From One-Fit-All to Perspective Aware Models: A Thesis Proposal

Anonymous ACL submission

Abstract

Variation in human perspectives has drawn increasing attention in natural language processing. The widespread human annotation disagreement challenges the traditional paradigm of a single "ground truth" and raises concerns about the limitations of conventional label aggregation methods and the uniform models built upon them, which often discard minority opinions and obscure valuable individual perspectives. This thesis proposal investigates three core dimensions of perspective-oriented research: (1) annotation formats that better capture the nuance and uncertainty of individual judgments; (2) modeling approaches that leverage socio-demographic features to improve prediction for underrepresented or minority viewpoints; and (3) personalized generation that tailor outputs to individual users' preferences and communicative styles. Through this work, we aim to advance methods that more faithfully reflect the diversity of human interpretation, enhancing both inclusiveness and fairness in language technologies.

1 Introduction

Understanding human perspectives and designing systems that cater to individual needs is an important goal of natural language processing (NLP), particularly in tasks involving subjective interpretation such as hate speech, sentiment analysis, and toxicity classification. Traditional annotation pipelines often rely on aggregated annotations in the datasets and treat the result as ground truth for model training.

However, in recent years, the notion of a "single ground truth" in data annotation has been increasingly questioned by researchers in the field of NLP (Plank, 2022; Cabitza et al., 2023; Sap et al., 2022; Frenda et al., 2024), and related disciplines, such as legal domain (Braun and Matthes, 2024; Xu et al., 2023), medical domain (Miñarro-Giménez et al., 2018) or music annotation (Koops et al., 2019). Growing evidence suggests that annotator perspectives are shaped by complex, context-dependent factors, including individual beliefs, their demographic backgrounds, text ambiguity or interpretive uncertainty, that are not fully captured by conventional annotation practices with aggregated majority labels. Studies (Braun, 2024) also highlighted that human annotators frequently provide different but equally valid labels, challenging the assumption that there is always one correct answer. This shift in perspective calls for a deeper investigation into human variation or human perspective inclusion in annotations (Plank, 2022), modeling (Uma et al., 2021; Mostafazadeh Davani et al., 2022; Mokhberian et al., 2024) and evaluation frameworks (Basile et al., 2021; Rizzi et al., 2024) in order to improve the alignment of models with human perspectives.

This proposal aims to advance a perspective-aware approach in NLP by providing insights into annotation methodologies that better capture the complexity of human perspectives, evaluating the influence of socio-demographic factors on annotator perspective modeling, and exploring methods to leverage persona information for more accurate prediction and personalized language generation. Three tasks are elaborated from Section 3 to Section 5.

Annotation Format: This task explores different formats of annotation types in representing perspectives: binary labels vs. continuous or Likert scale values. We assess whether continuous values or Likert scales, rather than discrete labels, better capture the intensity and uncertainty of annotators' perspectives. These insights will inform improved annotation practices and enhance model alignment with human judgment, ultimately leading to more refined methods for capturing the subtleties of diverse annotator perspectives.

Perspective Modeling and Pattern Learning: This task investigates the extent to which socio-demographic features can account for annotator

perspectives or annotation patterns. Specifically, we examine the reliability of predicting an individual's annotations based on their socio-demographic attributes in application domains that have not yet been explored.

Personalized Generation: This task explores the use of persona-based modeling for generating individualized textual outputs that reflect users' preferences and communication styles. We incorporate structured persona information, such as socio-demographic features, sentiment orientation, and linguistic preferences, along with historical dialogue or user-generated content to condition language generation. The objective is to produce responses or texts that are not only contextually appropriate but also tailored in terms of demographic groups, sentiment, and language complexity.

2 Related Studies

Recent studies have increasingly recognized the presence of human disagreement and diverse perspectives in annotation tasks. Various terms have been used to describe this phenomenon, including subjectivity (Reidsma and Carletta, 2008), human uncertainty (Peterson et al., 2019), perspectivism (Cabitza et al., 2023), perspectivist (Frenda et al., 2024), and label variation (Plank, 2022). Among these, the widely accepted term "human label variation" (Plank, 2022) encompasses the notion that multiple plausible labels can be assigned to the same instance by different annotators. Moreover, an increasing number of studies have released datasets (Wang et al., 2023; Kumar et al., 2021; Frenda et al., 2023; Passonneau et al., 2012; Dumitrache et al., 2018) annotated by multiple individuals, in contrast with the single label from the traditional majority-vote aggregation or score averaging.

Prior research (Plank et al., 2014; Sheng et al., 2008; Guan et al., 2018; Fornaciari et al., 2021; Xu et al., 2024) has demonstrated that incorporating labels from multiple annotators can enhance model performance by improving the model generalization ability. Specifically, this includes a cost-sensitive approach, where the loss of each instance is weighted based on label distribution such as Sheng et al. (2008) and Plank et al. (2014), as well as soft-loss approaches (Peterson et al., 2019; Lalor et al., 2017; Uma et al., 2020; Fornaciari et al., 2021). Furthermore, researchers have explored leveraging additional metadata, such as socio-demographic features (Goyal et al., 2022; Gordon et al., 2022), annotator IDs (Mokhberian et al., 2024), and partial annotation histories (Milkowski et al., 2021), to characterize individual annotation patterns and refine learning models.

The alignment of large language models (LLMs) with human annotation has also gained increasing attention under the context of embracing human disagreement, particularly in evaluating their ability to capture diverse perspectives and which groups' perspective that LLMs reflect (Hu and Collier, 2024; Beck et al., 2024; Salemi et al., 2024; Muscato et al., 2024). In the generation domain, MORPHEUS (Tang et al., 2024) introduces a three-stage framework to model roles from dialogue history. It compresses persona information into a latent codebook, enabling generalization to unseen roles through joint training. Lu et al. (2023) disentangle multi-faceted attributes in the latent space and use a conditional variational auto-encoder to align responses with user traits.

3 Task 1: Annotation Formats for Perspective Representation

This task explores two different annotation formats (binary classification or Likert-scales ratings) for representing perspectives and investigates their influence on modeling effectiveness. The goal is to provide guidance for future dataset construction by identifying annotation formats that best support model learning and more accurately capture the nuance of human perspectives.

3.1 Motivation and Research Hypothesis

Previous research (Plank, 2022; Mostafazadeh Davani et al., 2022) has primarily focused on label variation using discrete labels. Many studies, particularly in domains such as hate speech and offensive language detection, rely on binary annotations (Mostafazadeh Davani et al., 2022; Akhtar et al., 2020). In some cases, ordinal Likert-scale ratings are converted into binary labels (Orlikowski et al., 2023).

Ovesdotter Alm (2011) argues that acceptability is a more meaningful concept than rigid "right" or "wrong" labels. Human annotators exhibit varying degrees of uncertainty for specific items, and some tasks inherently involve continuous variation, such as the level of emotional arousal (Lee et al., 2022). Simple binary classes can obscure important nuances in annotation data. It may risk over-

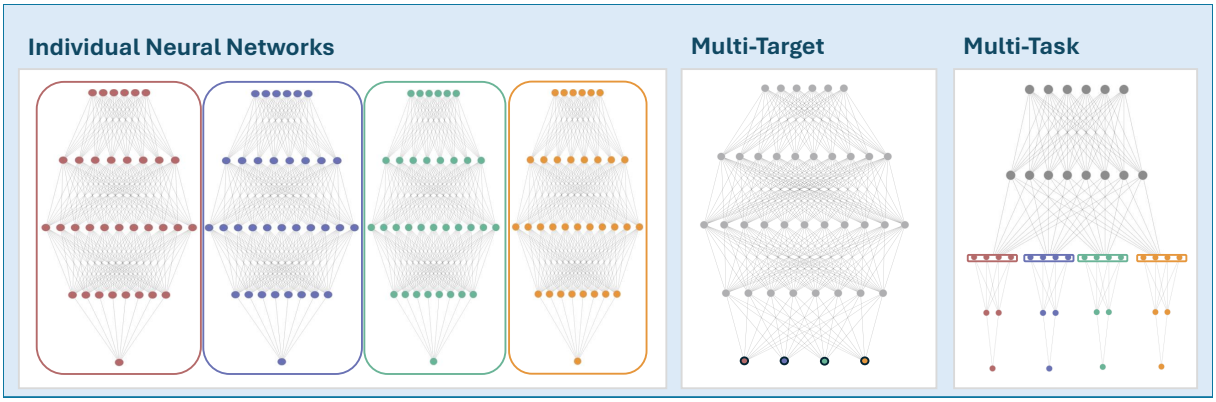

Figure 1: Neural Network Architectures for Individualized Modeling

simplifying the granularity of human perspectives, ultimately impacting model reliability and the interpretability of annotator uncertainty.

We hypothesize that continuous values or Likert scales provide a more effective source of capturing annotation variation. From the perspective of machine learning, incorporating finer-grained annotations may help align better with human judgment and enhance model robustness by smoothing the decision boundary. In addition, leveraging these finer-grained annotation formats may improve model interpretability by investigating predicted ratings with a degree specified rather than a binary decision.

3.2 Methodology

This study undertakes interdisciplinary research to investigate annotation variation across multiple domains, including hate speech detection, offensive language detection and sentence similarity[1]. By examining diverse datasets and modeling techniques, we aim to assess whether adopting finer-grained annotation scales improves the representation and learning of annotators' perspectives in a cross-domain context.

Data Construction: Two types of datasets will be used for this purpose. First, for datasets with Likert scales or continuous values, we will model using the original values and also transform them into binary labels for comparison. Second, for datasets originally with discrete labels, such as natural language inference, where three labels (entailment, contradiction, and neutrality) exist, we will annotate with an additional scale representing human uncertainty of the label selection to capture the uncertainty inherent in human judgment.

Modeling framework: To test the hypothesis (numerical values better represent human perspectives than binary labels, and models based on values show better effectiveness in machine learning), we will implement the three modeling architectures (see Figure 1) from Mostafazadeh Davani et al. (2022) to compare the results of two types of targets (binary encoding vs. continuous values):

- Individual Annotator Modeling: Each annotator's annotations will be modeled separately using distinct neural networks to capture individual perspectives.
- Multi-target Methods: A shared neural network will be trained with all annotators' annotations represented as target vectors, allowing the model to learn patterns across annotators.
- Multi-Task Learning: A partially shared neural network will be employed, with shared layers capturing common understanding and annotator-specific layers or heads capturing individualized annotation tendencies.

Evaluation and Result Analysis: Model performance will be evaluated using both traditional metrics based on aggregated labels and specialized analyses designed to assess the individualized prediction and advantages of finer-grained annotations. Since direct comparison between binary classification and regression outputs is inherently challenging, we propose two complementary evaluation strategies to facilitate a meaningful comparison:

- Binary Label Conversion: Continuous regression outputs will be converted into binary labels using a predefined threshold (consistent

[1]These tasks are known that human annotation variation exists and with relatively richer datasets annotated by multiple individuals, seen Wang et al. (2023); Akhtar et al. (2020); Waseem (2016); Jiang and de Marneffe (2022); Huang and Yang (2023) and Gruber et al. (2024).

with the threshold used during training for label derivation[2]). We will then compute standard classification metrics such as F1 score and accuracy to evaluate the alignment between the binarized predictions and the target.

- Ranked Correlation Comparison: While classifier outputs do not offer the same level of granularity as regression values, the predicted probabilities or logits can serve as proxies for prediction confidence or intensity (e.g., degree of toxicity). These values enable a ranking-based comparison with the ground truth labels. We will compute the Spearman rank correlation (r) between the model predictions and the true target values, allowing us to compare the correlation strength across both classifiers and regressors.

4 Task 2: Perspective Modeling with Demographic Features

This task investigates the extent to which socio-demographic features, such as age, gender, education level, political affiliation, and domain expertise contribute to explaining and modeling variation in human annotation. While prior research has explored this question in some subjective NLP tasks, findings remain inconclusive. For instance, Orlikowski et al. (2023) report that incorporating group-level socio-demographic features does not significantly improve predictive performance in toxicity classification tasks, especially when compared to randomly assigned groups. In contrast, Gordon et al. (2022) highlight a stronger alignment between annotator perspectives and their socio-demographic backgrounds, suggesting these features may meaningfully inform model learning.

These conflicting results raise a questions: in which application domains do socio-demographic features act as reliable indicators of perspective? Can modeling the conditional distribution $P(prediction|persona)$ yield better outcomes than assuming an undifferentiated $P(prediction)$? We aim to explore whether socio-demographic traits enhance our ability to predict annotation behavior, particularly in domains that have received limited attention in previous research.

Extending beyond traditional subjective tasks such as hate speech or sentiment classification, we apply this perspective modeling framework to the domain of business or economics to investigate the interpretation of business trends or sentiment toward economic statements. We explore how personal background and professional expertise may play a critical role in shaping interpretation. Specifically, we address the following research questions: (1) To what extent do socio-demographic attributes and domain expertise account for variation in annotator judgments in business-related tasks? (2) Which specific attributes, if any, serve as reliable predictors of annotation variation? and (3) Which modeling methods show advantages in modeling patterns of various socio-demographic groups?

4.1 Methodology

In this task, we will improve the modeling methods from the prior research (Orlikowski et al., 2023) in order to model socio-demographic features more efficiently. The following modeling methds will be explored.

- Socio-Demographic Embedding Learning: Embedding layers will be used to encode socio-demographic attributes, enabling the model to capture correlations and interactions among variables such as gender, nationality, and political orientation. This embedding-based model will be compared against a baseline where these features are randomly shuffled to assess their true contribution to performance.

- Partial Annotation Representation: We will incorporate a small subset of each annotator's historical annotations as input features, allowing the model to learn a latent representation of annotation style from observed data.

- Leveraging Large Language Models (LLMs): We will experiment with prompt-based approaches to incorporate persona information into LLM predictions. Specifically, we will encode demographic and stylistic attributes into prompts using structured key-value formats or natural language descriptions.

- Lightweight Fine-Tuning of LLMs: To further enhance performance, we will apply parameter-efficient fine-tuning techniques such as prefix tuning (Li and Liang, 2021). These methods enable personalization without extensive retraining, making them suitable for incorporating socio-demographic signals.

[2]Different threshold values can be explored to assess robustness.

Finally, we propose a comparative evaluation of two modeling paradigms: (1) socio-demographic enriched learning, which uses population-level features to inform predictions, and (2) individual-level modeling, which treats annotations from each annotator as separate outputs, as elaborated in Section 3. This comparison will shed light on whether generalizable demographic factors or personalized modeling better account for variation in annotation or human perspectives.

5 Task 3: Personalized Generation

Building on the exploration of annotation variation and perspective modeling in the previous tasks (Section 3 and Section 4), this task extends the research to personalized text generation. The goal is to generate language that aligns with individual users' backgrounds, preferences, and communication styles. This includes conditioning generation on persona-related factors such as socio-demographic attributes, historical dialogue context, and language preferences. Personalized generation aims to adapt to user needs and enhance user engagement and satisfaction.

5.1 Methodology

The proposed approach involves two main stages: (1) Persona Retrieval or Representation and (2) Generation with Alignment to Individual Preferences. Persona information can be constructed from both explicit features (e.g., age, gender, education level, profession) provided additionally during the data construction phase and implicit cues derived from historical text. It requires a preliminary persona prediction or persona representation learning. We will explore two main methods of generation based on persona representation:

Prompt-Based Personalization: Key persona information and dialogue textual features will be extracted and formulated into structured prompts, e.g., define a certain user role. It can be incorporated into the model input to guide generation in a controlled and personalization guided manner.

Latent Representation Learning and LLM Fine-tuning: Following approaches such as MORPHEUS (Tang et al., 2024) and MIRACLE (Lu et al., 2023), we can encode multi-faceted user attributes into latent embeddings and use learned embeddings for condition text generation, allowing for fine-grained control over personalization dimensions such as sentiment, formality, or linguistic complexity. Lightweight fine-tuning techniques (e.g., prefix tuning Li and Liang, 2021, LoRA Hu et al., 2022) will be explored to incorporate personalized signals into generation.

5.2 Evaluation

Evaluating personalized generation poses additional challenges beyond conventional evaluation of generation quality metrics. We will adopt multiple evaluation strategies to assess generation performance:

- Standard Generation Metrics: Including BLEU, ROUGE, and METEOR to assess content quality, coherence, and relevance.
- Persona-Based Metrics: We will evaluate the alignment between generated outputs and persona information by measuring the overlap or differences between generated texts and persona sentences in datasets like PersonaChat (Jandaghi et al., 2023). To assess whether generated texts reflect target attributes, we will use classification or clustering-based evaluations, measuring whether the generated texts reflect certain persona attributes.
- Human Evaluation: For a subset of outputs, human annotators will be used to rate the relevance, fluency, and personalization of responses with respect to the provided persona profiles.

6 Conclusion

This proposal advances a perspective-aware research in natural language processing by addressing three key components: annotation format design, perspective modeling by leveraging socio-demographic features, and personalized generation. First, it investigates how finer-grained annotation formats, such as Likert scales, better capture the nuances of human perspectives compared to traditional binary labels. Second, it examines the extent to which socio-demographic features influence annotation variation, particularly in relatively less explored domains of business and economics. Finally, it proposes methods for personalized generation that align output with user-specific attributes, including historical texts and socio-demographic features. Together, these tasks aim to enhance the inclusivity and fairness of NLP systems by recognizing and modeling the diversity of human perspectives.

Limitations

This proposal does not aim to comprehensively resolve all challenges associated with human annotation variation and annotator perspectives, particularly given its cross-domain nature. In addition, the availability of suitable datasets for certain tasks, especially those that include detailed annotator background information required for certain modeling and generation tasks, poses challenges to this research. To address this, the study may involve the construction of new datasets or the design of additional annotation tasks tailored to the specific research questions proposed.

Ethical Considerations

Research involving socio-demographic attributes and personal perspectives inherently carries ethical risks, particularly concerning the privacy and potential misuse of annotators' personal information. This study will take careful measures to protect the identities and privacy of all participants. All collected and analyzed data will be fully anonymized and handled in accordance with privacy-preserving protocols.

Special attention will be given to the ethical challenges of persona inference and demographic modeling. Minority and underrepresented viewpoints, which are essential to the study's objectives, will be treated with care and used solely for academic purposes to prevent any harm or stigmatization. Moreover, in the analysis and presentation of findings, efforts will be made to use neutral, respectful language and to avoid reinforcing harmful stereotypes or generalizations associated with specific demographic groups.

References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2020. Modeling annotator perspective and polarized opinions to improve hate speech detection. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, volume 8, pages 151–154.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, Alexandra Uma, and 1 others. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st workshop on benchmarking: past, present and future*, pages 15–21. Association for Computational Linguistics.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615, St. Julian's, Malta. Association for Computational Linguistics.

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

Daniel Braun and Florian Matthes. 2024. Agb-de: A corpus for the automated legal assessment of clauses in german consumer contracts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10389–10405.

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.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018. Capturing ambiguity in crowdsourcing frame disambiguation. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 6, pages 12–20.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, Massimo Poesio, and 1 others. 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*. 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.

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. 2023. EPIC: Multi-perspective annotation of a corpus of irony. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13844–13857, Toronto, Canada. Association for Computational Linguistics.

Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Nitesh Goyal, Ian D Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. Is your toxicity my toxicity? exploring the impact of rater identity on toxicity annotation. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–28.

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

Melody Guan, Varun Gulshan, Andrew Dai, and Geoffrey Hinton. 2018. Who said what: Modeling individual labelers improves classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

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

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

Pegah Jandaghi, XiangHai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. 2023. Faithful persona-based conversational dataset generation with large language models. *Preprint*, arXiv:2312.10007.

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

Hendrik Vincent Koops, W Bas De Haas, John Ashley Burgoyne, Jeroen Bransen, Anna Kent-Muller, and Anja Volk. 2019. Annotator subjectivity in harmony annotations of popular music. *Journal of New Music Research*, 48(3):232–252.

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

John P Lalor, Hao Wu, and Hong Yu. 2017. Soft label memorization-generalization for natural language inference. *arXiv preprint arXiv:1702.08563*.

Lung-Hao Lee, Jian-Hong Li, and Liang-Chih Yu. 2022. Chinese emobank: Building valence-arousal resources for dimensional sentiment analysis. *Transactions on Asian and Low-Resource Language Information Processing*, 21(4):1–18.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Zhenyi Lu, Wei Wei, Xiaoye Qu, Xian-Ling Mao, Dangyang Chen, and Jixiong Chen. 2023. Miracle: Towards personalized dialogue generation with latent-space multiple personal attribute control. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5933–5957, Singapore. Association for Computational Linguistics.

Piotr Milkowski, Marcin Gruza, Kamil Kanclerz, Przemyslaw Kazienko, Damian Grimling, and Jan Kocon. 2021. Personal bias in prediction of emotions elicited by textual opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 248–259, Online. Association for Computational Linguistics.

José Antonio Miñarro-Giménez, Catalina Martínez-Costa, Daniel Karlsson, Stefan Schulz, and Kirstine Rosenbeck Gøeg. 2018. Qualitative analysis of manual annotations of clinical text with snomed ct. *Plos one*, 13(12):e0209547.

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

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

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

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

Cecilia Ovesdotter Alm. 2011. Subjective natural language problems: Motivations, applications, characterizations, and implications. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 107–112, Portland, Oregon, USA. Association for Computational Linguistics.

Rebecca J Passonneau, Vikas Bhardwaj, Ansaf Salleb-Aouissi, and Nancy Ide. 2012. Multiplicity and word sense: evaluating and learning from multiply labeled word sense annotations. *Language Resources and Evaluation*, 46:219–252.

Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. 2019. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9617–9626.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Dennis Reidsma and Jean Carletta. 2008. Reliability measurement without limits. *Computational Linguistics*, 34(3):319–326.

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.

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

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 614–622.

Yihong Tang, Bo Wang, Dongming Zhao, Jinxiaojia Jinxiaojia, Zhangjijun Zhangjijun, Ruifang He, and Yuexian Hou. 2024. Morpheus: Modeling role from personalized dialogue history by exploring and utilizing latent space. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7664–7676.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2020. A case for soft loss functions. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 173–177.

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

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

Zeerak Waseem. 2016. Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Jin Xu, Mariët Theune, and Daniel Braun. 2024. Leveraging annotator disagreement for text classification. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 1–10, Trento. Association for Computational Linguistics.

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