McMaster at LeWiDi-2025: Demographic-Aware RoBERTa

Aadi Sanghani†, Sarvin Azadi†, Virendra Jethra, Charles Welch
McMaster University
{sanghana,azadis2,jethrav,cwelch}@mcmaster.ca

Abstract

We present our submission to the Learning With Disagreements (LeWiDi) 2025 shared task. Our team implemented a variety of BERT-based models that encode annotator meta-data in combination with text to predict soft-label distributions and individual annotator labels. We show across four tasks that a combination of demographic factors leads to improved performance, however through ablations across all demographic variables we find that in some cases, a single variable performs best. Our approach placed 4th in the overall competition.

1 Introduction

The shift in natural language processing toward more perspectivist approaches has been positive, in that it allows us to incorporate a variety of viewpoints for subjective tasks and construct models that are more aligned with, and useful for individuals. The number of available disaggregated corpora is small but growing, allowing us to test more techniques in annotator modeling. While the number of available corpora has increased, the amount and type of meta-data about annotators has not significantly changed.

Sociodemographic variables are sometimes collected with annotations for analysis or modeling of the annotators. Without this information, we are often left with only the set of annotations themselves from which to learn patterns. While these demographic variables are not sufficient to represent people or populations and their diverse viewpoints, they give us a starting point to building annotator models that can be expanded in future work as more relevant information becomes available.

With this in mind, we developed a demographic-aware RoBERTa model for the shared task competition. We chose RoBERTa as the transformer of choice as it is well established, and well finetuned which offers strong baseline results, with relatively easy to finetune. With RoBERTa also widely used for NLP tasks, it increased our speed of iteration, and allowed to focus more on demographic adaptions. The shared task we submitted our system to is the 3rd edition of the Learning with Disagreements (LeWiDi) competition (Casola et al., 2025). Our system uses embeddings of demographic features and encodings of text together to predict annotator labels. We evaluate using both soft-label and perspectivist metrics, showing that our model outperforms several baselines, including the Mistral-7b large language model (LLM). Mistral-7b was specifically choosen as it is lightweight enough to run on our hardware, and has demonstrated strong performance across benchmarks. We further perform ablations, exploring the significance of individual demographic variables and discuss directions for future work.

2 Background

The learning with disagreements shared task is motivated by recent efforts in annotator modeling, pluralistic alignment, and data perspectivism. We first describe work along these directions and follow this with an in-depth description of the shared task.

2.1 Related Work

The past decade of work in natural language processing has seen a shift from understanding ground truth as an absolute to be uncovered through annotation, to a subjective value that varies across individuals with different backgrounds and perspectives (Aroyo and Welty, 2015; Frenda et al., 2024). Majority voting can take voices away from underrepresented groups, e.g. older crowdsource workers (Díaz et al., 2019). This kind of aggregation removes perspectives of sociodemographic groups and makes it difficult to discern causes of model underperformance (Prabhakaran et al., 2021).

† Denotes equal contribution.

Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP (NLPerspectivists 2025), pages 208–218

Many recent works have begun releasing disaggregated labels, supporting perspectivist work (Cabitza et al., 2023). These can be used to model annotators using a variety of approaches. Works have used disagreements in Bayesian models to identify unreliable annotations in single ground-truth scenarios (Hovy et al., 2013) and in corpora with differing labels across subpopulations (Ivey et al., 2025). Others have examined the most efficient way to label data, requesting more labels from more uncertain annotators to more efficiently model a spectrum of viewpoints (Golazizian et al., 2024). Perspectivism and personalization have been applied simultaneously in cases where extra annotator information is available (Plepi et al., 2022) with extensions from classification to perspectivist generation (Plepi et al., 2024).

Fornaciari et al. (2021) predicted soft-label distributions for all annotators and found that their model was more robust and higher performing even on the aggregated labels (through majority vote comparison). Mostafazadeh Davani et al. (2022) implemented models with varying degrees in the number of shared parameters across annotators, with some fully independent models, or only shared layers, showing improved performance. They also showed how models that predict multiple labels can be used to measure uncertainty. Mokhberian et al. (2023) proposed a similar approach, which compares multi-task models to a model that embeds individual annotators. These approaches are possible when the set of annotators is not disjoint across the train and test splits.

Deng et al. (2023) studied annotator modeling on eight datasets, finding that demographics correlated with annotation patterns but only explained a fraction of the variance in annotations. While demographic factors are not adequate predictors of differences in opinion, an individuals lived experience can be viewed as a form of expertise which informs their annotation (Fleisig et al., 2024). There is a more meaningful connection between model performance and individual annotator perception than with sociodemographic factors (Orlikowski et al., 2025).

2.2 Shared Task

Our system was built to address the shared task for the 2025 Learning with Disagreements (LeWiDi) competition (Casola et al., 2025). This task invited submissions to build classifiers for tasks not previously addressed in earlier versions of the shared task, including natural language inference, irony detection, and sarcasm detection.

A variety of distributional and information-theoretic metrics have been proposed for modeling human label distributions (Kurniawan et al., 2025). Previous versions of the LeWiDi shared task used cross-entropy and other soft evaluation metrics (Rizzi et al., 2024). This shared task similarly uses soft label predictions for evaluation, where the system outputs the distribution over labels and uses Manhattan distance to measure distance between distributions. It also requires a perspectivist evaluation, where performance is measured as the percentage of correct instances classified at the individual annotator level.

3 System Description

We fined tuned RoBERTa-Large (Liu et al., 2019) to develop a general model and apply this to all datasets through finetuning. The model architecture consisted of many different layers. This model encodes a variable-length text sequence as input and produces embeddings of each token, and a sequence embedding, represented with the [CLS] token.

For all datasets we used all available demographics. We embedded these demographics as follows. To simplify age, we binned the ranges into groups of: $18-24$, $25-35$, $35-44$, $45-55$, $55+$, for the datasets that provided age. Other demographic variables had predefined sets of categorical values from their original work. These are further listed in the demographic breakdowns for each dataset in the Appendix. For each field, a learnable embedding matrix is created, and the text embedding and the demographic embedding are concatenated into a single feature vector. This vector is then normalized using LayerNorm, regularized with dropout and also passed through a linear classifier to produce the logits for classification.

The MP and CSC corpora had many annotators with no instances annotated by all annotators, whereas the Par and VarErr NLI datasets each had only four annotators who annotated all instances. This allows for a slightly different approach. For the first two corpora, we predict each annotators label individually and aggregate them afterward to compute evaluation metrics. In this case, there are no parameters that are specifically designated to any individual annotator. For the latter two corpora, we take a different approach, predicting all anno-

tator labels at the same time. This is similar to the multi-label model described in Mostafazadeh Davani et al. (2022). This approach is tractable due to the very small number of annotators in these corpora. The training time substantially increases with the number of annotators.

We also compared our approach to the Mistral-7b model (Jiang et al., 2023). This is a large language model shown to outperform similar sized models across reasoning, mathematics, and code generation tasks using several recent optimization techniques. This is an instruction-tuned model that is more receptive to prompting.

CSC. For the CSC (Conversational Sarcasm Corpus) dataset (Jang and Frassinelli, 2024), only age and gender were provided as demographic meta-data for the annotator model. A notable difference in CSC compared to other datasets is the presence of a context situation paired with a generated "response" from a speaker. The corpus consists of 7k pairs. To preserve their distinct roles in sarcasm, we concatenated the context and response fields into a single input string, delimiting each section with special tokens. This allowed the model to better understand the situation (context) and interpret the reply (response), helping it detect the mismatch or ironic twist between them. Unlike other datasets, the goal for this dataset was to predict the provided sarcasm ratings, which ranged from 0 (not sarcastic at all) to 6 (extremely sarcastic).

MP. Specifically, for the MultiPico (MP) dataset, all of the demographic information wasn't used. The following wasn't used for the final submission: country_birth, nationality, and student status. In preliminary experiments, we found that performance decreased when using all demographic variables. We found that using a combination of country_birth, nationality, country_residence, and ethnicity decreased performance, perhaps due to the inclusion of a redundant but noisy signal. The final model we submitted used Age, Gender, Ethnicity, Country_residence, and Employment as the embeddings. Similarly, the student status meta-data didn't provide any valuable information during preliminary tests and was also omitted. This dataset contains multilingual social media data with post-reply pairs (Casola et al., 2024). The posts are labeled for irony using 0s and 1s, where 1 means that the response is ironic.

Par. The paraphrase detection dataset (Par), consists of question pairs from Quora. We implemented an approach that significantly enhanced the general model architecture. We incorporated SBERT embeddings as a layer alongside RoBERTa-Large to capture semantic similarities between paraphrase pairs more effectively. Specifically, we used the pretrained "all-MiniLM-L6-v2" SBERT model as a frozen feature extractor, concatenating its 384-dimensional embeddings with RoBERTa's 1024-dimensional [CLS] token representation. The model architecture for Par consisted of three main embedding components: RoBERTa-Large embeddings (1024 dimensions), SBERT embeddings (384 dimensions), and demographic embeddings. We used a reduced set of demographic fields (age, gender, nationality, and education) rather than the full available set, as this improved performance by reducing noise from redundant features. Age was binned into discrete ranges, and each demographic field was embedded using learnable 8-dimensional vectors. The final concatenated representation (totaling 1424 dimensions plus demographic embeddings) was processed through LayerNorm and dropout for regularization before being passed to a linear classifier for the 11-class Likert scale prediction (-5 to +5). This approach allowed the model to leverage both syntactic patterns from RoBERTa and semantic similarities from SBERT while accounting for individual annotator perspectives through demographic embeddings.

VarErr NLI. For the Variable Error Natural Language Inference, (NLI) dataset, our approach closely followed the general model architecture. Where it differed is in the output distribution. Each annotator can assign more than one label, making each output a prediction of all three labels for each of the four annotators. This dataset consists of around 1.9k explanations and 7.7 validity judgments of NLI labels (Weber-Genzel et al., 2024). The dataset presented natural language inference tasks with context-statement pairs, where annotators classified relationships as entailment, contradiction, or neutral. We maintained the standard RoBERTa-Large text encoding approach, concatenating the context and statement using the separator token and extracting the [CLS] token representation. Similar to the Par dataset, we predicted the labels of all four annotators at the same time, using separate labels in the output layer. The demographic information available in VariErr NLI included gender, age, nationality, and education level for four annotators. We utilized all avail-

Demographic	CSC	MP	Par	NLI	Values
Age	✓	✓	✓	✓	5
Gender	✓	✓	✓	✓	3
Nationality		✓	✓	✓	33
Education			✓	✓	2
Ethnicity		✓			6
Co. Birth		✓			48
Co. Residence		✓			23
Student Status		✓			2
Emp. Status		✓			7

Table 1: Inclusion of each demographic feature across datasets, showing for which datasets the metadata is present and the number of possible discrete values associated with that feature. Co. stands for *country of* and Emp. for *employment*.

able demographic features without reduction, as the limited number of annotators and demographic diversity made each feature valuable for capturing annotator-specific biases. Age was handled using the same binning strategy as other datasets, and each demographic field was embedded using 8-dimensional learnable vectors. The model produced soft label distributions across the three NLI classes (entailment, contradiction, neutral) rather than hard classifications, allowing it to capture the inherent disagreement and uncertainty in human annotations.

4 Experimental Setup

CSC. The CSC model was trained using soft label cross-entropy loss based on the annotator distributions. We optimized the model using the AdamW optimizer with a learning rate of $2e-5$, weight decay of 0.01, and applied a linear learning rate scheduler with warm-up.

The primary evaluation metric was Manhattan Distance, ranging from 0 to 1, with lower values indicating better performance. We also calculated Absolute Distance (Mean Absolute Error) as a secondary metric to assess the degree of convergence between the annotators' labels and the predicted mean label.

To test model robustness, we experimented with alternative architectures, such as Mistral large language model. However, RoBERTa consistently outperformed these alternatives across both evaluation metrics. Therefore, we carried out an extensive hyperparameter tuning process to further enhance performance, testing with factors including batch size, weight decay, dropout rate, and the number of frozen layers in RoBERTa. After determining an effective value for one of the parameters, we tuned the others while maintaining the same value.

The prompt used for Mistral is as follows:

```
You are a sarcasm detection expert.

Given the following conversation, rate
how sarcastic, the response is on a
scale from 1 (not sarcastic at all) to
6 (extremely sarcastic). Respond only
with a number between 1 and 6.

Context:
{context}

Response:
{response}

How sarcastic is the response?
```

To create this final prompt, we applied prompt engineering techniques. First, we specified the role ("You are a sarcasm detection expert") to encourage analytical reasoning. We then constrained the output ("Respond only with a number") for machine-readability. Next, we separated Context and Response to highlight their distinct roles in sarcasm interpretation. Finally, we ended with a direct question to focus the model. These changes improved clarity, reduced variability, and ensured consistent outputs.

MP. Training for the MP model was based on soft-label cross-entropy loss using annotator distributions with AdamW optimization (learning rate of $2e-5$, and weight decay of 0.01). Similarly, the primary evaluation was Manhattan Distance between the predicted and the true probability distributions, where 0 is the best possible score. The submitted model had a Manhattan Distance of 0.442.

During training, we used plots of the loss in training and validation, learning rate schedule, and performance metrics to inform our tuning of hyperparameters. Parameters were tuned individually, including the dimension of demographic embeddings of size 8, weight decay of 0.01, warm-up ratio of 0.1 and dropout of 0.3 on the concatenated feature vector.

We used prompting the same way as in the MP task. The prompt used for Mistral is as follows:

```
Analyze this social media conversation for
irony:
    Post: "{post}"
```

```
    Reply: "{reply}"
    Is the reply ironic? Consider:
    - Does it say something positive
      about a negative situation?
    - Does it use obvious exaggeration
      or contradiction?
    - Does it mean the opposite of
      what it literally says?

Answer with ONLY a number:
    0 = Not ironic/sarcastic
    1 = Ironic/sarcastic
```

Par. Training for the Par model utilized cross-entropy loss with hard labels rather than soft distributions, as the paraphrase ratings were converted to discrete classes on the Likert scale (-5 to +5, mapped to 11 classes). We used AdamW optimization with a learning rate of $1e-5$, weight decay of 0.01, batch size of 16, and a maximum of 15 training epochs. The learning rate scheduler employed a warmup ratio of 0.15 followed by linear decay. The primary evaluation metric was Manhattan Distance between predicted and true probability distributions, calculated after converting logits to softmax probabilities. Early stopping was implemented with a patience of 5 epochs to prevent overfitting. We also employed gradient clipping (max norm of 0.5) and a dropout rate of 0.3 for regularization. During training, we generated comprehensive analysis plots for each epoch including: prediction vs target scatter plots, prediction distribution comparisons, error distribution histograms, and error vs target relationships. These visualizations helped track model performance and identify potential issues like prediction bias. Key hyperparameters that we tuned included the demographic embedding dimension (8), SBERT embedding dimension (384), dropout rate (0.3), and the specific set of demographic fields used. The reduced demographic field strategy improved performance over using all available features.

The prompt for the Mistral model is as follows:

```
You are an expert at determining semantic
similarity between question pairs. Rate
how similar these questions
are on a scale from -5 to +5, where:
    -5 = Completely different meanings
    -4 = Very different meanings
    -3 = Somewhat different meanings
    -2 = Slightly different meanings
    -1 = Minor differences in meaning
     0 = Neutral/unclear relationship
    +1 = Minor similarities in meaning
    +2 = Slightly similar meanings
    +3 = Somewhat similar meanings
    +4 = Very similar meanings
    +5 = Identical or nearly identical
    meanings

Examples:
Question 1: "How do I learn Python?"
Question 2: "What's the best way to
study Python programming?"
Rating: +4 (Very similar meanings)

Question 1: "What is machine learning?"
Question 2: "How do I bake a cake?"
Rating: -5 (Completely different meanings)

Now rate this pair:
Question 1: "{question1}"
Question 2: "{question2}"
Rating:
```

VariErr NLI. Training for the VariErr NLI model followed a similar approach to other datasets, using soft-label cross-entropy loss based on the three-class probability distributions (entailment, contradiction, neutral). We maintained the AdamW optimizer configuration with appropriate hyperparameters for the NLI task structure. The evaluation was primarily based on Manhattan Distance between predicted and ground truth soft label distributions across the three NLI classes. This metric effectively captured the model's ability to predict not just the most likely class, but the full distribution of annotator disagreement. The model's performance was assessed by how well it could reproduce the uncertainty and variability inherent in human NLI judgments. Given the limited number of annotators, and the importance of capturing individual perspectives in NLI tasks, we utilized all available demographic features without reduction. The hyperparameter tuning focused on balancing the model's capacity to learn individual annotator patterns while maintaining generalization across the three-class output space. The perspectivist approach was particularly important for this dataset, as legitimate disagreement between annotators is common in natural language inference tasks where context interpretation can vary based on background knowledge and reasoning patterns (Pavlick and Kwiatkowski, 2019).

The prompt for the Mistral model is as follows:

```
You are an expert at natural language
inference. Given a context and a
statement, determine the logical
relationship.
Choose from:
- ENTAILMENT: The statement is definitely
true given the context
- CONTRADICTION: The statement is
definitely false given the context
- NEUTRAL: The statement might be true or
false; can't be determined from context

Examples:
Context: "The cat is sleeping on the couch."
Statement: "There is an animal on the
furniture."
Answer: ENTAILMENT

Context: "All birds can fly."
Statement: "Penguins cannot fly."
Answer: CONTRADICTION

Context: "John went to the store."
Statement: "John bought milk."
Answer: NEUTRAL

Now analyze:
Context: "{context}"
Statement: "{statement}"
Answer:
```

5 Results

For all tasks we evaluated using multiple different architectures to understand the impact of various ways of find an optimal model. The summary of results can be found in Table 2, with the comparison against the Majority Baseline. Our main approach, which incorporates the *Demographic Embeddings* for the annotators performs well for the given tasks. This row represents our submission to the shared task competition, which landed us in fourth place when results were computed using the grand average. This scoring approach assigned a rank the same as the random baseline for any particular dataset for which a team performed below that baseline or did not submit any results. Demographic embeddings generally improved model performance. Our model outperformed the simple baseline, the RoBERTa base model, and the Mistral LLM model. The Mistral LLM was prompted to generate responses for each instance in each corpus.

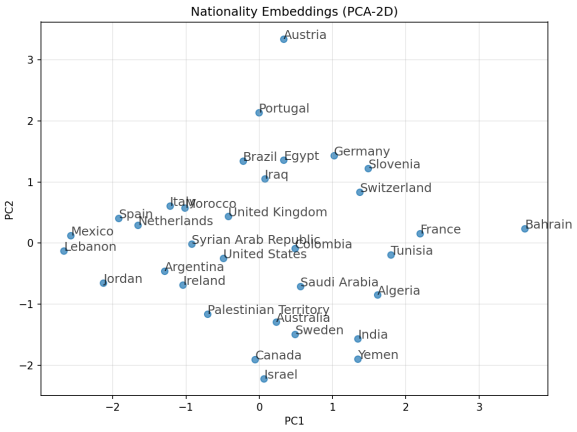

Figure 1: PCA plot showing similarity of embeddings of nationality for the MP task.

We found that even though neither the RoBERTa-base nor Mistral models incorporated annotator-specific features, the LLM performed much worse than RoBERTa.

We performed an ablation by each demographic factor, including only one piece of information at a time. We found that some variables have a much more significant impact on the model than others. The nationality/ethnicity variables appeared to perform best. Gender performed best for the Par and VariErr NLI corpora on the perspectivist evaluation. Surprisingly, we found that some of the single demographic models outperformed our submission to the shared task, showing that even better performance with a demographic-aware RoBERTa model is possible. The VariErr NLI task was the most difficult for our model, as our model underperformed on the soft evaluation and was close to the baseline on the perspectivist evaluation. Future work should explore these relationships in more detail.

6 Discussion

We noted that the LLM performance was substantially worse than the RoBERTa-based models. It is possible that the LLM could perform better with more effort put into prompt-tuning, though this remains to be shown. The added computational overhead and tuning efforts pose barriers to their practical use, over much more readily high-performing, and smaller BERT-based models.

The much smaller RoBERTa models were successful in this task, placing high on the leaderboard and showing greater improvement in our subsequent ablation experiments. Where a person is from, which is partially covered by four

Method	Soft Eval. ↓				Perspectivist Eval. ↓			
	CSC	MP	Par	VariErr NLI	CSC	MP	Par	VariErr NLI
Majority Baseline	1.169	0.518	3.23	**0.590**	0.238	0.316	0.360	0.340
Demographic Embeddings	**0.803**	0.439	1.610	0.640	**0.213**	0.311	0.200	0.340
- Age Only	0.809	0.443	1.118	0.635	0.216	0.314	0.190	0.335
- Country of Residence Only	-	0.470	-	-	-	0.329	-	-
- Country of Birth Only	-	0.442	-	-	-	0.309	-	-
- Employment Only	-	0.442	-	-	-	0.313	-	-
- Ethnicity Only	-	**0.435**	-	-	-	0.311	-	-
- Gender Only	0.811	0.444	1.145	0.633	0.215	0.310	**0.188**	**0.333**
- Education Only	-	-	1.114	0.650	-	-	0.250	0.400
- Nationality Only	-	**0.435**	**1.063**	0.630	-	**0.307**	0.270	0.380
- Student Only	-	0.449	-	-	-	0.315	-	-
RoBERTa Base	0.821	0.450	1.64	0.645	0.225	0.318	0.380	0.350
LLM (Mistral)	1.020	0.536	2.300	0.680	0.352	0.326	0.450	0.360

Table 2: Breakdown of results for the majority baseline, our submission to the LeWiDi competition, the RoBERTa base model, the large language model Mistral, and an ablation for all demographics. Empty cells mean the demographic is not available for that dataset according to Table 1. Results are shown for both the soft and perspectivist evaluations. Lowest (best) results for each column are shown in bold.

different demographic variables, appeared to have the strongest effect. As participants in the studies which collected the four datasets come from many different countries (see Appendix for details), it makes sense that this would be a variable that correlates strongly with differences in viewpoints or opinion. A PCA plot of the embeddings learned by our best RoBERTa model is shown in Figure 1, showing some regional clusters.

Sarumi et al. (2025) found that of the datasets for this shared task, VariErr NLI had the lowest annotator agreement measured by Krippendorff's alpha, $\alpha = 0.06$, while Par agreement was $\alpha = 0.09$, MP $\alpha = 0.26$ and CSC $\alpha = 0.34$. The low agreement for VariErr NLI, coupled with the low number of annotators may contribute to our lower performance on this task.

As noted in previous work, it is important to emphasize that demographics do not and cannot tell the full story (Fleisig et al., 2024). Given the historical context in which data as been collected and annotated for building NLP models, it is often the case that no meta-data is available for annotators, and when data is available it is often in the form of a handful of demographic variables. This provides us a rough starting point for beginning to explore annotator modeling, but future work must find ways to gather or infer more individual annotation patterns or those that do not directly align with sociodemographic factors.

7 Conclusion

We developed a demographic-aware RoBERTa model for annotator modeling on four tasks, including irony detection, sarcasm detection, paraphrase detection, and NLI. We found that our model could outperform baselines including a large language model; Mistral-7b. In an ablation of demographic factors, we found that nationality and ethnicity led to the biggest performance increases. We note that although demographics provide a starting point to exploring annotator modeling approaches, more individualized approaches will be needed to fully capture differences in annotation patterns.

Limitations

Our experiments with LLMs used only one type of model, which limits the generalizability of the findings, but nonetheless provides a point-of-reference for future exploration. Furthermore, our budget for hyperparameter tuning and further optimization was relatively low given our time constraints and higher performance of the BERT-based models is likely achievable as well.

Importantly, while demographics show that we can improve the model to some extent, they do not provide the full picture. We believe that more individualized approaches will be necessary to improve performance on perspectivist NLP tasks. Applications and developers should not assume that demographics are a sufficient proxy for modeling stakeholders in any scenario. Doing so poses risks to users, the severity of which depend on the specific application, but include both harms of representation and allocation (Blodgett et al., 2020).

References

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

Su Lin Blodgett, Solon Barocas, Hal Daumé Iii, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

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, Elisabetta Fersini, Diego Frassinelli, Hyewong Jang, Elisa Leonardelli, Maja Pavlovic, Siyao Peng, Massimo Poesio, and Giulia Rizzi. 2025. Learning with disagreements (LeWiDi) 3rd edition. In *Proceedings of the 4rd Workshop on Perspectivist Approaches to NLP (NLPerspectives)*. Association for Computational Linguistics.

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.

Naihao Deng, Xinliang Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You are what you annotate: Towards better models through annotator representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12475–12498, Singapore. Association for Computational Linguistics.

Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2019. Addressing age-related bias in sentiment analysis. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*.

Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024. The perspectivist paradigm shift: Assumptions and challenges of capturing human labels. 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 2279–2292, Mexico City, Mexico. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2024. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*, pages 1–28.

Preni Golazizian, Alireza S Ziabari, Ali Omrani, and Morteza Dehghani. 2024. Cost-efficient subjective task annotation and modeling through few-shot annotator adaptation. *arXiv preprint arXiv:2402.14101*.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Jonathan Ivey, Susan Gauch, and David Jurgens. 2025. Nutmeg: Separating signal from noise in annotator disagreement. *arXiv preprint arXiv:2507.18890*.

Hyewon Jang and Diego Frassinelli. 2024. Generalizable sarcasm detection is just around the corner, of course! 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 4238–4249, Mexico City, Mexico. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Kemal Kurniawan, Meladel Mistica, Timothy Baldwin, and Jey Han Lau. 2025. Training and evaluating with human label variation: An empirical study. *arXiv preprint arXiv:2502.01891*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements:

Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Matthias Orlikowski, Jiaxin Pei, Paul Röttger, Philipp Cimiano, David Jurgens, and Dirk Hovy. 2025. Beyond demographics: Fine-tuning large language models to predict individuals' subjective text perceptions. *arXiv preprint arXiv:2502.20897*.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Joan Plepi, Béla Neuendorf, Lucie Flek, and Charles Welch. 2022. Unifying data perspectivism and personalization: An application to social norms. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7391–7402, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Joan Plepi, Charles Welch, and Lucie Flek. 2024. Perspective taking through generating responses to conflict situations. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6482–6497, Bangkok, Thailand. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, Punta Cana, Dominican Republic.

Giulia Rizzi, Elisa Leonardelli, Massimo Poesio, Alexandra Uma, Maja Pavlovic, Silviu Paun, Paolo Rosso, and Elisabetta Fersini. 2024. Soft metrics for evaluation with disagreements: an assessment. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 84–94, Torino, Italia. ELRA and ICCL.

Olufunke O. Sarumi, Charles Welch, and Daniel Braun. 2025. NLP-ResTeam at LeWiDi-2025: Performance Shifts in Perspective Aware Models based on Evaluation Metrics. In *Proceedings of the 4th Workshop on Perspectivist Approaches to NLP (NLPerspectives)*. Association for Computational Linguistics.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating annotation error from human label variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269, Bangkok, Thailand. Association for Computational Linguistics.

Appendix

The following tables in this appendix describe the demographic breakdowns for all datasets used in the shared task.

Table 3: Age distribution of MP dataset annotators

Age Group	Count	Percentage
18–24	133	26.3
25–34	219	43.3
35–44	88	17.4
45–54	42	8.3
55+	24	4.7

Table 4: Gender distribution of MP dataset annotators

Gender	Count	Percentage
Male	274	54.2
Female	230	45.5
<UNK>	1	0.2

Table 5: Ethnicity distribution of MP dataset annotators

Ethnicity	Count	Percentage
White	315	62.3
Other	66	13.0
Mixed	64	12.6
Asian	44	8.7
Black	13	2.6
<UNK>	4	0.8

Table 6: Country of residence distribution of MP dataset annotators

Country	Count	Percentage
United States	66	13.0
United Kingdom	54	10.7
Germany	43	8.5
Spain	38	7.5
Canada	37	7.3
Portugal	36	7.1
Netherlands	34	6.7
France	31	6.1
Italy	30	5.9
Mexico	27	5.3
Austria	25	4.9
Switzerland	21	4.2
Australia	20	4.0
Ireland	18	3.6
Hungary	12	2.4
South Africa	5	1.0
Sweden	2	0.4
Israel	2	0.4
Poland	1	0.2
New Zealand	1	0.2
Belgium	1	0.2
Greece	1	0.2
Czech Republic	1	0.2

Table 7: Nationality distribution of MP dataset annotators

Nationality	Count	Percentage
United States	42	8.3
India	39	7.7
Canada	27	5.3
Germany	27	5.3
Netherlands	27	5.3
France	25	4.9
Austria	25	4.9
Portugal	25	4.9
Mexico	25	4.9
Colombia	25	4.9
Italy	24	4.7
Brazil	24	4.7
Spain	24	4.7
Argentina	24	4.7
Switzerland	21	4.2
United Kingdom	18	3.6
Australia	15	3.0
Ireland	15	3.0
Egypt	14	2.8
Syrian Arab Republic	8	1.6
Lebanon	6	1.2
Morocco	5	1.0
Jordan	4	0.8
Palestinian Territory	4	0.8
Saudi Arabia	3	0.6
Algeria	2	0.4
Israel	2	0.4
Slovenia	1	0.2
Bahrain	1	0.2
Tunisia	1	0.2
Sweden	1	0.2
Iraq	1	0.2
Yemen	1	0.2

Table 8: Employment status distribution of MP dataset annotators

Employment Status	Count	Percentage
Full-Time	178	35.2
<UNK>	109	21.5
Part-Time	74	14.6
Unemployed (and job seeking)	74	14.6
Other	36	7.1
Not in paid work (e.g. homemaker, retired)	24	4.7
Due to start a new job within next month	11	2.2

Table 9: Student status distribution of MP dataset annotators

Student Status	Count	Percentage
No	260	51.4
Yes	165	32.6
<UNK>	81	16.0

Table 10: Country of birth distribution of MP dataset annotators

Country of Birth	Count	Percentage
India	34	6.7
United States	31	6.1
Mexico	27	5.3
Colombia	26	5.1
Germany	25	4.9
Austria	25	4.9
Portugal	25	4.9
Netherlands	24	4.7
Brazil	24	4.7
Spain	24	4.7
Argentina	24	4.7
Canada	23	4.5
Italy	23	4.5
France	23	4.5
United Kingdom	17	3.4
Switzerland	17	3.4
Ireland	15	3.0
Egypt	14	2.8
Australia	11	2.2
Syrian Arab Republic	10	2.0
Lebanon	9	1.8
<UNK>	7	1.4
Morocco	7	1.4
Jordan	5	1.0
Saudi Arabia	5	1.0
UAE	4	0.8
Algeria	3	0.6
Togo	2	0.4
Israel	2	0.4
Iraq	2	0.4
Haiti	1	0.2
New Zealand	1	0.2
Hong Kong	1	0.2
South Africa	1	0.2
Dominican Republic	1	0.2
Martinique	1	0.2
Bosnia and Herzegovina	1	0.2
Romania	1	0.2
China	1	0.2
Nicaragua	1	0.2
Chile	1	0.2
Puerto Rico	1	0.2
Kuwait	1	0.2
Bahrain	1	0.2
Somalia	1	0.2
Tunisia	1	0.2
Palestinian Territory	1	0.2
Yemen	1	0.2

Table 11: Age distribution of CSC dataset annotators

Age Group	Count	Percentage
18–24	134	16.5
25–34	273	33.6
35–44	217	26.7
45–54	106	13.0
55+	83	10.2

Table 12: Gender distribution of CSC dataset annotators

Gender	Count	Percentage
Male	418	49.8
Female	397	47.3
<UNK> (Nan, Data_expired, Consent_revoked)	17	2.9

Table 13: Age distribution of Paraphrase dataset annotators

Age Group	Count	Percentage
25–34	3	75.0
35–44	1	25.0

Table 14: Gender distribution of Paraphrase dataset annotators

Gender	Count	Percentage
Male	2	50.0
Female	2	50.0

Table 15: Nationality distribution of Paraphrase dataset annotators

Nationality	Count	Percentage
Chinese	3	75.0
German	1	25.0

Table 16: Education distribution of Paraphrase dataset annotators

Education	Count	Percentage
Master student	4	100.0

Table 17: Age distribution of VariErrNLI dataset annotators

Age Group	Count	Percentage
18–24	1	25.0
25–34	2	50.0
35–44	1	25.0

Table 18: Gender distribution of VariErrNLI dataset annotators

Gender	Count	Percentage
Male	2	50.0
Female	2	50.0

Table 19: Nationality distribution of VariErrNLI dataset annotators

Nationality	Count	Percentage
Chinese	3	75.0
German	1	25.0

Table 20: Education distribution of VariErrNLI dataset annotators

Education	Count	Percentage
Master student	3	75.0
Postdoc	1	25.0