# Modeling Human Perspectives with Socio-Demographic Representations

**Leixin Zhang**
University of Tübingen
leixin.zhang@uni-tuebingen.de

**Çağrı Çöltekin**
University of Tübingen
cagri.coeltekin@uni-tuebingen.de

## Abstract

Humans often hold different perspectives on the same issues. In many NLP tasks, annotation disagreement can reflect valid subjective perspectives. Modeling annotator perspectives and understanding their relationship with other human factors, such as socio-demographic attributes, have received increasing attention. Prior work typically focuses on single demographic factors or limited combinations. However, in real-world settings, annotator perspectives are shaped by complex social contexts, and finer-grained socio-demographic attributes can better explain human perspectives. In this work, we propose Socio-Contrastive Learning, a method that jointly models annotator perspectives while learning socio-demographic representations. Our method provides an effective approach for the fusion of socio-demographic features and textual representations to predict annotator perspectives, outperforming standard concatenation-based methods. The learned representations further enable analysis and visualization of how demographic factors relate to variation in annotator perspectives. Our code is available at GitHub.[1]

## 1 Introduction

Recent studies have shown that human disagreement is widespread across many annotation tasks. Some instances lack a single ground truth and annotation variation reflects diverse but valid perspectives (Plank, 2022; Cabitza et al., 2023; Leonardelli et al., 2023; Wang et al., 2023; Pham et al., 2023).

This phenomenon is particularly evident in subjective tasks, such as hate speech and offensive content detection (Sang and Stanton, 2022; Waseem, 2016). Some studies also suggest that individuals' perspectives are associated with their lived experiences and cultural contexts. For example, Sap et al. (2022) find that conservative annotators are more likely to rate African American English as toxic. Larimore et al. (2021) show that white and non-white annotators differ significantly in their ratings of racist language.

Despite these advances, pitfalls remain in analyzing the role of socio-demographic features in perspective modeling. Prior research often examines annotator subgroups using a single demographic attribute at a time, such as gender, and analyzes the statistical difference between its subcategories, such as female and male. However, this approach is limited, as a single attribute cannot fully capture the complex interplay between social contexts and the formation of human perspectives. Instead, richer combinations of socio-demographic attributes are more likely to reflect the nuanced experiences and social context, and to better capture the diversity of annotators' perspectives. In this work, we adopt a machine learning approach to analyze annotators' socio-demographic attributes and address the following research questions:

- RESEARCH QUESTION 1: To what extent do socio-demographic features contribute to perspective prediction in hate speech and toxic content tasks, and which features are most predictive of annotators' perspectives?
- RESEARCH QUESTION 2: How should socio-demographic features be encoded and fused with textual representations to improve model performance?

As the key contribution of this work, we propose **Socio-Contrastive Learning**, a novel architecture that jointly learns (i) annotators' socio-demographic representations inferred from their labeling behavior and (ii) socio-demographic-specific label predictions. Our method outperforms existing models on label prediction while simultaneously producing annotators' socio-demographic representations that enable analysis of their association with perspective variation.

[1] https://github.com/Leixin-Zhang/Socio_Contrastive_Learning



*Findings of the Association for Computational Linguistics: ACL 2026*, pages 40781–40794

## 2 Related Studies

This section presents prior work on modeling individual perspectives. Section 2.1 focuses on methods for predicting individual annotators' labels without using socio-demographic features, while Section 2.2 focuses on approaches that incorporate socio-demographic information into the learning process.

### 2.1 Individual Perspective Modeling

Modeling each annotator's labels directly, rather than aggregating via majority vote, has been shown to improve performance (Mostafazadeh Davani et al., 2022; Uma et al., 2021; Mokhberian et al., 2024; Zhang and Çöltekin, 2026).

Kanclerz et al. (2021) propose personalized approaches that leverage annotator label embeddings derived from their partial annotations, showing that even a small number of annotations on highly controversial data can significantly outperform generalized models. Mostafazadeh Davani et al. (2022) introduce three methods for modeling annotation variation: an *ensemble* approach that aggregates predictions from annotator-specific models, a *multi-label* approach that represents annotators' labels as a target vector, and a *multi-task* approach that assigns a separate prediction head to each annotator. Mokhberian et al. (2024) use an embedding layer to represent individual annotators for annotator-specific prediction.

### 2.2 Socio-Demographics Enriched Modeling

Some prior studies show that certain socio-demographic features of annotators are associated with their annotated labels. Huang and Yang (2023) identify differences in culture-related natural language inference judgments between annotators from the United States and India. In the context of socio-demographic enriched learning, *Jury Learning* (Gordon et al., 2022) models annotator-specific labels by concatenating socio-demographic features and annotator IDs with text representations. Final predictions are aggregated via a sampling process with a predefined demographic composition, allowing practitioners to control which groups' perspectives are reflected in the final outputs and in what proportions.

Orlikowski et al. (2023) investigate the influence of socio-demographic features by grouping annotators according to a single attribute and employing group-specific layers for each subcategory (e.g., Age: 25–34, 35-44, etc.). The results do not show performance improvements over a baseline model without socio-demographic features and randomly shuffled attribute groups, and they conclude that annotation patterns cannot be explained by socio-demographic attributes. However, using the same dataset for toxic content tasks, our study arrives at a conclusion different from that in Orlikowski et al. (2023). We find that incorporating richer combinations (instead of an individual attribute in isolation) reveals a substantial contribution of socio-demographic information to modeling and explaining annotation variation.

## 3 Methodology

To address the first research question: the extent to which socio-demographic features contribute to perspective prediction, we compare model architectures in the following two categories:

**Category 1:** Models that do not use socio-demographic features:

- **Simple Model:** This model uses only text embeddings to predict aggregated labels. When a text item is annotated by multiple annotators, their labels are aggregated into a single ground-truth label via majority voting.
- **Multi-Task Model:** This model uses text embeddings as input but predicts each annotator's label in a multi-task setup as Mostafazadeh Davani et al. (2022). The architecture consists of shared layers for all annotators, followed by separate output heads for each annotator.

**Category 2:** Models that leverage socio-demographic features. We experiment with two commonly used methods and introduce a novel approach for encoding annotators' socio-demographic attributes.

- **Multi-Hot Encoding:** Annotators' socio-demographic features are encoded as multi-hot vectors and concatenated with text embeddings as model input.
- **Socio-Demographic Embedding:** Socio-demographic features are encoded using the same pretrained model as the text encoder, producing representations in a shared vector space for concatenation.

- **Socio-Contrastive Learning (Our Method):** As shown in Figure 1, multi-hot socio-demographic encodings are first projected into a learnable space (Projection Layer) and optimized with a contrastive loss to capture annotator-specific patterns. The resulting representations are then concatenated with text embeddings for label prediction.

## 4 Socio-Contrastive Learning

The primary objective of the model is to predict annotator labels. Meanwhile, the model refines socio-demographic representations using a contrastive loss based on each annotator's labels, enabling them to capture socio-demographic–relevant patterns.

### 4.1 Motivation

We hypothesize that, for certain subjective tasks, an individual's perspective—reflected in their annotations on text items—is associated with the socio-demographic groups to which they belong. In other words, prediction accuracy of annotator labels is expected to improve when fine-grained socio-demographic attributes are available. Formally, let $\hat{y} = \arg\max_y P(y \mid \cdot)$ denote the predicted label and $y$ the ground-truth label. We posit that:

$$\mathbb{P}(\hat{y} = y \mid \text{text}, \text{socio-demo}) > \mathbb{P}(\hat{y} = y \mid \text{text})$$

We expect that socio-demographic attributes are not equally predictive of annotators' perspectives. Certain attributes may be more closely associated with specific judgment patterns, while others may contribute weaker signals. To amplify the most informative attributes, we apply contrastive learning to refine socio-demographic representations based on annotation similarity, encouraging representations to be closer for annotators with similar annotation patterns and farther apart otherwise.

### 4.2 Model Design

The model input consists of two components: a multi-hot encoding of the annotator's socio-demographic attributes and a text representation encoded by a pretrained model. Before concatenation, the multi-hot socio-demographic vector is projected into a learnable space and optimized with a contrastive loss to capture annotator-specific patterns. The primary task of the model is to predict each annotator's label, using a cross-entropy loss.

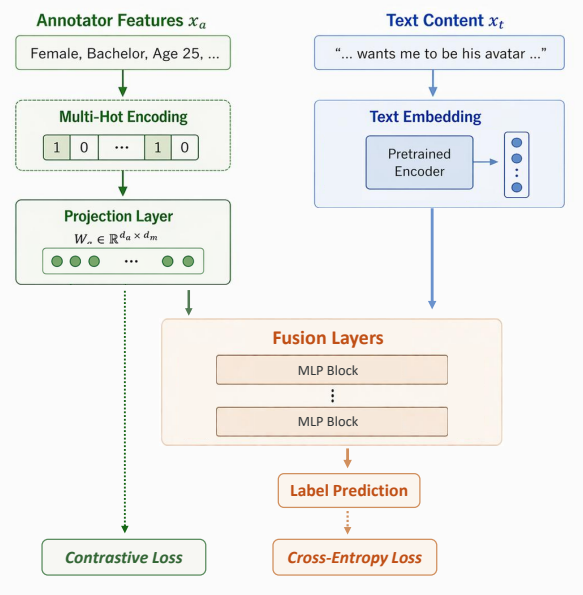


Figure 1: Socio-Contrastive Model Architecture

In parallel, the model also learns to optimize the socio-demographic representations through a contrastive loss, which guides the representations to capture annotation patterns.

### 4.3 Contrastive Representation Learning

We apply contrastive loss to learn annotators' socio-demographic representations. For a given text, annotators who provide the same label are treated as **positive cases**, and annotators who provide different labels serve as **negative cases**.

Standard contrastive learning methods, such as the InfoNCE loss, typically rely on predefined positive and negative cases and assume a fixed number of such examples (e.g., pairs or triplets). However, crowdsourced datasets pose several challenges. Texts are annotated by a variable number of annotators, ranging from a few to over 100, resulting in inconsistent numbers of positive and negative pairs across instances. In addition, items with perfect agreement contain no negative pairs. Excluding such instances would discard useful data and reduce the effectiveness of perspective modeling.

To address these issues, we design a flexible contrastive learning scheme tailored for socio-demographic representation learning. Our approach performs contrastive learning within each batch, prioritizing samples that share the same text ID (the same text annotated by multiple annotators) into the same batch. If the number of annotations for a given text is smaller than the batch size, other texts are included to fill the batch; if it exceeds the

| | Data Split | Text | Uniq. Text | Ann. | Labels |
|---|---|---|---|---|---|
| HATESPEECH | Train (70%) | 22,942 | 6,227 | 2,315 | 10,179 H<br>12,763 N |
| | Test (30%) | 10,359 | 2,670 | 2,263 | 4,415 H<br>5,944 N |
| TOXIC | Train (60%) | 25,556 | 4,638 | 3,517 | 10,895 T<br>14,661 N |
| | Test (40%) | 16,151 | 3,092 | 2,656 | 7,210 T<br>8,914 N |

**H**: Hate Speech; **T**: Toxic; **N**: Not Hate Speech or Toxic.

**Ann.**: Annotator Size.

Train set and test set splits are performed using unique text IDs to prevent leakage between training and evaluation.

Table 1: Dataset statistics for the hate speech and toxic content classification tasks.

batch size, the remaining annotations are processed in subsequent batches. During training, only annotations of the same text are considered in the contrastive loss, while those from different texts are masked out.

Within a batch, for negative pairs (annotators who disagree on a label), the loss is weighted by the similarity of their socio-demographic representations, pushing them apart. For positive pairs (annotators who agree on a label), the loss is weighted by the dissimilarity of their representations, pulling them closer. This scheme allows the model to learn socio-demographic representations that reflect systematic differences and agreements in annotation patterns. A detailed description of the algorithm is provided in Appendix A.

## 5 Experiments

This section presents the datasets and implementation details used in our experiments.

### 5.1 Datasets

We conduct our study using two crowd-annotated datasets that include rich socio-demographic metadata about the annotators: a hate speech dataset (Kennedy et al., 2020) and a toxicity dataset (Kumar et al., 2021). Prior work (Gordon et al., 2022; Orlikowski et al., 2023) has documented substantial annotation disagreement in both tasks.

Both datasets provide the following socio-demographic attributes of annotators: education, political ideology, age, gender, race, and sexuality. Additionally, the toxicity dataset includes the annotator's income range and self-reported importance of religions. The hate speech dataset includes whether the annotator is a parent. These variables serve as the socio-demographic features used in our modeling.

For reliability, we remove items with too few annotators (fewer than 2 for hate speech, and fewer than 4 for the toxic dataset) as well as annotations from annotators who contributed an insufficient number of labels (fewer than 20 for both datasets). After preprocessing, the hate speech dataset contains 6,227 unique texts and the toxicity dataset contains 4,638 unique texts. On average, each text is annotated by ~4 annotators for hate speech and ~6 annotators for toxic detection. Both datasets were originally annotated using Likert scales ratings. For direct comparison with other baseline models, we convert them to binary labels by mapping a score of 0 to the non-hate or non-toxic class and any score above 0 to the hate or toxic class. Table 1 presents detailed statistics for both datasets.

### 5.2 Implement Details

All models are implemented with PyTorch (Paszke et al., 2019).

**Text Representations.** Text inputs are encoded under three settings with three pretrained encoders: (i) Sentence-BERT (Reimers and Gurevych, 2019); (ii) BERT (Devlin et al., 2019) and (iii) RoBERTa (Liu et al., 2019).

**Model Architecture and Training.** Two dense layers with dropout are applied across all models. Cross-entropy serves as the primary training objective. For the Socio-Contrastive model, a customized contrastive loss is combined with cross-entropy, with both objectives weighted equally. We further conduct an ablation study by setting the contrastive loss weight to 0 for comparison. Model parameters are optimized using the Adam optimizer (Kingma, 2014).

**Evaluation Protocol.** Each model is trained under multiple hyperparameter configurations to determine the optimal setup (details are provided in Appendix B). After selecting the optimal hyperparameters, we train each model for six independent runs and report the mean performance and standard deviation across these runs. Given the presence of annotation disagreement and our focus on modeling divergent annotator perspectives, we evaluate models based on their ability to predict individual

| Model | Hate Speech | | | Toxic | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F1 | Precision | Recall | F1 |
| Simple Model | 0.438 ± 0.046 | 0.395 ± 0.065 | 0.415 ± 0.052 | 0.453 ± 0.005 | 0.525 ± 0.041 | 0.486 ± 0.019 |
| Multi-Task | 0.670 ± 0.028 | 0.565 ± 0.108 | 0.608 ± 0.074 | 0.614 ± 0.005 | 0.487 ± 0.012 | 0.543 ± 0.006 |
| Socio Multi-Hot | 0.759 ± 0.009 | 0.629 ± 0.042 | 0.687 ± 0.026 | 0.626 ± 0.011 | 0.607 ± 0.023 | 0.616 ± 0.009 |
| Socio Embedding | 0.750 ± 0.013 | 0.655 ± 0.031 | 0.699 ± 0.015 | 0.666 ± 0.015 | 0.596 ± 0.036 | 0.628 ± 0.013 |
| Socio Contrastive (Ours) | 0.729 ± 0.037 | 0.727 ± 0.068 | **0.725** ± 0.018 | 0.625 ± 0.013 | 0.667 ± 0.027 | **0.645** ± 0.008 |
| Ablation (w/o $\mathcal{L}_{contrastive}$) | 0.743 ± 0.018 | 0.631 ± 0.051 | 0.681 ± 0.030 | 0.649 ± 0.009 | 0.621 ± 0.024 | 0.634 ± 0.009 |

Table 2: Performance comparison of five models on hate speech and toxicity classification (mean ± standard deviation). The best $F_1$ scores are highlighted in **bold**. Models above the dashed line do not use socio-demographic features, and models below the dashed line incorporate socio-demographic features. Ablation results (w/o $\mathcal{L}_{contrastive}$) indicate the removal of the contrastive loss from our method.

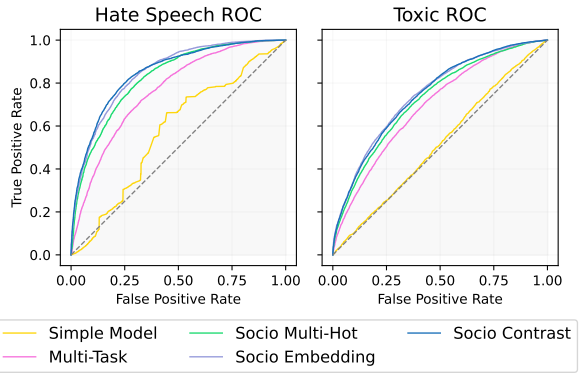


Figure 2: ROC curve: the prediction performance of five models

annotator labels rather than aggregated or majority-vote labels. A threshold of 0.5 on the sigmoid outputs is used to assign binary labels (hate vs. non-hate, toxic vs. non-toxic) and compute precision, recall, and $F_1$. We additionally present the AUC-ROC, which assesses the model's ranking ability by measuring how well it separates positive and negative instances across all possible decision thresholds.

## 6 Results and Discussion

Models that leverage socio-demographic information consistently outperform those that rely on textual input alone, as evidenced by the $F_1$ scores in Table 2 and ROC curves in Figure 2.[2]

[2]We report results using the best-performing pre-trained encoder in the main text. Sentence-BERT outperforms BERT and RoBERTa for text encoding, which aligns with the results from other studies (Reimers and Gurevych, 2019; Zhang and Çöltekin, 2024; Zhang et al., 2024). Results with BERT and RoBERTa encoders are provided in Appendix D.

### 6.1 Text-Only Models

Simple model uses aggregated labels as training targets, perform poorly in predicting individual annotator labels. This result highlights the importance of modeling unaggregated annotations to capture diverse perspectives. The multitask architecture proposed by Mostafazadeh Davani et al. (2022), which predicts each annotator's labels via a separate output head, underperforms socio-demographic–enriched models. This stems from the need for sufficient per-annotator data to effectively train annotator-specific heads. However, crowdsourced datasets typically involve a large number of annotators, each contributing only a small number of annotations (often around 20 or fewer after data splitting). Under such conditions, the multitask architecture is not well suited to settings with many annotators but limited per-annotator data. This limitation also explains the findings of Orlikowski et al. (2023) who report no performance gains from incorporating socio-demographic features. Their approach adopts a similar per-annotator head design, which, under limited per-annotator data conditions, constrains the model from learning meaningful annotation patterns. In addition, assigning an independent output head to each annotator introduces substantial computational overhead, particularly given that both datasets contain over 2,000 annotators.

### 6.2 Socio-Demographic Enriched Models

Our results suggest that richer combinations of socio-demographic attributes provide more informative signals for predicting hate speech and toxic content opinions, particularly under the Socio-Contrastive Learning setting.

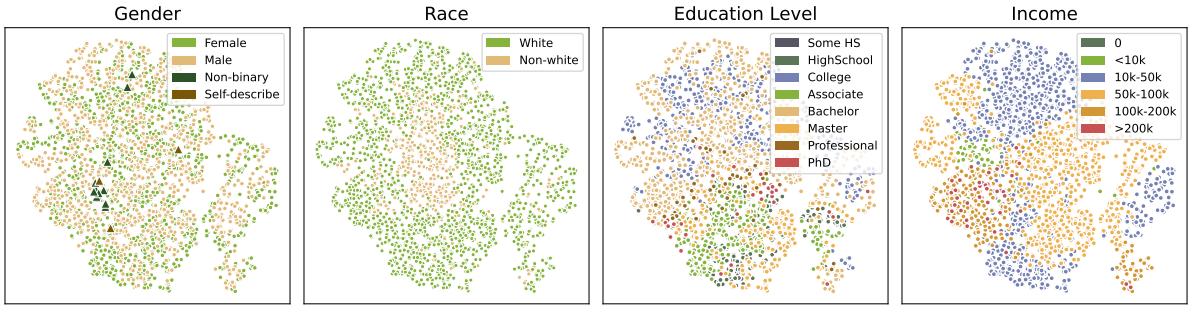


Figure 3: Visualization of Contrastively Learned Socio-Demographic Representations for **Hate Speech** Task

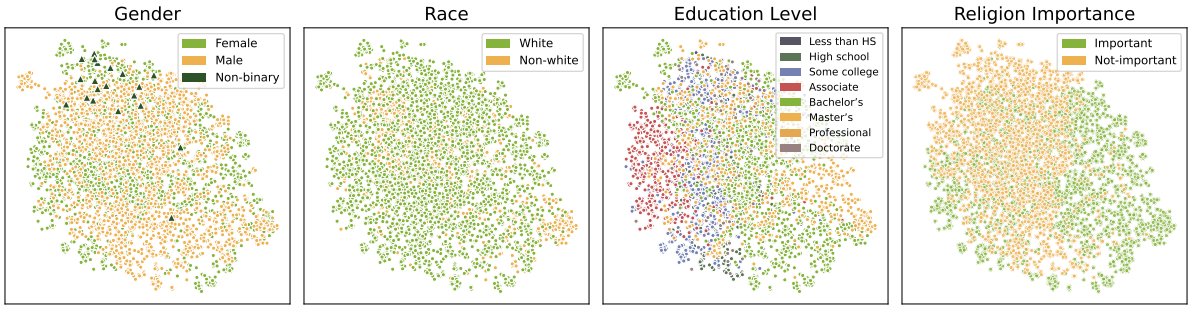


Figure 4: Visualization of Contrastively Learned Socio-Demographic Representations for the **Toxic** Task

The contrastively trained socio-demographic model achieves the best overall performance, reaching an $F_1$ score of 0.725 on the hate speech dataset and 0.645 on the toxicity dataset. Ablation results further demonstrate the effectiveness of contrastive learning, with performance drops by 4.4% and 1.1%, respectively, when the contrastive loss is removed. This can be explained by the fact that the contrastive strategy strengthens socio-demographic signals that are more closely associated with annotation patterns and improves the interaction between socio-demographic information and textual representations. It produces dense socio-demographic representations that better capture annotation patterns and are effectively adapted to the task through fusion with text representations.

The socio-embedding model, which encodes socio-demographic attributes using pretrained Sentence-BERT (Table 2) or RoBERTa (Table 6 in Appendix D), outperforms the corresponding multi-hot encoding model using the same encoder. This suggests that, for these two pre-trained models, mapping socio-demographic attributes into the aligned text embedding space better captures their interaction with textual content. However, for BERT-encoded texts, multi-hot encoding outperforms BERT-based socio-demographic representations (Table 5 in Appendix D).

## 7 Socio-Demographic Contributions

Our Socio-Contrastive model offers an additional advantage by learning the representations of the full combination of an annotator's socio-demographic features. As these vectors are optimized based on annotators' labeling behavior, annotators who consistently show disagreement are pushed farther apart in the representation space during training, while those with similar annotation patterns are brought closer together. The resulting distances between annotator vectors serve as an interpretable signal of *perspective divergence*, enabling analysis of how specific socio-demographic attributes contribute to differences in annotation behavior. To examine these contributions, we employ two approaches: (1) visualization of the learned representations in Section 7.1, and (2) statistical analysis of distances across socio-demographic groups in Section 7.2.

### 7.1 Representation Visualization

After the contrastive model is trained, we obtain learned vector representations for each unique annotator in the dataset (2,316 in the hate speech data and 4,408 in the toxic data).

| Attribute | Hate Speech | | | Toxic Content | | |
|---|---|---|---|---|---|---|
| | Observed | Random | Ratio | Observed | Random | Ratio |
| Education | 0.631 ± 0.007 | 0.244 ± 0.005 | 2.590 ± 0.045 | 0.702 ± 0.005 | 0.246 ± 0.004 | 2.857 ± 0.045 |
| Political Ideology | 0.415 ± 0.006 | 0.162 ± 0.002 | 2.566 ± 0.037 | 0.794 ± 0.004 | 0.347 ± 0.002 | 2.285 ± 0.019 |
| Income | 0.710 ± 0.006 | 0.343 ± 0.005 | 2.070 ± 0.024 | - | - | - |
| Age Group | 0.466 ± 0.005 | 0.232 ± 0.004 | 2.015 ± 0.030 | 0.678 ± 0.005 | 0.251 ± 0.004 | 2.701 ± 0.035 |
| Religion Importance | - | - | - | 0.882 ± 0.003 | 0.506 ± 0.002 | 1.743 ± 0.008 |
| Gender | 0.827 ± 0.004 | 0.503 ± 0.004 | 1.645 ± 0.013 | 0.823 ± 0.004 | 0.496 ± 0.001 | 1.659 ± 0.008 |
| Parental Status | - | - | - | 0.820 ± 0.003 | 0.501 ± 0.001 | 1.636 ± 0.007 |
| Race | 0.827 ± 0.005 | 0.685 ± 0.010 | 1.208 ± 0.012 | 0.829 ± 0.005 | 0.589 ± 0.009 | 1.409 ± 0.018 |
| Sexuality | 0.903 ± 0.003 | 0.777 ± 0.010 | 1.162 ± 0.013 | 0.926 ± 0.003 | 0.762 ± 0.009 | 1.216 ± 0.012 |
| Transgender | 0.968 ± 0.002 | 0.974 ± 0.005 | 0.994 ± 0.003 | - | - | - |

*Note:* Values are presented as mean ± standard deviation. "Observed" = the observed probability that nearest neighbors share the same attribute. "Random" = the expected probability by chance. "Ratio" = Observed / Random. Above the dashed line: Ratio > 1; Below the dashed line: Ratio < 1.

Table 3: Analysis of Annotator Socio-Demographics

UMAP (McInnes et al., 2018)[3] is used for dimensionality reduction, projecting annotator vectors into a two-dimensional space.

**Hatespeech Dataset** As shown in Figure 3, several socio-demographic attributes exhibit meaningful geographical patterns. Race displays a noticeable degree of separation, with "While" and "Non-white" annotators forming distinct regions in the embedding space. In the plot for education level, the "Associate" subgroup forms a more compact and distinguishable cluster compared to other categories. Income groups exhibit partial separation as well, suggesting that socio-economic background is correlated with differences in annotators' perspectives on hate speech in detectable ways.

**Toxic Dataset** In Figure 4, annotators who consider religion to be important form a more compact region in the vector space. Education level also shows a clear pattern, with annotators holding Associate, Some College, Bachelor's, and Master's degrees arranged in a roughly left-to-right progression. Parental status is also associated with annotators' perspectives on toxic content as shown in Figure 6 in Appendix C.

### 7.2 Statistical Analysis of Representation

While the visualizations provide an intuitive view of how socio-demographic attributes relate to the learned representations, dimensionality reduction inevitably discards some information. Moreover, for features with many subcategories (e.g., age groups, Figure 5 and Figure 6 in Appendix C), the resulting plots become difficult to interpret. To complement the visualization results, we quantitatively analyze the structure of socio-demographic representations.

For each annotator, we measure the probability that its nearest neighbors share the same socio-demographic attribute as the selected annotator vector, computing the ratio of (i) the probability observed in the learned representation space, and (ii) the probability expected by chance. We apply bootstrap sampling with replacement (1,000 iterations) over annotators and obtain the statistics.

**Observed Probability** For each annotator vector $i$, we retrieve its $k = 50$ nearest neighbors and compute the proportion that share the same socio-demographic attribute. We then average this value across all annotators, expressed as:

$$P_{\text{obs}} = \frac{1}{N} \sum_{i=1}^{N} \frac{\left|\{j \in \text{neighbors of } i : \text{attr}_j = \text{attr}_i\}\right|}{k} \quad (1)$$

where $N$ is the total number of annotators.

**Expected Probability by Chance** (McPherson et al., 2001) If annotator positions are random in the space, the probability that sampled annotators belong to the same category $c$ is determined solely by its relative frequency. The expected probability

[3] https://umap-learn.readthedocs.io/en/latest/

that two randomly selected annotators fall into the same category by chance is:

$$P_{\text{chance}} = \sum_{c \in C} \Pr(\text{annotator in category } c)^2 \quad (2)$$

where $C$ is the set of socio-demographic categories, and $c$ indexes each subcategory of $C$.

**Homophily Ratio** To quantify the degree of socio-demographic pattern in the learned space, we compute:

$$\text{Ratio} = \frac{P_{\text{obs}}}{P_{\text{chance}}}. \quad (3)$$

A Ratio $> 1$ indicates that the same socio-demographic attributes are more clustered than expected by chance, while a Ratio $\approx 1$ indicates random mixing of subcategories of a socio-demographic attribute.

As shown in Table 3, education exhibits the highest ratio in both datasets, followed by ideology, income, and age group, all with ratios greater than 2. This indicates that these socio-demographic attributes are strongly associated with variation in annotator perspectives. Most socio-demographic features have ratios above 1, suggesting that they contribute to modeling hate speech and toxicity judgments, albeit to varying degrees. Among all features, the observed probability for "transgender" is lower than expected by chance. This may be due to the limited number of transgender annotators (fewer than 5% of the annotator pool), resulting in patterns that are not well captured.

## 8 Discussion: Socio-Demographic Specific Modeling for Bias Mitigation

Prior studies (Cabitza et al., 2023; Zhang, 2025; Feng et al., 2024) argue that standard models trained on a single aggregated label ignore minority perspectives and encourage an averaged opinion derived from the training data. In this section, we discuss whether the proposed socio-contrastive method mitigates minority-ignorance bias and improves the fairness of representing opinions from underrepresented groups. Appendix E presents results across socio-demographic groups.

At first glance, there is no significant performance difference between the majority and minority groups under our method. For example, the $F_1$ scores for the female (0.735), male (0.711), and non-binary (0.711) groups on the hate speech task are comparable. On the toxicity task, the non-binary group achieves an $F_1$ score of 0.685, exceeding those of the female (0.640) and male (0.651) groups. Similarly, the transgender group achieves a score of 0.904, outperforming the non-transgender group (0.637).

However, we caution against overinterpreting these results and do not attribute this phenomenon to the proposed socio-demographic–enriched method, although the model has the potential to improve fairness and mitigate bias arising from the underrepresentation of minority opinions. The test sizes for minority groups are very small, as minority groups constitute a smaller proportion of the annotator pool. In both datasets, each annotator labels only around 20 items on average. The number of test instances annotated by non-binary annotators is 64 for the hate speech task and 86 for the toxicity task, compared to over 10,000 instances for the female and male groups combined. As a result, these findings may be statistically unreliable. For example, test items annotated by non-binary annotators may happen to be easy (clearly hateful or non-hateful) or difficult to predict.

On the other hand, a specific socio-demographic group can be particularly sensitive to items that target or attack their own group. This may not be sufficiently represented in the current setting due to random item assignment in crowdsourced annotation. To more reliably assess whether minority group opinions are represented to a similar extent as majority groups, a better annotation procedure is required. All annotators should be asked to label a shared set of items containing potentially disputed cases.

## 9 Conclusion

This study indicates that incorporating socio-demographic attributes improves model performance in hate speech and toxic content classification. Our proposed Socio-Contrastive Learning outperforms models that do not use socio-demographic features and models that incorporate them via concatenation of socio-demographic and textual representations. Furthermore, we analyze the relationship between the learned socio-demographic representations and annotator perspectives. The results show that several socio-demographic factors are associated with variations in hate speech and toxicity-related perspectives to different extents.

## Limitations

Our investigation into the role of socio-demographic features in perspective modeling is subject to several constraints. First, large-scale datasets that are annotated by diverse populations and include rich, reliable socio-demographic metadata remain scarce. This prevents us from experimenting on broader and more diverse datasets. Second, our analysis is restricted by the set of socio-demographic attributes provided by the original datasets. Although these attributes provide an initial lens on variation across groups, additional socio-demographic dimensions remain unexplored if they are potentially relevant to perspective differences. Finally, for many data collection efforts, particularly those involving more "objective" tasks, socio-demographic information is typically not collected under the assumption that such tasks are unaffected by demographic factors. As a result, we are unable to systematically compare how the contribution of socio-demographic features to perspective modeling varies across task types.

## References


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

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.

Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. 2024. Modular pluralism: Pluralistic alignment via multi-LLM collaboration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4151–4171, Miami, Florida, USA. 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.

Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7591–7609.

Kamil Kanclerz, Alicja Figas, Marcin Gruza, Tomasz Kajdanowicz, Jan Kocoń, Daria Puchalska, and Przemyslaw Kazienko. 2021. Controversy and conformity: from generalized to personalized aggressiveness detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5915–5926.

Chris J Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*.

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

Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing toxic content classification for a diversity of perspectives. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 299–318.

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. Association for Computational Linguistics.

Elisa Leonardelli, Alexandra Uma, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, and Massimo Poesio. 2023. Semeval-2023 task 11: Learning with disagreements (LeWiDi). *arXiv preprint arXiv:2304.14803*.

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*.

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

Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444.

Negar Mokhberian, Myrl Marmarelis, Frederic Hopp, Valerio Basile, Fred Morstatter, and Kristina Lerman. 2024. Capturing perspectives of crowdsourced annotators in subjective learning tasks. In *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, Mexico City, Mexico. Association for Computational Linguistics.

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.

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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. PyTorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Dong Pham, Xanh Ho, Quang Thuy Ha, and Akiko Aizawa. 2023. Solving label variation in scientific information extraction via multi-task learning. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 243–256, Hong Kong, China. 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.

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.

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

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 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 N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Yuxia Wang, Shimin Tao, Ning Xie, Hao Yang, Timothy Baldwin, and Karin Verspoor. 2023. Collective human opinions in semantic textual similarity. *Transactions of the Association for Computational Linguistics*, 11:997–1013.

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, Austin, Texas. Association for Computational Linguistics.

Leixin Zhang. 2025. Proposal: From one-fit-all to perspective aware modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1016–1025.

Leixin Zhang, David Burian, Vojtěch John, and Ondřej Bojar. 2024. Unveiling semantic information in sentence embeddings. In *Proceedings of the Fifth International Workshop on Designing Meaning Representations @ LREC-COLING 2024*, pages 39–47, Torino, Italia. ELRA and ICCL.

Leixin Zhang and Çağrı Çöltekin. 2024. Tübingen-CL at SemEval-2024 task 1: Ensemble learning for semantic relatedness estimation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1019–1025, Mexico City, Mexico. Association for Computational Linguistics.

Leixin Zhang and Çağrı Çöltekin. 2026. Quantifying and predicting disagreement in graded human ratings. In *Proceedings of the Fifth Workshop on Perspectivist Approaches to NLP (NLPerspectives) at Language Resources and Evaluation Conference (LREC) 2026*. ELRA.

# A Contrastive Loss in a Batch

**Algorithm 1**
Contrastive Loss for Socio-Demographic Representation

**Input:** Socio-Demo Embeddings $\mathbf{E} \in \mathbb{R}^{B \times d}$,
Annotator Labels $\mathbf{y} \in \mathbb{R}^{B}$,
Text Identifiers $\mathbf{t} \in \mathbb{R}^{B}$,
Temperature $\tau$.

**Output:** Loss value $\mathcal{L}$

$B \leftarrow$ batch size

Compute similarity matrix of socio-demo embeddings:

$$\mathbf{S} \leftarrow \frac{\mathbf{E}\mathbf{E}^{\top}}{\tau}$$

Compute text-matching mask:

$$\mathbf{M}_{\text{text}}[i,j] = \begin{cases} 1 & t_i = t_j \\ 0 & \text{otherwise} \end{cases}$$

Positive mask (same text, same label):

$$\mathbf{M}_{\text{pos}} \leftarrow \mathbf{M}_{\text{text}} \cdot \mathbb{I}(y_i = y_j)$$

Remove diagonal:

$$\mathbf{M}_{\text{pos}} \leftarrow \mathbf{M}_{\text{pos}} \cdot (1 - \mathbf{I})$$

Loss for positive cases:

$$\mathcal{L}_{\text{pos}} \leftarrow -\frac{\sum \log \text{Softmax}(\mathbf{S}) \cdot \mathbf{M}_{\text{pos}}}{\max(\sum \mathbf{M}_{\text{pos}}, 1)}$$

Negative mask (same text, different label):

$$\mathbf{M}_{\text{neg}} \leftarrow \mathbf{M}_{\text{text}} \cdot \mathbb{I}(y_i \neq y_j)$$

Loss for negative cases:

$$\mathcal{L}_{\text{neg}} \leftarrow \frac{\sum \text{Softmax}(\mathbf{S}) \cdot \mathbf{M}_{\text{neg}}}{\max(\sum \mathbf{M}_{\text{neg}}, 1)}$$

**return** $\mathcal{L} = \mathcal{L}_{\text{pos}} + \mathcal{L}_{\text{neg}}$

# B Model Parameters

| Hyperparameter | Value |
|---|---|
| **Model Configuration** | |
| Hidden dimension (layer 1) | 512 |
| Hidden dimension (layer 2) | 256 |
| Socio-Contrastive Layers | (64, 128) |
| Dropout rate | 0.2 |
| Activation function | ReLU |
| **Training Configuration** | |
| Learning rate | 0.01 |
| Batch size | 32 |
| Number of epochs | 7 |
| Optimizer | Adam |
| **Evaluation Settings** | |
| Decision threshold | 0.5 |
| F1 averaging | Binary |

Table 4: Hyperparameter settings.

## C Visualization of Additional Socio-demographic Attributes

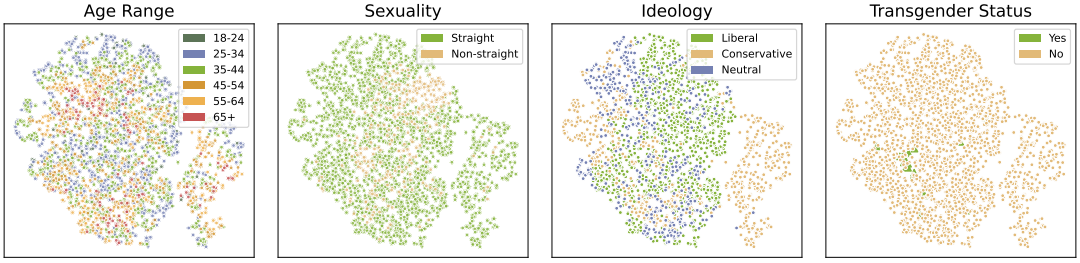


Figure 5: Visualization of Socio-Demographic Representation for the Hate Speech Dataset

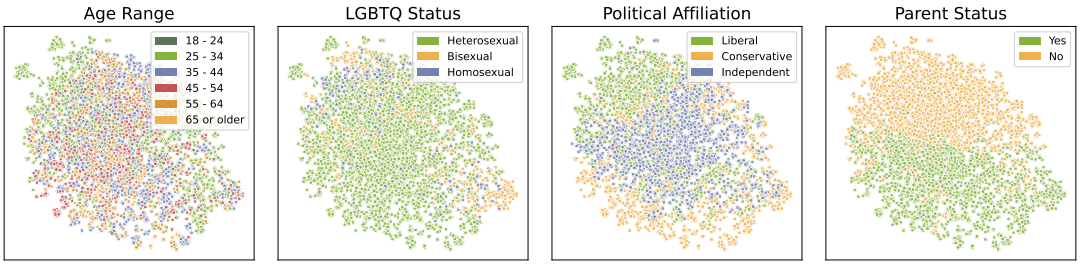


Figure 6: Visualization of Socio-Demographic Representation for the Toxic Dataset

## D Additional Results

| **BERT Encoder** | **Hate Speech** | | | **Toxic** | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F1 | Precision | Recall | F1 |
| Simple Model | 0.668 $\pm$ 0.028 | 0.735 $\pm$ 0.065 | 0.698 $\pm$ 0.031 | 0.565 $\pm$ 0.015 | 0.628 $\pm$ 0.061 | 0.593 $\pm$ 0.019 |
| Socio Multi-Hot | 0.736 $\pm$ 0.011 | 0.654 $\pm$ 0.030 | 0.692 $\pm$ 0.019 | 0.646 $\pm$ 0.011 | 0.562 $\pm$ 0.031 | 0.601 $\pm$ 0.014 |
| Socio Embedding | 0.740 $\pm$ 0.013 | 0.629 $\pm$ 0.033 | 0.679 $\pm$ 0.016 | 0.655 $\pm$ 0.013 | 0.546 $\pm$ 0.036 | 0.595 $\pm$ 0.016 |
| Socio Contrastive (Ours) | 0.690 $\pm$ 0.038 | 0.736 $\pm$ 0.092 | 0.708 $\pm$ 0.026 | 0.652 $\pm$ 0.012 | 0.608 $\pm$ 0.039 | 0.629 $\pm$ 0.016 |

Table 5: Performance comparison of models (mean ± std). Text representations are encoded using BERT.

| **RoBERTa Encoder** | **Hate Speech** | | | **Toxic** | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F1 | Precision | Recall | F1 |
| Simple Model | 0.683 $\pm$ 0.017 | 0.724 $\pm$ 0.049 | 0.702 $\pm$ 0.017 | 0.569 $\pm$ 0.017 | 0.670 $\pm$ 0.063 | 0.613 $\pm$ 0.019 |
| Socio Multi-Hot | 0.742 $\pm$ 0.009 | 0.610 $\pm$ 0.059 | 0.668 $\pm$ 0.033 | 0.650 $\pm$ 0.018 | 0.598 $\pm$ 0.055 | 0.621 $\pm$ 0.023 |
| Socio Embedding | 0.749 $\pm$ 0.008 | 0.650 $\pm$ 0.031 | 0.696 $\pm$ 0.017 | 0.647 $\pm$ 0.009 | 0.606 $\pm$ 0.026 | 0.625 $\pm$ 0.009 |
| Socio Contrastive (Ours) | 0.677 $\pm$ 0.074 | 0.768 $\pm$ 0.112 | 0.710 $\pm$ 0.029 | 0.645 $\pm$ 0.013 | 0.646 $\pm$ 0.029 | 0.645 $\pm$ 0.010 |

Table 6: Performance comparison of models (mean ± std). Text representations are encoded using RoBERTa.

## E Results by Socio-Demographic Group Divisions

| Group Division (n = Test Size) | Simple Model | Multi-Task | Socio Multi-Hot | Socio Embedding | Socio Contrastive |
|---|---|---|---|---|---|
| **Gender** | | | | | |
| female (n=6018) | 0.427 | 0.614 | 0.698 | 0.706 | 0.735 |
| male (n=4225) | 0.398 | 0.601 | 0.675 | 0.688 | 0.711 |
| non-binary (n=64) | 0.418 | 0.637 | 0.596 | 0.744 | 0.711 |
| **Age Group** | | | | | |
| 26-35 (n=3565) | 0.405 | 0.593 | 0.672 | 0.669 | 0.707 |
| 36-45 (n=2642) | 0.417 | 0.604 | 0.672 | 0.686 | 0.722 |
| 46-55 (n=1702) | 0.415 | 0.634 | 0.721 | 0.738 | 0.758 |
| 56-65 (n=1013) | 0.426 | 0.637 | 0.722 | 0.754 | 0.758 |
| 0-25 (n=1001) | 0.424 | 0.574 | 0.622 | 0.646 | 0.662 |
| 66+ (n=419) | 0.435 | 0.629 | 0.767 | 0.756 | 0.778 |
| **Education Level** | | | | | |
| college_grad_ba (n=3909) | 0.415 | 0.604 | 0.666 | 0.675 | 0.716 |
| some_college (n=2517) | 0.415 | 0.594 | 0.669 | 0.687 | 0.720 |
| college_grad_aa (n=1506) | 0.433 | 0.624 | 0.732 | 0.733 | 0.741 |
| masters (n=999) | 0.395 | 0.611 | 0.680 | 0.695 | 0.719 |
| high_school_grad (n=946) | 0.393 | 0.617 | 0.729 | 0.751 | 0.761 |
| professional_degree (n=244) | 0.425 | 0.660 | 0.745 | 0.752 | 0.757 |
| phd (n=149) | 0.381 | 0.602 | 0.655 | 0.715 | 0.652 |
| some_high_school (n=89) | 0.500 | 0.641 | 0.731 | 0.637 | 0.676 |
| **Sexuality Straight** | | | | | |
| Yes (n=9020) | 0.413 | 0.608 | 0.690 | 0.700 | 0.727 |
| No (n=1339) | 0.424 | 0.608 | 0.667 | 0.693 | 0.715 |
| **Race White** | | | | | |
| Yes (n=8295) | 0.418 | 0.609 | 0.696 | 0.707 | 0.731 |
| No (n=2064) | 0.402 | 0.605 | 0.653 | 0.667 | 0.699 |
| **Political Ideology** | | | | | |
| liberal (n=2507) | 0.406 | 0.609 | 0.690 | 0.698 | 0.730 |
| neutral (n=1838) | 0.409 | 0.608 | 0.679 | 0.699 | 0.720 |
| slightly_liberal (n=1637) | 0.425 | 0.607 | 0.669 | 0.684 | 0.704 |
| conservative (n=1321) | 0.413 | 0.618 | 0.699 | 0.704 | 0.747 |
| slightly_conservative (n=1321) | 0.439 | 0.590 | 0.677 | 0.703 | 0.712 |
| extremely_liberal (n=1185) | 0.393 | 0.601 | 0.678 | 0.689 | 0.721 |
| extremely_conservative (n=315) | 0.427 | 0.632 | 0.747 | 0.728 | 0.765 |
| **Income Range** | | | | | |
| 10k-50k (n=4264) | 0.415 | 0.607 | 0.679 | 0.689 | 0.718 |
| 50k-100k (n=4034) | 0.425 | 0.614 | 0.695 | 0.719 | 0.732 |
| 100k-200k (n=1299) | 0.410 | 0.595 | 0.693 | 0.664 | 0.730 |
| <10k (n=555) | 0.371 | 0.611 | 0.719 | 0.718 | 0.751 |
| >200k (n=200) | 0.318 | 0.577 | 0.513 | 0.564 | 0.598 |

Table 7: F1 Scores for Socio-Demographic Group Divisions on the Hate Speech Task.

| Group Division (n = Test Size) | Simple Model | Multi-Task | Socio Multi-Hot | Socio Embedding | Socio Contrastive |
|---|---|---|---|---|---|
| **Gender** | | | | | |
| Female (n=8515) | 0.485 | 0.527 | 0.618 | 0.623 | 0.640 |
| Male (n=7428) | 0.485 | 0.562 | 0.615 | 0.634 | 0.651 |
| Prefer not to say (n=114) | 0.523 | 0.511 | 0.535 | 0.534 | 0.624 |
| Nonbinary (n=86) | 0.564 | 0.500 | 0.663 | 0.706 | 0.685 |
| **Age Group** | | | | | |
| 25 - 34 (n=6246) | 0.500 | 0.582 | 0.641 | 0.647 | 0.668 |
| 35 - 44 (n=4205) | 0.471 | 0.528 | 0.585 | 0.604 | 0.609 |
| 45 - 54 (n=2453) | 0.487 | 0.526 | 0.605 | 0.611 | 0.646 |
| 18 - 24 (n=1425) | 0.470 | 0.503 | 0.635 | 0.640 | 0.664 |
| 55 - 64 (n=1155) | 0.485 | 0.480 | 0.601 | 0.623 | 0.627 |
| 65 or older (n=609) | 0.464 | 0.495 | 0.575 | 0.589 | 0.588 |
| Prefer not to say (n=58) | 0.558 | 0.645 | 0.628 | 0.746 | 0.653 |
| **Education** | | | | | |
| Bachelor's degree in college (n=6199) | 0.488 | 0.576 | 0.621 | 0.630 | 0.656 |
| Some college but no degree (n=3282) | 0.458 | 0.461 | 0.582 | 0.584 | 0.600 |
| Master's degree (n=2436) | 0.538 | 0.655 | 0.719 | 0.736 | 0.738 |
| Associate degree in college (2-year) (n=1997) | 0.466 | 0.427 | 0.546 | 0.548 | 0.596 |
| High school graduate (n=1500) | 0.455 | 0.486 | 0.559 | 0.578 | 0.560 |
| Professional degree (JD, MD) (n=251) | 0.526 | 0.440 | 0.542 | 0.571 | 0.603 |
| Doctoral degree (n=250) | 0.466 | 0.426 | 0.451 | 0.537 | 0.525 |
| Less than high school degree (n=138) | 0.488 | 0.576 | 0.661 | 0.643 | 0.653 |
| Prefer not to say (n=60) | 0.509 | 0.647 | 0.641 | 0.619 | 0.665 |
| Other (n=38) | 0.544 | 0.553 | 0.302 | 0.665 | 0.625 |
| **Sexuality** | | | | | |
| Heterosexual (n=13730) | 0.478 | 0.522 | 0.594 | 0.605 | 0.626 |
| Bisexual (n=1305) | 0.551 | 0.713 | 0.775 | 0.790 | 0.789 |
| Homosexual (n=611) | 0.524 | 0.580 | 0.670 | 0.696 | 0.687 |
| Prefer not to say (n=331) | 0.471 | 0.536 | 0.633 | 0.652 | 0.645 |
| Other (n=126) | 0.451 | 0.343 | 0.493 | 0.486 | 0.540 |
| **Transgender** | | | | | |
| No (n=15709) | 0.482 | 0.534 | 0.607 | 0.619 | 0.637 |
| Yes (n=252) | 0.635 | 0.851 | 0.893 | 0.921 | 0.904 |
| Prefer not to say (n=190) | 0.543 | 0.534 | 0.675 | 0.654 | 0.701 |
| **Political Affiliation** | | | | | |
| Liberal (n=6551) | 0.482 | 0.490 | 0.577 | 0.587 | 0.620 |
| Conservative (n=4383) | 0.497 | 0.612 | 0.660 | 0.677 | 0.679 |
| Independent (n=4197) | 0.479 | 0.533 | 0.612 | 0.613 | 0.641 |
| Prefer not to say (n=661) | 0.507 | 0.552 | 0.694 | 0.716 | 0.688 |
| Other (n=359) | 0.454 | 0.556 | 0.537 | 0.598 | 0.569 |
| **Religion Importance** | | | | | |
| Not important (n=5848) | 0.428 | 0.424 | 0.511 | 0.505 | 0.554 |
| Very important (n=4803) | 0.525 | 0.629 | 0.676 | 0.697 | 0.702 |
| Somewhat important (n=3354) | 0.511 | 0.587 | 0.654 | 0.680 | 0.685 |
| Not too important (n=1865) | 0.490 | 0.463 | 0.596 | 0.566 | 0.617 |
| Prefer not to say (n=281) | 0.486 | 0.550 | 0.661 | 0.660 | 0.635 |
| **Is Parent** | | | | | |
| Yes (n=8268) | 0.510 | 0.581 | 0.652 | 0.668 | 0.672 |
| No (n=7701) | 0.459 | 0.494 | 0.571 | 0.578 | 0.615 |
| Prefer not to say (n=182) | 0.402 | 0.497 | 0.531 | 0.456 | 0.561 |

Table 8: F1 Scores for Socio-Demographic Group Divisions on the Toxic Content Task.