Demographic Features for Annotation-Aware Classification

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

Abstract

This paper revisits the use of annotator demographics as interpretable meta-information for modeling such variation. We adapt a lightweight attention mechanism, Annotation-Wise Attention Network (AWAN), to condition predictions on demographic features, enabling per-annotator modeling. Experiments on the EXIST sexism dataset show that AWAN improves classification performance over standard baselines, especially in cases of high annotator disagreement.

1 Introduction

Annotation tasks for subjective NLP problems often reveal disagreement between annotators. While traditionally viewed as noise, recent emerging research recognizes annotation variation as a meaningful signal and explicitly models diversity (Wan et al., 2023; Sap et al., 2022; Fornaciari et al., 2021).

This paper explores using demographic features to model annotation disagreement. Demographic features correlate with differences in annotation behavior according to Sap et al. (2019, 2022); Gordon et al. (2022); Mokhberian et al. (2024). We empirically re-evaluate the value of incorporating annotator demographic features into model training. To do this, we adapt the Label-Wise Attention Network (LWAN) into an Annotation-Wise Attention Network (AWAN), a simple and interpretable architecture that allows us to isolate and assess the impact of demographic information.

We introduce **Annotation-Wise Attention Network (AWAN)**, a method for modeling demographic variation. Inspired by label-wise attention (Mullenbach et al., 2018), AWAN transforms token embeddings into feature-specific embeddings, attending to all the annotations' demographic bundles. Feature-specific embeddings feed into specialized classifiers to similarly predict feature-specific labels, allowing us to explicitly model annotation variation in subjective NLP tasks.

We hope that by incorporating annotator features into the classification process, AWAN enhances robustness, reliability, and traceability for subjective tasks that display high annotation diversity.

2 Related Work

Work on annotator variation either relies solely on the labels given by each annotator[1] or it adds meta-information about them, such as demographics.

2.1 Incorporating Annotations

Uma et al. (2021) review studies that use annotations only. Mostafazadeh Davani et al. (2022) train a multi-task model with a shared encoder and separate classification heads for each label. Mokhberian et al. (2024) combine annotation-related embeddings with text embeddings to learn annotation-specific representations. Both approaches showed that modeling annotation diversity outperforms relying solely on gold-standard labels.

2.2 Incorporating Meta-information

Prior work has integrated demographic and attitudinal metadata to model annotator disagreement. Jury Learning (Gordon et al., 2022) simulates individual annotators via juries, while others use prompting (Jiang et al., 2024) or demographic tokens as additional input tokens (Tahaei and Bergler, 2024). However, modeling demographic groups independently yields limited gains (Orlikowski et al., 2023). In contrast, we show that combining demographic features in a lightweight attention model improves sexism detection. Rather than simulating annotators, we isolate group-level demographic signals.

[1] We refer to each label given by an annotator as *annotation* in this paper. Demographic features are called *meta-information* in this paper.

Proceedings of Recent Advances in Natural Language Processing, pages 1232–1236
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-142

2.3 LWAN: Label-Wise Attention Network

Originally developed for medical multi-label classification, LWAN assigns attention weights to input tokens based on their relevance to each label (Mullenbach et al., 2018), producing label-specific representations. We adapt LWAN to a multi-task setting where attention selects tokens relevant to each demographic bundle, combining labels and demographics into annotation-aware representations.

2.3.1 LWAN Method

Transformers (Vaswani et al., 2017) produce contextualized token embeddings $H = [h_1, \ldots, h_n] \in \mathbb{R}^{n \times d}$, where n is the number of tokens and d the embedding size.

LWAN computes label-specific attention as:

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

Here, $W \in \mathbb{R}^{d \times l}$ is a learnable label query matrix, and $U \in \mathbb{R}^{n \times l}$ holds attention weights per token and label.

The resulting matrix $Z \in \mathbb{R}^{l \times d}$ contains label-specific embeddings for classification.

3 EXIST Task and Dataset

We use the EXIST dataset (Plaza et al., 2023), which contains 6,920 training tweets (English and Spanish), each annotated for sexism by six annotators, totaling 41,520 annotations. Each annotator is associated with demographic metadata. We use three demographic features: gender (male, female), age group (18–22, 23–45, 46+), and ethnicity (8 categories including Black or African American, Asian, Asian Indian, Hispano or Latino, White or Caucasian, Multiracial, Middle Eastern, and others). These are encoded as scalar indices for input to the attention layer..

We address the binary classification task of detecting sexism, including explicit expressions and criticism of sexism.

The overall Fleiss' κ is 0.37. We also compute individual Cohen's κ scores against the majority vote, with a large number of annotators fall within the 0.45–0.75 range, suggesting a moderate level of agreement with the majority label. The presence of a long tail of lower scores highlights that some annotators are out of sync with the consensus , which is not unexpected in a task involving subjective judgment. These discrepancies suggest that while the majority vote may serve as a proxy for an 'average annotator,' it may mask underlying disagreement within the annotator pool.

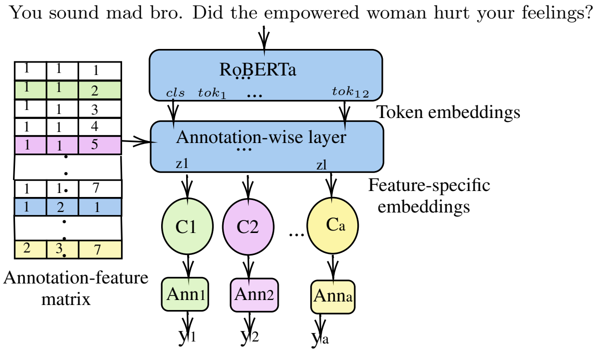

Figure 1: RoBERTa encodes tweets, which are combined with demographic matrix χ in AWAN to produce annotation-specific predictions.

4 AWAN: Annotation-Wise Attention Network

Our Annotation-Wise Attention Network (AWAN) uses a meta-information matrix χ of size $a \times f$, where a is the number of annotations per sample and f the number of demographic features (see Figure 2 and Section 4.1).

AWAN refines token representations H (from RoBERTa) using χ:

$$Q = W_q \chi, \quad K = W_k H \qquad (2)$$

where W_q and W_k are linear projections, randomly initialized.

Attention weights and contextual representations are computed as:

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

Here, $U \in \mathbb{R}^{a \times n}$ represents the weighted sum of the input sequence for a particular annotation, and $Z \in \mathbb{R}^{a \times d}$ holds annotation-specific representations. Each row of Z is passed to a classifier predicting the corresponding label, trained via binary cross-entropy.

4.1 Meta-information Representation

We encode meta-information as scalar features for matrix χ, using two initialization strategies (Figure 2):

Full: All $2 \times 3 \times 8$ combinations of demographic values (2 genders, 3 age groups, and 8 ethnicities) define 48 rows, each a hypothetical annotator. For each tweet, six rows are populated. The remaining rows, corresponding to unavailable demographic combinations, are masked using a dummy label in the 3-class setup.

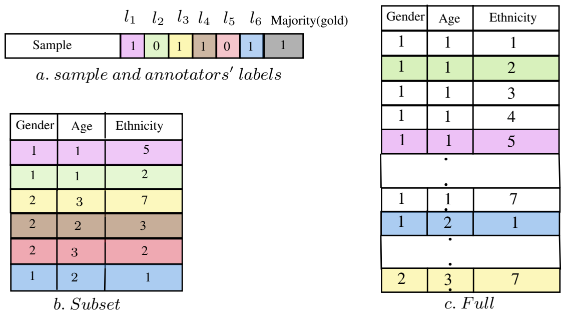

Figure 2: (a) Annotator labels and majority vote (b) *Subset* χ initialization (c) *Full* χ initialization (see Section 4.1)

Subset: Contains only six rows per tweet, representing actual annotations with their demographic values.

4.2 AWAN Variants

Unaggregated-Vote Trains on individual annotation labels and predicts them during training and evaluation, capturing diverse perspectives (Fornaciari et al., 2021; Jiang et al., 2024).

Majority-Vote Learns from annotation labels but evaluates against the majority vote, leveraging label variation to enhance overall performance. (Uma et al., 2021).

4.3 Baseline Models

Single-task Standard classification predicting majority vote, no annotation meta-information used, and tied samples excluded.

Multi-task Adapts Mostafazadeh Davani et al. (2022) by fixing six classifiers for the six annotations per sample (instead of one per annotator), sharing an encoder but no demographic features. This modification reduces sparsity and improves performance.

4.3.1 Experimental Setup

We used the 'cardiffnlp/twitter-roberta-base-sentiment-latest' model (Loureiro et al., 2022), a RoBERTa variant fine-tuned for sentiment analysis (Wolf et al., 2020). Other models (RoBERTa-XLM, Multilingual BERT) showed minimal gains, so we prioritized efficiency with RoBERTa Base.

Models were trained for 10 epochs (batch size 1, learning rate 5×10^{-6}) using Adam and binary cross-entropy (BCE)[2], which preserved annotator-level variation. Cross-entropy loss, which averages predictions, degraded performance by ignoring disagreement.

We tested two classifier setups: multiple heads (one per annotator) and a shared head predicting only valid demographic combinations. The shared head performed better and is used in all results.

We report macro F1, averaged over five fixed random seeds, with mean and variance on the test set. Since EXIST includes train and dev sets, we re-split the original train set into new train/dev splits for finetuning and evaluate on the official dev set.

5 Results

Table 1 shows results for AWAN variants compared to baselines. Incorporating annotation-specific information is impactful on the classification performance.

Both AWAN models outperform the baseline Single-task model on gold labels. The *Subset* AWAN achieves the best performance (0.83 F1), 3% higher than the Single-task model. The *Full* variant scores slightly lower (0.81 F1), likely due to its sparse representation of meta-information.

The *Full* AWAN creates a 48×3 demographic matrix, but only up to six rows are populated per sample. This sparsity weakens signal strength. In contrast, the *Subset* AWAN focuses attention on a targeted set of features, allowing better generalization and fewer distractions.

AWAN and the Multi-task model perform similarly under Unaggregated-Vote settings, suggesting the improvement stems from multi-label supervision rather than the demographics alone. However, AWAN's attention mechanism allows predictions conditioned on demographic profiles, supporting flexible modeling of annotator perspectives.

Our findings contrast with Orlikowski et al. (2023), who found no significant benefit from using demographic features independently. In contrast, AWAN jointly encodes demographics within attention, yielding gains in a subjective task.

These results suggest that demographic signals, when modeled jointly and contextually, can meaningfully enhance predictions in subjective annotation settings.

Per-Class Analysis by Agreement Rate We grouped test samples into three agreement bands[3]: High (6 annotators agree), Low (1 disagrees), and

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

[3] "Tied" cases (3 vs. 3) were excluded per task guidelines.

		Majority-Vote			Unaggregated-Vote		
		P	R	F1	P	R	F1
Base	Single-task	$.82_{\pm .01}$	$.78_{\pm .01}$	$.80_{\pm .009}$	–	–	–
	Multi-task	–	–	–	$.75_{\pm .004}$	$.75_{\pm .004}$	$.75_{\pm .004}$
AWAN	Subset	$.83_{\pm .009}$	$.82_{\pm .01}$	$\mathbf{.83}_{\pm .009}$	$.74_{\pm .008}$	$.74_{\pm .006}$	$.74_{\pm .003}$
	Full	$.83_{\pm .002}$	$.79_{\pm .002}$	$.81_{\pm .001}$	$.75_{\pm .004}$	$.75_{\pm .004}$	$.75_{\pm .004}$

Table 1: Precision (P), Recall (R), and Macro-F1 for both Majority-Vote and Unaggregated-Vote

Mid-level (2 disagree). For each band, we applied majority vote over AWAN-Subset's six per-annotator predictions to produce a single binary prediction and evaluated against the true majority label (Table 2).

AWAN-Subset consistently outperforms the Single-task baseline across all bands, with the largest gains in Mid-level disagreement. It trades a small drop in precision (2–5 pp) for notable recall gains (8–11 pp), improving F1. In the High-agreement band, it sacrifices 3.3 pp precision (95% vs. 98%) for 8.6 pp higher recall (88.3% vs. 79.8%). In Low-disagreement, it gives up about 4 pp precision for over 8 pp recall gain. In the Mid-level band, precision is comparable (72% vs. 74%), but AWAN-Subset improves recall by 11 pp (60% vs. 49%), yielding the biggest F1 gain.

These results show that AWAN's demographic attention helps most in ambiguous cases, where human disagreement increases—highlighting its value for subjective classification tasks like sexism detection.

Disagree	N	Prec.		Recall		F_1	
		Sub	ST	Sub	ST	Sub	ST
High (0–1)	366	0.95	0.98	0.88	0.80	0.91	0.88
Low (1/6, 5/6)	260	0.81	0.85	0.76	0.68	0.78	0.76
Mid (2–4)	308	0.72	0.74	0.60	0.49	0.65	0.58

Table 2: Performance by disagreement level: Sub = AWAN-Subset, ST = Single-task.

Per-Language Analysis The dataset includes tweets in English and Spanish, so we evaluated performance separately by language. mBERT, pre-trained on 100+ languages, offers strong non-English baselines. In contrast, `cardiffnlp/twitter-roberta-base-sentiment-latest` ("Cardiff") is fine-tuned on English tweets, excelling at informal English text.

As shown in Table 3, both models perform better on English than Spanish, reflecting their pretraining. However, mBERT shows smaller performance drops on Spanish (F1: 0.79→0.74) compared to Cardiff (0.81→0.73), demonstrating the benefit of multilingual pretraining.

Across both models, AWAN-Subset improves F1 scores in both languages. On English, it raises Cardiff's F1 by 3 points (0.81→0.84) and mBERT's by 2 (0.79→0.81). On Spanish, AWAN yields even larger gains: +5 for Cardiff (0.73→0.78) and +6 for mBERT (0.74→0.80). These gains reflect improved generalization from leveraging annotator meta-information, especially for low-resource languages.

		English			Spanish		
Model	Variant	P	R	F1	P	R	F1
Cardiff	Single	.81	.82	.81	.77	.74	.73
	Subset	.85	.84	**.84**	.79	.79	.78
mBERT	Single	.80	.80	.79	.74	.74	.74
	Subset	.82	.81	.81	.78	.79	**.80**

Table 3: Precision (P), Recall (R), and Macro-F1 by language and model.

6 Conclusions

This paper contributes to growing efforts to explicitly model annotator disagreement in NLP. We show that incorporating meta-information, specifically demographic features, can improve performance on subjective classification tasks. Our findings highlight the importance of how such information is represented and used during training, suggesting that learning representations conditioned on demographic profiles helps capture diverse annotator perspectives. While our implementation is a proof of concept, the approach offers a path forward for developing NLP systems that better reflect human diversity. This has potential applications in domains such as content moderation, education, and healthcare, where high levels of annotator disagreement are common and personalized or culturally sensitive interpretations are essential.

References

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

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*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Aiqi Jiang, Nikolas Vitsakis, Tanvi Dinkar, Gavin Abercrombie, and Ioannis Konstas. 2024. Re-examining sexism and misogyny classification with annotator attitudes. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15103–15125, Miami, Florida, USA. Association for Computational Linguistics.

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

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.

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

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.

Laura Plaza, Jorge Carrillo-de Albornoz, Roser Morante, Enrique Amigó, Julio Gonzalo, Damiano Spina, and Paolo Rosso. 2023. Overview of EXIST 2023 – learning with disagreement for sexism identification and characterization. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 316–342, Cham. Springer Nature Switzerland.

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

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.

Narjes Tahaei and Sabine Bergler. 2024. Analysis of annotator demographics in sexism detection. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 376–383, Bangkok, Thailand. Association for Computational Linguistics.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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

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