Modeling Human Perspectives with Socio-Demographic Representations

Anonymous ACL submission

Abstract

Recent studies show that many NLP tasks contain instances with annotation disagreement and multiple valid perspectives. Modeling these perspectives and understanding their association with socio-demographic attributes has also received growing attention. In this work, we propose a novel architecture that jointly models annotator perspectives and learns annotators' socio-demographic representations from their annotation patterns. Our approach not only improves the performance of hate speech and toxic content prediction but also produces meaningful socio-demographic representations. These representations enable further analysis and visualization of the relationships between socio-demographic attributes and variations in annotators' perspectives.

1 Introduction

Recent studies have shown that human annotation disagreement, and the underlying variation in annotators' perspectives are widespread across many NLP tasks. A growing body of work (Plank, 2022; Uma et al., 2021; Cabitza et al., 2023; Leonardelli et al., 2023; Wang et al., 2023; Pham et al., 2023; Zhang, 2025) emphasizes that many annotation instances do not have a single "ground truth", as humans often hold different perspectives on the same task. This phenomenon occurs not only in subjective tasks, such as hate speech and offensive content detection (Sap et al., 2022; Sang and Stanton, 2022; Waseem, 2016; Anand et al., 2024), irony detection (Casola et al., 2024) and natural language inference (Jiang and de Marneffe, 2022; Huang and Yang, 2023; Gruber et al., 2024; Lee et al., 2023), but also in tasks traditionally considered less subjective, such as part-of-speech tagging (Plank et al., 2014) and in legal domain annotation (Xu et al., 2023; Braun, 2024).

Numerous studies highlight that individuals' perspectives are influenced by their experiences and cultural contexts (Larimore et al., 2021; Huang and Yang, 2023; Luo et al., 2020). Accordingly, the relationship between annotators' socio-demographic characteristics and their annotation behavior has received increasing attention. More recently, research has extended perspective modeling to the domain of LLM alignment, investigating how large language models respond to or emulate different socio-demographic viewpoints (Salemi et al., 2024; Lee et al., 2023; Schäfer et al., 2024; Muscato et al., 2024; Santurkar et al., 2023; Hu and Collier, 2024; Feng et al., 2024).

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 socio-demographic attribute at a time, such as gender, and analyzes the statistical difference between its subcategories (e.g., Female and Male). This approach can be problematic because a single attribute alone cannot capture the complex interplay of experiences and social context that shapes an annotator's perspective. Instead, richer combinations of socio-demographic attributes are more likely to reflect the nuanced experiences and social environments, and capture the diversity of 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:** Do socio-demographic features contribute to perspective prediction in subjective NLP tasks? Specifically, does incorporating socio-demographic information into model inputs improve prediction performance compared to text-only models?
- **Research Question 2:** How should socio-demographic features be encoded and integrated with textual representations to enhance model performance?

As a key contribution of this work, we propose Socio-Contrastive Modeling (SCM), a novel architecture that jointly learns (1) annotators' socio-demographic representations inferred from their labeling behavior and (2) annotator-specific label predictions. SCM outperforms existing architectures while simultaneously producing annotators' socio-demographic representations that facilitate analysis of their association with perspective variation.

2 Related Studies

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

2.1 Individual Perspective Modeling

To predict the perspectives or opinions of individual annotators, 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).

Kanclerz et al. (2021) employ personalized approaches, such as embeddings derived from partial text annotations. They show that even a small number of annotations on highly controversial data is sufficient for their personalization methods to significantly outperform generalized models. Mostafazadeh Davani et al. (2022) introduce three methods for modeling annotation variations. The *ensemble* method models aggregated predictions from each annotator, where individual neural networks independently predict each annotator's response. The *multi-label* approach presents all annotators' labels as a target vector, with each dimension representing a specific annotator's label. In the *multi-task* approach, fully connected layers are connected to separate heads, with each head corresponding to an annotator's label. Mokhberian et al. (2024) uses an embedding layer for socio-demographic representation training, which is summed with text embeddings to improve prediction on subjective tasks.

2.2 Socio-Demographics Enriched Modeling

Many 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. Larimore et al. (2021) report significant differences between white and non-white annotators in their ratings of racist language.

In the context of socio-demographic enriched learning, *Jury Learning* (Gordon et al., 2022) models individual annotators. Final predictions are aggregated via a sampling process with a predefined demographic composition. It enables practitioners to control which groups' perspectives are reflected in the final predictions and in what proportion.

Orlikowski et al. (2023) investigate the influence of socio-demographic features by grouping annotators according to a single attribute (e.g., age) and employing group-specific layers for each subcategory (e.g., Age 25–34, Age 35-44, etc.). Each group layer is connected to the output heads of the annotators within that group. Their results do not show performance improvement compared to the baseline model which does not use socio-demographic features, as well as randomly mixed attribute groups, and they conclude that annotation patterns cannot be explained by socio-demographic attributes. However, using the same dataset, 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) of socio-demographic attributes reveals a substantial contribution of socio-demographic information to modeling and explaining annotation variation.

3 Methodology

This section describes the methods used to answer two research questions proposed in Section 1. To answer the first question: whether socio-demographic features play a role in perspective prediction, we experiment with model architectures in the following two categories:

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

- **Simple Model:** This model uses only the text embedding as input and employs a multilayer perceptron (MLP) to predict labels. When a text item is annotated by multiple annotators, their annotations are aggregated to a single label.

- **Multi-Task Model:** This model also 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 evaluate two commonly used methods and introduce a novel approach for encoding annotators' socio-demographic features. The resulting socio-demographic representations are concatenated with text embeddings to predict each annotator's labels.

- **Multi-Hot Encoding:** Annotators' socio-demographic features are represented as multi-hot vectors, with each socio-demographic feature encoded as a set of one-hot encoding and combined into a single multi-hot representation. This representation is then concatenated with text embeddings and used as input to the model.
- **Socio-Demographic Embedding:** Each socio-demographic feature is encoded as a sentence embedding using a pretrained sentence embedding model (as illustrated in Appendix A), the same model used to generate text embeddings. It embeds socio-demographic features and textual representations into a common vector space, enabling interaction between socio-demographic attributes and textual content.
- **Contrastive Representation (Our Method):** As shown in Figure 1, multi-hot socio-demographic encodings are first projected into a learnable space and tuned via a contrastive loss (see Section 4 for model details) to capture annotator-specific patterns. The resulting representations are concatenated with text embeddings to predict labels.

4 Socio-Contrastive Modeling

We propose a novel model architecture: Socio-Contrastive Modeling. Its primary objective is to predict annotator labels. Simultaneously, the model refines socio-demographic representations with contrastive loss based on each annotator's labels, allowing the representation to capture annotation patterns.

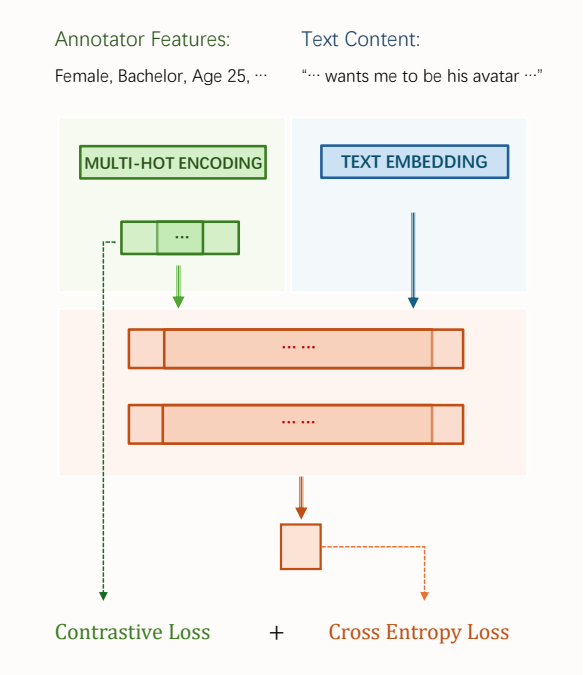

Figure 1: Socio-Contrastive Model Architecture

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, the probability of correctly predicting an annotator's label is higher when their detailed socio-demographic attributes are available. Formally, let $\hat{y} = \arg\max_y P(y \mid \cdot)$ denote the predicted label and y the true label. We posit that:

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

We hypothesize 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 input to the model consists of two components: a multi-hot encoding of the annotator's socio-demographic features, and a pretrained text embedding. Before concatenation, the multi-hot socio-demographic vector is projected to additional

fully connected layers to obtain a dense representation. The primary task of the model is to predict each annotator's label, using a cross-entropy loss. 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, such as the standard InfoNCE loss, typically operates by predefining positive and negative pairs. Its training often assumes a fixed number of positive and negative cases (often in pairs or triples). Crowdsourced datasets, however, pose several challenges. Texts are annotated by a variable number of annotators, ranging from a few to more than 100. The number of positive and negative cases often differs across items. For instance, items with perfect agreement have no negative pairs. Excluding such items would discard useful information and reduce the effectiveness of perspective modeling.

To address these issues, we design a flexible contrastive learning scheme tailored for socio-demographic representation learning. In our approach, the contrastive learning is performed within each batch. We prioritize data items with the same text ID (same text annotated by multiple annotators) into the same batch. If the number of annotators for a text is smaller than the batch size, other texts are included to fill the batch; if it exceeds the batch size, the remaining annotations are processed in subsequent batches. During training, only annotations for the same text are considered for the contrastive loss; annotations from other texts are masked.

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. A detailed description of the algorithm is presented in Appendix B.

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

H = Hate Speech; T = Toxic; N = Non HateSpeech / Toxic.

Ann. = Annotator Size. For the hate speech dataset, 2262 annotators in the testset overlap with the train set. For the toxic dataset, 1765 annotators in the testset overlap with trainset.

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

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.

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-style scales. To evaluate our socio-demographic contribution and direct comparison with baseline models, we convert them to bi-

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

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, while models below the dashed line incorporate socio-demographic features.

nary 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. Train set and test set splits are performed using unique text IDs to prevent text-level leakage between training and evaluation. Table 1 presents detailed statistics for both datasets.

5.2 Implement Details

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

Input Representations. Text inputs are encoded using the pretrained sentence-transformer model `all-MiniLM-L6-v2`[1] (Reimers and Gurevych, 2019). This model is also used to convert socio-demographic attributes into embeddings in the Socio-Embedding Model.

Loss and Optimization. For hate speech and toxicity prediction, we use cross-entropy as the objective function. For our contrastive model, we combine our customized contrastive loss with the cross-entropy loss, weighting both objectives equally. Model parameters are optimized using the Adam (Kingma, 2014) optimizer.

Training Procedure. Each model is trained under multiple hyperparameter configurations and different training epochs to determine the optimal setup. After selecting the optimal hyperparameters, we train each model for six independent runs and report the mean performance and standard deviation across these runs.

Evaluation Metrics. Given the presence of annotation disagreement and our focus on modeling individual annotator perspectives, we evaluate models on their ability to predict labels for each annotator rather than aggregated or majority-vote labels. We use a threshold of 0.5 on the sigmoid outputs 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.

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

6 Results and Discussion

Our results indicate that incorporating annotators' socio-demographic features significantly improves the prediction of individual labels for both the hate speech and toxicity datasets. As shown in Table 2, models that leverage socio-demographic information consistently outperform those that rely on textual input alone.

The socio-embedding model, which encodes socio-demographic attributes as embeddings from a pre-trained model, achieves higher predictive accuracy than the multi-hot model. Notably, our contrastively trained socio-demographic model attains the best overall performance, achieving an F_1 score of 0.725 on the hate speech dataset and 0.645 on the toxicity dataset.

Figure 2 presents the ROC curves of all five models. The trend observed in ROC analysis aligns with the F_1 results: models without socio-demographic features (simple model and multi-task model) show limited capability in classifying hate speech and toxic content. Among models incorporating socio-demographic information, embedding-based models and our contrastive model demonstrate comparable ROC performance. Both approaches outperform the multi-hot encoding model, suggesting that dense embedding representations more effectively capture the influence of annotator characteristics.

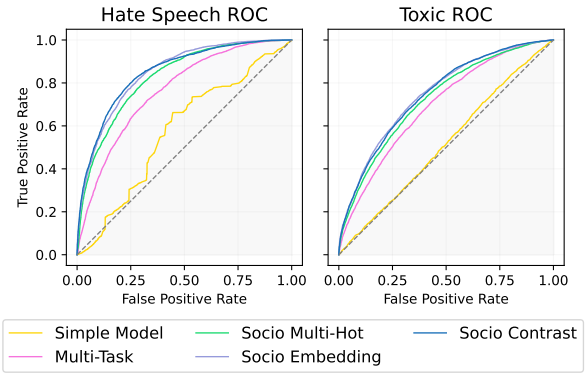

Figure 2: ROC curve: the prediction performance of five models

6.1 Analysis of Model Architecture

For models that do not incorporate socio-demographic features, like the simple model, which 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) predicts each annotator's labels via a separate output head, underperforms compared to socio-demographic enriched models. The multi-task model requires sufficient data for each annotator in order to effectively train annotator specific head. However, datasets collected from crowdsourcing platforms typically involve many annotators, with each contributing only a small number of annotations (often around 20 or fewer after noise filtering and train-test splitting). Under these conditions, the multitask architecture is not suitable for the setting with many annotators but limited data per annotator. Moreover, assigning an independent output head to each annotator is computationally expensive, particularly given that both datasets contain over 2,000 annotators.

This limitation also explains one of the reasons that Orlikowski et al. (2023) reported no performance gains from incorporating socio-demographic features. Their approach employed a similar per-annotator head architecture, which, under conditions of limited training data per annotator, restricts both the baseline and proposed models from learning meaningful annotation patterns. Furthermore, annotation behavior is rarely determined by a single socio-demographic factor. Orlikowski et al. (2023) model annotators using group-specific layers that represent individual attributes independently (e.g., gender in isolation). Our findings indicate that richer combinations of socio-demographic attributes provide more informative signals: even a simple multi-hot encoding of eight socio-demographic attributes yields measurable improvements in predictive performance.

6.2 Analysis of Socio-demographic Encoding

Compared to multi-hot encoding, socio-demographic embeddings demonstrate better results. This improvement is likely because embeddings capture the interaction between socio-demographic attributes and text content, by encoding and projecting socio-demographic attributes into a vector space that aligns with the text embeddings.

Our approach leverages contrastively learned socio-demographic representations. This strategy reduces the redundant semantic information as in socio-demographic embeddings obtained from a pretrained model, while amplifying socio-demographic signals that are more closely associated with annotation patterns. By doing so, it enhances the efficiency of interaction between socio-demographic information and textual representations.

7 Socio-Demographic Contributions

Our model architecture offers an additional advantage by representing each unique combination of socio-demographic features with a vector. Because these vectors are optimized contrastively based on annotators' labeling behavior, annotators who consistently disagree on the same items are pushed farther apart in the representation space during training. 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 complementary approaches: (1) visualization of the learned representations, and (2) statistical analysis of distances across socio-demographic groups.

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 hatespeech data

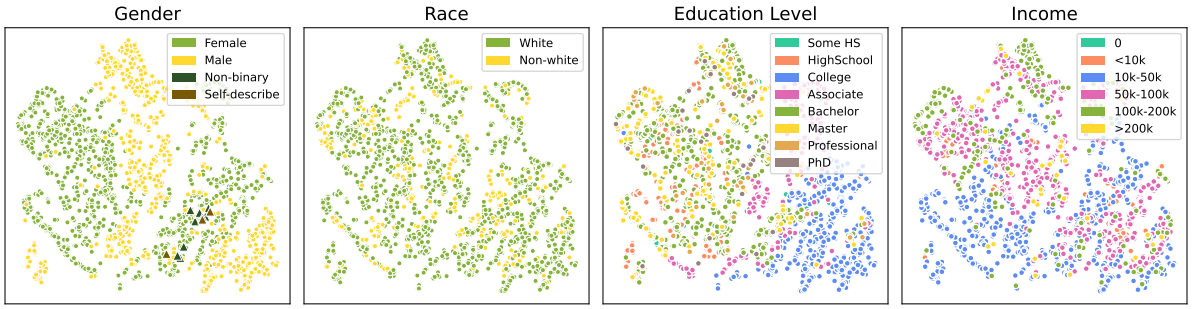

Figure 3: Visualization of Contrastively Learned Socio-Demographic Representations for the **Hatespeech** Dataset

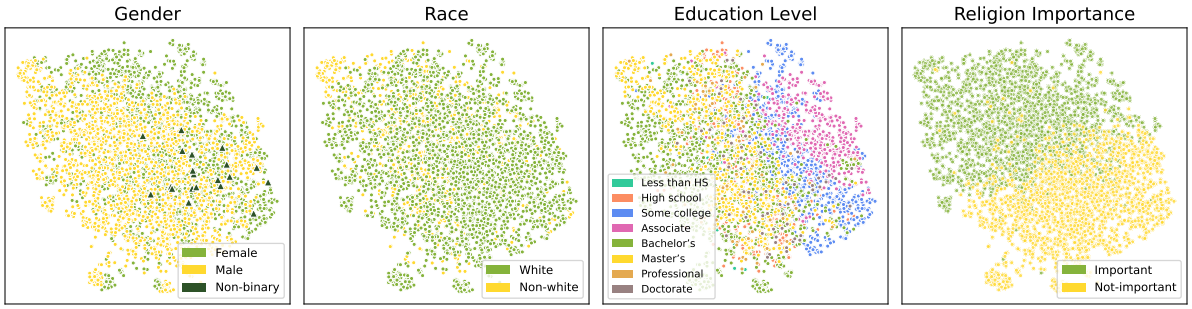

Figure 4: Visualization of Contrastively Learned Socio-Demographic Representations for the **Toxic** Dataset

and 4,408 in the toxic data). We apply UMAP[2] (McInnes et al., 2018) for dimensionality reduction and project the annotator vectors into a two-dimensional space.

Hatespeech Dataset As shown in Figure 3, several socio-demographic attributes exhibit meaningful geographical patterns. Gender displays a noticeable degree of separation, with "Female" and "Male" annotators forming distinct regions in the embedding space. In the plot for education level, the "College" 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 subtle but detectable ways.

Toxic Dataset In Figure 4, annotators who consider religion to be important versus not important form clearly separated regions in the vector space, indicating a strong correlation between attitudes toward religion and their perspectives on toxic content. Education produces a clear structure as well, with Bachelor's, Master's, Some College, and Associate-level annotators arranged in an orderly left-to-right pattern. This pattern indicates that annotators with similar educational backgrounds tend to share more aligned annotation styles or perspectives on toxic content.

Both Datasets Among all features visualized in Figure 3 and Figure 4, race consistently shows the least interpretable clustering: white and non-white annotators are highly intermixed in both datasets. This suggests that race contributes minimally to the perspective variation captured by the learned representations, at least within the contexts of hatespeech and toxic annotation tasks.

7.2 Statistical Analysis of Representation

While the visualizations provide an intuitive view of how socio-demographic attributes relate to the learned representation space, 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 (1) the probability observed in the learned representation space, and (2)

[2] https://umap.org.cn/en/latest/

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

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 were random in the space, the probability that sampled annotators belong to the same category c is determined solely by its relative frequency. Thus, the expected probability 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 subcategories of a socio-demographic attribute, and c indexes each subcategory.

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, indicating that these two socio-demographic attributes are most strongly associated with variation in annotator perspectives. Most socio-demographic features exhibit ratios greater than 1, indicating that they effectively contribute to modeling hate-speech or toxicity judgments, though 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 Conclusion

This study demonstrates that incorporating socio-demographic attributes can enhance models' performance of hatespeech and toxicity classification. The method to encode these attributes also affects model performance. Our proposed contrastive representation outperforms both multi-hot encoding and socio-demographic embeddings. Furthermore, we analyze the correlation of socio-demographic attributes with annotator perspectives. The results show that several socio-demographic factors are associated with toxic and hate speech perspectives to varying degrees, with education and political ideology as particularly strong indicators of perspective variation.

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

Abhishek Anand, Negar Mokhberian, Prathyusha Naresh Kumar, Anweasha Saha, Zihao He, Ashwin Rao, Fred Morstatter, and Kristina Lerman. 2024. Don't blame the data, blame the model: Understanding noise and bias when learning from subjective annotations. *arXiv preprint arXiv:2403.04085*.

Daniel Braun. 2024. I beg to differ: how disagreement is handled in the annotation of legal machine learning data sets. *Artificial intelligence and law*, 32(3):839–862.

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.

Silvia Casola, Simona Frenda, Soda Marem Lo, Erhan Sezerer, Antonio Uva, Valerio Basile, Cristina Bosco, Alessandro Pedrani, Chiara Rubagotti, Viviana Patti, and Davide Bernardi. 2024. MultiPICo: Multilingual perspectivist irony corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16008–16021, Bangkok, Thailand. 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.

Cornelia Gruber, Katharina Hechinger, Matthias Assenmacher, Göran Kauermann, and Barbara Plank. 2024. More labels or cases? assessing label variation in natural language inference. In *Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language*, pages 22–32.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307. Association for Computational Linguistics.

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

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

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.

Noah Lee, Na Min An, and James Thorne. 2023. Can large language models capture dissenting human voices? *arXiv preprint arXiv:2305.13788*.

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*.

Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020. Detecting stance in media on global warming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3296–3315. Association for Computational Linguistics.

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.

Benedetta Muscato, Chandana Sree Mala, Marta Marchiori Manerba, Gizem Gezici, Fosca Giannotti, and 1 others. 2024. An overview of recent approaches to enable diversity in large language models through aligning with human perspectives. In *In Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives)@ LREC-COLING 2024*, pages 49–55. European Language Resources Association (ELRA).

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.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751, Gothenburg, Sweden. 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.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7370–7392, Bangkok, Thailand. 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.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.

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.

Johannes Schäfer, Aidan Combs, Christopher Bagdon, Jiahui Li, Nadine Probol, Lynn Greschner, Sean Papay, Yarik Menchaca Resendiz, Aswathy Velutharambath, Amelie Wührl, and 1 others. 2024. Which demographics do llms default to during annotation? *arXiv preprint arXiv:2410.08820*.

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.

Shanshan Xu, Santosh T.y.s.s, Oana Ichim, Isabella Risini, Barbara Plank, and Matthias Grabmair. 2023. From dissonance to insights: Dissecting disagreements in rationale construction for case outcome classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9558–9576, Singapore. 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.

A Pretrained Embedding Model Usage

```python
from sentence_transformers import
    SentenceTransformer

sentence_bert = SentenceTransformer("all
    -MiniLM-L6-v2")

text = "gender:Female"

# Encode the text into a vector
embedding = sentence_bert.encode(text)

print(embedding)
```

The output is a numerical vector of fixed length (384 dimensions). In the Socio-Embedding Model, each socio-demographic attribute (e.g., "gender: Female") is encoded as an individual vector. The combination of socio-demographic attributes is represented by concatenating these individual vectors. Text inputs are encoded using the same pretrained model.

B 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 τ.

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

C Visualization of Additional Socio-demographic Attributes

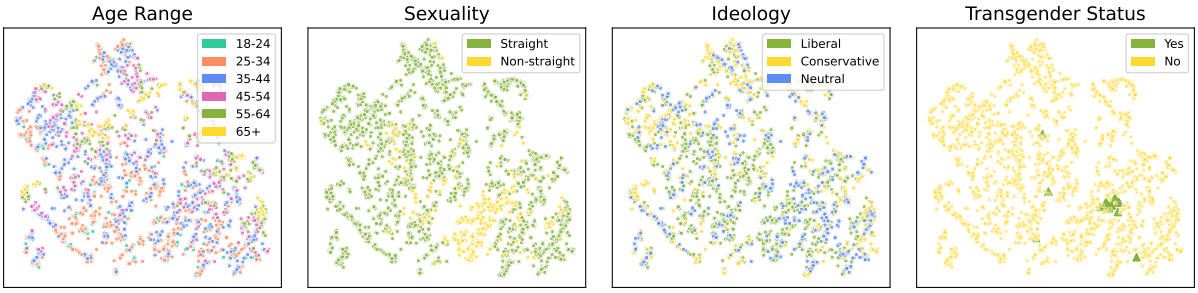

Figure 5: Visualization of Socio-Demographic Representation on the Hatespeech Dataset

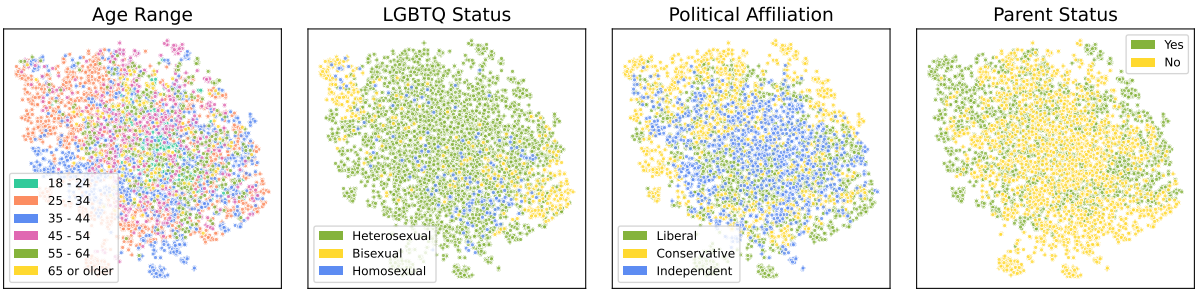

Figure 6: Visualization of Socio-Demographic Representation on the Toxic Dataset