Beyond Consensus: Use of Demographics for Datasets that Reflect Annotator Disagreement

Narjes Tahaei, Sabine Bergler
CLaC Lab
Concordia University, Montreal
n_tahaei@encs.concordia.ca, sabine.bergler@concordia.ca

Abstract

Annotator disagreement in subjective NLP tasks often reflects meaningful differences in perspective tied to demographic identity. To model this variation, we propose the Annotation-Wise Attention Network (AWAN), a demographic-aware model that learns to predict individual annotations using annotator meta-information. AWAN conditions token-level attention on demographic bundles to generate perspective-specific representations. We evaluate AWAN on two datasets, (EXIST (sexism detection) and EPICORPUS (irony detection)), showing consistent improvements over single- and multi-task baselines. We further explore how different combinations of demographic features affect performance, finding that simple, well-represented features (in the EPICorpus dataset *employment, nationality*) yield strong results, while imbalanced features (in EXIST: *study level*) can reduce model effectiveness. Our results show the promise of incorporating demographic context to model subjective variation in annotation.

1 Introduction

Supervised classification tasks in natural language processing (NLP) have long relied on the assumption that a single *gold* label can represent the correct interpretation of each example. However, this approach often hides the diversity of annotator perspectives, especially in tasks involving subjectivity, such as hate speech, sexism, or irony detection. In such contexts, annotator disagreement is not simply noise but a reflection of differing personal experiences, social positions, and demographic identities. Recent work has called attention to the risks of turning annotation diversity into a singular ground truth, particularly when focusing on underrepresented viewpoints. Recognizing this, a growing body of research has begun to explore how annotation disagreement can be leveraged rather than discarded (Aroyo & Welty, 2015; Wan et al., 2023b; Fornaciari et al., 2021; Plank et al., 2014). Annotator behavior often correlates with demographic characteristics such as age, gender, nationality, or ethnicity (Sap et al., 2019; Gordon et al., 2022; Mokhberian et al., 2024) or depends on attitudes, beliefs, or social position (Curry et al., 2024; Chulvi et al., 2023; Jiang et al., 2024). When this is the case and how to make use of this type of meta-information when it is available, is an active research area.

In this work, we investigate the integration of annotator demographic features into NLP models to better capture annotation variability in subjective tasks. Our goal is to improve classification performance and to enhance interpretability by tracing model predictions back to specific demographic perspectives. We assess our approach using two datasets, one focused on sexism detection, the other on irony detection, both tasks with high variability in human judgments.

To achieve this, we emulate the multi-task learning framework of Mostafazadeh Davani et al. (2022) by incorporating annotation-specific metadata through an attention-based mechanism. Specifically, we introduce an Annotation-Wise Attention Network (AWAN), which begins with a shared encoder to generate a general representation of the text. This representation is then refined using attention over demographic features and label inputs, yielding feature-

specific embeddings. Each embedding is routed through its own classification head to predict the corresponding annotation's label. This approach builds on related methods such as the Label-Wise Attention Network (LWAN) (Mullenbach et al., 2018), which was developed for multi-label document classification. While LWAN focuses on generating label-specific token representations, AWAN focuses on demographics-aware representations, allowing us to model how different social groups/identities interpret content.

Our findings demonstrate that incorporating demographics-informed token representations enhances performance in tasks marked by high inter-annotator disagreement. Beyond improved performance, this approach also increases transparency by enabling model predictions to be traced back to the annotator perspectives that influenced them. Furthermore, we show that well-represented annotator meta-information can significantly contribute to performance gains, a consideration that should inform future dataset collection and annotation strategies.

2 Related Work

Manually annotated datasets are foundational to the success of NLP systems. However, for tasks involving subjective interpretation, such as hate speech detection, irony recognition, or offensive language classification, annotator disagreement is common. A growing body of work has challenged the assumption that such disagreement is merely noise, advocating instead for modeling variation as a signal, which has motivated new strategies that retain individual annotations during training, enabling models to capture the diversity of perspectives that shape subjective judgments (Aroyo & Welty, 2015; Plank, 2022; Uma et al., 2021).

Annotator Demographics in Annotation

Several studies have investigated the influence of annotator identity and characteristics on labeling behavior. For example, Sap et al. (2022) and Almanea & Poesio (2022) report strong associations between annotators' backgrounds, such as *gender, religion*, and *cultural context*, and their judgments in toxicity and hate speech tasks. Similarly, Jiang et al. (2021) show that perceptions of harmful language vary substantially across annotators from different countries. Building on this line of work, researchers have begun to explore various forms of meta-information. Wan et al. (2023a), for instance, demonstrate that features such as *gender, ethnicity, education*, or *employment status* help predict not only individual annotations but also patterns of disagreement. However, the question of which demographic attributes are most informative remains open.

Empirical findings are mixed. While Orlikowski et al. (2023) found limited benefits when incorporating demographic groupings into models for toxicity detection, Fleisig et al. (2023) raise concerns about the practical gains of demographic-aware pretraining for downstream NLP performance. In contrast, other studies, including Mokhberian et al. (2024) and Jiang et al. (2024), report that when demographic metadata is involved, it can enhance both fairness and predictive accuracy.

Our work contributes to this conversation by evaluating which combinations of demographic features most effectively improve classification performance in two subjective tasks. We investigate what demographic data is most predictive in classifiers for these tasks. This is relevant in guiding future dataset collection: collecting only demographic features that provide consistent benefit can reduce privacy risks.

Demographics as a Predictive Signal

Several modeling efforts have proposed methods that integrate annotator-level information into training. Some approaches focus on improving the prediction of (aggregated) consensus labels by leveraging meta-information about the annotators (Sap et al., 2019; Akhtar et al., 2020). These methods assume that annotation variation can help disambiguate difficult examples or explain variance in annotation. For example, Mostafazadeh Davani et al. (2022)

and Mokhberian et al. (2024) show that modeling individual-level annotators' labels during training outperforms models that train on a single majority-vote label.

Other methods retain the full distribution of annotations at training and inference time, learning to model inter-annotator variation directly (Jiang et al., 2024; Mokhberian et al., 2024; Orlikowski et al., 2023). These approaches aim to answer how different individuals might assign different labels.

In our study, we adopt the latter approach: rather than discarding annotation disagreement, we leverage it by conditioning attention on demographic features through the Annotation-Wise Attention Network. By doing so, we aim to learn demographic-aware representations that improve prediction while providing transparency into how different social profiles influence model outputs. Using two distinct datasets, we demonstrate that when demographic features are diverse and well-distributed, they serve as strong predictors of labeling patterns. These findings support the use of demographic metadata as a meaningful modeling signal in subjectivity-rich NLP tasks.

3 Datasets

3.1 EXIST

The EXIST dataset (Plaza et al., 2023) labels sexist expressions in tweets. Each tweet is annotated by six annotators, identified only by their demographic data (*id, age, gender, level of study, country, ethnicity*). The training set contains 6,920 tweets (3,660 in Spanish, 3,260 in English), each annotated by 6 annotators from a pool of 725 from 45 countries, yielding 41,520 annotated instances. Each sample, therefore, has 6 individual annotations and the gold label based on a majority vote as well as meta-information (demographics) on each annotator (see Table 1).

Feature	**Range (count)**
Gender	M: 20760, F: 20760
Age	18–22: 13840, 23–45: 13840, >46: 13840
Levels of study	bachelor's degree: 20794, high school degree or equivalent: 12483, master's degree: 6635, less than a high school diploma: 684, doctorate: 639
Ethnicity	White or Caucasian: 26221, Hispano or Latino: 11742, Black or African American: 2348, Multiracial: 468, Asian: 342, Middle Eastern: 171, Asian Indian: 0
Language	Spanish: 21960 , English: 19560
Country	...

Table 1: Demographic meta-information on annotators and its frequency in the EXIST dataset. Number of annotations in total: 41520.

Our study addresses EXIST Subtask 1, a binary classification task that determines whether a tweet exhibits sexist expressions or behaviors. A sample can be classified as sexist if it directly expresses sexism, describes a sexist situation involving discrimination against women, or criticizes sexist behavior. Figure 1 illustrates an example from the dataset.

3.2 EPICorpus

The EPIC dataset (Frenda et al., 2023) was developed to support research on irony detection, with a particular focus on the subjective nature of irony perception. EPIC contains about 4500 short conversational pairs, each consisting of a post and its reply, collected from Twitter and Reddit. Each conversation pair was annotated for irony by various annotators from one up to 8, drawn from a diverse pool of 74 total annotators across five major English-speaking countries, with dominant representation from the UK, Ireland, Australia, the United States, and India. Declared annotator age ranges from 19 to 64, which we groupe into bins ($\leq$24, 25–44, 45+). In addition to irony judgments, annotators provided self-reported demographic metadata, including *age*, *sex*, *ethnicity*, *country of birth*, *country of residence*, *nationality*, *student*

status, and *employment status* (Table 2). The number of annotators per sample ranges from 2 to 8, with the majority of samples annotated by 4 or 5 individuals. To ensure consistency, we retain only those samples that have between 3 to 5 annotations.

Feature	Range (count)
Country of Birth	UK: 2539, Ireland: 2463, USA: 2433, Australia: 1653, India: 1236, Others (7 countries): 1162
Country of Residence	USA: 2743, UK: 2641, Australia: 2624, Ireland: 2457, Others (5 countries): 1021
Employment Status	Full-Time: 3799, Expired: 3557, Unemployed: 1849, Part-Time: 1598, Others (2 values): 683
Ethnicity	White: 7594, Asian: 2539, Others (4 values): 1436
Nationality	Ireland: 2457, UK: 2385, USA: 2358, Australia: 2330, India: 1956
Sex	Male: 6355, Female: 5131
Student Status	No: 7252, Yes: 2113, Expired: 2121
Age	25-44: 6999, 45+: 2895, <25: 1427, Expired: 165

Table 2: EPICorpus dataset: Demographic summary of samples with 3–5 annotators. Minority values are summed and reported as *Others*. *Expired* refers to values not assigned by annotators. Number of annotations in total: 11,486

4 AWAN: Annotation-Wise Attention Network

Developed for multi-label classification in medical document analysis, LWAN (Label-Wise Attention Network) generates label-specific medical document representations by assigning varying attention weights to input tokens based on their relevance to specific labels (Mullenbach et al., 2018), demonstrating the potential of using attention to bridge between token embeddings and meta-information by generating specialized representations for distinct aspects of a task.

We emulate LWAN by generating instead *annotation-specific* representations. Our model uses demographics and annotator labels to produce demographics-aware token representations for subjective classification.

4.1 LWAN: Label-Wise Attention Network

LWAN (Mullenbach et al., 2018) generates label-specific representations by applying attention over contextualized token embeddings. Each label receives its own attention distribution for the sample, allowing the model to highlight tokens most relevant to it. Formally, given a token representation matrix $H \in \mathbb{R}^{n \times d}$, LWAN computes label-wise representations via:

$$U = \text{softmax}(HW), \quad Z = U^\top H \tag{1}$$

where $W \in \mathbb{R}^{d \times l}$ is a learnable label query matrix, and $Z \in \mathbb{R}^{l \times d}$ contains label-specific embeddings used by dedicated classifiers.

4.2 AWAN Method

The **AWAN** model generalizes this mechanism using annotation meta-information (annotator demographics and labels). Each sample includes a demographic matrix (called avatar matrix) $\chi \in \mathbb{R}^{a \times f}$, where a is the number of demographic combinations and f the number of demographic features. We project χ and token embeddings H into query and key spaces:

$$Q = W_q \times \chi, \quad K = W_k \times H \tag{2}$$

and compute attention as:

$$U = \text{softmax}(QK^\top), \quad Z = UH \tag{3}$$

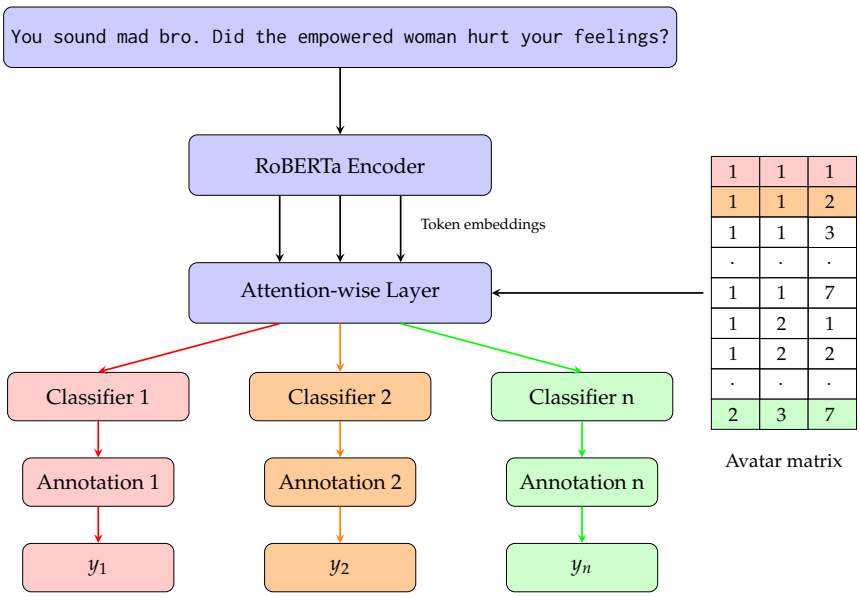

Figure 1: AWAN - from top to bottom: tweet input to RoBERTa produces fine-tuned embeddings which feed together with the Avatar matrix χ into the Annotation-wise layer producing a feature-specific embeddings that feed into a classifiers *Ann-i*.

Here, $U \in \mathbb{R}^{a \times n}$ assigns attention weights over tokens for each of the a demographic bundles. The resulting matrix $Z \in \mathbb{R}^{a \times d}$ holds demographic-aware representations, which are fed into separate binary classifiers, one for each demographic bundle. Training minimizes binary cross-entropy for each classifier, enabling the model to learn specific patterns informed by demographic context.

4.3 Avatar Matrix

χ, referred to as the *avatar matrix*, encodes demographic configurations for use in the AWAN layer. To construct this matrix, we first convert the demographic feature bundles available in the dataset into scalar representations. An example of this encoding for the EXIST dataset is shown in the left panel of Figure 2. We then explore two strategies for constructing χ, illustrated in the right panel of the same figure:

Full In this approach, χ is defined as a fixed matrix where each row corresponds to a unique combination of demographic feature values. Each row represents an *avatar*, a hypothetical annotator profile covering the full space of possible demographic configurations.

Since only a small number of annotators are associated with each sample, only a subset of rows of the full matrix is active per sample. Therefore, we extend the binary classification task to a three-class setting by introducing a dummy label '2' for unassigned rows.

Subset In this approach, χ is constructed dynamically, including only the rows corresponding to the actual annotators for a given sample. This method is particularly suited to datasets like EXIST, where samples are labeled by a distinct set of annotators. Each row in the matrix captures the demographic features of one annotator.

In summary, the avatar matrix can be initialized in one of two ways: as a complete, fixed-size matrix covering all possible demographic profiles (*Full*), or as a compact, instance-specific matrix reflecting only the observed annotators per sample (*Subset*), as illustrated in Figure 2.

Feature	Range	Code
Gender	Male	1
	Female	2
Age	18–22	1
	23–45	2
	>46	3
Ethnicity	Black/African American	1
	Hispano or Latino	2
	White or Caucasian	3
	Multiracial	4
	Asian	5
	Asian Indian	6
	Middle Eastern	7
	Others	8

	l_1	l_2	l_3	l_4	l_5	l_6	Majority(gold)
sample	1	0	1	1	0	1	1

a. sample and annotators' labels

Annotator	Age	Gender	Ethnicity
ann1	1	1	5
ann2	1	1	2
ann3	2	3	7
ann4	2	2	3
ann5	3	6	2
ann6	1	2	1

b. Subset

Age	Gender	Ethnicity
1	1	1
1	1	2
1	1	3
.	.	.
1	1	7
1	2	1
1	2	2
.	.	.
2	3	7

c. Full

Figure 2: Left: Three demographic features from EXIST with numeric encodings. Right: (a) Labels from six annotators with majority vote (input), (b) *Subset* initialization of χ, where the first column (*Annotator*) is not part of χ but is shown only to illustrate how the subset is extracted from the dataset, and (c) *Full* initialization of χ.

4.4 AWAN Learning Setup

Our model is trained on *unaggregated labels*, meaning it directly learns to predict the individual annotations associated with each sample, rather than relying on the consensus label. During fine-tuning, the loss is computed only with respect to these individual annotation labels not using the *gold* label. At inference time, however, model predictions are evaluated against the majority vote label, following prior work that combines annotation-level learning with aggregate-label evaluation by Uma et al. (2021) and Mokhberian et al. (2024).

This setup allows the model to capture the diversity of annotator perspectives during training while retaining compatibility with conventional evaluation metrics.

4.5 Baseline Models

To evaluate the effectiveness of our approach, we compare it against two baseline methods:

Single-task A standard classification model trained on majority vote labels. Tied annotations are excluded[1], and neither annotation labels nor demographic information is used.

Multi-task Following Mostafazadeh Davani et al. (2022), this approach treats each annotator as a separate task with a shared encoder and individual classification heads. Since our dataset includes hundreds of annotators, with each sample labeled by only six, assigning a separate classifier to each annotator results in extreme sparsity. To address this, we adapt the method by fixing the number of classifier heads to the number of annotations per sample, regardless of annotator identity. This adjustment significantly improves performance.

4.6 Experimental Setup

We employed the `cardiffnlp/twitter-roberta-base-sentiment-latest` model of Loureiro et al. (2022), a RoBERTa variant fine-tuned for sentiment analysis, available via the HuggingFace Transformers library (Wolf et al., 2020). While we also tested other models such as RoBERTa-XLM and Multilingual BERT, the performance gains were minimal. We selected the RoBERTa Base model for its balance of efficiency and adequate performance.

[1]This is the competition setting of the EXIST shared task (Plaza et al., 2023)

Training was conducted for 10 epochs with a batch size of 1 and a learning rate of 5×10^{-6}, using the Adam optimizer. We adopted binary cross-entropy (BCE)[2] as the loss function, computed independently for each annotation's label to preserve perspective-specific supervision. We also tested PyTorch's standard cross-entropy loss, but found it suboptimal for our setting. Since it averages predictions across annotators before comparison with labels, it effectively ignores inter-annotator variation. This led to diminished performance by reducing the model's sensitivity to disagreement. In contrast, BCE preserved label variance across annotators, enabling the model to better capture diverse perspectives.

We used a single shared classifier whose output dimensionality matched the number of possible demographic combinations. However, predictions were computed only for the relevant demographic bundles present in each sample.

Model performance is evaluated using the macro-averaged F1 score (Macro-F1). Each experiment is repeated across five runs with fixed random seeds, and we report the mean and variance of the Macro-F1 scores on the test set.

For the EXIST dataset, which includes predefined training and development sets, we partitioned the original training set into new training and validation subsets for fine-tuning. Final evaluation is conducted on the official EXIST development set.

For the EPICORPUS dataset, which does not provide predefined splits, we divided the data into three subsets: training, development, and test. All evaluations are reported on the held-out test set. Each instance in EPICORPUS consists of a post and its corresponding reply. For our experiments, we concatenate these into a single input text.

		EXIST			EPICorpus		
		P	R	F1	P	R	F1
Base	Single-task	.82$_{\pm.01}$	.78$_{\pm.01}$	.80$_{\pm.009}$	.66$_{\pm.025}$	.61$_{\pm.017}$	.63$_{\pm.017}$
	Multi-task	.80$_{\pm.012}$	.81$_{\pm.009}$	.81$_{\pm.007}$	.70$_{\pm.013}$	.65$_{\pm.018}$	.66$_{\pm.016}$
AWAN	Subset	.82$_{\pm.009}$	.82$_{\pm.01}$	.82$_{\pm.01}$	.71$_{\pm.013}$	.68$_{\pm.019}$	.69$_{\pm.008}$
	Full	.83$_{\pm.002}$	.82$_{\pm.002}$	**.83**$_{\pm.006}$	.72$_{\pm.005}$	.69 $_{\pm.014}$	**.70**$_{\pm.008}$

Table 3: Precision (P), Recall (R), and Macro-F1 for both datasets

5 Results

We evaluate the effectiveness of AWAN using macro-averaged F1 (Macro-F1) on two datasets: EXIST and EPICORPUS. Table 3 shows the performance of our model in comparison to baseline methods, using *age*, *gender*, and *ethnicity* as inputs for constructing the avatar matrix. Additionally, we analyze how varying demographic configurations affect classification performance.

5.1 Performance Across Datasets

On the EXIST dataset, the AWAN model outperforms both the single-task and multi-task baselines. The single-task model, trained on majority-vote labels, achieves a Macro-F1 of 0.80, while the multi-task model improves slightly to 0.81. AWAN achieves the best performance, with 0.82 using the SUBSET avatar matrix and 0.83 with the FULL variant, indicating the benefit of incorporating demographic-aware attention.

On the EPICORPUS dataset, the single-task and multi-task models achieve Macro-F1 scores of 0.63 and 0.66, respectively. AWAN, using the FULL avatar matrix, obtains a significantly higher score of 0.70. This demonstrates AWAN's ability to capture annotation-specific labeling behavior, particularly in more demographically diverse datasets.

[2] https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html

5.2 Demographic Analysis: EXIST Dataset

To examine the effect of different demographic combinations on model performance, we experimented with several configurations of the SUBSET avatar matrix on the EXIST dataset. Across combinations (e.g., *age+gender+ethnicity*, *age+gender+study*), the performance varied only slightly between 0.82 and 0.83. Although the performance is slightly above the baselines, this modest variation is consistent with the dataset's demographic structure: each sample is annotated by a balanced set of annotators in terms of *age* and *gender*. However, *ethnicity* and *study level* are heavily imbalanced, with approximately 91% of annotations identifying as either *White or Caucasion* or *Hispano or Latino* and about 80% either *bachelor's degree* or *high school degree or equivalent*.

This imbalance impacts the utility of the avatar matrix. Many hypothetical demographic combinations in the FULL matrix do not correspond to any real annotations in the training data, and thus receive no gradient updates. Only a limited portion of the demographic space of the matrix is actively learned, constraining the potential gains from demographic-aware modeling.

5.3 Demographic Analysis: EPICorpus Dataset

In contrast, the EPICORPUS dataset includes a broader and more balanced range of demographic attributes, including *country of birth, country of residence, employment status, student status*, and *ethnicity* (see Table 2). To investigate the impact of this demographic diversity, we constructed multiple SUBSET avatar matrices using varying combinations of features, ranging from single features to all pairs, feature-triples, and feature-quadruples.

Top Performing	**Macro-F1**	**Precision**	**Recall**
sex, ethnicity, nationality, employment	**0.691**$_{\pm 0.014}$	0.72	0.67
sex, country of birth	0.688 $_{\pm 0.021}$	0.72	0.67
ethnicity, country of birth	0.688$_{\pm 0.015}$	0.71	0.67
ethnicity, residence	0.684$_{\pm 0.016}$	0.71	0.67
Low Performing			
sex, nationality, student	0.662$_{\pm 0.026}$	0.71	0.65
nationality, country of birth, student	0.660$_{\pm 0.030}$	0.72	0.65
age, sex, student	0.658$_{\pm 0.021}$	0.70	0.64
age, ethnicity, student	**0.654**$_{\pm 0.025}$	0.72	0.64

Table 4: Top and low performing demographic combinations based on Macro-F1 scores for EPICorpus. Results include F1 and standard deviation across five runs. *country of residence*: `residence`, *student status*: `student`. Singletons are not shown in the table

A subset of the results is shown in Table 4. Macro-F1 scores ranged from 0.65 to 0.69, confirming that model performance is sensitive to the choice of demographic features. The best-performing configuration, which included *age, nationality, ethnicity*, and *employment status*, achieved a Macro-F1 of 0.69. These results show that AWAN is capable of learning fine-grained distinctions in annotation patterns (avatar matrix) when diverse demographic features are well represented. The larger number of active rows in the avatar matrix during training allows for more comprehensive learning across the demographic space, resulting in larger performance gains than those observed in the EXIST setting, where demographic imbalance result in sparse updates for many demographic bundles, thereby limiting the model's ability to generalize across underrepresented groups.

To better understand the influence of demographic complexity, we averaged results based on the number of features used in each avatar matrix (Table 5). On average, single-feature configurations yielded the highest performance, with a mean Macro-F1 of 0.68 and the top score of 0.69. In contrast, three-feature combinations performed the worst overall, both in terms of mean and minimum scores. Pairwise and four-feature combinations achieved

intermediate performance, with relatively small differences between them. This suggests that simpler demographic representations, especially single features such as employment status or nationality, can be highly predictive. Adding more features introduces sparsity or noise in the avatar matrix, which reduces performance[3].

The demographics in the dataset can be categorized into three groups: basic identity attributes (*age, sex, ethnicity*), residential information (*nationality, country of residence, country of birth*), and cultural or socioeconomic factors (*student status, employment status*). The results suggest that all three demographic groups (identity, residential, and cultural) can contribute to predictive performance when there is sufficient variation of their values in the dataset. As shown in Table 4, the top-performing combinations often paired identity attributes (e.g., *sex, ethnicity*) with either *employment status* or *residential* information, highlighting the value of combining distinct demographic perspectives. In contrast, the lowest-performing configurations consistently involved student status[4]. This aligns with its distribution in the dataset, where the feature is notably imbalanced. As a result, student status provides limited predictive signal and may even introduce noise when combined with other features, decreasing performance.

# Features	Mean Macro-F1	Min	Max	Count
1 (single)	**0.686**	0.682	**0.69**	8
2 (pairs)	0.676	0.66	0.69	28
3 (triplets)	0.669	**0.65**	0.68	56
4 (quads)	0.674	0.66	0.69	70

Table 5: A summary of Macro-F1 scores of 162 runs grouped by number of demographic features used in the avatar matrix of EPICorpus. Count is the number of combinations in each group. Single avatar matrix perform best on average, triplets perform the worst.

6 Conclusions and Future Work

We show that conditioning attention on demographic bundles allows models to better capture annotation variability and consistently outperforms standard baselines. Our experiments on the EXIST and EPICORPUS datasets reveal that certain demographic features, particularly those that are diverse and well-represented, are strong predictors of labeling behavior, while imbalanced features can introduce noise. Nonetheless, incorporating demographic information consistently improves performance across settings.

For future work, we plan to explore dynamic representations of annotator profiles beyond fixed categorical encodings. This would allow the model to learn which demographic attributes are most informative during training, potentially enhancing both performance and interpretability.

References

Sohail Akhtar, Valerio Basile, and Viviana Patti. Modeling annotator perspective and polarized opinions to improve hate speech detection. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8(1):151–154, Oct. 2020.

Dina Almanea and Massimo Poesio. ArMIS - the Arabic misogyny and sexism corpus with annotator subjective disagreements. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 2282–2291, Marseille, France, June 2022.

Lora Aroyo and Chris Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24, 2015.

[3]This mirrors the dominance of uni-gram models over larger n-grams due to sparsity effects.

[4]Student status is temporary and does not usually demarcate a break with opinions afterwards.

Berta Chulvi, Lara Fontanella, Roberto Labadie-Tamayo, and Paolo Rosso. Social or individual disagreement? Perspectivism in the annotation of sexist jokes. In *Proceedings of the 2nd CEUR Workshop on Perspectivist Approaches to NLP*, 2023.

Amanda Cercas Curry, Gavin Abercrombie, and Zeerak Talat. Subjective *Isms*? on the danger of conflating hate and offence in abusive language detection. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pp. 275–282, 2024.

Eve Fleisig, Rediet Abebe, and Dan Klein. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, December 2023. Association for Computational Linguistics.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. Beyond black & white: Leveraging annotator disagreement via soft-label multitask learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2591–2597, 2021.

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

Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.

Aiqi Jiang, Nikolas Vitsakis, Tanvi Dinkar, Gavin Abercrombie, and Ioannis Konstas. Reexamining sexism and misogyny classification with annotator attitudes. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15103–15125, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

Jialun Aaron Jiang, Morgan Klaus Scheuerman, Casey Fiesler, and Jed R. Brubaker. Understanding international perceptions of the severity of harmful content online. *PLOS ONE*, 16(8):1–22, 08 2021.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-collados. TimeLMs: Diachronic language models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 251–260, Dublin, Ireland, May 2022.

Negar Mokhberian, Myrl Marmarelis, Frederic Hopp, Valerio Basile, Fred Morstatter, and Kristina Lerman. Capturing perspectives of crowdsourced annotators in subjective learning tasks. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7337–7349, Mexico City, Mexico, June 2024.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 2022.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. Explainable prediction of medical codes from clinical text. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1101–1111, New Orleans, Louisiana, June 2018.

Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. The ecological fallacy in annotation: Modeling human label variation goes beyond sociodemographics. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1017–1029, Toronto, Canada, July 2023.

Barbara Plank. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 10671–10682, Abu Dhabi, United Arab Emirates, December 2022.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 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*, pp. 742–751, April 2014.

Laura Plaza, Jorge Carrillo-de Albornoz, Roser Morante, Enrique Amigó, Julio Gonzalo, Damiano Spina, and Paolo Rosso. Overview of EXIST 2023 – learning with disagreement for sexism identification and characterization. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Cham, 2023. Springer Nature Switzerland.

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

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 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*, pp. 5884–5906, 2022.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. Learning from disagreement: A survey. *The Journal of Artificial Intelligence Research*, 72:1385–1470, 2021.

Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. Everyone's voice matters: quantifying annotation disagreement using demographic information. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence AAAI-23*, AAAI Special Track on AI for Social Impact, 2023a.

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

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