Everyone's Voice Matters: Quantifying Annotation Disagreement Using Demographic Information

Ruyuan Wan[1*], **Jaehyung Kim**[2*], **Dongyeop Kang**[3]
[1]University of Notre Dame, [2]KAIST, [3]University of Minnesota
rwan@nd.edu, jaehyungkim@kaist.ac.kr, dongyeop@umn.edu

Abstract

In NLP annotation, it is common to have multiple annotators label the text and then obtain the ground truth labels based on the agreement of major annotators. However, annotators are individuals with different backgrounds, and minors' opinions should not be simply ignored. As annotation tasks become subjective and topics are controversial in modern NLP tasks, we need NLP systems that can represent people's diverse voices on subjective matters and predict the level of diversity. This paper examines whether the text of the task and annotators' demographic background information can be used to estimate the level of disagreement among annotators. Particularly, we extract disagreement labels from the annotators' voting histories in the five subjective datasets, and then fine-tune language models to predict annotators' disagreement. Our results show that knowing annotators' demographic information, like gender, ethnicity, and education level, helps predict disagreements. In order to distinguish the disagreement from the inherent controversy from text content and the disagreement in the annotators' different perspectives, we simulate everyone's voices with different combinations of annotators' artificial demographics and examine its variance of the fine-tuned disagreement predictor. Our paper aims to improve the annotation process for more efficient and inclusive NLP systems through a novel disagreement prediction mechanism. Our code and dataset are publicly available.

Introduction

Supervised AI systems are trained on annotated datasets with labels determined by consensus among multiple annotators. The subjective opinions of different annotators often bring annotation disagreement in the decision of the final labels. Most commonly, this disagreement is addressed by ignoring highly-disagreed cases and only including those whose opinions were voted on by the majority as the final label. When the labeling tasks become more subjective and require the annotator's own interpretation and judgment, such as detecting offensiveness and judging social dilemmas (Reidsma and op den Akker 2008), this majority-based aggregation often fails to learn the true distribution of annotators' voices. The increasing subjectivity of NLP problems in modern NLP will cause the annotators' disagreement not only due to the potential random error in the process but also because annotators may interpret the text with different views and make judgments based on their own connotations.

[*]This work was done while RW and JK were at the Minnesota NLP lab.

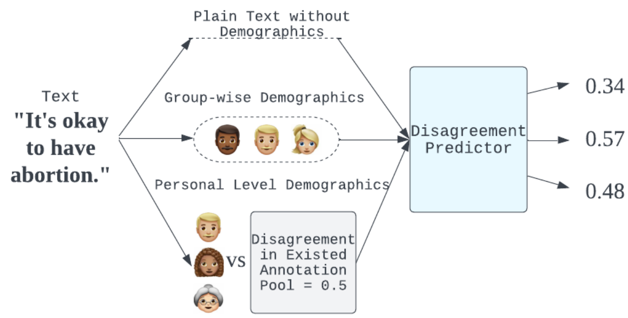


Figure 1: Disagreement prediction predicts disagreement only from input text or input text with (group-wise or individual against the majority) annotators' demographic information.

Different demographics, cultural backgrounds, and living experiences influence how people receive and interpret information. This difference is more visible in subjective tasks. For example, Sap et al. found more consecutive annotators who have higher scores on racist beliefs are more likely to label African American English as toxic rather than label anti-Black language as toxic (Sap et al. 2022). In these cases, the aggregated singular labels can bring bias by using less inclusive and *societal-representative labels* to accommodate everyone's voices in subjective studies.

This paper assumes that annotators' disagreement potentially comes from the *limited representations of the annotator group assigned* or *controversy of the text in nature*. This study focuses on exploring the relationship between the annotator group and natural controversy in text by developing a **disagreement predictor** with and without the information about the annotator group, as depicted in Figure 1. In particular, we analyze annotators' disagreement from five subjective task datasets to answer the following research questions:

1. Is it possible to predict the level of annotators' disagreement with text using language models? Does knowing annotators' identities, like demographic information in

addition to text, help predict annotation disagreement?

2. Is the disagreement caused by the natural controversy of the text or by the biased distribution of the assigned annotators?

Our research demonstrates that disagreement is predictable in subjective annotation tasks by using Roberta model (Liu et al. 2019) to predict both hard disagreement (binary label) and soft disagreement (continuous label) [1]. We further design two demographic augmentation experiments and find that bringing the annotator-level demographic information can significantly improve disagreement prediction performance. Finally, based on the findings, we simulated artificial annotators' backgrounds to predict disagreement to check whether the disagreement score will be changed in a wider annotator population. In short, we propose a disagreement measurement that can efficiently suggest the optimal number of annotators and assign an appropriate demographic group of annotators per text, possibly helping improve the fairness and quality of subjective annotation.

Related Works

Tasks like toxicity detection (Sap et al. 2020; Yu 2022), sentiment analysis (Potts et al. 2021), and social, ethical labeling (Forbes et al. 2020; Hendrycks et al. 2021) are highly subjective and controversial. One can think one post is offensive, but others may consider it acceptable. There is no one objective ground truth. Röttger et al. summarized three key challenges of descriptive annotation in subjective NLP tasks: interpretation of disagreement, label aggregation, and representativeness of annotators.

Due to the absence of absolute ground truth, the interpretation of disagreement becomes complicated (Alm 2011). For instance, the disagreement may result in considerably different reliabilities: whether the annotators disagree on the most critical or least crucial instances (Foley 2018). In addition, researchers commonly use some notion of the agreement to measure the task's subjectiveness, such as using inter-annotator agreement metrics Cohen's Kappa (Cohen 1960) or Fleiss' Kappa (Fleiss 1971) to measure annotations' reliability. But when presenting the final results in the downstream task, people usually use the aggregated labels that can conceal informative disagreement and evaluation metrics that are unaware of the task's subjectiveness (Röttger et al. 2022).

Rather than 'correct' or 'wrong,' Alm pointed out the concept of acceptability. There might be multiple acceptable answers in subjective tasks. However, aggregating labels through major voting will increase the risk of discarding minority voices. To address the problem, Davani, Díaz, and Prabhakaran treats predicting each annotator's judgments as separate subtasks, which achieved the same or better performance than aggregating labels in the data before training. They also further evaluate the model uncertainty using the variance of the predicted annotation label. However, this still concerns that recognizing the aggregated major votes as the final targets does not always represent all acceptable answers. On the other hand, Uma, Almanea, and Poesio used posterior calibration with a soft-loss approach to learning from data containing disagreement. They noticed that temperature scaling only functions with data where disagreements are caused by label overlap and not with data where disagreements are caused by annotator subjective judgment or language ambiguity. This aligns with Foley 's finding for tasks with subjective labels: without collecting additional labels, models reach the ceiling of performance given the small dataset size and the inherent disagreement between annotators on which documents are controversial.

In previous research, annotators' demographics have shown importance in improving the annotation quality in subjective tasks. For example, Prabhakaran, Davani, and D'iaz demonstrated that the agreement scores could be very significant among different socio-demographic groups annotators identified when certain individual annotators disagree with the majority labels. Further, Gordon et al. proposed jury learning, a recommender system approach defining which people or groups, in what proportion, determine the classifier's prediction. For instance, a jury learning model would recommend women and black jurors for online hate speech detection, who are mainly targeted in online harassment. However, many public datasets didn't collect annotators' demographics with their annotations. Further, the datasets that reported the annotator's demographics also have imbalanced representative concerns. For example, the race is often skewed, and dominant with the white race (Sap et al. 2020; Forbes et al. 2020; Hendrycks et al. 2021; Sap et al. 2022).

As shown above, researchers have implicitly resolved the label disagreement using majority votes, annotator selection, and the soft-loss approach(Uma et al. 2021). Different from the interrater disagreement resolution, which defines disagreement as a sign of poor quality or mistakes to be resolved(Oortwijn, Ossenkoppele, and Betti 2021), our research explicitly quantifies disagreement as our task target and further distinguishes the nuance among various socio-demographic groups.

Methods

This section presents our method for quantifying subjective annotation disagreements. Our main idea is modeling the annotation disagreement using demographic information of each annotator as additional inputs, with the pre-trained language model, *e.g.,* RoBERTa (Liu et al. 2019). In Section , we first introduce the mathematical notations. Then, we elaborate on the details of the proposed method in Section . Finally, we provide a way to simulate the annotators' demographic information in Section .

Preliminaries

We first describe the problem setup of our interest under a text classification scenario with K classes. Let $\mathcal{D} = (\mathcal{X}, \mathcal{Y})$ denote the given annotated dataset where $\mathcal{X}$ is a set of texts and $\mathcal{Y}$ is the annotation matrix of $\mathcal{X}$. Specifically, each entry of $\mathcal{Y}$, $y_i(\mathbf{x}) \in \{1, \dots, K\}$, represents ith an-

[1]Our code and dataset are publicly available. https://github.com/minnesotanlp/Quantifying-Annotation-Disagreement

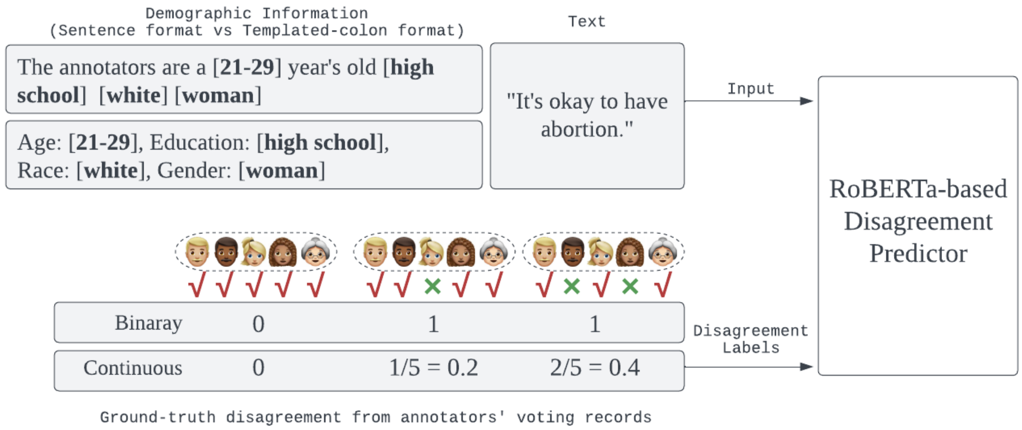


Figure 2: Our proposed disagreement predictor that takes the task sentence and/or (group or person) demographic information as input and ground-truth disagreement among annotators as labels. The demographic information is concatenated to the task sentence either in sentence format or templated-colon format. The ground-truth labels are aggregated from the annotators' voting records as binary labels with a threshold (i.e., 3/5) or continuous labels as they are.

notation assigned to text $\mathbf{x} \in \mathcal{X}$.[2] We assume that there are N different annotations for each text $\mathbf{x}$ and $\mathbf{y}(\mathbf{x}) = [y_1(\mathbf{x}), \ldots, y_N(\mathbf{x})]$ denotes all annotations assigned to $\mathbf{x}$. Then, $r_k(\mathbf{x}) = \sum_{i=1}^{N} 1[y_i(\mathbf{x}) = k]/N$ denotes the agreement rate of $\mathbf{x}$ to the label k where $\sum_{k=1}^{K} r_k(\mathbf{x}) = 1$. In addition, we assume that T different demographic information of all N annotators is available such as gender, age and race[3], and denote it as $\mathbf{d}^{(\mathtt{t})}(\mathbf{x}) = [d_1^{(\mathtt{t})}(\mathbf{x}), \ldots, d_N^{(\mathtt{t})}(\mathbf{x})]$ where $t = 1, \ldots, T$. Remarkably, majority voting, which is a popular common practice of assigning the label from the multiple annotations $\mathbf{y}(\mathbf{x})$ to the maximally agreed label, can be represented as $y_{\mathtt{maj}}(\mathbf{x}) := \arg\max_k r_k(\mathbf{x})$.

Binary vs Continuous disagreement labels. From the agreement rate $r_k(\mathbf{x})$, we first compute a binary disagreement label $\bar{r}_{\mathtt{b}}(\mathbf{x}) = 1[r_{y_{\mathtt{maj}}}(\mathbf{x}) \neq 1]$, which indicates if there are different opinions among the annotators for this instance $\mathbf{x}$. We further define a continuous disagreement label $\bar{r}_{\mathtt{c}}(\mathbf{x}) = 1 - r_{y_{\mathtt{maj}}}(\mathbf{x})$, that has the scale of 0 (everyone agrees with the same annotation result) to 1 (a significant number of people holding different opinions on the annotation results). Namely, the binary label $\bar{r}_{\mathtt{b}}$ indicates the existence of at least some different opinions, and the continuous label $\bar{r}_{\mathtt{c}}$ measures the degree of disagreement among the annotators. Without loss of generality, we refer both types of disagreement as $\bar{r}$. The text with highest disagreement means annotators hold different opinions, and this text content is very controversial.

Disagreement Prediction with Demographic Information

Our goal is to predict the disagreement $\bar{r}(\mathbf{x})$ of given text $\mathbf{x}$ because it provides an effective way to understand which content is controversial or not. To this end, our first idea is to utilize the pre-trained language model, e.g., RoBERTa (Liu et al. 2019), for training a predictor f_θ of the disagreement of given text. Specifically, we train the model by minimizing a mean square error (MSE) loss as follow:

$$\underset{\theta}{\text{minimize}}\, \mathcal{L}_{\mathtt{MSE}}(f_\theta(\mathbf{x}), \bar{r}(\mathbf{x})) \tag{1}$$

However, the annotators' disagreement is not only from the controversy of the text in nature but also from the limited representations of the assigned annotator group. Hence, more than just using text as input is needed to capture the disagreement fully.

Incorporation of demographics: Group vs Personal. To this end, our key idea is incorporating the demographic information of annotators $\{\mathbf{d}^{(\mathtt{t})}(\mathbf{x})\}_{t=1}^{T}$ to train the model f_θ. Intuitively, it is expected to encode the valuable information of the disagreement of the text $\mathbf{x}$, especially related to limited representations of the annotator group assigned. To be specific, we propose two different ways to incorporate the demographic information: (1) *Text with group demographic information* and (2) *Text with personal demographic information*.

Text with group demographic information $\widetilde{\mathbf{x}}_{\mathtt{group}}$ is constructed by listing all N annotators' information $\mathbf{d}^{(\mathtt{t})}(\mathbf{x})$ in one string and then concatenating with the targeted text $\mathbf{x}$:

$$\widetilde{\mathbf{x}}_{\mathtt{group}} = \text{Combine}[\mathbf{d}^{(1)}(\mathbf{x}), \ldots, \mathbf{d}^{(T)}(\mathbf{x}), \mathbf{x}] \tag{2}$$

Therefore, the group demographics supplemented text also has the same number of instances as the original dataset.

On the other hand, text with personal demographic information $\widetilde{\mathbf{x}}_{\mathtt{person}}$ is constructed by concatenating only one annotator's demographic with text:

$$\widetilde{\mathbf{x}}_{\mathtt{person}} = \text{Combine}[d_j^{(1)}(\mathbf{x}), \ldots, d_j^{(T)}(\mathbf{x}), \mathbf{x}] \tag{3}$$

[2]We clarify that the ith annotation could be labeled by different annotators between different texts in $\mathcal{X}$.

[3]This assumption will be relaxed in Section

Datasets	Text	Annotation Distribution	Disagreement Label
SBIC	"Abortion destruction of the nuclear family contraceptives feminism convincing women to wait for children damaging economy so youth cannot leave the nest ramping up tensions between sexes all serves one primary goal to lower the population."	A1 (age: 32, politics: liberal, race: white, gender: woman) votes for inoffensive A2 (age: 34, politics: liberal, race: white, gender: woman) votes for inoffensive A3 (age: 29, politics: mod-liberal, race: hispanic, gender: woman) votes for offensive ⟶ Aggregated Label: **inoffensive**	Binary: 1 Continuous: 1/3
SChem101	"It's okay to have abortion."	A1 (age: 30-39, education: high school, race: white, gender: woman) votes for people ocassional think this A2 (age: 40-49, education: grad, race: white, gender: man) votes for controversial A3 (age: 30-39, education: bachelor, race: white, gender: man) votes for common belief A4 (age: 21-29, education: high school, race: white, gender: woman) votes for controversial A5 (age: 30-39 , education: bachelor, race: hispanic, gender: woman) votes for controversial ⟶ Aggregated Label: **controversial**	Binary: 1 Continuous: 2/5
Dilemmas	1st action: "refusing to do a survey on the credit card reader while paying with cash at the Office Max." 2nd action: "saying my bf has no right to dictate who I tell about my abortion."	1 annotator votes for the first action is less ethical while 4 others vote the second action is less ethical ⟶ Aggregated Label: **2nd action is less ethical**	Binary: 1 Continuous: 1/5
Dynasent	"Had to remind him to toast the sandwich."	4 annotators believe it's negative while one think it is neutral ⟶ Aggregated Label: **negative**	Binary: 1 Continuous: 1/5
Politeness	"Where did you learn English? How come you're taking on a third language?"	5 annotators politeness scores are 5, 13, 9, 11, 11 with the maximum of 25. ⟶ Aggregated Label: **impolite**	Binary: 0 Continuous: 0

Table 1: Examples from the five disagreement datasets used in this paper. A stands for annotator.

where $j = 1, \dots, N$ and hence it results in N times larger dataset with N different annotators.

Format: Templated vs Sentence. For combining the demographic information and text, we further propose two different ways with specific templates: (1) *Templated format* and (2) *Sentence format*. Templated format represents the category and value of each demographic information in a separate sentence, then concatenate all of them with the given text. For example, if one annotator is 36 years white woman, this demographic information is converted to *"Age: 36, Color: white, Gender: women"*, then concatenated with the original sentence in case of the text with person demographic. On the other hand, sentence format represents the demographic information with a natural sentence, *e.g., the annotator is a 36 years old white woman.*, then concatenate it with the original sentence.

With these demographics supplemented text $\widetilde{\mathbf{x}}$ ($\widetilde{\mathbf{x}}_{\texttt{group}}$ or $\widetilde{\mathbf{x}}_{\texttt{person}}$), we train our model similar to the case with the original sentence $\mathbf{x}$ in Equation (1):

$$\underset{\theta}{\text{minimize}}\, \mathcal{L}_{\texttt{MSE}}(f_\theta(\widetilde{\mathbf{x}}), \bar{r}(\mathbf{x})) \qquad (4)$$

An illustration of the proposed demographic-based disagreement predictor is presented in Figure 2.

Simulation of Demographic Information

In addition, we propose a simulation of demographic information, which is a novel approach to analyze *how the different annotator groups impact disagreement prediction*. It is expected to separately reveal the inherent disagreement of annotators from the controversy of the text in nature. Specifically, instead of ground-truth $\{\mathbf{d}^{(t)}(\mathbf{x})\}_{t=1}^{T}$, we combine the artificial demographic information $\{\bar{\mathbf{d}}^{(t)}(\mathbf{x})\}_{t=1}^{T}$ with the given text $\mathbf{x}$ and annotations $\mathbf{y}(\mathbf{x})$, to simulate the scenario with different annotators. Such as, the gender demographic type has *four* possible options: woman, man, transgender, non-binary; and ethnicity with *seven* options: white, black or African American, American Indian or Alaska Native, Asian, Native Hawaiian or other pacific islanders, Hispanic, or some other race. Overall, we have a total $28 = 4 \times 7$ different combinations of the annotator's demographic information for the simulation, while the ground-truth demographic information is one of them; hence, it offers an opportunity to explore the more extensive range of demographic information with the increased number of instances. Then, we obtain a predicted disagreement using f_θ, which is trained with $\mathbf{x}$ and $\{\mathbf{d}^{(t)}(\mathbf{x})\}_{t=1}^{T}$ as introduced in Section .

Then, we evaluate whether the predicted disagreement is easily or hard to be changed among the simulated demographic profiles so that we can distinguish whether the disagreement comes from the controversy of text or uncertainty from annotators for the disagreement label. For example, if the variation of predicted disagreement among the simulated combinations is high and the average change of the predicted disagreement between the simulated combinations and real disagreement is large, it might reveal that disagreement is highly related to the uncertainty of annotators. In contrast, the lower variation and smaller change between real disagreements indicate the disagreement is based on the controversy in the text, which is stable disagreement among various kinds of people.

Experiments

Benchmark Datasets

To obtain the annotators' disagreement, we choose the following five datasets of subjective tasks that include annotators' voting records in the raw format.[4]

Social Bias Inference Corpus (SBIC) (Sap et al. 2020) contains 150k structured annotations of social media posts. Each post has three different annotators. Annotators indicated whether the post could be considered "offensive to anyone." The offensiveness is a categorical variable with three possible answers (yes, maybe, no).

Social Chemistry 101 (SChem101) (Forbes et al. 2020) is a corpus of cultural norms via free-text rules-of-thumb created by crowd workers. A rule-of-thumb is a judgment of action which is further broken down into 12 theoretically-motivated dimensions of people's judgments. Our study focuses on the anticipated agreement category. It reflects workers' opinion on what portion of people probably agree with the judgment given the action. The category has five possible answers: almost no one believes, people occasionally think this, controversial, common belief, universally true. Each rule of thumb is annotated by five workers.

Scruples-dilemmas (Lourie, Bras, and Choi 2021) is a resource for normative ranking actions. Each instance pairs two unrelated actions and identifies which action crowd workers found less ethical. Each instance is annotated by five different annotators.

Dyna-Sentiment (Potts et al. 2021) is an English language benchmark task for ternary sentiment analysis. Each Yelp review is validated by five crowd workers into three possible sentiment results: positive, negative, and neutral.

Wikipedia Politeness (Danescu-Niculescu-Mizil et al. 2013) is a collection of requests from Wikipedia Talk pages, annotated with politeness. Each Wikipedia request is annotated by five annotators on a 1 to 25 scale. As Danescu-Niculescu-Mizil et al. ignored neutral cases for politeness prediction, we extracted the disagreement between the binary classes of request, i.e., polite and impolite.

Disagreement Label Distributions The Figure 3 shows the distributions of disagreement scores among five datasets. For dynasent dataset, since the majority of the dataset has

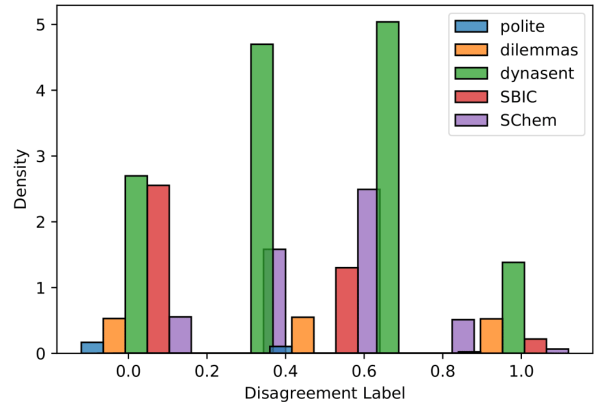


Figure 3: Disagreement distributions for five datasets

	Binary Label		Continuous Label	
Datasets	F1 (↑)	MSE (↓)	F1 (↑)	MSE (↓)
SBIC	61.5	0.309	**66.0**	**0.086**
SChem101	0.0	0.905	**52.3**	**0.056**
Dilemmas	0.0	0.330	**34.2**	**0.165**
DynaSent	**74.9**	0.361	11.8	**0.114**
Politeness	55.9	0.490	**56.8**	**0.110**

Table 2: Evaluation results of vanilla (RoBERTa) classifiers only with text input on the five datasets with disagreement.

disagreement between 0.3 to 0.6. The prediction concentrate around 0.4 to 0.5. The comparison among multiple datasets reflects that the subject topics influence the crowd annotators' disagreement. For example, most texts regarding offensiveness had consensus opinions from the annotators, while most annotators disagreed regarding sentiment.

Experimental Details

All the experiments are conducted by fine-tuning RoBERTa-base (Liu et al. 2019) using Adam optimizer (Kingma and Ba 2015) with a fixed learning rate 1e-5 and the default hyperparameters of Adam. For the text classification tasks, the model is fine-tuned with batch size 8 for 15 epochs.

To the best of our knowledge, we couldn't find any existing disagreement predictors to be used as baselines. As a result, we compare our predictors with different input types and disagreement labeling setups. Different versions of pretrained language models were tested, but RoBERTa always performed better. For the evaluation of the performance of the trained disagreement predictor, we use both 1) hard score F1 and 2) soft score Mean Square Error (MSE), and compare the measurement effect of binary disagreement label and continuous disagreement rate.

Main Results

Disagreement prediction only with text From Table 2, we notice that continuous disagreement achieves better prediction than binary disagreement for most of the datasets. Among the datasets, the disagreement prediction models

[4]Note that only the SBIC and SChem101 datasets report annotators' demographic information, so we used these two datasets to evaluate the effect of including demographic information in disagreement prediction.

			Group of demographics		Personal level demographics	
Datasets	**Input Setup**	**Label Type**	F1 (↑)	MSE (↓)	F1 (↑)	MSE (↓)
SBIC	Demographics(sentence) ; Text	Continuous	65.4	0.086	**85.6**	**0.033**
	Demographics(templated) ; Text	Continuous	64.6	0.087	**85.6**	**0.033**
SChem101	Demographics(sentence) ; Text	Continuous	14.6	0.065	**48.8**	**0.009**
	Demographics(templated) ; Text	Continuous	16.9	0.073	**50.7**	**0.009**

Table 3: Evaluation results of vanilla (RoBERTa) classifiers with text and demographics inputs on the SBIC and Social Chemistry datasets. Note that only these two datasets among the five include the demographics of the annotators.

Type	**Input**	**Predicted Disagreement**	**Predicted Disagreement with Demographics**	**Collected Disagreement**
(1)	age: 22, politics: liberal, race: white, gender: woman. [SEP] what is the most common name for a black dude king kong.	0.003	0.015	0
(2)	age: 25, politics: liberal, race: asian, gender: woman. [SEP] swear to god they are the same people that objectify women and reify rape culturemisogyny. More importantly y do they feel so threatened.	1.034	0.971	1
(3)	age: 18, politics: other, race: asian, gender: man. [SEP] two strong men needed to some furniture.	0.223	0.096	0
(4)	age: 29, politics: liberal, race: white, gender: woman. [SEP] 13 reasons why is making me wish i went to a real fucking highschool.	0.332	0.01	0.5

Table 4: Examples of four types of predictions: (1) the predictions are the same as the collected disagreement, (2) both predictions are the same but differ from the collected disagreement. When the prediction given only text differ from the given text with demographics, (3) the former is close to collected disagreement, and (4) the latter is close to collected disagreement.[5]

work the best in the SBIC dataset. The binary label prediction are close to continuous prediction for SBIC and Politeness datasets. But SChem and Dilemmas have 0 F1 scores which only give 0 outputs. That means the binary label is not reliable for the two datasets.

For Dynasent, the binary label has an inconsistent performance based on hard score F1 and soft score MSE. We think one potential reason is that the binary disagreement is highly unbalanced while converting a continuous prediction to categorical labels like 0, 0.33, 0.67, and 1 is easy to accidentally assign an intermediate value to a wrong group. Therefore, even though we used both F1 and MSE metrics, they are used to have a parallel comparison between the binary label and continuous label setup. Among the binary classification, we consider F1 as the metric of model goodness, on the opposite, we use MSE to evaluate the regression fitness.

Disagreement prediction with text and demographic information Further, by comparing different experiment setups for disagreement with demographic information in Table 3, we focus on the different effects of a group of demographics or the personal level of demographics. The results show that personal-level demographics improve the disagreement prediction more than group-level demographics. One potential reason is that the annotator's level of demographics may imitate the annotation process that each annotator labels the text without knowing each other. And also because concatenating personal level demographics can be considered as oversampling that group-level setup can not.

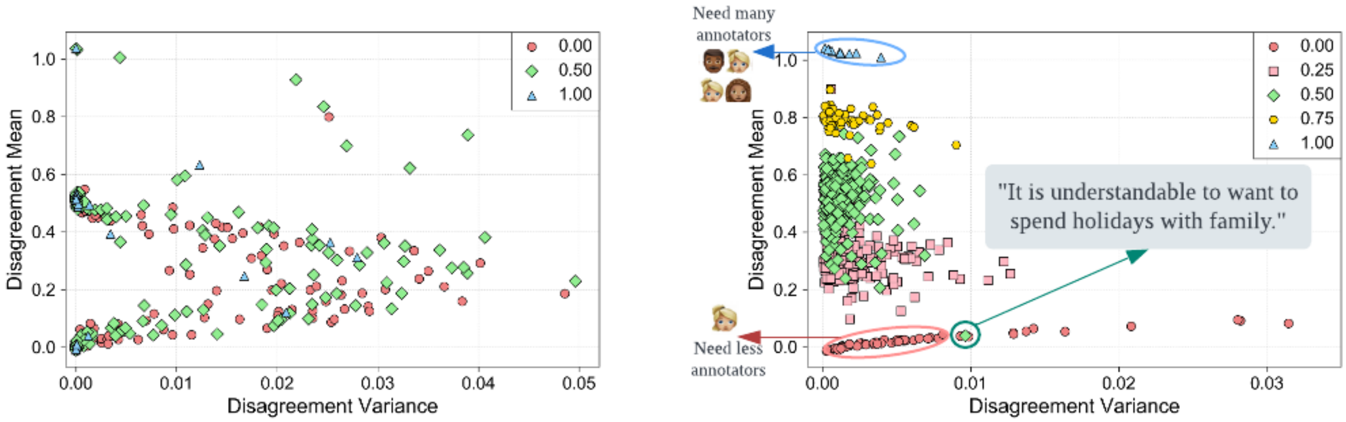
Qualitative Results Analysis Lastly, we categorize prediction into four types and provide an example per each in Table 4. We found demographics have been used in prediction with the text.

Simulation of Everyone's Voices with Artificial Demographics

One remaining question is how to reflect everyone's diverse opinions on such subjective and socially sensitive annotation tasks. To explore this aspect, we run additional experiments with the simulated demographics introduced in Section . Namely, we simulate a different combination of all possible *artificial* demographic groups, rather than using the real annotators' demographics used in model training (Section). Then, the disagreement of the simulated demographic information and the text is predicted using the fine-tuned disagreement predictor introduced in Section .

Our study is motivated by the Intersectionality theory (Crenshaw 1990), assuming that people's perspectives are shaped by the intersection of all available demographic categories. We set four gender types, seven ethnicity types, and five age ranges (see appendix for details[6]), and thus we have 140 ($4\times7\times5$) artificial annotators' unique demographic characteristics. Since we only trained our disagreement predictor with demographic information on SBIC and SChem101 datasets, the simulation experiments are also applied to these two datasets. We randomly sampled 600s text

[6]The technical appendices are available in our paper repository in arXiv. https://arxiv.org/pdf/2301.05036.pdf



(a) SBIC Dataset

(b) SChem 101 Dataset

Figure 4: Disagreement prediction with simulated demographic information on (a) SBIC and (b) Schem101 datasets, respectively. Different shapes and colors indicate the different disagreement labels as denoted in the legend. Best viewed in color.

instances in each dataset and concatenated them with 140 artificial annotators' demographic information in the colon template to predict continuous disagreement.

To visualize the simulation result of 140 artificial annotators per text, we made a scatter plot based on the mean and variance of 140 disagreement prediction as shown in Figure 4. The color and shape denoted at the legend shows the text's disagreement label in the original dataset. The higher points in the plot means higher predicted disagreement rate. The more rightward point implies a greater variance in the disagreement prediction among the 140 artificial annotators. The difficulty of disagreement prediction is related to the dataset's topic, quality etc. SBIC is collected from social media data while SChem is created by crowdsourcing, which might explain why the clusters are more clear in the Figure 4(b) than in the Figure 4(a). From 4(b), most text are predicted into corresponding disagreement clusters. But some outliers are predicted to be more controversial or agreeable. For example, the circled outlier has an original 0.5 disagreement label but ends up with a 0.04 disagreement prediction among 140 artificial annotators. The text is "It is understandable to want to spend holidays with family." Those outliers in the simulation experiment show the disagreement rate would change if the annotator change. Other than the outliers, the disagreement clusters shows they are less influenced by annotator change. With this simulation, we can distinguish disagreements caused by the natural controversy of the text or by the biased distribution of the assigned annotators.

Discussion and Future Work

We could think of potential applications in NLP data annotation pipeline using our disagreement prediction model:

Annotator number estimation. We could potentially use the predicted disagreement score in order to decide the appropriate number of annotators in a cost-efficient manner, e.g., we may not need three or five annotators for the text being predicted zero disagreements. For instance, we may need one or two annotators if a text is predicted to have lower disagreement scores. Other than that, we can assign five or even more annotators to those texts being predicted as highly disagreeable.

Annotator group assignment. Additionally, we suggest considering the annotation disagreement as a critical factor in finding the optimal group of annotator pools. This can be used as a novel annotator assignment supporting system for the data annotation pipeline. In the current annotator recruiting process, there is usually some uncontrollable randomness from annotators, either from skewed representatives or individual variations. We present a low-cost approach to simulate as diverse as possible artificial annotation pools to identify the controversial samples that maximize the disagreement. Thus, we avoid ignoring human bias and listening to opinions from a more diverse group of people to avoid polarized analysis. We hope our study can evoke others' attention in designing a more fair and representative annotation pipeline.

Potential risk of using demographic information. Last but not least, though our research shows that annotators' demographics help disagreement prediction, we should be careful about collecting private and personal information. Also, we admit that NLP or AI systems trained on demographic information might make another bias toward certain demographic groups.

Conclusion

Overall, we propose a disagreement prediction framework that measures annotators' disagreement in subjective tasks, predicts disagreement with/without demographic information and simulates 140 artificial annotators to build a relatively fair annotation pool. Our results show that the annotators' disagreement could be fairly predictable from the text, even better performs when we know the demographic information of the annotators. With our disagreement predictor, we believe we could shed light on various applications of data annotation in a more effective and inclusive manner.

Acknowledgments

We thank Dr. Maxwell Forbes for sharing the demographic information data for Social Chemistry 101 dataset. We also thank the anonymous reviewers and Minnesota NLP members for their insightful comments and suggestions.

References

Alm, C. O. 2011. Subjective Natural Language Problems: Motivations, Applications, Characterizations, and Implications. In *ACL*.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1): 37–46.

Crenshaw, K. 1990. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stan. L. Rev.*, 43: 1241.

Danescu-Niculescu-Mizil, C.; Sudhof, M.; Jurafsky, D.; Leskovec, J.; and Potts, C. 2013. A computational approach to politeness with application to social factors. In *ACL*.

Davani, A. M.; Díaz, M.; and Prabhakaran, V. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10: 92–110.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5): 378.

Foley, J. 2018. Explainable Agreement through Simulation for Tasks with Subjective Labels. *ArXiv*, abs/1806.05004.

Forbes, M.; Hwang, J. D.; Shwartz, V.; Sap, M.; and Choi, Y. 2020. Social Chemistry 101: Learning to Reason about Social and Moral Norms. *ArXiv*, abs/2011.00620.

Gordon, M. L.; Lam, M. S.; Park, J. S.; Patel, K.; Hancock, J. T.; Hashimoto, T.; and Bernstein, M. S. 2022. Jury Learning: Integrating Dissenting Voices into Machine Learning Models. *CHI Conference on Human Factors in Computing Systems*.

Hendrycks, D.; Burns, C.; Basart, S.; Critch, A.; Li, J. Z.; Song, D. X.; and Steinhardt, J. 2021. Aligning AI With Shared Human Values. *ArXiv*, abs/2008.02275.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lourie, N.; Bras, R. L.; and Choi, Y. 2021. Scruples: A Corpus of Community Ethical Judgments on 32, 000 Real-Life Anecdotes. In *AAAI*.

Oortwijn, Y.; Ossenkoppele, T.; and Betti, A. 2021. Interrater disagreement resolution: A systematic procedure to reach consensus in annotation tasks. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, 131–141.

Potts, C.; Wu, Z.; Geiger, A.; and Kiela, D. 2021. DynaSent: A Dynamic Benchmark for Sentiment Analysis. *ArXiv*, abs/2012.15349.

Prabhakaran, V.; Davani, A. M.; and D'iaz, M. 2021. On Releasing Annotator-Level Labels and Information in Datasets. *ArXiv*, abs/2110.05699.

Reidsma, D.; and op den Akker, R. 2008. Exploiting 'Subjective' Annotations. In *COLING 2008*.

Röttger, P.; Vidgen, B.; Hovy, D.; and Pierrehumbert, J. B. 2022. Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks. In *NAACL*.

Sap, M.; Gabriel, S.; Qin, L.; Jurafsky, D.; Smith, N. A.; and Choi, Y. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. *ArXiv*, abs/1911.03891.

Sap, M.; Swayamdipta, S.; Vianna, L.; Zhou, X.; Choi, Y.; and Smith, N. A. 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*, 5884–5906. Seattle, United States: Association for Computational Linguistics.

Uma, A.; Almanea, D.; and Poesio, M. 2022. Scaling and Disagreements: Bias, Noise, and Ambiguity. *Frontiers in Artificial Intelligence*, 5.

Uma, A.; Fornaciari, T.; Dumitrache, A.; Miller, T.; Chamberlain, J. P.; Plank, B.; Simpson, E.; and Poesio, M. 2021. SemEval-2021 Task 12: Learning with Disagreements. In *SEMEVAL*.

Yu, X. 2022. Hate Speech and Counter Speech Detection: Conversational Context Does Matter. *ArXiv*, abs/2206.06423.