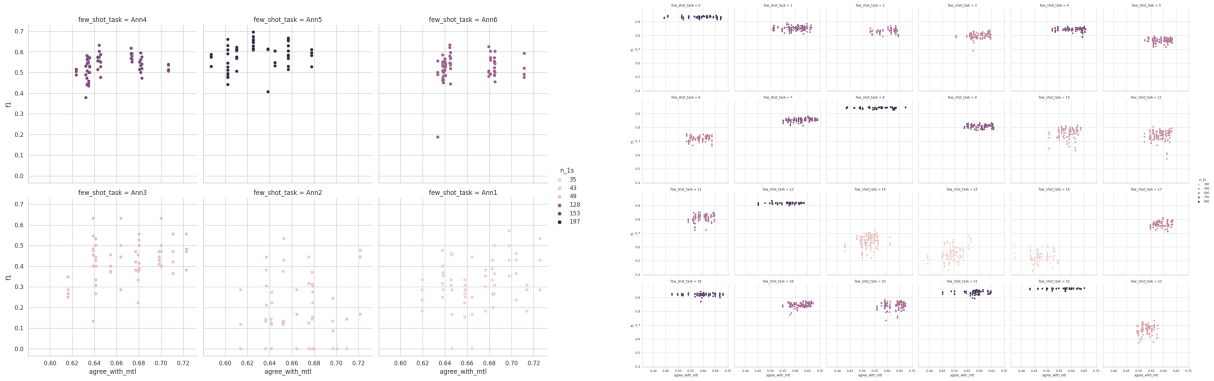
In Figure 8 we demonstrate the impact of agreement (as a measure of similarity) between the first and second-stage annotators ($\mathcal{A}_{mtl}$ and $\mathcal{A}_{fs}$) on the performance of the model for the second stage annotators. Importantly, we do not observe performance degradation as the agreement between the two sets decreases.

F Mathematical Symbols

Table 11 provides a directory of mathematical symbols used in our paper, along with their respective meanings, to facilitate ease of understanding for the reader.

$metric = F_1^{overall} \uparrow$		Brexit				MFSC			
		50%	66%	83%	100%	25%	50%	75%	100%
$\frac{X\% \times \lvert D_{a_i} \rvert}{\text{MTL}}$		$50\%\lvert D_{a_i}\rvert$	$66\%\lvert D_{a_i}\rvert$	$83\%\lvert D_{a_i}\rvert$	$\lvert D_{a_i}\rvert$	$25\%\lvert D_{a_i}\rvert$	$50\%\lvert D_{a_i}\rvert$	$75\%\lvert D_{a_i}\rvert$	$\lvert D_{a_i}\rvert$
		0.417(0.049)	0.449(0.027)	0.418(0.018)	0.431(0.014)	0.763(0.016)	0.773(0.011)	0.772(0.015)	0.776(0.004)
$\frac{X\% \times \lvert \mathcal{A} \rvert}{}$		$50\%\lvert\mathcal{A}\rvert$	$66\%\lvert\mathcal{A}\rvert$	$83\%\lvert\mathcal{A}\rvert$		$25\%\lvert\mathcal{A}\rvert$	$50\%\lvert\mathcal{A}\rvert$	$75\%\lvert\mathcal{A}\rvert$	
$k = 16$	$\mathcal{S}_{bal}$	0.443(0.005)	0.454(0.015)	0.457(0.015)		0.777(0.002)	0.779(0.0)	0.78(0.003)	
	$\mathcal{S}_{dis}$	0.421(0.01)	0.44(0.011)	0.457(0.008)		0.784(0.01)	0.787(0.004)	0.782(0.005)	
	$\mathcal{S}_{mv}$	0.426(0.008)	0.441(0.019)	0.455(0.019)		0.789(0.004)	0.786(0.004)	0.783(0.006)	
	$\mathcal{S}_{rand}$	0.422(0.012)	0.44(0.025)	0.455(0.007)		0.795(0.009)	0.79(0.006)	0.785(0.005)	
$k = 32$	$\mathcal{S}_{bal}$	0.449(0.008)	0.458(0.009)	0.458(0.008)		0.779(0.002)	0.78(0.001)	0.78(0.003)	
	$\mathcal{S}_{dis}$	0.423(0.008)	0.44(0.015)	0.457(0.016)		0.786(0.01)	0.788(0.004)	0.783(0.005)	
	$\mathcal{S}_{mv}$	0.424(0.017)	0.444(0.02)	0.458(0.011)		0.791(0.004)	0.787(0.003)	0.783(0.007)	
	$\mathcal{S}_{rand}$	0.428(0.006)	0.447(0.019)	0.452(0.016)		0.795(0.01)	**0.791**(0.006)	0.785(0.004)	
$k = 64$	$\mathcal{S}_{bal}$	0.453(0.003)	0.458(0.016)	0.459(0.011)		0.78(0.003)	0.781(0.003)	0.781(0.004)	
	$\mathcal{S}_{dis}$	0.436(0.01)	0.455(0.016)	0.468(0.01)		0.787(0.01)	0.789(0.004)	0.783(0.005)	
	$\mathcal{S}_{mv}$	0.427(0.007)	0.439(0.026)	0.459(0.013)		0.791(0.005)	0.788(0.003)	0.784(0.007)	
	$\mathcal{S}_{rand}$	0.433(0.012)	0.451(0.015)	0.456(0.013)		0.797(0.009)	0.791(0.006)	0.785(0.004)	
$k = 128$	$\mathcal{S}_{bal}$	**0.471**(0.002)	**0.474**(0.018)	**0.468**(0.014)		0.781(0.002)	0.781(0.002)	0.782(0.003)	
	$\mathcal{S}_{dis}$	0.45(0.008)	0.461(0.019)	0.466(0.016)		0.788(0.009)	0.789(0.003)	0.783(0.005)	
	$\mathcal{S}_{mv}$	0.434(0.015)	0.445(0.022)	0.458(0.016)		0.793(0.005)	0.788(0.004)	0.784(0.007)	
	$\mathcal{S}_{rand}$	0.439(0.015)	0.457(0.012)	0.455(0.011)		**0.798**(0.008)	0.791(0.005)	**0.786**(0.004)	

Table 10: Overall F_1 results on Brexit and MFSC datasets for different budgets of annotation (B), with various few shot sampling strategies; mean and standard deviation calculated over repeated runs.



(a) BREXIT (b) MFSC

Figure 8: Each plot demonstrates the effect of a single annotator's agreement with the initial set of annotators used for MTL training (A_{mtl}), on its F_1 score performance, when adopted as a few-shot task. The x-axis represents the agreement measure, and the y-axis represents the F_1 score. The darker color of the scatter plot corresponds to a higher number of positive labels provided by the respective annotator.

Symbol	Meaning
$\mathcal{A}_{fs}$	Annotators in MTL model
$\mathcal{A}_{mtl}$	Annotators adopted as few shot task
$\mathcal{S}_{mv}$	Sampling based on majority vote
$\mathcal{S}_{bal}$	Sampling based on balanced samples across classes
$\mathcal{S}_{dis}$	Sampling based on high disagreement of annotaions
$\mathcal{S}_{rand}$	Random sampling
B	Budget
D	All annotations for a dataset
F_1^{fs}	Avg. F_1 scores of the few-shot model for $\mathcal{A}_{fs}$
F_1^{mtl}	Avg. F_1 scores of the multi-task model for $\mathcal{A}_{mtl}$

Table 11: Mathematical notations used throughout the paper with their explanations