Capturing Perspectives of Crowdsourced Annotators in Subjective Learning Tasks

Negar Mokhberian[1], **Myrl G. Marmarelis**[1], **Frederic R. Hopp**[2],
Valerio Basile[3], **Fred Morstatter**[1], **Kristina Lerman**[1]
{nmokhber, myrlm, fredmors, lerman}@isi.edu, f.r.hopp@uva.nl, valerio.basile@unito.it
[1]Information Sciences Institute, University of Southern California
[2]School of Communication Research, University of Amsterdam
[3]Computer Science Department, University of Turin

Abstract

Supervised classification heavily depends on datasets annotated by humans. However, in subjective tasks such as toxicity classification, these annotations often exhibit low agreement among raters. Annotations have commonly been aggregated by employing methods like majority voting to determine a single ground truth label. In subjective tasks, aggregating labels will result in biased labeling and, consequently, biased models that can overlook minority opinions. Previous studies have shed light on the pitfalls of label aggregation and have introduced a handful of practical approaches to tackle this issue. Recently proposed multi-annotator models, which predict labels individually per annotator, are vulnerable to under-determination for annotators with few samples. This problem is exacerbated in crowdsourced datasets. In this work, we propose **Annotator Aware Representations for Texts** (**AART**) for subjective classification tasks. Our approach involves learning representations of annotators, allowing for exploration of annotation behaviors. We show the improvement of our method on metrics that assess the performance on capturing individual annotators' perspectives. Additionally, we demonstrate fairness metrics to evaluate our model's equability of performance for marginalized annotators compared to others.[1]

1 Introduction

Supervised machine learning tasks often train on data provided by human annotators. To improve the quality of ground truth data, the data samples are labeled by multiple annotators, whose independent decisions are then aggregated, often by taking the majority label or average value. This approach works well in some tasks, like image recognition, where "objective" ground truth exists. Objective judgments should not be affected by individual differences among annotators and in practice, most people would perceive and classify them similarly, except for obvious errors or oversight (Basile et al., 2021a).

However, some annotation tasks are subjective by nature, and traditional approaches to aggregating annotations that ignore context and subjectivity may lose important information (Plank, 2022). For instance, individuals may differ about what they consider to be toxic or hate speech on social media, depending on their background and attitudes (Sap et al., 2022). As a result, not all annotators will agree when a particular message constitutes hate speech. Disagreement among annotators in subjective tasks is expected because, for each item, the annotation outcome depends on (1) the annotator's perspective and lived experience, (2) the item being evaluated, (3) the time and circumstances surrounding the annotation task, and (4) the clarity of the designed prompts or instructions given to annotators (Basile et al., 2021b; Denton et al., 2021). In addition, annotators vary in how much effort they put into the task, with prolific annotators labeling many samples and sparse annotators labeling few samples. A classifier that does not take variety in human labels into account risks overlooking minority perspectives by ignoring potentially informative differences between annotators and integrating the biases of a particular annotator pool.

Historically, most public datasets did not provide labels at the individual annotator level. This limited the ability of machine learning researchers to develop methods that take advantage of differences in annotator perspectives in subjective learning tasks. Recently, researchers have acknowledged the importance of incorporating this information and have released raw annotations, often accompanied by demographic information about the annotators. It has now become commonplace to include multiple annotations per item in published datasets (Frenda

[1] Our code and implementation are publicly available at https://github.com/negar-mokhberian/aart.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 7337–7349
June 16-21, 2024 ©2024 Association for Computational Linguistics

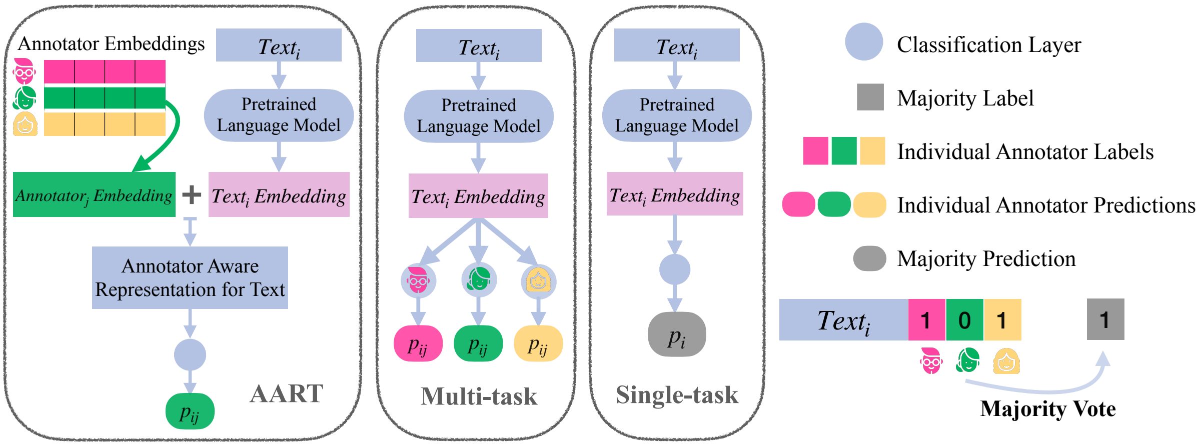


Figure 1: The architecture of AART compared to the multi-task and single-task approaches. Whereas a multi-task model incorporates annotator subjectivity by training entirely separate classification heads per annotator, AART represents annotator-specific biases as embeddings that are added to the text embeddings prior to classification. This new approach constrains the degrees of freedom and allows a common head across annotators, lending helpful inductive biases to sparse annotators.

et al., 2023; Sap et al., 2022; Prabhakaran et al., 2021; Denton et al., 2021). Recent studies have revealed that the aggregation of annotations exhibits detrimental biases toward specific groups of annotators (Prabhakaran et al., 2021; Aroyo et al., 2024).

Instead of adhering to the conventional notion that labels should be aggregated across all stages of AI pipelines, recent research advocates for leveraging the inherent diversity in human-assigned labels by pioneering innovative approaches to accommodate multiple votes per item at different stages of AI pipelines. This can be applied across domains such as model architecture design (Weerasooriya et al., 2023; Mostafazadeh Davani et al., 2022; Hayat et al., 2022; Gordon et al., 2022; Deng et al., 2023; Gordon et al., 2021), performance evaluation metrics (Basile et al., 2021b; Dumitrache et al., 2019; Gordon et al., 2021), or label aggregation, surpassing simplistic methods like averaging or majority voting (Aroyo and Welty, 2014; Dumitrache et al., 2018).

Despite some progress, few researchers have systematically tackled the problem of learning from multiple annotations. We address this gap by proposing a new approach for modeling annotator perspectives in subjective classification tasks via **Annotator Aware Representations for Texts (AART)**. Our approach learns annotator embeddings that represent their perspectives in downstream classification tasks (see §3 for more details). We demonstrate empirically that AART affords a more accurate and context-sensitive interpretation of subjective data at the individual annotator level.

We compare our approach to a widely recognized method (Mostafazadeh Davani et al., 2022) that captures differences in annotator judgments using a multi-task architecture consisting of a separate trainable classification-head per annotator. We show that the multi-task learning suffers from having under-trained classification heads for *sparse* annotators that contribute few annotations. Although this work marks an advance in the study of annotations of subjective tasks, its difficulties in accounting for differences in perspectives of infrequent, sparse annotators limits its utility in practical applications.

In contrast, our method captures the semantics inherent in subjective classification tasks, accounting for the diverse perspectives of annotators, including those who contribute few annotations. The AART architecture combines text embedding with annotator embeddings and provides this joint information to the classification layer. We couple this design choice with using the L2 regularizer in the loss, to prevent annotator embeddings from overfitting. Additionally, we use the contrastive loss to aid in learning from the similarities and dissimilarities of label choices among annotators, a capability lacking in Multi-task. Furthermore, in Section 4, we examine the learned annotation behaviors.

By leveraging metrics from algorithmic fairness, we demonstrate that the viewpoints of sparse annotators—who either contribute few samples or select less popular labels—are accurately represented (See §4.5). This inclusivity is crucial for applications where the diversity of perspectives is essential for understanding patterns in subjective annotations. Practitioners can utilize this approach to anticipate how individuals and, by extension, subgroups would react to a particular input.

2 Related Work

Sources of Annotator Disagreement. Some previous research has shown aggregating the labels takes away the voice of minority socio-demographic groups (Prabhakaran et al., 2021). Al Kuwatly et al. (2020) conducted a study grouping annotators according to socio-demographic attributes and discovered biases in the annotations of personal attack corpora is linked to annotators' age, education, and first language. Similarly, (Sap et al., 2022) found that posts written in African American dialect were more likely to be labeled as offensive by conservative annotators. Furthermore, (Larimore et al., 2021) demonstrated that there are differing interpretations among white and non-white workers when annotating racially sensitive topics, indicating the importance of considering the perspectives of those affected when gathering annotations of racist language.

It is important to note that many public datasets lack information regarding the demographics of the annotators alongside their annotations. Moreover, even in cases where datasets do report annotator demographics, mostly there exists an imbalance in the representation of demographics among annotators (Orlikowski et al., 2023). For instance, while females constitute the majority of crowd workers in the US, they are significantly underrepresented in most other countries. Additionally, online platforms tend to attract a predominantly young demographic of annotators (Posch et al., 2022; Hara et al., 2019).

However, it is essential to recognize that relevant lived experience and mindset among annotators do not always perfectly align with demographic characteristics (Orlikowski et al., 2023). Waseem (2016) found that amateur annotators were more prone to mislabel content as racist or sexist compared to experts with backgrounds in feminism and anti-racism activism. This finding emphasizes the need for researchers to identify annotator behaviors beyond relying solely on demographic characteristics. Hence, some previous work has explored methods to identify groups of annotators with similar annotation behaviors. Wich et al. (2020) employed graph methods, and (Lo and Basile, 2023) conducted hierarchical clustering of annotators, while (Akhtar et al., 2020) and (Basile, 2021) utilized exhaustive-search algorithms to automatically group human annotators in order to maximize the inter-group differences in annotation patterns.

Modeling Annotator Disagreement. Examining annotation disagreement has been a subject of interest in previous research. One line of research treats disagreement as noise and attempts to minimize it in order to improve performance (Mokhberian et al., 2022; Hovy et al., 2013). However, another line of research argues in favor of embracing annotation disagreement, considering it as a valuable signal rather than entirely noise (Aroyo and Welty, 2015; Rodrigues and Pereira, 2018). Our paper explores the latter possibility.

Certain studies have demonstrated that the level of disagreement among annotators for each data point can be predicted based on the text itself (Kenyon-Dean et al., 2018). Additionally, incorporating demographic information of the annotators has shown even better performance (Wan et al., 2023).

Recent work for numerous tasks has proposed to approximate the full annotators' label distribution on each item, instead of aggregating, utilizing methods such as soft-label learning (Fornaciari et al., 2021; Uma et al., 2021, 2022).

Other research has developed approaches to model annotators for subjective tasks and to predict a personalized label for each annotator. Plepi et al. (2022) explore methods of personalization to model Reddit users and analyze the effect of their method on demographic groups. Akhtar et al. (2020) employ a comprehensive search approach to divide annotators into two groups based on their annotation patterns. They demonstrate the effectiveness of an ensemble model comprising two separate classifiers representing the perspectives of each group over the traditional single-task model, which merely observes aggregated labels. Mostafazadeh Davani et al. (2022) explored the effectively of three different multi-annotator architectures for predicting each annotator's labels. Deng et al. (2023) integrate annotator embeddings into their model design, yet they were unable to identify significant annotator patterns solely based on these embeddings without incorporating additional embeddings for representing annotation labels. Gordon et al. (2022) introduced the concept of jury learning, a recommender system approach that specifies the demographic composition of individuals, along with their respective proportions, tasked with determining the predictions of the classifier. For example, in the context of online hate speech detection, a jury learning model might recommend the inclusion of women and Black jurors, as they are often primary targets of online harassment. Drawing motivation from this line of work, our paper investigates a new approach to model the subjectivity of tasks through label variations among individual annotators. By training annotator-specific embeddings with a contrastive loss, we learn a latent space that captures similarities and differences in annotator behavior.

3 Methodology

This section presents the details of our proposed approach and baselines used for evaluation. Our proposed method ***AART*** (§3.2) learns separate representations for each annotator in an end-to-end fashion during the training phase of the main classification task. In the inference phase, we are able to predict personalized labels for each annotator based on the annotation behaviors observed during training. A commonly used method for learning subjective perspectives is ***multi-task*** model (§3.3) which trains a separate classification head for each of the annotators. Other than comparing our method to the multi-task model we also compare it to a ***single-task*** (§3.4) model which is fed the aggregated labels during training and is not aware of the variety of annotations for each item. The overall framework of these approaches is shown in Fig. 1.

3.1 Problem Definition

We consider an input corpus $\mathcal{D} = \{(x_i, a_j, y_{ij})\}$ that is composed of triplets, with elements coming from text items $X = \{x_i\}_{i=1}^{N}$, annotators $A = \{a_j\}_{j=1}^{M}$, and annotations $y_{ij} \in \{1, ..., Q\}$. Each (i, j) combination that appears (at most once) in $\mathcal{D}$ means that annotator a_j assigned label y_{ij} to the text item x_i. In most annotated datasets, many y_{ij} values are not available as annotators only label a portion of items.

Most traditional classification studies aim to predict the aggregated label that would be assigned to text x_i by the annotator pool $\{a_j\}_{j=1}^{M}$. This aggregated label is usually calculated via majority-voting or averaging. This fundamental model, referred to as the *single-task* model in this paper (See §3.4), solely considers and predicts the aggregated labels. In this study, however, we are interested in predicting the label that annotator a_j would assign to an unseen text item x_i, i.e., predict y_{ij} given a pair of annotator and item (x_i, a_j). We use the overarching term "multi-annotator models" to encompass these models, and we will describe our proposed multi-annotator model, *AART* (see §3.2), and the established *multi-task* (see §3.3) model found in prior literature.

All the approaches we investigate in this paper leverage pretrained transformer-based language models for encoding text (such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019)). Given a text item x_i, we extract its generic representation by taking the embedding of the [CLS] token from the final layer of the language model, denoted as $e(x_i) \in \mathbb{R}^d$ (d is the size of the representation vector defined in the language model). This representation is then fine-tuned in conjunction with the other parameters of the model architectures during training.

3.2 Annotator-Aware Representations for Texts (AART)

We present a novel approach for incorporating annotator information into text representations for the purpose of learning an annotator's label on a given text instance. Our approach utilizes M annotator embeddings, with one randomly initialized embedding per annotator, as learnable parameters with the same dimensionality as the text embeddings. Given a text instance x_i and annotator a_j, the annotator-aware text embedding $g(x_i, a_j)$ is calculated as the sum of the text embedding $e(x_i)$ and annotator embedding $f(a_j)$, yielding $g(x_i, a_j) = e(x_i) + f(a_j)$. We also tried concatenating a one-hot encoding for each annotator with text embeddings which resulted in lower performance compared to taking the sum. Our idea draws inspiration from the positional encoding used by transformers (Vaswani et al., 2017), where a positional embedding of the same dimensionality as the token embedding is added to the token embedding at each position to form the sequence representation. A similar approach has been found effective in dialogue act classifications by learning speaker-turn embeddings in a conversation (He et al., 2021). We feed the annotator-aware representations of text $g(x_i, a_j)$

into a fully connected classification layer.

To train AART we propose the following objective function:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \sum_{j} \|f(a_j)\|_2 + \alpha \sum_{j,j'} \mathcal{L}_C(j, j')$$

Where the first term is a cross-entropy loss for predicting the item-annotator labels y_{ij}, the second term is the regularizing L_2 norm for annotator embeddings, and the third term is contrastive loss. We calculate InfoNCE contrastive loss (Oord et al., 2018; Chen et al., 2020) on pairs of a_j and $a_{j'}$ that have both worked on the same text x_i. When two annotators a_j and $a_{j'}$ have assigned the same label to x_i we consider a_j and $a_{j'}$ as a positive pair and our loss will push them closer in the latent space. We consider a negative pair to be any two annotators that have assigned different labels to that item. The contrastive loss for a positive pair of annotators (j, j') is defined as below:

$$\mathcal{L}_C(j, j') = \\ -\log \frac{\exp\left\{-\|a_j - a_{j'}\|_2^2/\tau\right\}}{\sum_{k \in \mathcal{N}_j \cup \{j'\}} \exp\left\{-\|a_j - a_k\|_2^2/\tau\right\}},$$

in which τ is a temperature constant and $\mathcal{N}_j$ is the set of all negatives for annotator j.

3.3 Multi-task Approach

The multi-task approach learns each annotator's labeling behavior via individual classification tasks, utilizing shared language model layers to derive a consistent representation of the input text, $e(x_i)$, for all annotators (Mostafazadeh Davani et al., 2022). To capture annotator-specific labeling behavior, this method has a separate fully connected layer specifically optimized for each annotator. However, the text encoding layers generating $e(x_i)$ are refined based on the outputs of all annotator tasks. The optimization objectives are formulated separately for each annotator as a cross-entropy loss function, which takes into account each annotator's available labels for each item x_i.

3.4 Single-task Approach

The baseline model embodies the typical approach of using a single-task classifier to predict the aggregated label for each instance. This model is constructed by adding a fully connected classification layer to the text embedding $e(x_i)$. The fully connected layer performs a linear transformation in order to predict the majority label. Unlike the multi-annotator models detailed in §3.2 and §3.3, the baseline model does not incorporate the individual annotator labels y_{ij}; rather, it directly predicts the aggregated label $\bar{y}_i$ for each item x_i.

4 Experiments

4.1 Datasets

The Multi-Domain Agreement dataset ($\mathcal{D}_{MDA}$): (Leonardelli et al., 2021) created an dataset for the task of offensive language detection. The dataset consists of 9,814 English tweets from three domains (Black Lives Matter movement, Election 2020, and COVID-19 pandemic). Each tweet was annotated for offensiveness by 5 annotators via Amazon Mechanical Turk.

English Perspectivist Irony Corpus ($\mathcal{D}_I$): The irony detection corpus, as presented by Frenda et al. (2023), encompasses 3,000 Post-Reply pairs extracted from social media. These pairs are evenly sourced from Twitter and Reddit, originating from five English-speaking countries: Australia, India, Ireland, the United Kingdom, and the United States. For Twitter data, the authors utilized the API geolocation service to distinguish the five English varieties. In the case of Reddit, data was gathered from specific subreddits, assuming the texts' origins: r/AskReddit (United States), r/CasualUK (United Kingdom), r/britishproblems (United Kingdom), r/australia (Australia), r/ireland (Ireland), r/india (India). The 74 annotators were balanced in terms of both gender and nationality, with approximately 15 raters assigned to each of the mentioned nationalities. Each rater labeled around 200 instances, resulting in a corpus containing 14,172 annotations, with a median of 5 annotations per instance.

The Racial Bias Toxicity Detection Corpus ($\mathcal{D}_{RB}$): Sap et al. (2019) have studied the effect of annotators' biases on their toxicity perceptions. They have confirmed in their work that determining what is harmful is a subjective task and annotators' racial biases can transmit to their annotations. They have gathered toxicity annotations for tweets on Amazon Mechanical Turk asking workers to determine whether a tweet (a) is offensive to them, and (b) could be seen as offensive to anyone. We work with the former as personal perspectives are more relevant to this study.

Dataset	*Single-task*	*Multi-task*	*AART* ($\alpha = 0$)	*AART* ($\alpha > 0$)
Annotator-level F1				
$\mathcal{D}_{MDA}$	66.80 ± 0.7	61.87 ± 1.5	$\underline{69.64 \pm 1.3}$	$\mathbf{69.72 \pm 1.1}$
$\mathcal{D}_{I}$	$\underline{58.59 \pm 1.9}$	57.31 ± 0.9	56.45 ± 7.9	$\mathbf{59.67 \pm 0.9}$
$\mathcal{D}_{RB}$	68.61 ± 1.5	66.57 ± 3.1	$\mathbf{71.6 \pm 1.6}$	$\underline{71.1 \pm 3.2}$
Global-level F1				
$\mathcal{D}_{MDA}$	71.99 ± 0.6	74.19 ± 0.9	$\mathbf{77.40 \pm 0.6}$	$\underline{77.38 \pm 0.4}$
$\mathcal{D}_{I}$	60.23 ± 1.7	$\mathbf{66.78 \pm 0.9}$	63.64 ± 5.4	$\underline{66.16 \pm 1.4}$
$\mathcal{D}_{RB}$	71.97 ± 1.7	77.72 ± 1.8	$\underline{79.38 \pm 2.4}$	$\mathbf{79.96 \pm 1.9}$
Item-level Disagreement Correlations				
$\mathcal{D}_{MDA}$	NA	0.33 ± 0.03	$\mathbf{0.39 \pm 0.03}$	$\underline{0.37 \pm 0.04}$
$\mathcal{D}_{I}$	NA	$\mathbf{0.23 \pm 0.04}$	0.14 ± 0.10	$\underline{0.20 \pm 0.06}$
$\mathcal{D}_{RB}$	NA	0.47 ± 0.05	$\underline{0.48 \pm 0.12}$	$\mathbf{0.54 \pm 0.04}$

Table 1: This table presents experimental results for the classification tasks. AART is the new method introduced here. Single-task is a baseline that only sees and predicts aggregated labels. We repeated each experiment 10 times and report the mean and standard deviation. α is the coefficient for contrastive loss. The best performing model is indicated in bold and the second best is underlined. See section §4 for further details.

Table 2: Statistics of the datasets used in our experiments. $\mathcal{D}_{MDA}$ refers to the offensive speech detection dataset (Leonardelli et al., 2021), $\mathcal{D}_{I}$ refers to the irony detection corpus (Frenda et al., 2023), and $\mathcal{D}_{RB}$ to the racial bias corpus (Sap et al., 2019).

	$\mathcal{D}_{MDA}$	$\mathcal{D}_{I}$	$\mathcal{D}_{RB}$
# Annotators	819	74	334
# Annotations per Annotator	60 ± 131	192 ± 27	35 ± 51
# Unique Texts	9814	2884	1309
# Annotations per Text	5	5	9 ± 2

4.2 Evaluation Metrics

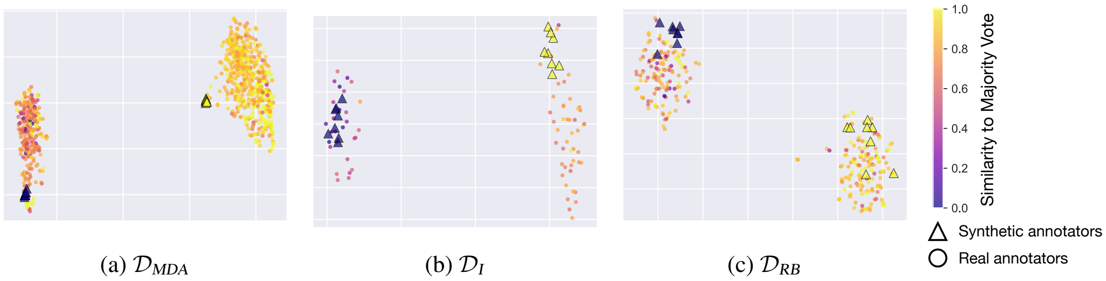
Annotator-level F1 This study is primarily motivated by the objective of safeguarding minority perspectives that might be overlooked when resolving annotator disagreements through the common practice of aggregating labels in machine learning. Consequently, it becomes crucial to demonstrate performance of model across all annotators, regardless of whether they have made substantial or minimal contributions to the annotated dataset. To address this, we introduce the metric of *Annotator-level F1* defined as the average macro F1 score across all annotators. The performance for each annotator a_j over items x_i appearing in the test split is calculated by comparing the existing ground truths y_{ij} to the model's predicted label for (x_i, a_j).

Global-level F1 The metric *Global-level F1* considers all pairs of item-annotator (x_i, a_j) appearing in the test set and calculates the macro F1 comparing their true label y_{ij} to the predicted label of the model for (x_i, a_j). In contrast to the Annotator-level F1, this metric implicitly gives preference to larger annotators, potentially overshadowing the performance of sparse annotators within the substantial contributions of prolific annotators.

Disagreement Measurement for Items In the context of multi-annotator models, a crucial consideration is whether the degree of disagreement in the predicted labels aligns with the actual degree of disagreement for each item. We follow the definitions from (Wan et al., 2023) for quantifying disagreement for each item. They introduce a continuous disagreement metric, quantifying the fraction of votes for an item that differs from the majority vote. For instance, if five votes for an item are recorded as 0,1,0,1,1, the disagreement is computed as 0.4. The disagreement scale ranges from 0 (complete unanimity in annotation results) to 1 (a significant divergence in opinions on annotation results). However, in the case of binary labels, the disagreement scale is confined to the range $[0, 0.5]$. Notably, we calculate predicted disagreement for each item exclusively based on annotators for whom ground truth information is available, aligning with the set of annotators used to calculate the ground truth disagreement per item.

4.3 Quantitative Analysis of Performance

Table 1 presents the numerical outcomes based on the metrics outlined in §4.2. The experimental setup is described in §A. Given that the primary motivation in this paper is to achieve equitable performance across all annotators, we initially focus on the *Annotator-level F1*. Across all three datasets, one of the two variants of the AART model (with or without the use of contrastive loss) consistently outperforms both single-task and multi-task models. This underscores the AART model's capacity to excel across diverse annotators, irrespective of



(a) $\mathcal{D}_{MDA}$ (b) $\mathcal{D}_I$ (c) $\mathcal{D}_{RB}$

Figure 2: The UMAP projection visualizes the annotator embeddings, with the colorbar indicating the fraction of annotator votes aligning with the majority vote. The smaller circles show the real annotators. Larger triangles represent two distinct sets of synthetic annotators integrated into the annotator pool for each dataset. Within each group of eight synthetic annotators, the yellow triangles consistently assign the majority vote, while the purple triangles consistently diverge from the majority. A distinct separation is evident between annotators aligning more closely with the majority opinion and those deviating further from the majority vote in their annotation patterns.

their contribution counts or other characteristics.

It's noteworthy that incorporating contrastive loss with AART proves particularly beneficial for the Irony detection dataset ($\mathcal{D}_I$). Without contrastive loss, the AART model faces challenges in converging on this dataset, possibly due to the inherent difficulty of automatically recognizing irony in text. The generally lower overall performances on $\mathcal{D}_I$ compared to the other two datasets affirm the challenging nature of this particular dataset. Meanwhile, the disparities in AART's performance with or without the contrastive loss on the other two datasets are marginal, indicating the overall benefit of including the contrastive loss in the training objective.

The *Global-level F1* metric assesses how well the predicted label for an annotator-item pair aligns with the true label. Notably, whenever the AART approach outperforms the two baselines on *Annotator-level F1*, it also excels on *Global-level F1*. This observation indicates that AART achieves a robust *Annotator-level F1* without merely overfitting to sparse annotators, maintaining a comprehensive understanding of item-annotator relationships across all pairs in the dataset.

The Pearson correlations between item-level labels disagreements, derived from true labels versus predicted labels, are presented in the third section of Table 1. The capacity to capture disagreements among annotators at the item level is a major advantage of multi-annotator models over conventional single-task models. The results demonstrate the effectiveness of utilizing contrastive loss ($\alpha > 0$) in improving the correlation of item-level disagreements when compared to the *AART* model, which does not employ contrastive loss ($\alpha = 0$). This enhancement is observed in the case of both $\mathcal{D}_I$ and $\mathcal{D}_{RB}$ datasets. However, in the case of $\mathcal{D}_{MDA}$, there is a marginal decline in performance, although it still outperforms the multi-task model.

4.4 Annotation Patterns Captured in Embeddings

In this section, we aim to gain a deeper understanding of the annotation patterns captured in the annotator embeddings. We are particularly interested in exploring whether the embeddings are influenced by the annotators' tendency to align with the majority opinion. To investigate this, we added two sets of eight synthetic annotators into the annotator pool for each dataset. Each synthetic annotator was exposed to $1/8$ of the data items. Within each set, eight annotators consistently provided the majority vote (determined from true labels) for the items, while the remaining eight consistently assigned a label different from the majority vote.

Upon completion of the AART training, we illustrate the embeddings in a 2D plot using UMAP projection (McInnes et al., 2018). Figure 2 demonstrates that the two groups of synthetic annotators exhibit distinct separation from each other. Also, the real annotators aligning more closely with the majority opinion and those deviating further from the majority vote in their annotation patterns are showing divergence in the embeddings. This observation shows the similarity to the majority opinion indeed plays a significant role in shaping the annotator embeddings.

We encourage future research to investigate additional patterns of similarity among annotators. These similarities may stem from shared demographic features or the conditions in which annotators have been provided with prompts for the annotation task (Sap et al., 2019).

4.5 Fairness to Marginalized Annotators

Similarity to Majority			
	Single-Task	Multi-task	AART
$\mathcal{D}_{MDA}$	0.14 ± 0.02	0.07 ± 0.01	0.07 ± 0.01
$\mathcal{D}_{I}$	0.14 ± 0.02	0.12 ± 0.02	0.12 ± 0.02
$\mathcal{D}_{RB}$	0.20 ± 0.04	0.04 ± 0.03	0.05 ± 0.03
Count of Contributions			
	Single-Task	Multi-task	AART
$\mathcal{D}_{MDA}$	0.03 ± 0.02	0.14 ± 0.04	0.06 ± 0.02
$\mathcal{D}_{I}$	0.03 ± 0.02	0.03 ± 0.02	0.02 ± 0.02
$\mathcal{D}_{RB}$	0.07 ± 0.03	0.07 ± 0.05	0.03 ± 0.03

Table 3: Statistical Parity based on two sensitive attributes for the annotators: (1) Similarity to majority vote and (2) Count of Contributions. Mean and Standard deviations are reported based on running with 10 different random seeds.

In this section, we assess the model's ability to capture the perspectives of annotators who exhibit minority opinion, and those who contribute few annotations. We do this through the lens of Statistical Parity, a popular metric for algorithmic fairness (Mehrabi et al., 2021). Statistical parity measures the relative rate of a successful outcome for the majority group against that of the minority group. In separate experiments, we study the majority and minority groups from two perspectives: (1) the similarity of annotation selections compared to the majority vote and (2) the count of contributions. In (1), the grouping of annotators is determined by the fraction of annotations generated by a labeler that are equal to the majority vote of the items they annotated. The annotators that fall in the lowest 25th percentile in terms of similarity to the majority vote are considered in the minority group and the rest fall into the majority group. In (2), the minority group of annotators is determined by the lowest 25th percentile in terms of the number of contributions. We compare Statistical Parity across *Single-task*, *Multi-task*, and *AART*. Our interest lies in showing whether each method performs equally accurate for the sensitive group versus the rest of annotators.

Table 3 displays the results of disparity among the two groups of annotators. The lower disparity values indicates the predictions are more fair. Observing the grouping of annotators based on their similarity to majority behavior, we note that the Single-task approach which consistently predicts the majority vote exhibits the most significant disparity in accuracy between the two groups. Additionally, AART and Multi-task show similar levels of disparity, which is less compared to the Single-task approach. This observation confirms that aggregating labels causes unfairness in accuracy toward annotators whose opinions differ from the majority behavior. And motivates the use of unaggregated annotations and multi-perspective classification models instead of aggregated ground truths.

When we group annotators based on their count of contributions, Multi-task shows the most disparity among the sensitive group of annotators and others, while AART is more equitable than the other two methods. This observation confirms that Multi-task is limited for sparse annotators and performs poorly for them, whereas AART demonstrates more balanced performance among sparse and prolific annotators.

5 Discussion

Predicting a single label remains the most popular problem in classification applications. Previous studies have demonstrated that multi-annotator models can perform comparably to single-task models in predicting the aggregated labels (Mostafazadeh Davani et al., 2022; Hayat et al., 2022; Al Kuwatly et al., 2020). In the single-task model, the aggregated labels are computed directly, while in multi-annotator models, the aggregated label is derived by consolidating predictions across annotators. We assess the performance of predicting the aggregated label and present the results in Table 4. Our findings align with previous observations, indicating that all three models effectively predict single label outcomes.

An additional advantage of training multi-annotator models compared to single-task models is the ability to learn the entire label distribution rather than solely focusing on aggregated labels. This flexibility proves valuable in various real-world scenarios. As models are increasingly tailored for more personal, subjective use cases, a direct application of this work could be to predict personalized preferences for each annotator.

Furthermore, (Gordon et al., 2022) introduces an inference approach for trained multi-annotator models, advocating for the selection of a task-specific subset of annotators akin to a "Jury." For instance, for the task of toxicity detection in online posts, prioritizing votes from women and Black individuals—often targeted by online harassment—

Dataset	*Single-task*	*Multi-task*	*AART* ($\alpha = 0$)	*AART* ($\alpha > 0$)
		Aggregated Label Performance		
$\mathcal{D}_{MDA}$	78.58±0.9	78.99±1.3	79.51±1.0	79.66±1.2
$\mathcal{D}_{I}$	64.05±2.9	65.47±2.5	60.22±9.4	64.95±2.1
$\mathcal{D}_{RB}$	81.09±3.5	81.43±3.1	81.43±3.5	81.02±1.6

Table 4: In this table the performance (F1) of the three models on predicting the aggregated label is reported. The single-task model computes the majority labels directly. For both the multi-task model and our proposed method (AART), we determined the majority of predictions across labels for each annotator.

could prove advantageous. Their goal with jury learning is to assist practitioners in acknowledging and incorporating annotator disagreement within the classifier pipeline. To achieve this, they encourage practitioners to form a jury that specifies the individuals or groups their classifiers should emulate.

Rather than completely disregarding the diversity present in human labels, our paper proposes an effective approach to predict individual perspectives while maintaining competence in predicting the aggregated label. Instead of eliminating disagreement at the initial stages of machine learning, our aim is to learn from the various perspectives of annotators. By embracing the diversity of opinions through data-driven pipelines, which enhances data transparency, we are able to present it to end-users and human decision-makers. This empowerment allows them to take corrective actions and interventions by leveraging common sense and contextual information.

6 Conclusion

This paper tackles the challenges posed by annotator disagreements in subjective tasks, where establishing a single ground truth is not straightforward. We propose an approach named Annotator Aware Representations for Texts (AART) for modeling annotators' perspectives. Our methodology combines representations of individual annotators with those of text in the classification task. By learning the embeddings of annotators through Cross Entropy loss, contrastive loss, and regularization, we capture their individual labeling patterns. We compare our approach to a popular multi-task architecture, which is constrained by under-trained classification heads for annotators with limited annotations. Through several experiments, we demonstrate that AART, despite its simplicity, effectively captures the semantics of textual data for classification tasks while accommodating different annotators' perspectives. Moreover, we show that AART is fair and equitable when it comes to marginalized annotators, who may have few labeled samples or provide votes that often differ from the majority votes for samples.

Limitations

The performance of the models outlined in this paper may be adversely affected when dealing with noisy annotations. This noise can arise from unclear instructions or insufficient attention from the worker. A limitation of this study, as well as many others focusing on designing multi-annotator models, is the tendency to overlook the high likelihood of certain annotators being unreliable. Differentiating between labeling noise and natural disagreement is a challenge that remains unaddressed in this work, and in numerous other studies. Additionally, the degree of overlap among items annotated by multiple individuals can enhance the learning process for multi-annotator models by exposing both similarities and differences among the annotators. Nevertheless, crowdsourced datasets are inherently imperfect and susceptible to noise or sparsity. It is imperative for future research to examine the robustness of multi-annotator models in the face of sub-optimal dataset features.

Ethical Considerations

This work seeks to improve AI systems by incorporating perspectives beyond just the most prolific annotators, as is currently done in most AI systems. This will encourage developers of subsequent AI systems to recruit more sparse annotators, as the difference in perspective they bring will advance the performance of their resulting models, rather than hinder it. This advances the ACM and ACL ethics code's core tenet that "all people are stakeholders in computing," and to "take action not to discriminate."[2] In this case, the discrimination would be against those who do not provide a wealth of annotations, due to time or other constraints.

[2] https://www.acm.org/code-of-ethics

As with any approach, there is the possibility that it could cause ethical harm. For example, a coordinated and nefarious minority could have amplified sway in our model compared to one that only focuses on the most prolific annotators. Furthermore, research has shown that large language models, such as the ones we use in this work, capture various biases that can harm our proposed model. Thus, the perspective we learn for each annotator is through the lens of these models, and it might be different from reality. Nevertheless, we maintain that this potential harm is vastly outweighed by the potential of incorporating all voices.

Acknowledgements

We express gratitude to the anonymous reviewers for providing valuable feedback and offering suggestions for our project. This work was funded in part by Defense Advanced Research Projects Agency (DARPA) and Army Research Office (ARO) under Contract No. W911NF-21-C-0002 and by the *Multilingual Perspective-Aware NLU* project in partnership with Amazon Alexa.

References

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.

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Online. Association for Computational Linguistics.

Lora Aroyo, Alex Taylor, Mark Diaz, Christopher Homan, Alicia Parrish, Gregory Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. 2024. Dices dataset: Diversity in conversational ai evaluation for safety. *Advances in Neural Information Processing Systems*, 36.

Lora Aroyo and Chris Welty. 2014. The three sides of crowdtruth. *Human Computation*, 1(1).

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.

Valerio Basile. 2021. It's the end of the gold standard as we know it: Leveraging non-aggregated data for better evaluation and explanation of subjective tasks. In *AIxIA 2020–Advances in Artificial Intelligence: XIXth International Conference of the Italian Association for Artificial Intelligence, Virtual Event, November 25–27, 2020, Revised Selected Papers*, pages 441–453. Springer.

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

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021b. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

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.

Emily Denton, Mark Díaz, Ian Kivlichan, Vinodkumar Prabhakaran, and Rachel Rosen. 2021. Whose ground truth? accounting for individual and collective identities underlying dataset annotation. *arXiv preprint arXiv:2112.04554*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2019. A crowdsourced frame disambiguation corpus with ambiguity. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2164–2170, Minneapolis, Minnesota. Association for Computational Linguistics.

Anca Dumitrache, Oana Inel, Lora Aroyo, Benjamin Timmermans, and Chris Welty. 2018. Crowdtruth 2.0: Quality metrics for crowdsourcing with disagreement. *arXiv preprint arXiv:1808.06080*.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021.

Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online. Association for Computational Linguistics.

Simona Frenda, Alessandro Pedrani, Valerio Basile, Soda Marem Lo, Alessandra Teresa Cignarella, Raffaella Panizzon, Cristina Marco, Bianca Scarlini, Viviana Patti, Cristina Bosco, and Davide Bernardi. 2023. EPIC: Multi-perspective annotation of a corpus of irony. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13844–13857, Toronto, Canada. Association for Computational Linguistics.

Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Mitchell L Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein. 2021. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Benjamin V Hanrahan, Jeffrey P Bigham, and Chris Callison-Burch. 2019. Worker demographics and earnings on amazon mechanical turk: An exploratory analysis. In *Extended abstracts of the 2019 chi conference on human factors in computing systems*, pages 1–6.

Hassan Hayat, Carles Ventura, and Agata Lapedriza. 2022. Modeling subjective affect annotations with multi-task learning. *Sensors*, 22(14).

Zihao He, Leili Tavabi, Kristina Lerman, and Mohammad Soleymani. 2021. Speaker turn modeling for dialogue act classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2150–2157, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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, Atlanta, Georgia. Association for Computational Linguistics.

Kian Kenyon-Dean, Eisha Ahmed, Scott Fujimoto, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhanderi, Robert Belfer, Nirmal Kanagasabai, Roman Sarrazingendron, Rohit Verma, and Derek Ruths. 2018. Sentiment analysis: It's complicated! In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1886–1895, New Orleans, Louisiana. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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, Online. Association for Computational Linguistics.

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.

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*.

Soda Marem Lo and Valerio Basile. 2023. Hierarchical clustering of label-based annotator representations for mining perspectives. In *Proceedings of the 2nd Workshop on Perspectivist Approaches to NLP co-located with 26th European Conference on Artificial Intelligence (ECAI 2023), Kraków, Poland, September 30th, 2023*, volume 3494 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35.

Negar Mokhberian, Frederic R Hopp, Bahareh Harandizadeh, Fred Morstatter, and Kristina Lerman. 2022. Noise audits improve moral foundation classification. In *Proceedings of the 2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*.

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.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

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, Toronto, Canada. Association for Computational Linguistics.

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.

Joan Plepi, Béla Neuendorf, Lucie Flek, and Charles Welch. 2022. Unifying data perspectivism and personalization: An application to social norms. *EMNLP*.

Lisa Posch, Arnim Bleier, Fabian Flöck, Clemens M. Lechner, Katharina Kinder-Kurlanda, Denis Helic, and Markus Strohmaier. 2022. Characterizing the global crowd workforce: A cross-country comparison of crowdworker demographics. *Human Computation*, 9(1):22–57.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Filipe Rodrigues and Francisco C. Pereira. 2018. Deep learning from crowds. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Alexandra Uma, Dina Almanea, and Massimo Poesio. 2022. Scaling and disagreements: Bias, noise, and ambiguity. *Frontiers in Artificial Intelligence*, 5.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's voice matters: Quantifying annotation disagreement using demographic information. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):14523–14530.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

Tharindu Cyril Weerasooriya, Alexander Ororbia, Raj Bhensadadia, Ashiqur KhudaBukhsh, and Christopher Homan. 2023. Disagreement matters: Preserving label diversity by jointly modeling item and annotator label distributions with DisCo. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4679–4695, Toronto, Canada. Association for Computational Linguistics.

Maximilian Wich, Hala Al Kuwatly, and Georg Groh. 2020. Investigating annotator bias with a graph-based approach. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 191–199, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Experimental Setup

Common parameters: We conducted our experiments using a shared server in our institution with a GPU NVIDIA GeForce RTX 2080 Ti 11GB.

Across all three approaches we set max sequence length to 100, training batch size to 100, and learning rate to $5\mathrm{e}^{-5}$. The training steps use the Adam optimizer (Kingma and Ba, 2014) with a weight decay rate equal to 0.01. We use early stopping with 20 max number of epochs and we save the model parameters from the epoch with best annotator-level F1. For the language model we use “roberta-base" for $\mathcal{D}_{MDA}$ and $\mathcal{D}_{I}$ datasets. However, because the $\mathcal{D}_{RB}$ is a much smaller dataset (only contains 1309 text items) we use a language model pretrained on the task of offensiveness detection “cardiffnlp/twitter-roberta-base-offensive"[3]. We load the transformer models from HuggingFace Library (Wolf et al., 2020).

We initiate the splitting process with a train/dev/test split of 50/25/25 percent stratified on item-level disagreements, after which we transfer the items from dev and test splits annotated by an annotator which is unseen in the training set to augment the training data. We repeat the splitting with 10 random seeds.

AART parameters: We set $\tau = 0.07$ for contrastive loss and performed hyper-parameter search for $\alpha \in \{0.1, 0.2, 0.3\}$ and for $\lambda \in \{0.0, 0.1, 0.2\}$ which are respectively coefficients for contrastive loss and for L_2 norm regularization in the objective function.

[3] https://huggingface.co/cardiffnlp/twitter-roberta-base-offensive