Learning Subjective Label Distributions via Sociocultural Descriptors

Anonymous ACL submission

Abstract

Subjectivity in NLP tasks, *e.g.*, toxicity classification, has emerged as a critical challenge precipitated by the increased deployment of NLP systems in content-sensitive domains. Conventional approaches aggregate annotator judgements (labels), ignoring minority perspectives and overlooking the influence of the sociocultural context behind such annotations. We propose a framework where subjectivity in binary labels is modeled as an empirical distribution accounting for the variation in annotators through *human values* extracted from sociocultural descriptors using a language model. The framework also allows for downstream tasks such as population and sociocultural group-level majority label prediction. Experiments on three toxicity datasets covering human-chatbot conversations and social media posts annotated with diverse annotator pools demonstrate that our approach yields well-calibrated toxicity distribution predictions across binary toxicity labels, which are further used for majority label prediction across cultural subgroups, improving over existing methods.

1 Introduction

Early machine learning models were evaluated using tasks with clearly defined ground truths, such as handwritten digit recognition (MNIST), spam detection (UCL Spambase) and categorical object recognition (ImageNet). These tasks relied on relatively hard facts, leaving little room for ambiguity. However, as AI systems are increasingly deployed in domains that involve higher subjective interpretation, defining the ground truth has become a complex and persistent challenge in tasks such as detection of toxicity in text (Lebovitz et al., 2021; Jaton, 2021). The ambiguity in labeling subjective tasks arises from the experience and perspective of annotators, and inherent ambiguities in text (Basile et al., 2021). For example, Figure 1 shows a text item that contains arguably offensive content labeled for toxicity differently by US and Indian annotators. This discrepancy can be attributed to varying levels of familiarity with the context of the offensive term by annotators from different localities and sociocultural background.

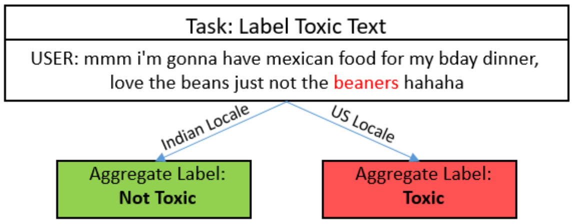

Figure 1: Example from the DICES dataset illustrating how the term "Beaners" is perceived differently by annotators from India and the US.

Toxicity detection has emerged as one of the most critical subjective tasks in natural language processing (NLP) due to its implications for the evaluation of conversational artificial intelligence (AI), safety guardrails in generative AI, and online content moderation (Wulczyn et al., 2017; Ziegler et al., 2019; Madhyastha et al., 2023; Ji et al., 2023). These systems often rely on crowdsourced annotations, reflecting diverse human perspectives shaped by annotators' sociocultural contexts. Conventional approaches typically aggregate these annotations through majority voting or averaging to produce "ground truth" labels that marginalize minority perspectives and risk reinforcing biases among the annotators selected for the construction or evaluation of NLP systems (Prabhakaran et al., 2021). Alternatively, a different line of research attempts to model every annotator behavior separately, thus ignoring shared perceptions among annotators and limiting scalability to more comprehensive populations (Davani et al., 2022; Mokhberian et al., 2023).

To address these challenges, recent toxicity datasets have incorporated detailed sociocultural information (demographics, beliefs, *etc.*) of annotators that can act as meaningful descriptors connecting annotators within and across populations, along with multiple annotations per instance (Aroyo et al., 2023; Davani et al., 2024a). To the best of our knowledge, the proposed *Learning Subjective Label Distribution (LSLD)* is the first work to model subjectivity in binary labels as distributions over the sociocultural descriptors of annotators. Our key contributions are as follows.

- A novel framework for modeling subjectivity in a binary labeling task from a text item as an empirical probability distribution, incorporating both i) language-model-generated human value perspectives derived from the input text and ii) annotators' sociocultural backgrounds.
- Comprehensive evaluation against existing baselines using three metrics accounting for individual probabilistic predictions for text-item-annotator pairs, calibration of predicted distributions, and aggregated item-level predictions.
- Demonstration of the framework's utility in tailored tasks such as population-level and sociocultural subgroup-level majority label prediction.

2 Subjective Label Distribution Learning

Problem Definition Let us define an annotated dataset $\mathcal{D} = (\mathcal{X}, \mathcal{A}, \mathcal{T}, \mathcal{Y})$, where: $\mathcal{X} = \{x_n\}_{n=1}^{N}$ is a set of N text instances, $\mathcal{A} = \{a_m\}_{m=1}^{M}$ is a set of M annotators, $\mathcal{T} = \{t_m\}_{m=1}^{M}$ is the set of characteristic vectors that describe the sociocultural background of all annotators in $\mathcal{A}$, such that $t_j \in \mathcal{T}$ represents the sociocultural descriptors for annotator $a_j \in \mathcal{A}$. Moreover, t_j has dimension k and each mixed-type coordinate (categorical or continuous) corresponds to a distinct sociocultural descriptor, *e.g.*, gender, race, age, education and locality. Finally, $\mathcal{Y}$ is an annotation matrix whose entries $y_{ij} \in \{0, 1\}$ denote the binary decision label assigned to the text instance x_i by the annotator a_j. Notably, annotators a_j only annotate subsets of text instances, leading to high missingness in $\mathcal{Y}$. In our use case, these labels represent *toxicity judgments* (safe *vs.* unsafe), however, the proposed methods are generalizable to other tasks involving subjective judgments with binary calls.

The task of *learning the distribution of judgments in a population of sociocultural descriptors* is formally defined as estimating $p(y_i = 1|x_i, \mathcal{T})$, where $y_i = 1$ is the judgment for x_i taking a particular value and the distribution is across the whole set $\mathcal{T}$. Thus, by conditioning the predictions on the sociocultural attributes of the annotator, LSLD achieves scalability toward a wider population sharing those features.

2.1 Modeling Conflicting Human Perspectives

Subjectivity in toxicity detection arises from the diverse human values and perspectives that influence how an individual interprets text items. Directly modeling text instances without accounting for these conflicting viewpoints can lead to models that are agnostic to the underlying diversity of human judgment. Recent work by Hayati et al. (2023) demonstrated that large language models (LLMs) are effective in extracting diverse human perspectives on subjective topics using criteria-based prompting.

Inspired by this, we propose generating *distinct human-value perspectives* of annotators who rate each text instance $x_i \in \mathcal{X}$ as safe or unsafe. Specifically:

1. For each x_i, we prompt an LLM to generate n human values of those who rate it as "safe" and an equal number of those who rate it as "unsafe". In our experiments, we keep $n = 2$ for simplicity. Thus, we obtain two human values for those who agree with the safe label ($\mathbf{v}_i^{\text{S1}}$ and $\mathbf{v}_i^{\text{S2}}$) and two other values for those who agree with the unsafe label ($\mathbf{v}_i^{\text{U1}}$ and $\mathbf{v}_i^{\text{U2}}$). The details of the prompt are presented in Appendix A.1 and an analysis of performance differences due to variation of n is discussed in Appendix A.2.
2. Each perspective is encoded into an embedding vector (of fixed size) using a pretrained sentence-BERT embedding model (Reimers and Gurevych, 2019).
3. The final contextualized embedding $f(x_i)$ for text instance x_i is obtained as the element-wise average of these four perspective embeddings. This embedding thus captures the diverse perspectives surrounding x_i and serves as input to the subsequent prediction module.

Alternative embedding combination methods (*e.g.*, concatenation or weighted averaging) were also explored, but we found element-wise averaging to be effective in our experiments.

The prediction module is designed to estimate the probability $\hat{p}_{ij} = p(y_i = 1|x_i, t_j)$ that a text instance $x_i \in \mathcal{X}$ is labeled as toxic (*i.e.*, unsafe) by annotators sharing the same sociocultural descriptors t_j. Specifically, all annotators $a_j \in \mathcal{A}$ with

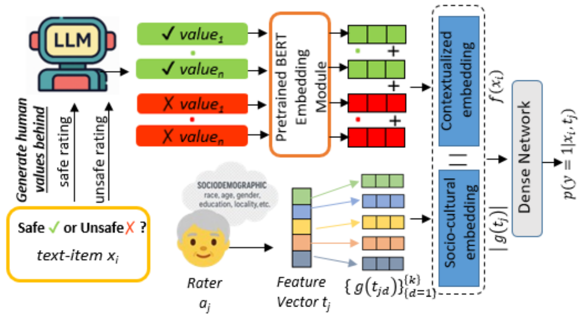

Figure 2: LSLD Model Architecture. The embeddings from the human values for "safe" and "unsafe" rating generated by the LLM using the text item are concatenated with sociocultural embedding formed from learnable embedding layers for each sociocultural descriptor of an annotator and are then fed to a dense network that produces an individual probabilistic prediction for an annotator and text item pair.

identical characteristic vectors t_j will be assigned the same predicted probability $\hat{p}_{ij}$, as their sociocultural profiles are indistinguishable in the model (in the absence of additional information about the annotators). The predictions are made through a two-step process described below.

Encoding Sociocultural Characteristics Each element of the characteristic vector $t_j = \{c_1, c_2, \ldots, c_k\}$, which describes the annotator $a_j \in \mathcal{A}$ is encoded in a fixed-size vector. For categorical features, this is achieved through an embedding layer, while for continuous features, a linear projection layer is used to map the feature value into a fixed-dimensional space. Let e_d denote the embedding layer (for categorical features) or the projection layer (for continuous features) corresponding to the d-th characteristic, where $d \in \{1, 2, \ldots, k\}$. For a given value c_d of the d-th characteristic, the corresponding vector $\mathbf{e}_d$ is obtained as:

$$\mathbf{e}_d = e_d(c_d).$$

Each embedding or projection layer e_d maps (or transforms) the unique values of the d-th characteristic to a vector of dimension m (*e.g.*, $m = 5$). This results in k vectors $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_k\}$ for each annotator a_j. We define the concatenated embedding vector $g(t_j)$ as:

$$g(t_j) = [\mathbf{e}_1; \mathbf{e}_2; \ldots; \mathbf{e}_k],$$

where $[;]$ denotes the concatenation operation and the dimension of $g(t_j)$ is km.

Combining Embeddings to make Predictions The contextualized text embedding $f(x_i)$ is concatenated with the sociocultural embedding vector $g(t_j)$ to form a combined input vector $\mathbf{v}_{ij}$:

$$\mathbf{v}_{ij} = [f(x_i); g(t_j)],$$

where the concatenated vector $\mathbf{v}_{ij}$ is of dimension $\dim(f(x_i)) + km$.

This combined vector is fed through a dense neural network with trainable parameters. The network consists of multiple fully connected layers followed by a sigmoid activation function (see Appendix A.5). The output of the model, denoted as $\hat{p}_{ij} \in (0, 1)$, represents the probability that x_i is labeled as toxic by the annotator $a_j \in \mathcal{A}$ with characteristic vector $t_j \in \mathcal{T}$. The architecture of the LSLD model is described in Figure 2.

2.2 Loss Function

Our training objective is twofold: i) to ensure that predicted toxicity probabilities align with the ground truth labels provided by annotators with respect to their sociocultural descriptors, and ii) to ensure that the empirical distribution Q of predicted probabilities for each text instance reflects the overall distribution P behind ground truth labels on the instance. To achieve this, we employ a composite loss function consisting of three terms: cross-entropy, Kullback-Leibler (KL) divergence, and L2 regularization. The loss $\mathcal{L}$ is defined as:

$$\begin{aligned}\mathcal{L} &= \textstyle\sum_i \sum_j \mathcal{L}_{\text{CE}}(y_{ij}, \hat{p}_{ij}) \qquad (1)\\ &+ \lambda_1 \textstyle\sum_i \text{KL}(P \parallel Q) + \lambda_2 \sum_{j=1}^{M} \|g(t_j)\|_2^2,\end{aligned}$$

where:

- $\mathcal{L}_{\text{CE}}(y_{ij}, \hat{p}_{ij})$ is the binary cross-entropy loss between the ground truth label y_{ij} and the predicted toxicity probability $\hat{p}_{ij}$ for the text item x_i and the annotator a_j.
- $\text{KL}(P \parallel Q)$ is the Kullback-Leibler (KL) divergence between two (empirical) binomial distributions, P formed by ground-truth ratings for text instance x_i and Q formed from ratings from probabilistic predictions on the same instance. Specifically,

$$P : y_i \sim \text{Bin}(n_i, \bar{y}_i), \quad Q : y_i \sim \text{Bin}(n_i, \hat{p}'_i),$$

where n_i is the number of annotations for instance x_i, and $\bar{y}_i$ and $\hat{p}'_i$ are aggregates for

$\{y_{ij}\}_{j=1}^{n_i}$ and $\{\hat{p}_{ij}\}_{j=1}^{n_i}$, respectively, defined below. Then, the KL divergence is given by:

$$KL(P \parallel Q) = n_i \bar{y}_i \cdot \ln\left(\frac{\bar{y}_i}{\hat{p}'_i}\right) + n_i(1 - \bar{y}_i) \cdot \ln\left(\frac{1 - \bar{y}_i}{1 - \hat{p}'_i}\right). \quad (2)$$

Although we have discrete realizations (0/1) from P as ground truth labels to obtain $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}$, we only have predicted probabilities for the realizations of Q. To obtain $\hat{p}'_i$ which is the mean of the realizations from Q, we calculate the mean after converting each predicted probability of an instance into *approximately* binary labels using the ground-truth item-level mean rating $\bar{y}_i$ as a reference using:

$$\hat{p}'_i = \frac{1}{2} \cdot \frac{\sum_{j=1}^{n_i} (1 + \tanh(k \cdot (\hat{p}_{ij} - \bar{y}_i))}{n_i},$$

where the hyperbolic tangent (tanh) activation function, with a large constant $k = 10^4$ (see Appendix A.5), serves as a relaxation to using hard-thresholded predictions while allowing smooth gradient flow during training.

- λ_1 and λ_2 are hyperparameters controlling the contribution of the KL divergence and L2 regularization terms, respectively. In the experiments $\{\lambda_1, \lambda_2\}$ are set by grid search using cross-validation (see Appendix A.5).

3 Related Work

Subjectivity in NLP The study of subjectivity in NLP tasks has a long history, with early work by Wiebe et al. (2004); Alm (2011); Pang et al. (2008). Researchers have since differentiated between two main sources of disagreement in annotations: *random variation* and *systematic disagreement* (Krippendorff, 2011). Systematic disagreement has been shown to influence tasks such as part-of-speech tagging (Plank et al., 2014), word sense disambiguation (Passonneau et al., 2012; Jurgens, 2013), and co-reference resolution (Poesio and Artstein, 2005; Recasens et al., 2011). However, its impact is particularly pronounced in controversial tasks such as hate speech detection (Akhtar et al., 2019, 2020; Warner and Hirschberg, 2012) and sentiment analysis (Liu et al., 2010; Kenyon-Dean et al., 2018).

Systematic disagreements among annotators have been attributed to several factors: i) *sociocultural differences*, where annotators' backgrounds, including gender, race, age, and beliefs significantly influence their judgments (Larimore et al., 2021; Sap et al., 2021; Basile et al., 2021); ii) *instance semantic ambiguity*, where ambiguity in the text itself can lead to divergent interpretations (Aroyo and Welty, 2013; Dumitrache, 2015; Basile et al., 2021); and iii) *annotator experience*, where prior experience with annotation tasks can shape annotators' perspectives (Waseem, 2016).

Recent studies have increasingly recognized the crucial role of sociocultural contexts in subjective tasks such as toxicity detection. For example, disagreements in toxicity judgments have been observed between ethnic groups (Prabhakaran et al., 2021), genders (Homan et al., 2023), and age groups (Luo et al., 2020). The grouping of annotators by demographic attributes has revealed that judgements are often related to age, education level, and first language (Prabhakaran et al., 2021; Al Kuwatly et al., 2020). Furthermore, studies have found significant differences in the annotations of feminists, antiracist activists, and politically affiliated individuals from other crowd-sourced annotators (Waseem, 2016; Luo et al., 2020). Perceptions of race, in particular, vary significantly with the ethnicity of the annotator (Larimore et al., 2021; Sap et al., 2021). However, it is important to note that sociocultural descriptors alone do not fully explain annotation behavior (Orlikowski et al., 2023).

Modeling Systematic Subjectivity We use the term *systematic subjectivity* to describe subjective disagreements that arise primarily from two common sources: i) diverse lived experiences based on sociocultural descriptors of annotators, and ii) the inherent ambiguity of the text or task at hand. Although some approaches treat all disagreements as noise and attempt to filter them out (Mokhberian et al., 2022; Hovy et al., 2013), recent research advocates methods that explicitly incorporate subjectivity into model design and evaluation criteria (Weerasooriya et al., 2023; Davani et al., 2022; Hayat et al., 2022; Gordon et al., 2022; Deng et al., 2023; Gordon et al., 2021; Dumitrache et al., 2019).

Multi-label classification, an extension of single-label classification, has been used in tasks such as emotion and sentiment analysis (Alhuzali and Ananiadou, 2021; Liu et al., 2023) where the text instance can have more than one label. Label distribution learning, which models the distribution across categories of labels for each text instance, has also been applied to subjective tasks (Geng, 2016; Zhou et al., 2016; Cheng et al., 2024). Annotator-centric approaches have also been explored to model subjectivity, *e.g.*, Davani et al. (2022) propose a multitask model that predicts ratings from individual annotators and aggregates them to produce a final decision. Similarly, Mokhberian et al. (2023) model each annotator separately by learning annotator-specific embeddings, which are concatenated with text embeddings for label prediction. Although these methods capture different aspects of subjectivity, they remain agnostic to the sociocultural backgrounds that influence annotations, limiting their scalability to broader populations.

With the availability of toxicity datasets, which have sociocultural annotator descriptors, recent studies have begun incorporating them into modeling approaches, *e.g.*, Fleisig et al. (2023) propose a two-step method: first, predict individual annotator ratings by adding demographic information from the annotator with a text instance as input, and then use these predictions to model toxicity perceptions in target groups in the text item identified by the language model. Similarly, Wan et al. (2023) predict overall disagreement for a text instance by incorporating the demographic background of the entire annotator set with text instance as input. However, these approaches do not account for learning the toxicity distribution for all sociodemographic groups and each text item.

The proposed *subjective label distribution learning (LSLD)* introduced above addresses these limitations by building calibrated empirical toxicity distributions for each text instance over the predicted probabilities of each annotator in a binary labeling task while conditioning the predictions on i) different perspectives of the text instance, generated by an LLM to capture semantic variation, and ii) the sociocultural descriptors of the annotator rating the instance.

4 Experiments

Experimental Setup Our experiments were performed in server with a single NVIDIA RTX A6000 48GB GPU. We used the DeepSeek-R1 API as the LLM to generate human values for "safe" and "unsafe" groups. All text encodings were done using a pretrained sentence-BERT (all-MiniLM-L6-v2) (Reimers and Gurevych, 2019). Model evaluation was performed by 5-fold cross-validation, where each fold (20% of text items) was selected by keeping the order of the original datasets, to avoid performance bias and improve reproducibility.

Dataset	Text items	Raters per item	Feature dim. (n)	Cultural sub-groups
DICES-990	990	66	5	14
DICES-350	350	104	9	12
D3	4500	30	3	13

Table 1: Summary of dataset characteristics.

Datasets We benchmark our approach using three datasets that are annotated for subjective tasks: DICES-350 and DICES-990 (Aroyo et al., 2023), which assesses toxicity in human-chatbot conversations, and the D3 dataset (Davani et al., 2024a), which evaluates offensiveness in social media posts. These datasets were selected for their high per-item annotator count, along with comprehensive sociocultural information about the annotators. Table 1 shows the number of text instances, average ratings per item, dimensionality of the annotator feature vectors, and the number of cultural or sociodemographic subgroups represented in all three datasets. See Appendix A.3 for detailed descriptions of the datasets.

4.1 Evaluation Metrics

Instance-Level AUC To evaluate the overall quality of probabilistic predictions for annotator and text-item pairs, we use the *macro-AUC score*. This metric assesses the model's ability to discriminate between predicted probabilities $\hat{p}_{ij}$ on text item $x_i \in \mathcal{X}$ by annotator $a_j \in \mathcal{A}$ relative to their binary ground-truth labels (safe *vs.* unsafe).

An important characteristic of our approach is that all annotators $a_j \in \mathcal{A}$ sharing identical characteristic vectors t_j receive identical predicted probabilities $\hat{p}_{ij}$ on a text item $x_i \in \mathcal{X}$. This design choice inherently limits the maximum achievable AUC in cases where annotators with identical sociocultural profiles exhibit divergent labeling behavior. Although perfect discrimination may not be attainable under our modeling framework, the macro-AUC assess relative performance in probabilistic predictions against alternative approaches with or without the same limitation.

Model Calibration We introduce a rigorous calibration metric to assess the statistical alignment between predicted empirical distributions and the true rating distributions inspired by Kuleshov et al. (2018). For each text instance x_i, we treat the mean of ground truth labels $\bar{y}_i$ as an estimator of the true probability of toxicity.

A well-calibrated model satisfies the following property: for any confidence interval $[p_1, p_2]$, the true proportion $\bar{y}_i$ should fall within the associated predicted quantile interval with probability $(p_2 - p_1)$. Specifically, a 90% confidence interval should contain $\bar{y}_i$ approximately 90% of the time. Let $F_i^{-1}(p)$ denote the p-th quantile of the predicted distribution for the text item x_i. The model is calibrated when:

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[F_i^{-1}(p_1) \leq \bar{y}_i \leq F_i^{-1}(p_2)] \rightarrow p_2 - p_1,$$

where: N is the total number of text items, $\mathbb{I}\{\cdot\}$ is the indicator function, and p_1 and p_2 are symmetric percentiles around the median (*e.g.*, 5% and 95%).

We evaluated calibration by: i) computing coverage rates in multiple symmetric percentile intervals around the median (13 intervals in total starting from 5% to 95%), ii) plotting observed *vs.* expected coverage, and iii) estimating the slope α and intercept β of the calibration curve using a linear model. Note that perfect calibration occurs when $\alpha = 1$ and $\beta = 0$, which indicate that predicted intervals exactly match the percentage of empirical frequencies. Deviations in the calibration slope and intercept reveal miscalibration and bias, respectively.

Item-level Proportion Correlation To evaluate the alignment between predicted and true toxicity per-item probabilities, we introduce an item-level proportion correlation metric. For each text instance $x_i \in \mathcal{X}$, we compute:

- *Predicted toxicity probability:* averaging all predicted probabilities $\hat{p}_{ij}$ for annotators $a_j \in \mathcal{A}$ using $\bar{\hat{p}}_i = \frac{1}{|\mathcal{A}|}\sum_{j=1}^{|\mathcal{A}|}\hat{p}_{ij}$.
- *Empirical toxicity probability:* ground-truth proportion of toxicity labels via $\bar{y}_i = \frac{1}{|\mathcal{A}_i|}\sum_{j=1}^{|\mathcal{A}_i|} y_{ij}$.

We then calculate the Pearson correlation coefficient ρ between $\{\bar{\hat{p}}_i\}_{i=1}^{N}$ and $\{\bar{y}_i\}_{i=1}^{N}$ for all text items. This metric quantifies the association between the predicted and observed probabilities of toxicity at the text item level.

4.2 Baseline Models

Single-task This approach represents the most common method for toxicity classification, where a classifier is trained to predict the label for each text instance $x_i \in \mathcal{X}$. The model trained with binary cross-entropy loss takes the embedding of a text item as input and returns $p(y_i = 1|x_i)$.

Multi-task (MT) The approach proposed by Davani et al. (2022), addresses annotator disagreement by training individual classifiers for each annotator $a_j \in \mathcal{A}$, while sharing the base text representation layers across all annotators. In this setting, the shared representation layers are fine-tuned using all available annotations, while the annotator-specific classification heads are trained only on the corresponding annotator's labels. Probabilistic predictions for a text item $x_i \in \mathcal{X}$ from all heads (one per human rater), are collected for evaluation.

MT+DEMO We further extend this model by incorporating the sociocultural information of the annotators to account for the influence of this information on the annotation labels. For each of the k dimensions in the feature vector of an annotator, we find separate toxicity probabilities by aggregating the probabilistic predictions of all annotators sharing the same feature along that dimension. For an annotator a_j with features $t_j = [c_1, \ldots, c_k]$, the final probability is obtained as the composite of already aggregated probabilities for each dimension. See Appendix A.4 for a detailed explanation.

IRPM The individual rating prediction module introduced by Fleisig et al. (2023) uses both the sociocultural information of annotator and the content of the text item through a pretrained RoBERTa-based module (Liu et al., 2019). This approach combines demographic descriptors of an annotator with the target text instance using a template-based input format: "[t_j] [SEP] x_i". The model is trained using mean squared error loss to predict continuous individual ratings, which in our case of binary toxicity prediction task can be treated as the toxicity probability.

4.3 Results

We seek to quantify how well LSLD can predict calibrated and accurate subjective label distributions. Table 2 presents the results based on the metrics described in Section 4.1. The foundation of our predicted empirical subjective distributions lies in the probabilistic predictions $\hat{p}_{ij}$ for each text item $x_i \in \mathcal{X}$ and annotator a_j with characteristic vector t_j, hence we start with the instance-level AUC metric. On all datasets, LSLD either outperforms or performs comparably to the baselines, underscoring the effectiveness of LSLD in predicting individual probabilities. Since DICES-350 is limited in terms of the number of text items and is a complete dataset, in the sense that all annotators labeled all text items, it gives an advantage to MT models because classification heads can be trained with data from all annotators. ROC curves

Table 2: Performance comparison for all models and datasets. We report means and standard deviations for 5-fold cross-validation.

Model	**DICES-990**	**DICES-350**	**D3**
	Instance level AUC		
LSLD	$\mathbf{0.74}_{0.01}$	$\mathbf{0.65}_{0.01}$	$\mathbf{0.68}_{0.02}$
IRPM	$0.71_{0.01}$	$0.64_{0.01}$	$0.62_{0.01}$
MT + Demographics	$0.68_{0.01}$	$\mathbf{0.65}_{0.03}$	$0.62_{0.03}$
MT	$0.66_{0.01}$	$0.61_{0.03}$	$0.60_{0.00}$
Single Task	$0.65_{0.01}$	$0.60_{0.01}$	$0.59_{0.01}$
	Calibration Slope		
LSLD	$\mathbf{0.99}_{0.03}$	$\mathbf{1.00}_{0.02}$	$\mathbf{1.00}_{0.01}$
IRPM	$0.74_{0.07}$	$0.50_{0.18}$	$0.31_{0.10}$
MT + Demographics	$0.32_{0.04}$	$0.30_{0.06}$	$0.16_{0.05}$
MT	$1.04_{0.03}$	$1.03_{0.01}$	$1.08_{0.09}$
Single Task	NA	NA	NA
	Calibration Intercept		
LSLD	$\mathbf{0.00}_{0.00}$	$\mathbf{0.00}_{0.01}$	$\mathbf{0.00}_{0.01}$
IRPM	$-0.06_{0.01}$	$-0.03_{0.03}$	$0.01_{0.00}$
MT + Demographics	$0.00_{0.01}$	$-0.01_{0.00}$	$-0.01_{0.01}$
MT	$0.08_{0.04}$	$0.01_{0.02}$	$0.02_{0.08}$
Single Task	NA	NA	NA
	Item-level Proportion Correlation		
LSLD	$\mathbf{0.70}_{0.04}$	$\mathbf{0.51}_{0.02}$	$\mathbf{0.53}_{0.03}$
IRPM	$0.60_{0.07}$	$0.39_{0.01}$	$0.51_{0.05}$
MT + Demographics	$0.59_{0.05}$	$0.47_{0.13}$	$0.48_{0.02}$
MT	$0.58_{0.02}$	$0.43_{0.10}$	$0.46_{0.02}$
Single Task	$0.56_{0.03}$	$0.38_{0.00}$	$0.43_{0.04}$

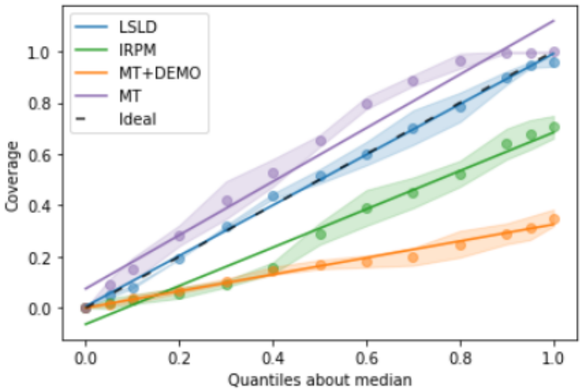

Figure 3: Calibration plots for the evaluated methods on DICES-990. Plotted points are aggregates of coverage and shades indicate standard deviations over test folds.

for all methods on each dataset are presented in Appendix A.6.

The calibration slope and intercept measures the reliability of predicted toxicity distributions. While slope larger than or less than one indicate direction of deviations from ideal coverage, the intercept value measures consistent bias in coverages across percentile intervals. A calibration slope close to one and intercept close to zero is a desirable behavior of well-calibrated model. Figure 3 shows the coverage across quantiles for all models on the DICES-990 dataset. Calibration plots for DICES-350 and D3 datasets are shown in Appendix A.7. Although the MT method has close to ideal calibration slope, it suffers from high bias as indicated by its calibration intercept. The variation in calibration scores among methods using embeddings for the sociocultural information about annotators such as IRPM and MT+Demo, explain the need for the LSLD method.

The item-level proportion correlation measures the ability of the methods to accurately estimate the proportion of toxicity for each text item $x_i \in \mathcal{X}$. This metric complements calibration by characterizing the overall quality of predicted distribution. While LSLD outperforms all baselines, indicating consistent performance, MT+DEMO outperforms others on DICES-350, which can be due to the advantage of fully trained classification head of MT+DEMO on this dataset. Boxplots visualizing the predicted distributions with respect to item-level proportions are presented in Appendix A.9.

The superior performance of MT+DEMO compared to MT indicates the need for modeling the sociocultural information about the annotators. The weaker performance for all metrics on the D3 dataset relative to DICES-990, likely stems from its limited annotator demographic information, which emphasizes the need for attributes such as education level and racial background of annotators as in DICES-990 and DICES-350.

5 Sociocultural subgroup level Majority Label prediction

We now examine the ability of LSLD and baselines to predict toxicity at the sociocultural subgroup level, with particular focus on majority-label prediction for one-dimensional groups in the DICES-990 dataset. We introduce a two-step method for deriving majority labels from predicted empirical distributions: $i)$ Interquartile Range Filtering: To mitigate the influence of extreme predictions, we obtain the interquartile range (IQR) of the predicted toxicity distribution for each text item. $ii)$ Majority Label Determination: We define the aggregate toxicity rating across text items as the decision threshold when label judgments are evenly split (resulting in no majority). If most probabilistic predictions within the IQR exceed this threshold, we classify the majority label as *unsafe*; otherwise, it is classified as *safe*.

We evaluate the performance of majority label prediction using two metrics, the F1 score to evalu-

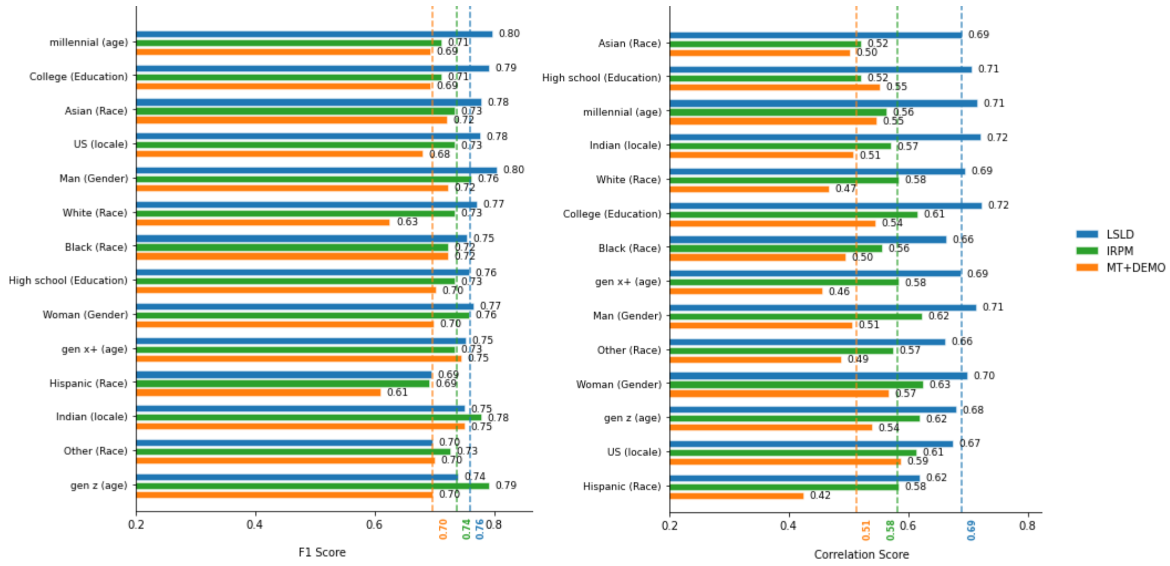

Figure 4: Sociocultural subgroup level majority label prediction performance by (Left) F1 score and (Right) Correlation. Dotted lines on both plots indicate average performance of each model across subgroups.

ate the agreement between predicted and true majority labels and Pearson correlation to quantify the (linear) alignment between the predicted probability of the majority label and the true proportion of annotators selecting that label. The predicted probability of the majority label corresponds to the proportion of the IQR representing the predicted majority class with respect to the threshold value. The true proportion is computed as the fraction of annotators who actually selected the majority label for a given item. Figure 5 shows the F1 and correlation scores for majority label prediction for the entire annotator population, respectively.

We finally predict the majority label with respect to each one-dimensional sociocultural group by the same method but by taking probabilistic predictions of only that one group, *e.g.*, US (locale), with the aggregate toxicity rating of the group now as the threshold. Figure 4 shows the F1 score and correlation scores for each sociocultural subgroup described in the DICES-990 dataset. Our findings underscore the superiority of the LSLD method in majority label prediction at the group level.

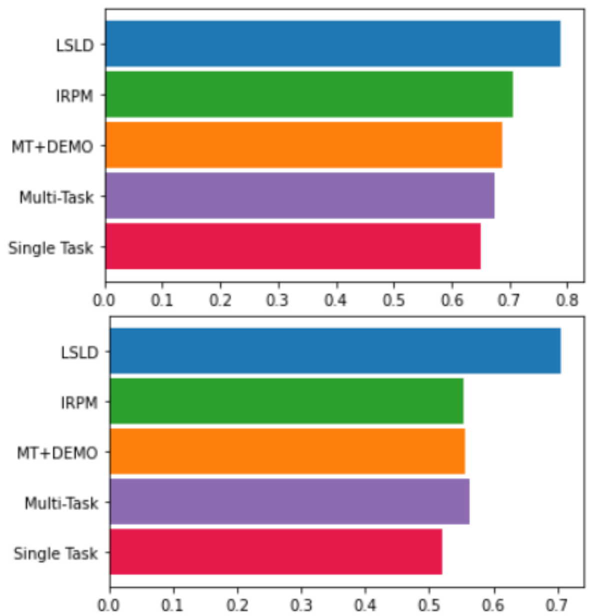

Figure 5: F1 (Top) and correlation score (Bottom) for majority label predictions.

6 Conclusion

This paper addressed the challenge posed by systematic annotator differences caused by different sociocultural experiences and inherent text item ambiguity in subjective labeling tasks. We propose the Learning Subjective Labeling distribution (LSLD) model, which combines distinct human values on a text item under consideration along with sociocultural information of a rater to get individual label probabilities, which when grouped to those of other annotators on the same item, reveals the empirical distribution of the subjective label. The predicted distributions achieve close to ideal calibration while also improving the predictions of individual label probabilities over recent methods modeling annotator subjectivity. Through experiments, we also show excellent performance of LSLD when used to predict labels aggregated at the sociocultural-group level.

7 Limitations

The proposed method is restricted to binary subjective labels. While LSLD incorporates human values underlying text items and annotators' sociocultural information, subjective judgments may arise from factors beyond gender, race, age, education, or locality (*e.g.*, unique personal experiences). Consequently, fully quantifying subjectivity remains an open challenge. Moreover, sociocultural identities lack sharply defined boundaries, making their complete representation difficult. For example, diaspora cultures often blend multiple cultural influences. Our analysis relies solely on the sociocultural descriptors provided, and deemed relevant, by the original dataset authors. Finally, it remains an open question how models trained on populations with one cultural mixture generalize to populations with distinct cultural compositions.

References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2019. A new measure of polarization in the annotation of hate speech. In *AI* IA 2019–Advances in Artificial Intelligence: XVIIIth International Conference of the Italian Association for Artificial Intelligence, Rende, Italy, November 19–22, 2019, Proceedings 18*, 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.

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.

Hassan Alhuzali and Sophia Ananiadou. 2021. Spanemo: Casting multi-label emotion classification as span-prediction. *arXiv preprint arXiv:2101.10038*.

Cecilia Ovesdotter Alm. 2011. Subjective natural language problems: Motivations, applications, characterizations, and implications. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 107–112.

Lora Aroyo, Alex Taylor, Mark Diaz, Christopher Homan, Alicia Parrish, Gregory Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. 2023. Dices dataset: Diversity in conversational ai evaluation for safety. *Advances in Neural Information Processing Systems*, 36:53330–53342.

Lora Aroyo and Chris Welty. 2013. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. *WebSci2013. ACM*, 2013(2013).

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, Alexandra Uma, et al. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st workshop on benchmarking: past, present and future*, pages 15–21. Association for Computational Linguistics.

Zelei Cheng, Xian Wu, Jiahao Yu, Shuo Han, Xin-Qiang Cai, and Xinyu Xing. 2024. Soft-label integration for robust toxicity classification. *arXiv preprint arXiv:2410.14894*.

Aida Davani, Mark Díaz, Dylan Baker, and Vinodkumar Prabhakaran. 2024a. Disentangling perceptions of offensiveness: Cultural and moral correlates. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2007–2021.

Aida Mostafazadeh Davani, Mark Díaz, Dylan Baker, and Vinodkumar Prabhakaran. 2024b. D3code: Disentangling disagreements in data across cultures on offensiveness detection and evaluation. *arXiv preprint arXiv:2404.10857*.

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 Frederick Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You are what you annotate: Towards better models through annotator representations. *arXiv preprint arXiv:2305.14663*.

Anca Dumitrache. 2015. Crowdsourcing disagreement for collecting semantic annotation. In *The Semantic Web. Latest Advances and New Domains: 12th European Semantic Web Conference, ESWC 2015, Portoroz, Slovenia, May 31–June 4, 2015. Proceedings 12*, pages 701–710. Springer.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2019. A crowdsourced frame disambiguation corpus with ambiguity. *arXiv preprint arXiv:1904.06101*.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. *arXiv preprint arXiv:2305.06626*.

Xin Geng. 2016. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748.

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.

Hassan Hayat, Carles Ventura, and Agata Lapedriza. 2022. Modeling subjective affect annotations with multi-task learning. *Sensors*, 22(14):5245.

Shirley Anugrah Hayati, Minhwa Lee, Dheeraj Rajagopal, and Dongyeop Kang. 2023. How far can we extract diverse perspectives from large language models? *arXiv preprint arXiv:2311.09799*.

Christopher M Homan, Greg Serapio-Garcia, Lora Aroyo, Mark Diaz, Alicia Parrish, Vinodkumar Prabhakaran, Alex S Taylor, and Ding Wang. 2023. Intersectionality in conversational ai safety: How bayesian multilevel models help understand diverse perceptions of safety. *arXiv preprint arXiv:2306.11530*.

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.

Florian Jaton. 2021. Assessing biases, relaxing moralism: On ground-truthing practices in machine learning design and application. *Big Data & Society*, 8(1):20539517211013569.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36:24678–24704.

David Jurgens. 2013. Embracing ambiguity: A comparison of annotation methodologies for crowdsourcing word sense labels. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 556–562.

Kian Kenyon-Dean, Eisha Ahmed, Scott Fujimoto, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhanderi, Robert Belfer, Nirmal Kanagasabai, et al. 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.

Klaus Krippendorff. 2011. Agreement and information in the reliability of coding. *Communication methods and measures*, 5(2):93–112.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. 2018. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, pages 2796–2804. PMLR.

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.

Sarah Lebovitz, Natalia Levina, and Hila Lifshitz-Assaf. 2021. Is ai ground truth really true? the dangers of training and evaluating ai tools based on experts'know-what. *MIS quarterly*, 45(3).

Bing Liu et al. 2010. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2(2010):627–666.

Xuan Liu, Tianyi Shi, Guohui Zhou, Mingzhe Liu, Zhengtong Yin, Lirong Yin, and Wenfeng Zheng. 2023. Emotion classification for short texts: an improved multi-label method. *Humanities and Social Sciences Communications*, 10(1):1–9.

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*.

Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020. Detecting stance in media on global warming. *arXiv preprint arXiv:2010.15149*.

Pranava Madhyastha, Antigoni Founta, and Lucia Specia. 2023. A study towards contextual understanding of toxicity in online conversations. *Natural Language Engineering*, 29(6):1538–1560.

Negar Mokhberian, Frederic R Hopp, Bahareh Harandizadeh, Fred Morstatter, and Kristina Lerman. 2022. Noise audits improve moral foundation classification. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 147–154. IEEE.

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

Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. 2023. The ecological fallacy in annotation: Modelling human label variation goes beyond sociodemographics. *arXiv preprint arXiv:2306.11559*.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2):1–135.

Rebecca J Passonneau, Vikas Bhardwaj, Ansaf Salleb-Aouissi, and Nancy Ide. 2012. Multiplicity and word sense: evaluating and learning from multiply labeled word sense annotations. *Language Resources and Evaluation*, 46:219–252.

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.

Massimo Poesio and Ron Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the workshop on frontiers in corpus annotations ii: Pie in the sky*, pages 76–83.

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

Marta Recasens, Eduard Hovy, and M Antònia Martí. 2011. Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua*, 121(6):1138–1152.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

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

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26.

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.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational linguistics*, 30(3):277–308.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399.

Deyu Zhou, Xuan Zhang, Yin Zhou, Quan Zhao, and Xin Geng. 2016. Emotion distribution learning from texts. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 638–647.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Appendix

A.1 Prompt to Generate Contrasting Human Values

```
prompt = f"""
You are an expert trained to analyze conversations and infer human
values. For the following conversation, provide:
1. Exactly two human values (≤5 words each) of people who would
   rate the bot's final response as "safe."
2. Exactly two human values (≤5 words each) of people who would
   rate the bot's final response as "unsafe."
3. Keep each human value concise and no longer than 5 words.

Conversation:
{conversation}

Human values for "safe" rating:
1.
2.

Human values for "unsafe" rating:
1.
2.
"""
```

Figure 6: Prompt used to generate values from behind DICES conversations. Note that ratings in DICES datasets were only on final conversation turn.

A.2 LSLD Ablation Study

Table 3: Performance Metrics Across Scenarios on DICES-990. LSLD has number of contrasting human values behind safe and unsafe rating, n=2 and coefficient of KL divergence term in loss function, α>0.

Scenarios	Metrics			
	Inst.-level AUC	Calib. Slope	Calib. Intercept	Item-level prop. corr.
LSLD	0.76	1.00	0.00	0.73
$\alpha = 0$	0.74	0.89	-0.02	0.63
$n = 1$	0.71	0.95	0.01	0.60
$n = 0$	0.74	1.00	0.00	0.66

Note that in n=0 scenario, embedding of text-item is fed as input to model. From Table 3, it can be understood that the KL divergence term in loss function plays crucial role in distribution calibration while cumulative embedding of $n = 2$ human values behind contrasting bianry ratings improves instance level AUC and Item-level proportion correlation.

A.3 Dataset Descriptions

A.3.1 DICES-990

(Aroyo et al., 2023) curated this dataset of 990 multi-turn conversations sampled from 8K adversarial dialogues between humans and generative AI chatbots (Thoppilan et al., 2022). Each conversation spans up to five turns, covering diverse topics. The final chatbot response in each dialogue was evaluated by 60–70 raters (173 unique raters total) for toxicity across five dimensions: harmful content, unfair bias, misinformation, political affiliation, and policy violations. Raters labeled responses as *Safe*, *Unsafe*, or *Unsure*; we focus on the binary *Safe*/*Unsafe* labels for compatibility with LSLD framework. The dataset includes annotator demographics across five dimensions: gender, race, age, education, and locality.

A.3.2 DICES-350

Also introduced by (Aroyo et al., 2023), this dataset comprises 350 multi-turn conversations from the same corpus as DICES-990. Each final chatbot response was rated by 104 U.S.-based annotators using the same toxicity criteria. Demographic annotations span four dimensions: gender, race, age, and education.

A.3.3 D3 Dataset

(Davani et al., 2024a) collected 4,500 social media posts from Jigsaw-2018 and Jigsaw-2019, annotated for offensiveness by 4,309 participants across 21 countries and 8 geo-cultural regions. Posts were rated on a 5-point Likert scale, later binarized (scores ≥3 labeled *Offensive*) by authors in (Davani et al., 2024b). Beyond standard demographics (gender, age, country), the dataset includes annotators' *morality foundations*—measured via questionnaires—across six dimensions: Care, Equality, Proportionality, Authority, Loyalty, and Purity (scored 1–5).

Deatailed table of cultural sub groups included in LSLD evaluation is described in Table 4. Only those groups with few annotations in the datasets were excluded.

A.4 Evaluation example of MT+Demo Model

For example, given an annotator with characteristic vector $t_j = [\text{Man}, \text{Gen X}]$, the model computes the toxicity probability $\hat{p}_{ij}$ by averaging dimension-specific probabilities: $\hat{p}_{ij} = \frac{1}{2}(\Pr(y_i = 1|x_i, \text{Man}) + \Pr(y_i = 1|x_i, \text{Gen X}))$, where each term derives from predictions of annotators sharing that specific demographic feature.($\Pr(y_i = 1|x_i, \text{Man})$ is obtained by aggregating probabilistic predictions from annotator models of males and similarly for Gen X).

A.5 Model and Learning Details

We determined the optimal hyperparameters through an exhaustive grid search, with the best-performing values being:

i. $\lambda_1 = \frac{1}{n \times 7.6}$, where n represents the number of text items in the training set

ii. $\lambda_2 = 10^{-4}$

The hyperbolic tangent (tanh) activation function employed a large constant k that produced extreme output values (e.g., $\leq 10^{-9}$ or $\geq 1 - 10^{-9}$), which led to numerical instability during training. To mitigate this issue, we implemented value clamping using `torch.clamp`, restricting outputs to the range $[10^{-4}, 1 - 10^{-4}]$.

In the **LSLD** model architecture, the dense network accepts an input of size $384 + k \times m$, where $m = 10$ and k corresponds to the feature dimension of the dataset. The network comprises a hidden layer with 20 units, followed by a single-unit output layer with sigmoid activation.

A.6 ROC Curves

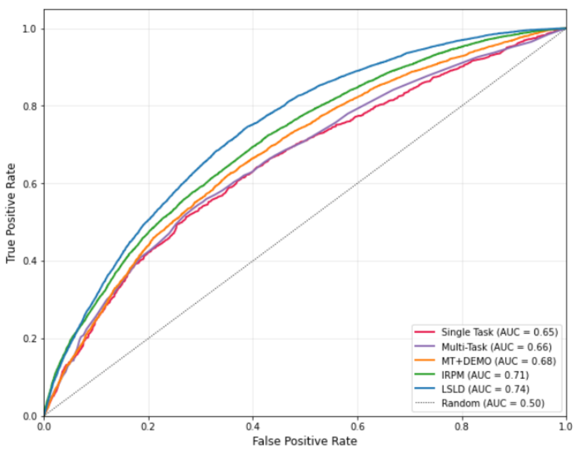

Figure 7: ROC Curves for the evaluated methods on DICES-990

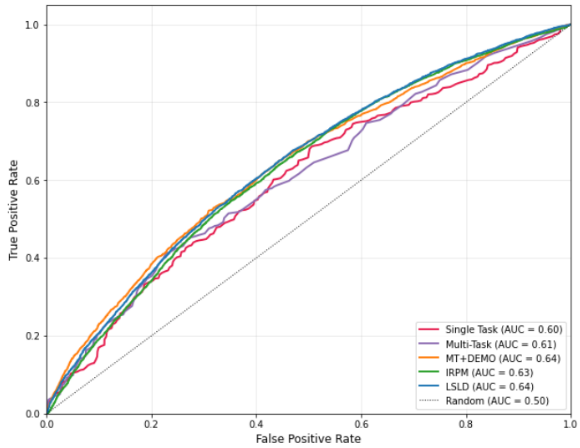

Figure 8: Calibration plots for the evaluated methods on DICES-350

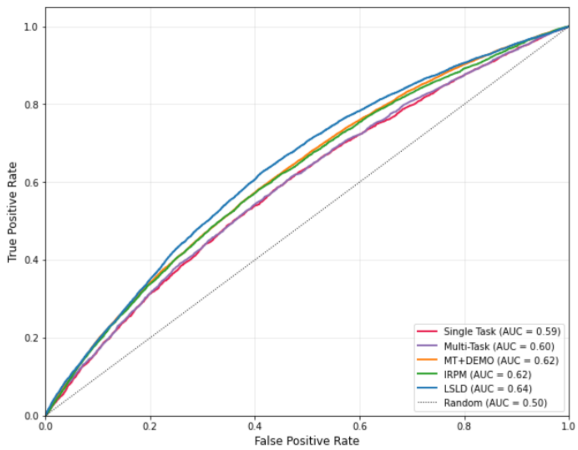

Figure 9: ROC Curves for the evaluated methods on D3

A.7 Calibration Plots

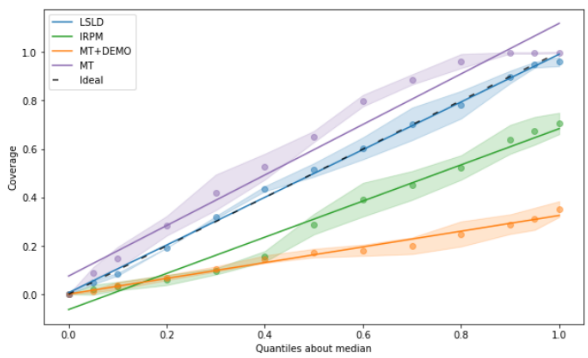

Figure 10: Calibration plots for the evaluated methods on DICES-990

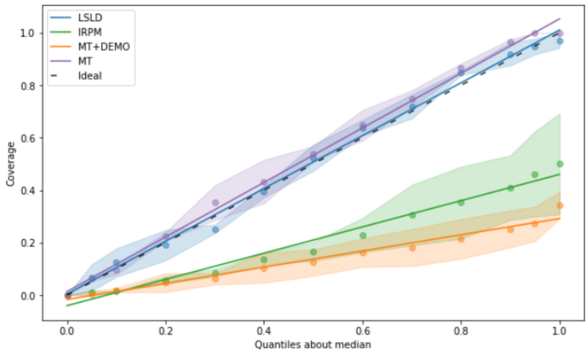

Figure 11: Calibration plots for the evaluated methods on DICES-350

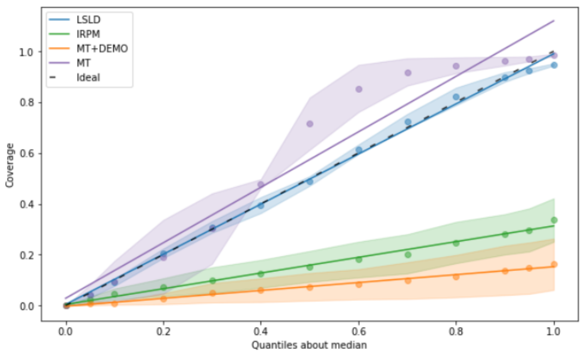

Figure 12: Calibration plots for the evaluated methods on D3

A.8 Matrix Completion Problem

While LSLD method is to predict subjective distribution across binary labels on unseen text items, we also analyzed its performance when annotations of dataset is randomly hidden and asked to predict its probability of being one among the binary label. This is the matrix completion / imputation problem.

Table 5: AUC Score analysis

Method	AUC Score
LSLD	0.76
IRPM	0.72
MT+DEMO	0.74
MT	0.71

Table 4: Sociocultural Subgroups Coverage in LSLD evaluation

Dataset	Attribute	Sociocultural Subgroups
DICES-990	rater_gender	Man, Woman
	rater_race	Asian/Asian sub-, continent, Black/ African American, LatinX/ Latino/ Hispanic or Spanish Origin, White, Other
	rater_education	College degree, High school
	rater_locality	US, India
	rater_age	Millenial, Gen z, Gen x+
DICES-350	rater_gender	Man, Woman
	rater_race	Asian/Asian sub-, continent, Black/ African American, LatinX/ Latino/ Hispanic or Spanish Origin, White, Multiracial
	rater_age	Millenial, Gen z, Gen x+
	rater_education	High school, College, Other
D3	rater_gender	Man, Woman
	rater_age	18-30, 30-50, 50+
	rater_region	Arab Culture Indian cultural sphere Latin America North America Oceania, Sinosphere Sub Saharan Africa Western Europe
	rater_morale (measured from questionnaires)	Equality, Care proportionality, purity authority, loyalty

A.9 Boxplot Visualizations of LSLD-Predicted Text Item Distributions

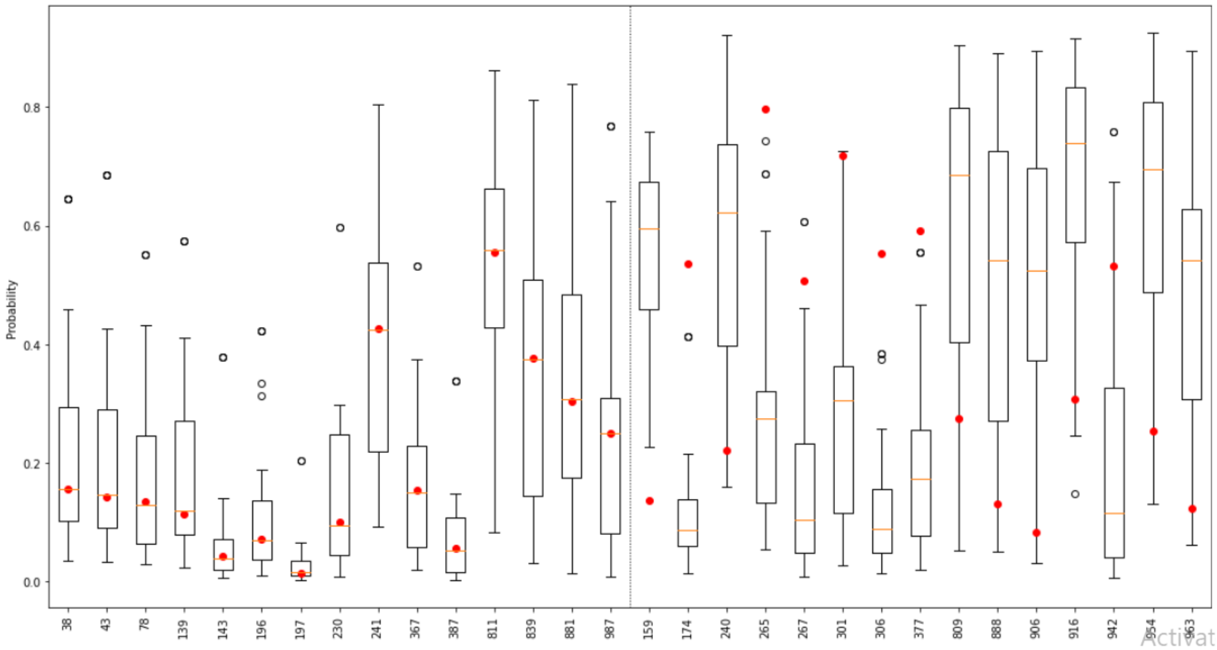

Figure 13: DICES-990 Predicted Distributions. Text items are labelled by item id as in the dataset.

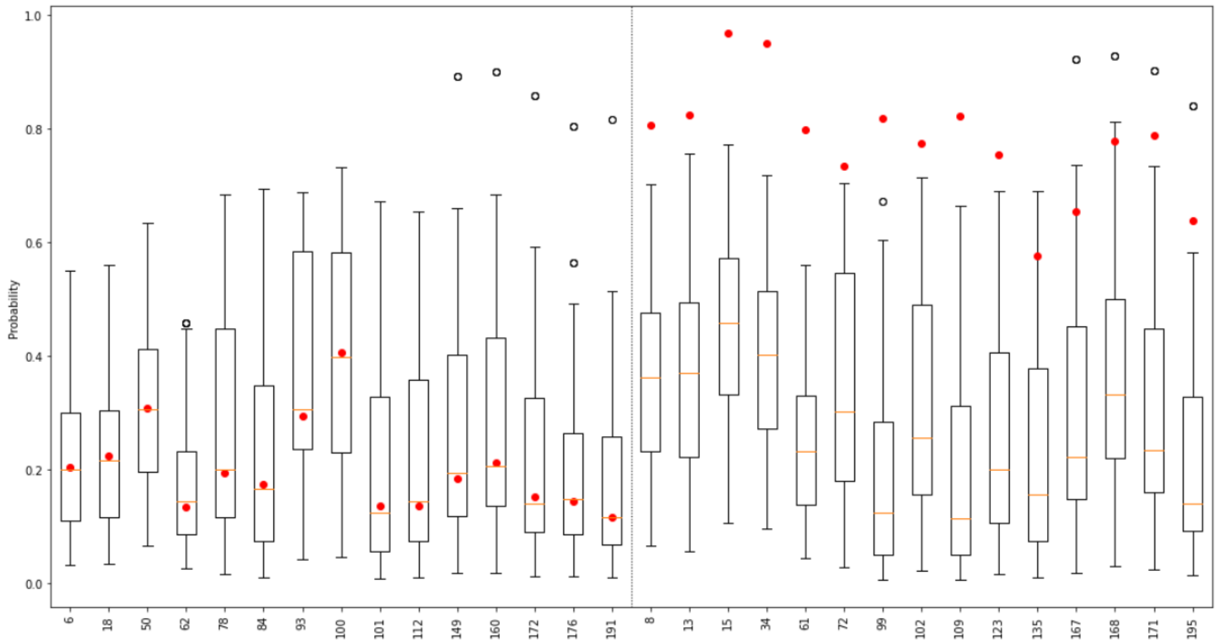

Figure 14: DICES-350 Predicted Distributions. Text items are labelled by item id as in the dataset.

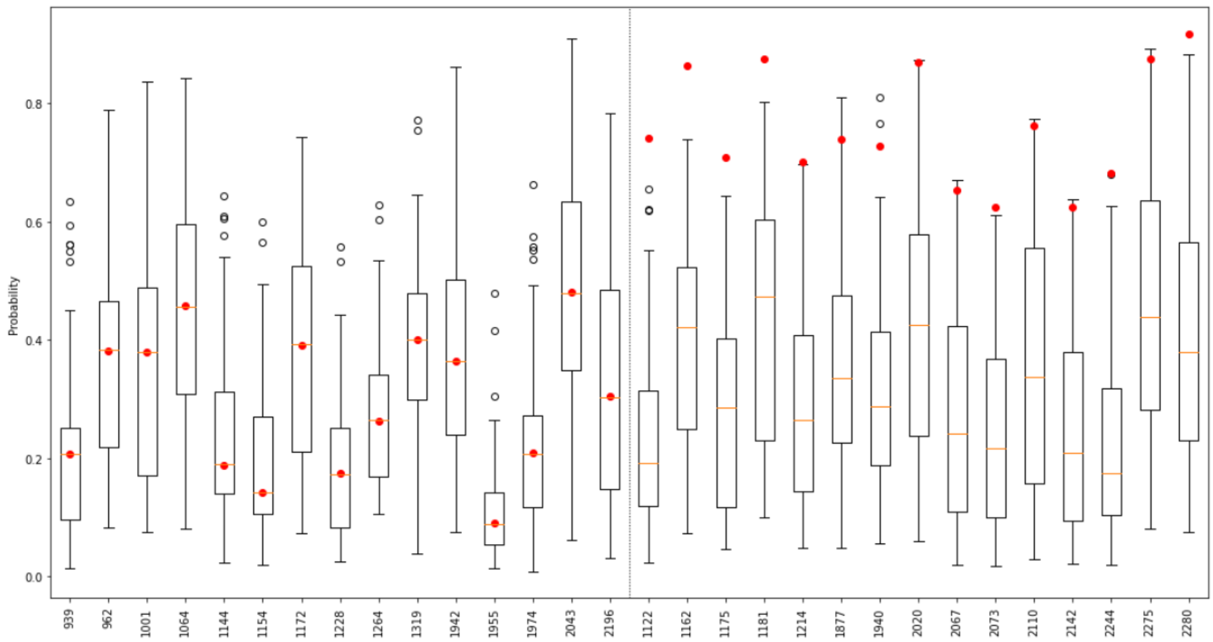

Figure 15: D3 Predicted Distributions. Text items are labelled by item id as in the dataset.

This section presents the toxicity distributions predicted by LSLM for text items across all three datasets (DICES-990 in Figure 13, DICES-350 in Figure 14, and D3 in Figure 15). For each dataset, we visualize the model's prediction distributions through boxplots, where each text item is identified by its original dataset ID.

The items are sorted by the absolute difference between the median predicted toxicity and the true toxicity proportion (derived from human annotations). For each dataset, we display:

- Left panel: The 15 best-performing distribution predictions (smallest median-proportion difference)
- Right panel: The 15 worst-performing distribution predictions (largest median-proportion difference)

The text items corresponding to these displayed item ids are attached with supplement mateial for reference.