Accurate and Data-Efficient Toxicity Prediction when Annotators Disagree

Harbani Jaggi* UC Berkeley
Kashyap Murali* UC Berkeley
Eve Fleisig UC Berkeley
Erdem Bıyık USC

Abstract

When annotators disagree, predicting the labels given by individual annotators can capture nuances overlooked by traditional label aggregation. We introduce three approaches to predict individual annotator ratings on the toxicity of text by incorporating individual annotator-specific information: a neural collaborative filtering (NCF) approach, an in-context learning (ICL) approach, and an intermediate embedding-based architecture. We also study the utility of demographic information for rating prediction. NCF showed limited utility; however, integrating annotator history, demographics, and survey information permits both the embedding-based architecture and ICL to substantially improve prediction accuracy, with the embedding-based architecture outperforming the other methods. We also find that, if demographics are predicted from survey information, using these imputed demographics as features performs comparably to using true demographic data. This suggests that demographics may not provide substantial information for modeling ratings beyond what is captured in survey responses. Our findings raise considerations about the relative utility of different types of annotator information and provide new approaches for modeling annotators in subjective NLP tasks.

1 Introduction

Disagreement among data annotators can reveal nuances in NLP tasks that lack a simple ground truth, such as hate speech detection. For instance, what one group of annotators deems acceptable might be considered offensive by another. The current standard for resolving such disagreement, aggregation via majority voting, casts aside variance in annotator labels as noise, when in subjective tasks this variance is key to understanding the perspectives that arise from the annotators' individuality and backgrounds.

To address this problem, recent research has explored alternatives to majority voting. Most notably, studies have taken the approach of predicting the ratings of individual annotators (Davani et al., 2022; Fleisig et al., 2023; Gordon et al., 2022). We aim to improve the prediction of rating behavior, guided by the following questions:

- **Does incorporating annotator information via collaborative filtering, embedding-based architecture, or in-context learning improve downstream rating predictions?**
- **What annotator information best informs toxicity rating predictions? Do demographics provide useful information beyond what survey information can provide?**

We proposed and tested a neural collaborative filtering (NCF) module, an embedding-based architecture, and an in-context learning (ICL) module for individual rating prediction. First, we incorporated NCF to the classification head of a RoBERTa-based model (Liu et al., 2019). Embedded annotator information[1] was combined with a separate embedding of annotators' rating history to predict individual annotator toxicity ratings. Secondly, we used embedding models to encode annotator information, then performed regression to predict toxicity ratings. Lastly, we prompted LLMs such as Mistral (Jiang et al., 2023) and GPT-3.5 (Brown et al., 2020) to study different ways of integrating annotator information.

Our findings indicate that while NCF does not outperform baseline models, ICL and our embedding-based architecture improve performance, with the embedding-based architecture significantly outperforming all other approaches tested. In addition, our research on the effectiveness of demographic information as a feature indicates that imputing demographics from survey data

[1] The annotator information used is a combination of demographic information, survey information, and annotator rating history.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 21910–21917
November 12-16, 2024 ©2024 Association for Computational Linguistics

performs similarly to using direct demographic inputs, suggesting that survey responses already capture the relevant demographic information for rating prediction. This suggests that, on this task, demographics have little predictive power beyond what survey information provides.

2 Motivation and Related Work

Our work is fundamentally motivated by the need for alternatives to majority-vote label aggregation in NLP tasks. Pavlick and Kwiatkowski (2019) find that disagreement among annotators is partially attributed to differences in human judgment. Basile et al. (2021) underscore the importance of the consideration of a system's output over instances where annotators disagree.

Newer work in this field aims to directly model individual annotator rating behavior. Davani et al. (2022) employ a multi-task based approach, where predicting each annotators' judgment is a subtask to their larger architecture. Fleisig et al. (2023) use a RoBERTa-based model to predict an individual annotators' ratings. Gordon et al. (2022) put together a jury of annotators, predicting individual judgments.

For the individual annotator rating prediction task, Deng et al. (2023) create individual annotator embeddings and annotation embeddings. This idea of learning embeddings based on user-specific data has been applied in various domains successfully, e.g., imitation learning (Beliaev et al., 2022) or recommendation systems (Biyik et al., 2023).

Collaborative filtering (CF) learns user embeddings based on their past behaviors (Bokde et al., 2015). He et al. (2017) show that neural collaborative filtering (NCF) offers better performance than more naive CF implementations. This motivates our NCF approach to learning annotator embeddings. Intuitively, this approach would be effective in learning deeply rooted preferences and behaviors of annotators. Thus, we hypothesized that this method would more accurately predict individual annotator ratings.

Several recent approaches use sociodemographic traits of individual annotators to learn for the rating prediction task (Fleisig et al., 2023; Davani et al., 2022), but Andrus et al. (2021) warn that legal and organizational constraints, such as privacy laws and concerns around self-reporting, often make collecting demographic data challenging. Gupta et al. (2018) suggest using semantically related features in the absence of sensitive demographic data. For instance, in the absence of gender information, (Zhao et al., 2019) use other demographic features – age, relation, and marital status – for their prediction task. This work motivates our objective of incorporating auxiliary annotator information (survey information and annotator history) in the prediction task.

Lastly, Orlikowski et al. (2023) challenge the utility of demographic information, since they do not find strong evidence that explicitly modeling demographics helps to predict annotation behavior. In concurrent work, Hu and Collier (2024) argue that there is an inherent limit to how much predictive power can be provided by demographics. Their findings indicate that while incorporating demographic variables can provide modest improvements in prediction accuracy, these gains are often constrained by the relatively low variance explained by these variables. This motivates our final objective, studying the efficacy of demographics as a useful mediating variable for rating prediction.

3 Approach

Our approach includes creating three separate modules based on neural collaborative filtering (NCF), an embedding-based architecture, and in-context learning (ICL). We evaluate each approach's efficacy in predicting annotator rating behavior. The latter two modules are used to investigate our second research question; we integrate different ablations of annotator information as input to the rating prediction models to study their effect on toxicity rating prediction.

We used Kumar et al. (2021)'s dataset to evaluate the performance of our rating prediction modules. This dataset consists of sentences rated for toxicity (0 = least toxic, 4 = most toxic). Each sentence has been labeled by 5 annotators and each annotator has labeled 20 distinct sentences. For each annotator, the dataset contains their rating behavior; demographic information (race, gender, importance of religion, LGBT status, education, parental status, and political stance); and survey information, e.g., their preferred forums, social media, whether they have seen toxic content, if they think toxic content is a problem, and their opinion on whether technology impacts peoples' lives.

For ablations, we took distinct combinations of annotator information (rating history, demographics, survey information) along with the text to be rated, assessing the impact of each on the model's performance. To study whether demographics are a necessary feature for predicting annotator ratings, we also used a separate model to predict annotator demographics using rating history and survey information and applied these predicted demographics as input for our ablations.

For all three methods, we used Mean Absolute Error (MAE) of predicting individual annotators' ratings as the evaluation metric, allowing us to quantify the performance of different model configurations.

3.1 Neural Collaborative Filtering

Our NCF method integrates textual and annotator-specific information to predict annotator ratings for the toxicity detection task (Figure 1). We aimed to create both a textual embedding and an annotator embedding for each (text, annotator) pair and capture latent interactions between both entities by using a hybrid neural architecture inspired by neural collaborative filtering. The goal was to learn more complex, non-linear relationships between annotator preferences and the text itself to more accurately predict an annotator's toxicity rating.

To create embedded representations of the textual information which has ranging levels of toxicity, we leveraged a RoBERTa model (Liu et al., 2019) fine-tuned on the Jigsaw Toxic Comment Classification Challenge dataset (cjadams et al., 2017) and the hate speech detection datasets introduced by Kumar et al. (2021). In parallel, we initialized and stored random embeddings for each annotator in the RoBERTa classification head. During training, these embeddings were concatenated with text embeddings and passed through 4 dense layers before predicting the rating.

In developing this hybrid model architecture, we explored variations in the dimensionality of the annotator embeddings, methods for integrating the sentence and annotator embeddings, and the impact of freezing the RoBERTa model (Appendix A describes variations tested).

3.2 Embedding-Based Architecture

We generated embeddings for the concatenated annotator information and the current text to be rated using two text embedding models, OpenAI's text-embedding-3-small and text-embedding-3-large. These embeddings then served as input for a custom regression model with multiple fully connected layers, which was trained to predict toxicity ratings based on the extracted features (Figure 2).

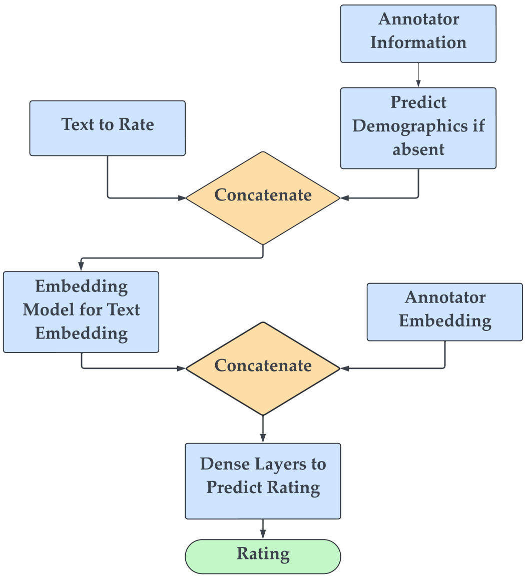
Annotator Information
Text to Rate
Predict Demographics if absent
Concatenate
Embedding Model for Text Embedding
Annotator Embedding
Concatenate
Dense Layers to Predict Rating
Rating

Figure 1: Design of our neural collaborative filtering (NCF) architecture. Annotator information and the text being rated were passed into an embedding model, then concatenated with the annotator embedding, and passed through a series of dense layers to predict the rating.

3.3 In-Context Learning

Our in-context learning architecture prompts a language model to process a range of combinations of annotator information. Each combination serves as input to the model (Mistral or GPT-3.5), enabling it to account for the specific context of the annotator when predicting toxicity ratings. The model was prompted to generate predictions based on the contextual information provided. This approach aims to enhance the model's ability to make informed predictions by integrating diverse sources of information relevant to the rating task. A sample prompt of this approach is shown in Figure 3.

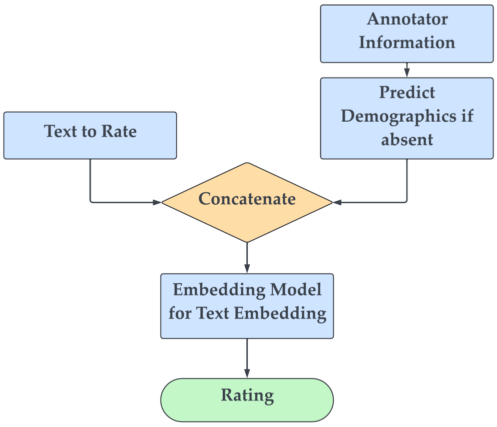
Annotator Information
Predict Demographics if absent
Text to Rate
Concatenate
Embedding Model for Text Embedding
Rating

Figure 2: Design of our embedding-based architecture.

4 Results

Our three approaches predicted annotators' toxicity ratings on a scale from 0 to 4, based on both textual data and various combinations of annotator-specific information (demographics, survey responses, rating history). We also examine how well these models handle predicted demographic data rather than using the ground truth demographic values for each annotator. This helps to assess the data efficiency and effect of demographics as an input to the rating prediction task.

For our ablations that studied the improvement on rating predictions, we compared our results to previous baselines that predicted ratings of annotators using the same dataset.

Q1: Does incorporating annotator information via collaborative filtering, the embedding-based architecture, or in-context learning improve downstream rating predictions?

Our embedding-based architecture outperformed all other experiments with an MAE of 0.61; the best ICL approach (with Mistral) reached an MAE of 0.69. Both the ICL approach and embedding-based architecture outperform the most recent baseline for the dataset (Fleisig et al., 2023) and the embedding-based architecture matches the best previous MAE on this dataset (Gordon et al., 2022). The best-performing models use all available annotator-specific information as input (annotator demographics, survey information, and historical rating data). At its best, our ICL configuration with Mistral had an MAE of 0.69 (using annotator demographics, survey information, and historical rating data). The NCF approach had consistently poorer results, with a best MAE of 0.79 when including all annotator-specific information.

When creating the NCF architecture, we tested several variations. We first created a baseline from which we compared different outputs of our NCF module. Evaluating the finetuned RoBERTa model with all annotator-specific information as input along with the text to be rated yielded a baseline MAE of 0.81. We experimented with integrating embeddings through dot product vs. concatenation, freezing RoBERTa during the training process, and placing the collaborative filtering task in different parts of the RoBERTa architecture. Our best performing model froze the pretrained RoBERTa model, used concatenation, and placed the collaborative filtering piece in the classification head. However, it was only able to achieve an MAE of 0.80, not significantly improving on our baseline.

Our embedding-based architecture consistently outperformed other approaches on every ablation, suggesting that a feature-extraction and regression hybrid approach most effectively uses annotator-specific information in rating predictions.

Q2: What annotator information best informs toxicity rating predictions? Do demographics provide useful information beyond what survey information can provide?

Incorporating demographic information improves performance over using only survey information, rating history, or both across ablations. However, we find that much of this gap can be compensated for by distilling demographic information out of survey information. Compared to the text-only baseline, incorporating predicted demographics with survey information and annotator history achieved MAE reductions of 10.26% with Mistral, 8.64% with GPT-3.5, 11.84% with text-embedding-3-small, and 12% with text-embedding-3-large. Replacing true demographic information with predicted demographic information results in nearly as strong performance for Mistral, GPT-3.5, and text-embedding-3-small.

Incorporating predicted demographics alongside survey information and annotator history notably improves accuracy. This occurs despite the fact that the accuracy of predicted demographics varies widely (highest for race and gender, but near-random for some demographics; see Table 4). Although the true demographics are somewhat helpful, annotator ratings can be effectively predicted without direct demographic data. This finding suggests that detailed demographic data may not be especially useful as a feature in individual rating prediction, beyond what can be inferred from individual preferences in survey responses.

Predicting Demographics. The performance of predicting demographics was evaluated across various configurations (Table 4). The baseline approach incorporating only survey information achieved the highest accuracies, with 47% for race and 63% for gender. Combining survey information with text slightly reduced the performance, potentially indicating the noise that the text to be rated added. The majority class approach is indicated as a baseline comparison to highlight the performance improvements for the different categories.

Our findings indicate that successively incorporating annotator demographics, rating history,

Model	Mistral	GPT 3.5	text-embedding-3-small	text-embedding-3-large
Text only	0.78	0.81	0.76	0.75
+ demo.	0.76	0.79	0.73	0.71
+ demo. + history	0.75	0.78	0.73	0.69
+ history	0.73	0.75	0.70	0.66
+ survey	0.73	0.75	0.70	0.70
+ demo. + survey	0.71	0.73	0.68	0.64
+ history + survey	0.70	0.73	0.67	0.69
+ predicted demo. + history + survey	0.70	0.74	0.67	0.66
+ demo. + history + survey	**0.69**	**0.72**	**0.66**	**0.61**

Table 1: Comparison of mean absolute error across different model configurations for the test set (with or without annotator demographics, rating history, and survey responses). Both ICL and embedding-based architectures improve on the baseline, with embedding-based architectures performing best.

and survey information improves performance for nearly all configurations tested (Table 1). Overall, the comprehensive model incorporating demographics, annotator history, and survey data consistently outperformed other configurations, demonstrating the value of integrating multiple data sources for demographic and rating predictions.

5 Conclusion

Leveraging the embedding-based architecture and ICL methods substantially improved toxicity rating predictions. NCF, by contrast, was not a competitive method for predicting ratings. Incorporating annotator information significantly enhances model performance. The best-performing embedding-based architecture achieved the lowest MAE of 0.61 by integrating demographics, annotator history, and survey data. This suggests that personalized predictions based on individual annotator preferences can lead to more accurate outcomes. Meanwhile, the ability to predict some demographics from survey information, and the fact that these imputed demographics nearly match performance with the true demographics, suggest that although demographics are helpful, individual annotator ratings can be predicted effectively without demographic data. This finding suggests that some differences in annotator opinions may be best captured by modeling individual preferences rather than demographic trends. In addition, the effectiveness of our embedding-based architecture suggests that it could help to inform future frameworks for annotator rating prediction.

6 Limitations

While our study advances the accuracy of annotator rating predictions, several limitations exist. The generalizability of our findings is limited to English text from the U.S. and Canada, which hinders applicability in other linguistic and cultural contexts. The integration of detailed annotator information poses ethical and privacy risks and can amplify existing biases in the data. Additionally, the complexity and computational demands of our models challenge scalability and interpretability. Future research should address these issues to enhance the robustness and fairness of predictive models in subjective NLP tasks. It should also focus on expanding these methods to other domains and exploring the ethical implications of incorporating inferred data for predictions. By continuing to refine these approaches, we can develop more accurate and reliable models that better capture the complexities of human behavior and preferences.

7 Ethical Considerations

We found that individual ratings can be predicted well without demographic information. This is helpful in that it permits individualized rating prediction without collecting demographic information. Unfortunately, that does not mean the ratings are predicted *independent* of demographic information: in fact, we also found that survey information is a close enough proxy that demographics can be predicted with substantially better than random accuracy, especially for race and gender, off of survey information responses. Incorporating these predicted demographics further improves accuracy. However, our finding thus uncovered the potential privacy issue that collecting seemingly innocuous survey information data carries the risk of revealing annotator demographics. This suggests that future research in this area must proceed with caution: collecting or inferring demographic information improves prediction accuracy, but risks tokenism (where opinions within a demographic group are assumed to be homogeneous). Instead, future research could identify survey information questions that help to improve rating prediction but do *not* risk revealing annotator demographics.

References

McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. 2021. What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 249–260.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, Alexandra Uma, et al. 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.

Mark Beliaev, Andy Shih, Stefano Ermon, Dorsa Sadigh, and Ramtin Pedarsani. 2022. Imitation learning by estimating expertise of demonstrators. In *International Conference on Machine Learning*, pages 1732–1748. PMLR.

Erdem Biyik, Fan Yao, Yinlam Chow, Alex Haig, Chih-wei Hsu, Mohammad Ghavamzadeh, and Craig Boutilier. 2023. Preference elicitation with soft attributes in interactive recommendation. *arXiv preprint arXiv:2311.02085*.

Dheeraj Bokde, Sheetal Girase, and Debajyoti Mukhopadhyay. 2015. Matrix factorization model in collaborative filtering algorithms: A survey. *Procedia Computer Science*, 49:136–146.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

cjadams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, nithum, and Will Cukierski. 2017. Toxic comment classification challenge.

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.

Naihao Deng, Xinliang Frederick Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You are what you annotate: Towards better models through annotator representations. *arXiv preprint arXiv:2305.14663*.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726.

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.

Maya Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Wang. 2018. Proxy fairness.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pages 173–182.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in llm simulations. *arXiv preprint arXiv:2402.10811*.

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.

Deepak Kumar, Patrick Gage, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing toxic content classification for a diversity of perspectives. In *SOUPS*. Usenix.

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*.

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.

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

Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J Gordon. 2019. Conditional learning of fair representations. *arXiv preprint arXiv:1910.07162*.

System Prompt

You are a model that predicts the toxicity rating of text from 0 to 4, where 0 is the least toxic and 4 is the most toxic.

User Prompt

The annotator has annotated these texts: "This is a harmless comment" is rated as 0, "You're an idiot" is rated as 3, "I respectfully disagree" is rated as 1 [SEP] The reader uses social media, news sites, and video sites. The reader has seen toxic comments, has been personally targeted by toxic comments, thinks technology has a positive impact on people's lives, and thinks toxic comments are a serious problem. [SEP] The reader is a 25-34 year old Asian female who has a Bachelor's degree, is politically liberal, is not a parent, and thinks religion is not important. [SEP] Annotate this text: "Why don't you go jump off a cliff?"

Figure 3: Sample prompt for toxicity prediction model. The system prompt (in teal) defines the model's role. The user prompt (in olive) provides historical annotations, survey results, demographic information, and the text to be rated.

A Appendix

Approaches Taken

1. Tried to cluster the annotator embeddings (PCA) – they weren't linearly separable based on demographics
2. Where to incorporate recommender systems
   (a) Classification head start – features
   (b) Later layer
   (c) before appending to 'features'
3. Tried to train plan RoBERTa on the entire dataset using the pretrained_multitask_demographic dataset
4. Different dimensions of annotator embeddings
   (a) Tried dim 8: little to no predictive power for annotator demographics
   (b) Changed to 512
   (c) Now using dim 768
5. Dual RoBERTa
   (a) Instead of randomly instantiating an embedding layer, we tried using RoBERTa to represent annotators based on their IDs.

Text Structure

For these predictions, the input is formatted as $h_1 \dots h_n$ `[SEP]` $s_1 \dots s_n$ `[SEP]` $d_1 \dots d_n$ `[SEP]` $w_1 \dots w_n$, where $h_1 \dots h_n$ represents the other texts reviewed and their ratings as provided by the annotator, $s_1 \dots s_n$ is a template string describing the annotator's survey information data, $d_1 \dots d_n$ is a template string containing the annotator's demographic information (e.g., "The reader is a 55-64 year old white female who has a bachelor's degree, is politically independent, is a parent, and thinks religion is very important. The reader is straight and cisgender"), $w_1 \dots w_n$ is the text being rated, and `[SEP]` is a separator token. We use a template string instead of categorical variables in order to best take advantage of the model's language pretraining objective (e.g., underlying associations about the experiences of different demographic groups).

Dataset Size

The dataset we used to evaluate the performance of our approaches – (Kumar et al., 2021) – has 3 splits: train, dev, and test. The training set has 488,100 samples, the dev set has 25,000 samples, and the test set also has 25,000 samples.

Model Information

For the collaborative filtering approach, we used a RoBERTa model that has 355 million trainable parameters, and it took 2 GPU hours per epoch when fine-tuned on 2 NVIDIA Quadro RTX 8000 GPUs.

For the ICL approach, we used an API version of OpenAI's text-embedding-3-large model, which we don't have access to, so as to determine its size, and infrastructure requirements.

Experimental Setup

We observed the best performance when having 4 dense layers after the embedding was outputted, which transformed the embedding from 3072 dimensions to 1024 dimensions, then keeps it at 1024 dimensions for another 2 layers after which the last layer is then shrunk to 5 dimensions.

Demographics Prediction Task Figure 1: Neural Collaborative Filtering

Model	Mistral	GPT 3.5	text-embedding-3-small	text-embedding-3-large
Text only	0.74	0.77	0.73	0.72
+ D	0.73	0.76	0.71	0.68
+ D + H	0.71	0.74	0.69	0.66
+ H	0.70	0.72	0.67	0.63
+ S	0.69	0.71	0.66	0.67
+ D + S	0.67	0.69	0.64	0.61
+ H + S	-	-	-	0.65
+ PD + H + S	-	-	-	0.62
+ D + H + S	**0.65**	**0.68**	**0.62**	**0.58**

Table 2: Comparison of mean absolute error across different model configurations (dev set results). Ablations that included both annotator history and survey information were only performed on the best-performing model. D refers to Annotator Demographics, H refers to other texts an annotator has rated, S refers to survey responses, PD refers to predicted demographics.

Experiment Description	Individual MAE
Initial training with Collaborative Filtering approach and RoBERTa	1.12
Adjusted annotation embedding dimensions from 8 to 512	0.89
Freezing RoBERTa after pre-training on (Kumar et al., 2021)	0.80

Table 3: Significant Experiments and Their Impact on Mean Absolute Error (MAE)

Generated Data	Race	Gender	Importance of Religion	LGBT Status	Education	Political Stance
Survey Info	47%	63%	37%	38%	57%	48%
Survey Info + Text	43%	60%	33%	34%	52%	44%
Majority Class	9%	52%	31%	81%	52%	40%

Table 4: Comparison of demographic prediction accuracy across different data configurations.