NUTMEG: Separating Signal From Noise in Annotator Disagreement

Jonathan Ivey
Johns Hopkins University
jivey6@jhu.edu

Susan Gauch
University of Arkansas
sgauch@uark.edu

David Jurgens
University of Michigan
jurgens@umich.edu

Abstract

NLP models often rely on human-labeled data for training and evaluation. Many approaches crowdsource this data from a large number of annotators with varying skills, backgrounds, and motivations, resulting in conflicting annotations. These conflicts have traditionally been resolved by aggregation methods that assume disagreements are errors. Recent work has argued that for many tasks annotators may have genuine disagreements and that variation should be treated as signal rather than noise. However, few models separate signal and noise in annotator disagreement. In this work, we introduce NUTMEG, a new Bayesian model that incorporates information about annotator backgrounds to remove noisy annotations from human-labeled training data while preserving systematic disagreements. Using synthetic and real-world data, we show that NUTMEG is more effective at recovering ground-truth from annotations with systematic disagreement than traditional aggregation methods, and we demonstrate that downstream models trained on NUTMEG-aggregated data significantly outperform models trained on data from traditionally aggregation methods. We provide further analysis characterizing how differences in subpopulation sizes, rates of disagreement, and rates of spam affect the performance of our model. Our results highlight the importance of accounting for both annotator competence and systematic disagreements when training on human-labeled data.

1 Introduction

NLP is largely dependent on labeled data to train and evaluate models. Typically, labels are generated by humans through an annotation process that involves aggregating the judgments of multiple individuals (Snow et al., 2008; Nowak and Rüger, 2010; Zheng et al., 2017). Given the cost of experts, many approaches opt for crowdsourcing labels from a large number of annotators who have

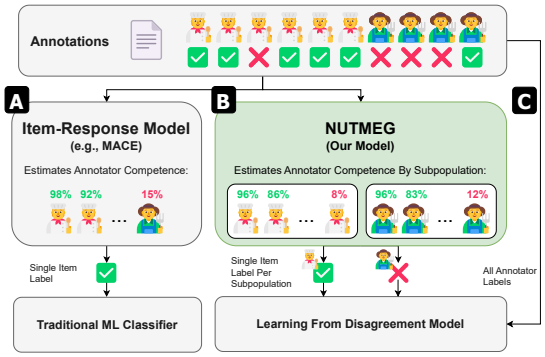

Figure 1: When aggregating annotators' labels, (A) traditional item response models like MACE (Hovy et al., 2013) ignore meaningful label variation in subpopulations to produce a single label, while (C) learning from disagreement models take all annotations as input, ignoring potential spam labels. Our approach (B), NUTMEG, extends the item response paradigm to infer labels per subpopulation, which can provide more accurate inputs when learning from disagreement.

varying levels of training and expertise in the task. In this setting, some labels are assumed to be errors due to mistakes, ambiguity in the task, or even adversarial behavior[1] by annotators (Hsueh et al., 2009; Aroyo and Welty, 2014; Jagabathula et al., 2017). As a result, there are many approaches for estimating an item's true label the from potentially-noisy collective annotations (Whitehill et al., 2009; Liu et al., 2012; Zheng et al., 2017; Paun et al., 2018; Goh et al., 2023; Bernhardt et al., 2022). While these models are effective, they assume that any deviation from the consensus label is a mistake. However, recent work has established that annotators from specific backgrounds may systematically differ in their judgments on an item, particularly for subjective tasks (Larimore et al., 2021;

[1]Here, adversarial behavior refers to instances where annotators intentionally provide incorrect labels, e.g., to sabotage the data collection process or to maximize payment for minimal effort.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 2874–2887
November 4-9, 2025

Sap et al., 2022; Pei and Jurgens, 2023; Wan et al., 2023; Mostafazadeh Davani et al., 2024). As a result, annotator groups who systematically disagree with the majority label are likely lost in aggregation. Here, we introduce a new Bayesian model for inferring ground truth labels that incorporates annotator backgrounds and allows for identifying systematic disagreement in labels (Figure 1).

Annotators disagree and two strands of research have proposed approaches to resolve these disagreements. One strand has framed the disagreement resolution as an unsupervised learning problem where a model simultaneously learns the probable ground truth label while also learning which labelers are more likely to give accurate answers (Zheng et al., 2017). Approaches such as MACE (Hovy et al., 2013) use Bayesian models to infer a single ground truth label per item, which is suitable for training most machine learning models. However, there are many subjective tasks, such as detecting hate speech, where the assumption of a single label can cause these models to ignore valid disagreements.

In contrast, a more recent strand has noted that some disagreements are meaningful and proposed new machine learning methods for *Learning from Disagreement* (Uma et al., 2021); such methods learn to predict from the original disaggregated data. While this latter branch is effective at incorporating diverse views—e.g., how different groups might view the same item—such models likely overweigh non-systematic disagreement, such as those due to mistakes or adversarial behavior.

Here, we introduce a Bayesian model for learning ground truth labels that is able to model systematic variation between subpopulations within the annotators. Our approach, NUTMEG (**N**uanced **U**nderstanding of anno**T**ation by **M**ultipl**E** **G**roups) estimates annotator competence and infers per-subpopulation labels for each item. In experiments on synthetic data, we demonstrate that (i) NUTMEG can accurately recover distinct labels for each subpopulation when they differ, while still recognizing when annotators are spamming[2] and (ii) NUTMEG is effective even for small numbers of annotations per subpopulation, making it readily amenable to use with crowdsourcing. Finally, in experiments with real data labeled with demographics, we show that by first reducing noise with NUTMEG, we can use subpopulation labels with learning from disagreement models to make more accurate predictions. We release NUTMEG and the synthetic data generation framework at https://github.com/jonathanivey/NUTMEG.

[2]Here, we follow common notation and refer to a deviation from the correct label as "spam." However, this label category reflects any type of disagreement, adversarial or not.

2 Modeling Annotator Disagreements

Annotators may disagree with each other for a variety of reasons—valid or not—and multiple branches of research have focused on understanding or resolving these disagreements to improve machine learning performance.

Modeling Annotator Backgrounds An individual's background (e.g., demographics, occupation) is known to systematically influence their annotation behavior, leading to disagreements in labeling (Lerner et al., 2024; Pei and Jurgens, 2023). While not focused on resolving these disagreements, recent work in NLP has focused on understanding how much of the disagreement can be attributed to an annotator's background; for example, showing that conservative annotators are less likely to rate anti-Black language as toxic (Sap et al., 2022). While earlier annotated data rarely included information about the annotators, more recent work has called for a responsible collection of this data (Santy et al., 2023; Davani et al., 2022), particularly for improving models by including diverse viewpoints (e.g., Fleisig et al., 2023; Orlikowski et al., 2023) and identifying biases in LLM behaviors (e.g., Santy et al., 2023; Deng et al., 2023). Our work fills a key gap by showing how to incorporate systematic diversity in groups' ratings in modeling while still accounting for noise and mistakes during the annotation process.

Inferring Annotator Competence Prior work has identified differences in annotator labels as label-noise and attempted to reduce it prior to model training (Dawid and Skene, 1979). Many of these methods use unsupervised probabilistic models of annotator behavior to identify incorrect labels, also known as spam, and estimate annotator competence (Whitehill et al., 2009; Liu et al., 2012; Hovy et al., 2013; Paun et al., 2018). More recent models have used information from classifiers for the same goal of estimating annotator competence or reducing noise (Goh et al., 2023; Bernhardt et al., 2022).

These methods rely on the assumption that disagreements between annotators indicate errors or a lack of competence; however, for many tasks there can be genuine disagreements between annotators (Plank et al., 2014). In this work, we create a new

model of annotator competence that retains genuine disagreements between annotators while also reducing label-noise.

Task Subjectivity There are many causes for annotator disagreements, also known as human label variation, including expertise, item-difficulty, and motivation (Sommerauer et al., 2020; Plank, 2022; Cabitza et al., 2023; Fleisig et al., 2024), but a particularly important cause is how an annotator's background interacts with the task objectives. Previous work has shown that in many subjective NLP tasks, like detecting offensiveness, politeness, and toxicity, annotator disagreements are correlated with backgrounds and social variables such as gender, race, age, and region (Larimore et al., 2021; Sap et al., 2022; Pei and Jurgens, 2023; Wan et al., 2023; Mostafazadeh Davani et al., 2024). This relationship is particularly important because current models of annotator competence treat consensus as correct—and disagreement as error—and as a result, valid label disagreements by minority subpopulations risk being omitted.

In this work, we introduce a new model to estimate different truth values for each relevant subpopulation for each item in a dataset. This multiple-truth modeling approach allows us to pass disagreement—in connection with its systemic causes—to downstream modeling applications and reduce the risk of omitting meaningful variation in labeling decisions.

Learning from Disagreement When given data with conflicting judgments, one line of research known as *Learning from Disagreement* has proposed treating disagreement as signal rather than noise. The simplest approach to treating disagreement as a signal is to not aggregate annotations at all and instead use the full distribution of responses for each item as the desired output of a model (Uma et al., 2021).

Other approaches choose to model every annotator response by training multi-task models (Mostafazadeh Davani et al., 2022; Mokhberian et al., 2024) or training on every annotator-item pair (Gordon et al., 2022; Fleisig et al., 2023; Weerasooriya et al., 2023). These methods are effective at incorporating annotator diversity, but they do not account for other causes of disagreement like mistakes and adversarial behavior that may reduce model performance.

Our work introduces a complementary method to enable noise reduction while retaining disagreement, which can be used in combination with these Learning from Disagreement approaches to reduce noise prior to training. Limited work has focused on both reducing noise and retaining disagreement in human labels. Weber-Genzel et al. (2024) designed models to separate annotation errors from valid disagreements; however, their work is designed for natural language inference and relies on a more-involved setting that collects and evaluates explanations by annotators to remove errors. Here, we use subpopulation information to identify systemic disagreement, and our method, which does not require the collection of annotator explanations, can be applied to existing datasets.

The most similarly themed work to ours are the CrowdTruth 2.0 metrics (Dumitrache et al., 2018), which are designed to capture media unit quality, worker quality, and annotation quality. Though CrowdTruth 2.0 does model disagreement, it does not model the origins of disagreement or produce training labels for downstream tasks. In this work, we use relevant information about annotators, through subpopulation divisions, to model the causes of systematic disagreement, allowing downstream stakeholders to attribute disagreement and train models tailored to different use-cases.

3 Methods

We introduce NUTMEG, a Bayesian model that estimates annotator competence and predicts item labels while retaining subpopulation-level disagreement (Figure 2). Our approach builds upon Bayesian model designs for items' and annotators' variation (e.g., Paun et al., 2018). NUTMEG most closely resembles MACE (Hovy et al., 2013) which attempts to simultaneously learn the spamming rates of annotators and the item's likely label.

One key assumption of prior Bayesian models is that there exists a single correct label for each item agreed upon by annotators. In NUTMEG, we relax this assumption so that there exists a single correct label *for each subpopulation* that is always given by an annotator in that subpopulation when they try to. While real data likely has additional within-group variation, this simplifying assumption allows our model to focus on capturing systematic variation that is most relevant to downstream applications.

To incorporate subpopulation identity, the generative step of our model works as follows (also shown in 3): First, for each item i and subpopulation k, we sample the true subpopulation label T_{ik} from a uniform prior. Then, for each annotator j,

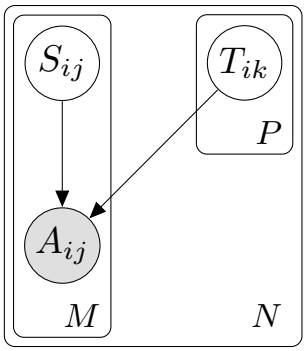

Figure 2: Plate diagram of NUTMEG. Annotator j from subpopulation k produces label A_{ij} on item i. The label choice depends on the item's true label for subpopulation k, T_{ik}, and whether j is spamming on i, modeled by binary variable S_{ij}. $N = |items|$, $M = |annotators|$, and $P = |subpopulations|$.

for $i = 1 \dots N$:
  for $k = 1 \dots P$:
    $T_{ik} \sim$ Uniform
  for $j = 1 \dots M$:
    $S_{ij} \sim$ Bernoulli$(1 - \theta_j)$
    if $S_{ij} = 0$:
      $A_{ij} = T_{ik}$
    else :
      $A_{ij} \sim$ Multinomial(ξ_j)

Figure 3: The generative process for NUTMEG. See the text for a full description of variables.

we sample a binary variable S_{ij} from a Bernoulli distribution with parameter $1 - \theta_j$. S_{ij} represents whether annotator j is spamming on item i. If the annotator is not spamming, then they use the *true label of their subpopulation* to produce annotation A_{ij}. Otherwise, their annotation A_{ij} is sampled from a multinomial with parameter ξ_j. The parameter θ_j represents the probability that annotator j is not spamming on a given item; this is a measure of their competence. The parameter ξ_j represents annotator j's individual behavior when they are spamming, which could produce the correct label, but only by chance.

To fit the model, we use Variational-Bayes training to maximize the probability of the observed data:

$$P(A; \theta, \xi) = \sum_{T,S} \left[\prod_{i=1}^{N} \prod_{k=1}^{P} P(T_{ik}) \cdot \prod_{i=1}^{N} \prod_{j=1}^{M} P(S_{ij}; \theta_j) \cdot P(A_{ij} | S_{ij}, T_{ik}; \xi_j) \right]$$

We train with symmetric Beta priors with parameters of 0.5 on θ_j and symmetric Dirichlet priors on ξ_j. As identified in MACE, these priors model the extremes of behavior common in annotation (i.e., either an annotator often gives the correct label or they rarely give the correct label). However, NUTMEG also supports adjusting these priors should an end-user desire a more informed prior.

Because not every subpopulation is guaranteed to label every item, NUTMEG must handle items where there is an unobserved subpopulation k_x for item i_x. In those cases, we calculate the estimated label for the observed subpopulations of i_x, then we identify the set of items that have the same estimated labels for those subpopulations and contain annotations from k_x. Finally, we take the average of the posterior probabilities for the items in this set to estimate the posterior of $T_{i_x k_x}$. For this process, we make the simplifying assumption that observed labels from different subpopulations are independent. Though this is often not the case, it allows us to use a larger number of items to estimate unobserved instances. Future work could explore more robust methods for estimating items' labels for unobserved subpopulations. We note that NUTMEG does not require the use of these imputed subpopulation truths, and for Experiment 3 (§6), we choose not to estimate unobserved samples to reduce the risk of introducing additional label noise.

NUTMEG requires that each annotator be associated with a subpopulation label. Given the NLP community's recent recognition of the importance of collecting information about the annotators themselves (Lerner et al., 2024; Santy et al., 2023; Mihalcea et al., 2025) and the recent uptick in the creation of such datasets (e.g., Kumar et al., 2021; Sap et al., 2022; Pei and Jurgens, 2023), we believe that annotator data, like demographics, will be increasingly important and available. NUTMEG can use this data to better separate signal from noise in annotator disagreement. However, we note that NUTMEG does not require that subpopulations be derived from additional annotator data, like those collected from questionnaires. Instead, it can be

combined with any unsupervised method for grouping annotators based on behavior (e.g., Vitsakis et al., 2024), and the subpopulations can be derived from the resultant inferred groups.

4 Experiment 1: Synthetic Data

To evaluate how effective NUTMEG is at reducing noise and recovering ground truth from multiple subpopulations, we first evaluate performance using synthetic data that precisely simulates annotator behavior in a setting with systematic disagreement. Using synthetic data allows us to compare NUTMEG's estimates to the true opinions of annotators that are not available in real annotated data.

4.1 Experimental Setup

To generate the synthetic data, we first create a set of 150 annotators and assign them randomly to one of two subpopulations, a majority and a minority, using an 80% and 20% split for this experiment. We then assign each annotator a spamming rate in [0,1]; these rates represent the probability that an annotator ignores their subpopulation's true label when labeling an item. The mean spamming rate across individuals indicates the overall level of spam in the data. After generating our annotators, we create a set of 500 items with two possible labels. We designate a proportion of the items as *divisive* according to a global divisiveness rate. For divisive items, the subpopulations will hold different true opinions, and for non-divisive items, they will hold the same true opinion. The divisiveness rate indicates the level of systematic disagreement in the data.

To simulate the annotation process for an item, we first decide whether each annotator is spamming based on their competence score. Spammers assign labels to the item randomly, while non-spammers provide the true opinion of their subpopulation. Finally, to simulate the availability of crowdsourced annotations we randomly sample from this dataset so that each item has 5 annotations and each annotator labels more than 20 items (average of 16.67 items per annotator). We will release this synthetic data and evaluation framework for future research on modeling subpopulation variation in annotation.

To evaluate how effective NUTMEG is at recovering ground truth, we use the above procedure to generate multiple synthetic datasets with different divisiveness rates varying from 0 to 1 and global spamming rates varying from 0 to 0.25. We compare performance against five models for estimating ground truth to show the effect of systematic variation by subpopulation. We include a majority vote, the original Dawid and Skene (1979) model (D&S), and its extension MACE, which is the closest comparison to our model. We also follow the model recommendations from the large survey by Zheng et al. (2017) and include the Learning from Crowds (LFC; Raykar et al., 2010) and Bayesian Classifier Combination (BCC; Kim and Ghahramani, 2012), which perform best for the type of nominal data used in our experiments. We fit all five models on each dataset and calculate an accuracy score by comparing a model's estimates for each subpopulation to its true label in the dataset.

4.2 Results

We find that as the rate of systematic disagreement increases, NUTMEG correctly identifies the true label for both the majority and minority subpopulations (Figure 4), despite the minority having four times less available data. Importantly, NUTMEG makes these improvements without reducing its ability to estimate the majority opinion. In contrast, while the other methods are accurate for recovering the majority's true label, they are increasingly inaccurate for the minority subpopulation's true label as the rate of divisiveness increases (as expected).

Note that while NUTMEG is able to recover most of the true labels, the accuracy for the minority subpopulation is still lower. This gap is primarily due to data sparsity. For some items, too few minority annotators may be assigned to accurately distinguish meaningful disagreement from spam. Further, as spam rates increase from 0 to 25%, we see an average 4.22% drop in minority accuracy. While a 25% spam rate is likely on the high side, NUTMEG's overall accuracy is still high regardless of how often the subpopulation disagrees. As experimenters may not know how often a particular group might disagree, this performance trend suggests NUTMEG can be accurately deployed even when divisive items are relatively rare.

How good is NUTMEG at distinguishing systematic disagreement from spam? In our simulations, all annotators are capable of spamming and thus, not all divergent labels by annotators in the minority subpopulation are meaningful. To assess whether NUTMEG recognizes these labels as spam, we compare NUTMEG's estimated proportion of disagreements to the true global rate of divisiveness. This gives an indication of how accurately

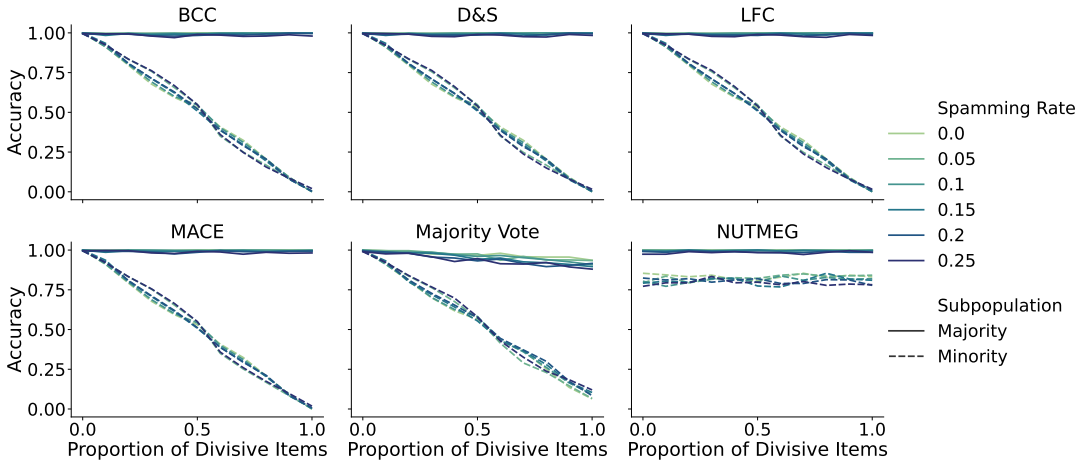

Figure 4: As the rate of disagreement increases between majority and minority populations for the subset of items, traditional models are increasingly less accurate at correctly inferring the subpopulation's true label (dashed line), while still being accurate for the majority subpopulation (solid line). By contrast, NUTMEG can effectively estimate the opinions of both the majority and minority subpopulations with varying rates of spam (shown with color).

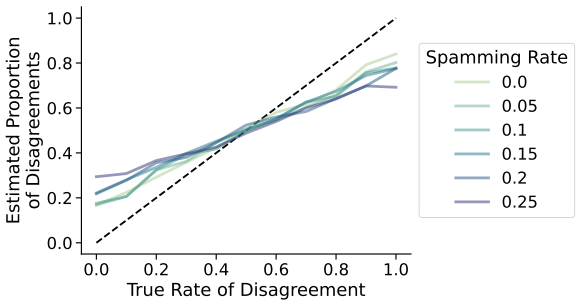

Figure 5: NUTMEG can effectively differentiate genuine disagreement from spam annotations. As the rate of spam increases, NUTMEG's estimate of the disagreement rate diverges from the true rate, especially at the extremes. The dotted line indicates perfect predictions.

NUTMEG can differentiate between genuine systemic disagreement and spam.

We find that NUTMEG can effectively differentiate genuine disagreement from spam annotations, as shown in Figure 5. However, as the rate of disagreement approaches the high and low extremes, NUTMEG's estimate of the rate of disagreement diverges from the true rate. At especially high rates of disagreement, the model underestimates and at especially low rates of disagreement, the model overestimates. We believe that this divergence is caused by the increased effect that contrary evidence has at extreme rates of disagreement. Even a single annotation that diverges from expectations may bring the model's estimates away from extremes, and this effect is exacerbated by spam, which provides increased contrary evidence.

We further find that NUTMEG is effective at assessing annotator competence. The average Pearson's correlation between NUTMEG's estimate of annotator competence θ_j and the annotator's true competence across all runs is 0.81. By comparison, MACE's average correlation is 0.58. These correlations do not significantly differ for members of the majority or minority subpopulations, which illustrates how using traditional item-response models can lead a practitioner to erroneously conclude that they have low-quality annotators when in fact they have high-quality annotators and systematic disagreement.

5 Experiment 2: Subpopulation Size

Practitioners labeling data often have limited control of how many annotators in specific groups are present in the annotator pool. Experiment 1 (§4) showed a decrease in accuracy for estimating the minority opinion as the rate of spam increases. This result leads to a natural question: how much data is enough to accurately represent a minority subpopulation given different rates of spam?

5.1 Experimental Setup

To test for the effect of minority subpopulation size during annotation, we repeat our synthetic data generation procedure from §4, but fix the global spam rate of 0.1 and a global divisiveness rate of 0.2. We then vary the size of the minority subpopulation from 10% to 50% and the total number of annotators sampled for each item from 3 to 15. Note that because annotators are randomly assigned to items, for settings with few annotations per item,

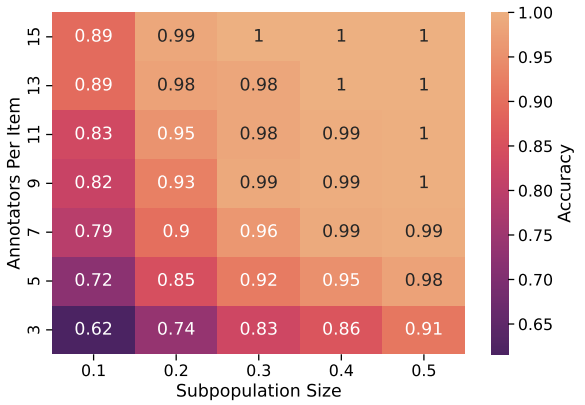

Figure 6: As the size of a subpopulation decreases, NUTMEG needs more annotations to ensure that the subpopulation is sufficiently represented and estimate its true opinions. This result highlights the importance of data collection strategies for representing minority subpopulations.

some items may not receive any annotators by a person of the minority subpopulation, which mirrors real world settings. Finally, we run NUTMEG on the datasets and calculate an accuracy score to compare its estimates to the true opinions of the minority subpopulation.

5.2 Results

We find that as the size of a subpopulation in a dataset gets smaller, NUTMEG requires a much larger number of annotations per item to maintain the same level of accuracy (Figure 6). With a subpopulation proportion of 0.3 (equivalent to 45 annotators), NUTMEG only requires 5 annotations to achieve 92% accuracy, but for a subpopulation proportion of 0.1 (equivalent to 15 annotators), NUTMEG would need more than 15 annotations per item to achieve the same performance. This result shows the importance of intentional data collection methods when representing the opinions of small subpopulations. If new datasets need an accurate estimate for multiple subpopulations, they should focus on having sufficient coverage for each subpopulation rather than simply increasing the total number of annotations collected. Alternatively, training sets can choose not to calculate estimates for sets items without sufficient labels from small subpopulations, which is the approach that we take for model training in §6.

6 Experiment 3: Downstream Modeling

The purpose of NUTMEG is to remove spam annotations in human-labeled datasets while retaining valid subpopulation-level disagreement. Experiments 1 and 2 have demonstrated this on synthetic data and, here, we evaluate whether NUTMEG-aggregated labels improve downstream modeling.

NUTMEG produces estimated true labels for each subpopulation and therefore, for predictive modeling, we adopt the learning from disagreement setting where a classifier is trained on multiple labels per annotation. We hypothesize that using the full distribution of labels (directly from annotators) introduces unnecessary noise, and in this experiment aim to answer the question: does removing spam annotations improve performance on downstream tasks using real-world data?

6.1 Experimental Setup

Dataset To effectively use NUTMEG, we require annotated data with metadata indicating which annotators belong to which subgroups. While NUTMEG can be used with inferred subpopulations that do not require additional data collection (as explained in §3), we opt to use demographics as a way to partition annotators to reduce potential confounds to the interpretation of the results introduced by clustering methods. Multiple works have noted meaningful variation by race and gender (Larimore et al., 2021; Sap et al., 2022; Pei and Jurgens, 2023; Wan et al., 2023). We use data from POPQUORN (Pei and Jurgens, 2023) which provides annotations on a 1–5 Likert scale for two classification tasks (i) offensiveness and (ii) politeness, and age, race, and education demographics of annotators. We replicate their preprocessing steps with two additions. We binarize the Likert ratings in their data at ≥ 3 to simplify the analysis and communication of our results. We also remove subpopulations with $\leq 5\%$ of annotations to ensure that the models are being trained with sufficient representation.

We note that though we choose to evaluate on data using annotator demographics to split subpopulations, NUTMEG does not necessarily require additional data collection as recent work has demonstrated the efficacy of clustering annotators into subpopulations based on annotator behaviors (Vitsakis et al., 2024).

Modeling For each task, we train a multi-task classification model, where a base model has separate classification heads for each subpopulation. This setup uses multiple tasks to represent salient disagreements and is a popular approach in learning from disagreement (e.g., Fornaciari et al., 2021; Mostafazadeh Davani et al., 2022; Mokhberian

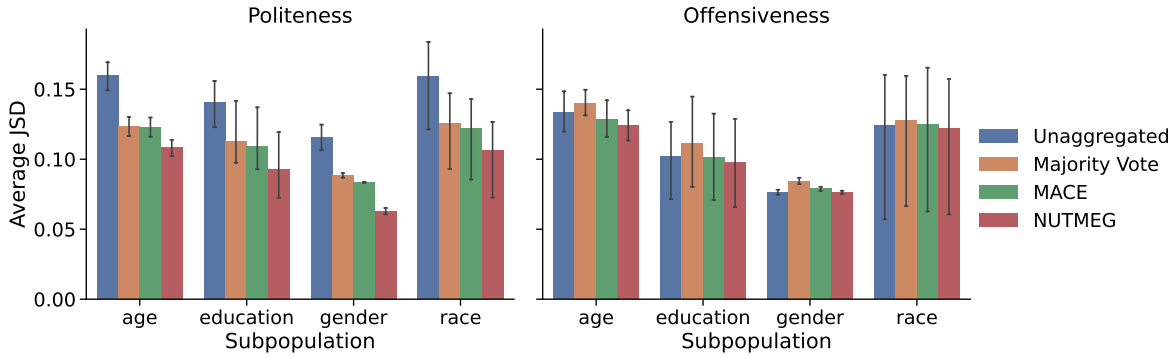

Figure 7: Performance at replicating the ground truth label distribution by subgroup (lower is better) using learning from disagreement models trained with different types of label aggregation. The bar height indicates the mean JSD across all subpopulations' label distribution in the test set, and error bars indicate the total range of JSD across all subpopulations. Our results show that any type of aggregation is helpful to reduce noise, but by jointly modeling subpopulation-preference with annotator competence, NUTMEG is better able to match the true label distributions.

et al., 2024; Wang and Plank, 2023). Prior to training, we run NUTMEG on the training and validation sets using separate runs for each demographic category (i.e., gender, age, race, and education). We then fine-tune ModernBERT (Warner et al., 2024) models with additional classification layers for each subpopulation in a demographic (e.g., Man and Woman in gender). In this case, our multiple tasks are predicting the true subpopulation labels.

As a baseline for standard aggregation methods, we also fine-tune single-task ModernBERT models on labels aggregated by either majority vote or MACE.[3] Finally as a baseline for training on the full distribution of disaggregated annotations, we follow the popular approach of training a multi-task model where each task is a different annotator in the dataset (Mostafazadeh Davani et al., 2022; Wang and Plank, 2023; Mokhberian et al., 2024). The final prediction for each subpopulation is then the average of the predicted probabilities for annotators in that subpopulation. Additional details on the training procedure are in Appendix B.

Evaluation Recognizing the many causes of human label variation (Plank, 2022), we measure model performance by comparing the predicted probabilities output by the model for each subpopulation (if it is multi-task) to the true distribution of labels provided by that subpopulation in the test set; in the single task setups, we use the same probability for each group. Following previous work, we quantify the similarity of these distributions with Jensen–Shannon divergence (JSD; Uma et al., 2021), where lower is better. Note that the set of annotators in the training and validation sets is entirely separate from the set of annotators in the test set, and we score using the label distribution of the full, disaggregated labels in the test set.

[3]Because it is commonly used in NLP and performed similarly to LFC and BCC in Experiment 1, we use only MACE as a baseline for simplicity.

6.2 Results

Politeness We find that for politeness detection, models trained on NUTMEG outperform ($p < 0.05$) both models trained on traditionally aggregated annotations and models trained on disaggregated annotations for all subpopulation splits except race (Figure 7, left). By learning from more-accurate aggregations that both reduce noise and highlight systematic disagreement, models are better able to predict the full label distribution for each item in the test set. Our results show aggregation generally helps, with even the naive but commonly-used majority voting often reducing noise in the data. Yet, the gap between MACE and NUTMEG highlights that there is additional benefit to modeling subpopulation-variation—and that for this politeness task, there is likely meaningful variation that can be modeled. This trend supports our hypothesis that learning directly from disaggregated annotations can introduce noise and NUTMEG's noise reduction improves performance on downstream tasks. It also demonstrates that accounting for relevant subpopulation-level variation when predicting annotator competence improves performance.

Offensiveness We find that for offensiveness detection (Figure 7, right), there are no significant differences between models trained with different aggregation methods or no aggregation. However, among the aggregation steps NUTMEG does as well

as or better than no-aggregation. This neutral behavior highlights it is at least not introducing additional noise, unlike majority vote which generally has higher JSD.

We interpret this trend as an example where there is no systematic variation by subpopulation in the data. After a manual review of the data, we found that the majority of examples exhibit high subjectivity. For example, some annotators may label an encouraging statement as offensive if it contains vulgar language, while others may label an unkind statement as inoffensive if there is implied sarcasm.

7 Discussion

In synthetic experiments, we show that NUTMEG is highly effective at distinguishing genuine systematic disagreement from noise in annotations. Then, through real-world experiments, we demonstrate that NUTMEG's ability to distinguish systematic disagreement can be combined with Learning from Disagreement models to improve performance on subjective classification tasks. In real scenarios, annotators may disagree for a variety of reasons (Sandri et al., 2023), not all of which relate to a subpopulation identity. However, our results show subpopulation modeling is a promising direction for improving the representation and evaluation for the diverse stakeholders of NLP models (Cabitza et al., 2023). This direction has many useful applications, like tailoring content moderation to differing community standards, helping social scientists identify differences in perceptions of bias based on political affiliation, or enabling mental health models to distinguish between humor and expressions of depression in differing age groups.

Though we consider these compelling use cases for NUTMEG, we also find that it does not always improve performance. For example, in the offensiveness detection task, we often encounter difficult cases that cannot be explained by the model. Some cases arise from inherent textual ambiguity, such as the statement "I'd almost forgotten that one. What a gem!", where it is unclear whether the author is being sincere or sarcastic, a well-known confound (cf. Basile et al., 2021; Sandri et al., 2023). Others depend on personal preferences not captured by our subpopulation splits, as in "You're a f—ing legend", where the decision relies on whether an annotator considers vulgar language offensive. Such Though, we note that in this task, NUTMEG does not hurt performance compared to other methods.

From these findings, we conclude that NUTMEG will be most effective when applied to tasks where there is a prior belief that subpopulations differ in their judgments or attitudes. Future work should consider how to balance the differing degrees of within-group variation, between-group variation, and inherent item difficulty in the diverse set of NLP tasks—as well as new approaches to automatically learn annotator subpopulations directly from annotations (e.g., Lo et al., 2023; Vitsakis et al., 2024).

8 Conclusion

In this work, we introduce NUTMEG, a Bayesian model that infers ground truth labels from annotations while accounting for systematic differences among annotator subpopulations. By extending item-response models, NUTMEG jointly estimates annotator competence and identifies when groups consistently diverge in their labeling decisions, addressing the limitations of traditional aggregation models that treat deviations as errors. Our experiments on synthetic and real-world data demonstrate that NUTMEG effectively recovers distinct subpopulation labels, mitigates spam annotations, and improves the performance of Learning from Disagreement models. By preserving meaningful disagreement, NUTMEG provides a more nuanced understanding of annotation data, particularly in subjective NLP tasks. Its data efficiency makes it well-suited for crowdsourcing settings, and its ability to model annotator variation contributes to more representative NLP models. We release NUTMEG and the accompanying synthetic data generation and evaluation libraries. This work highlights the importance of incorporating diverse perspectives in annotation modeling and encourages further research into principled approaches for handling disagreement in human-labeled data.

9 Limitations

NUTMEG works by identifying systematic disagreement in subpopulations of annotators. While we have demonstrated that NUTMEG can correctly identify such disagreement, most real data contains label variation beyond that due to subpopulations. Thus, while the method can potentially improve quality, it is not a universal panacea for noisy annotations. Indeed, we note that NUTMEG doesn't always improve downstream model performance (in the Offensiveness task) suggesting that even

when demographic labels are present and modeled, other sources of label variation in the data may more strongly influence performance; indeed, past work in NLP has found a mixed trend, where for some datasets, demographics explain little variation (Orlikowski et al., 2023), while for others, demographics explain substantial variation (Larimore et al., 2021; Sap et al., 2022; Pei and Jurgens, 2023; Wan et al., 2023).

NUTMEG produces estimates of the ground truth by subpopulation. We anticipate that these labels will be most useful for practitioners who use learning from disagreement models that are designed to model subpopulations. However, most of NLP still uses models that produce a single label. Our work could still potentially help in these settings by allowing practitioners to train models for separate populations (e.g., an offensive language detector optimized for the views of a particular subpopulation) and then deploy these strategically. Further, even when a traditional machine learning model is used, NUTMEG can still help identify meaningful disagreement in the data and raise awareness for the practitioner. Additionally, NUTMEG's design treats all truth variables as nominal. This allows to be applied for tasks with both nominal truth values or ordinal truth values; however, future iterations may improve performance on certain tasks by tailoring models for ordinal truth values.

Our experiments with synthetic and real data used a known, discrete subpopulation label for each annotator. However, for many datasets, this type annotator information is not present. While we note that recent work has pointed to the ability to cluster annotators to create inferred subpopulations (Vitsakis et al., 2024), we have not evaluated that strategy here in favor of first demonstrating that the method works with known demographics. Our experimental design limits the potential confounding influence of the clustering model on the potential benefits of NUTMEG. Nevertheless, we view this as a promising direction for future work to explore.

Finally, our experiments used a limited number of real-world datasets to demonstrate effectiveness. Demographically-labeled data is growing in NLP (cf. Santy et al., 2023), but still uncommon. The POPQUORN dataset is among the largest available dataset with multiple tasks, making it ideal for our study. However, we recognize that future work could evaluate NUTMEG with more datasets as they become available.

10 Ethical Considerations

NUTMEG requires that annotators be associated with a particular subpopulation. We anticipate that for many practitioners, this background will be based on demographics or other personal attributes. As a result, NUTMEG could potentially increase the collection of personal data, which needs to be responsibly stored. However, we view this risk as being outweighed by the benefits of having the different views of subpopulations better represented in models.

Acknowledgments

This work was supported in part by the National Science Foundation under Grant No. IIS-2143529 and Grant No. OIA-1946391. This work was also supported by a University of Arkansas Honors College Research Grant.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pages 2623–2631, New York, NY, USA. Association for Computing Machinery.

Lora Aroyo and Chris Welty. 2014. The three sides of crowdtruth. *Human Computation*, 1(1).

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.

Mélanie Bernhardt, Daniel C. Castro, Ryutaro Tanno, Anton Schwaighofer, Kerem C. Tezcan, Miguel Monteiro, Shruthi Bannur, Matthew P. Lungren, Aditya Nori, Ben Glocker, Javier Alvarez-Valle, and Ozan Oktay. 2022. Active label cleaning for improved dataset quality under resource constraints. *Nature Communications*, 13(1):1161. Publisher: Nature Publishing Group.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

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.

A. P. Dawid and A. M. Skene. 1979. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28. Publisher: [Royal Statistical Society, Oxford University Press].

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

Anca Dumitrache, Oana Inel, Lora Aroyo, Benjamin Timmermans, and Chris Welty. 2018. CrowdTruth 2.0: Quality Metrics for Crowdsourcing with Disagreement (short paper). *ArXiv*.

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, Singapore. Association for Computational Linguistics.

Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024. The perspectivist paradigm shift: Assumptions and challenges of capturing human labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2279–2292, Mexico City, Mexico. Association for Computational Linguistics.

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

Hui Wen Goh, Ulyana Tkachenko, and Jonas Mueller. 2023. CROWDLAB: Supervised learning to infer consensus labels and quality scores for data with multiple annotators. *arXiv preprint*. ArXiv:2210.06812 [cs].

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 *CHI Conference on Human Factors in Computing Systems*, pages 1–19, New Orleans LA USA. ACM.

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

Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. 2009. Data quality from crowdsourcing: A study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 27–35, Boulder, Colorado. Association for Computational Linguistics.

Srikanth Jagabathula, Lakshminarayanan Subramanian, and Ashwin Venkataraman. 2017. Identifying unreliable and adversarial workers in crowdsourced labeling tasks. *Journal of Machine Learning Research*, 18(93):1–67.

Hyun-Chul Kim and Zoubin Ghahramani. 2012. Bayesian classifier combination. In *Artificial Intelligence and Statistics*, pages 619–627. PMLR.

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.

Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. 2021. Reconsidering annotator disagreement about racist language: Noise or signal? In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 81–90, Online. Association for Computational Linguistics.

Emilia Agis Lerner, Florian E. Dorner, Elliott Ash, and Naman Goel. 2024. Whose preferences? differences in fairness preferences and their impact on the fairness of ai utilizing human feedback. In *Annual Meeting of the Association for Computational Linguistics*.

Qiang Liu, Jian Peng, and Alexander T Ihler. 2012. Variational Inference for Crowdsourcing. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Soda Marem Lo, Valerio Basile, et al. 2023. Hierarchical clustering of label-based annotator representations for mining perspectives. In *CEUR WORKSHOP PROCEEDINGS*, volume 3494, pages 1–10. CEUR-WS.

Rada Mihalcea, Oana Ignat, Longju Bai, Angana Borah, Luis Chiruzzo, Zhijing Jin, Claude Kwizera, Joan Nwatu, Soujanya Poria, and Thamar Solorio. 2025. Why ai is weird and shouldn't be this way: Towards ai for everyone, with everyone, by everyone. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28657–28670.

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 Diaz, Dylan K Baker, and Vinodkumar Prabhakaran. 2024. D3CODE: Disentangling disagreements in data across cultures on offensiveness detection and evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18511–18526, Miami, Florida, USA. 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.

Stefanie Nowak and Stefan Rüger. 2010. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the International Conference on Multimedia Information Retrieval*, MIR '10, page 557–566, New York, NY, USA. Association for Computing Machinery.

Matthias Orlikowski, Paul Röttger, Philipp Cimiano, Dirk Hovy Bielefeld University, University of Oxford, Computing Sciences Department, Bocconi University, Milan, and Italy. 2023. The ecological fallacy in annotation: Modeling human label variation goes beyond sociodemographics. In *Annual Meeting of the Association for Computational Linguistics*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. 2018. Comparing Bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585.

Jiaxin Pei and David Jurgens. 2023. When do annotator demographics matter? measuring the influence of annotator demographics with the POPQUORN dataset. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 252–265, Toronto, Canada. Association for Computational Linguistics.

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. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning from crowds. *Journal of machine learning research*, 11(4).

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Ježek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441.

Sebastin Santy, Jenny T Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. Nlpositionality: Characterizing design biases of datasets and models. *ArXiv*, abs/2306.01943.

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.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

Pia Sommerauer, Antske Fokkens, and Piek Vossen. 2020. Would you describe a leopard as yellow? evaluating crowd-annotations with justified and informative disagreement. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4798–4809, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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.

Nikolas Vitsakis, Amit Parekh, and Ioannis Konstas. 2024. Voices in a crowd: Searching for clusters of unique perspectives. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12517–12539, Miami, Florida, USA. Association for Computational Linguistics.

Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's Voice Matters: Quantifying Annotation Disagreement Using Demographic Information. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):14523–14530. Number: 12.

Xinpeng Wang and Barbara Plank. 2023. ACTOR: Active learning with annotator-specific classification heads to embrace human label variation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2046–2052, Singapore. Association for Computational Linguistics.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

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

Tharindu Cyril Weerasooriya, Alexander Ororbia, Raj Bhensadadia, Ashiqur KhudaBukhsh, and Christopher Homan. 2023. Disagreement matters: Preserving label diversity by jointly modeling item and annotator label distributions with DisCo. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4679–4695, Toronto, Canada. Association for Computational Linguistics.

Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. 2009. Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

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

Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and Reynold Cheng. 2017. Truth inference in crowdsourcing: Is the problem solved? *Proceedings of the VLDB Endowment*, 10(5):541–552.

A NUTMEG Usage

We provide the following example of how to use NUTMEG programmatically. More detailed information is available at https://github.com/jonathanivey/NUTMEG.

```
# import package
from nutmeg.nutmeg_cython import NUTMEG

# instantiate model
nutmeg = NUTMEG()

# fit to our data
nutmeg.fit(df)

# acccess model predictions
nutmeg.labels_    # labels
nutmeg.probas_    # label probabilities
nutmeg.spamming_ # annotator competence
```

B Modeling Details

NUTMEG runs entirely on CPU and can be run on any reasonably equipped computer.

We trained the ModernBERT models with 149M parameters on a single NVIDIA RTX A6000 GPU Hugging Face Transformers 4.48.1 (Wolf et al., 2020) and PyTorch 2.5.1 (Paszke et al., 2019) on a CUDA 12.4 environment. All models were trained for 12 epochs with a batch size of 192 and were tuned for a learning rate in the range $[1 \times 10^{-5}, 2 \times 10^{-3}]$ with Optuna (Akiba et al., 2019). We use the train, validation, and test splits provided by the original POPQUORN dataset (Pei and Jurgens, 2023). To ensure reproducibility, we set all random seeds in Python to be 42.