Quantifying and Predicting Disagreement in Graded Human Ratings

Leixin Zhang Çağrı Çöltekin
Universität Tübingen, Germany
leixin.zhang@uni-tuebingen.de cagri.coeltekin@uni-tuebingen.de

Abstract

It is increasingly recognized that humans do not always agree, and disagreement is inherent in many annotation tasks. However, not all items in a given task elicit the same level of opinion divergence. In this paper, we study the extent to which item-level annotation variation and variation structure can be captured from text features, focusing on inappropriate language detection, including offensive language, hate speech, and toxic language detection. We model annotation variation to assess whether the degree of annotation divergence can be predicted from item-level textual features. We also propose the Opposition Index, a metric that quantifies the extent of opposing stances among annotators based on their Likert ratings.

Keywords: annotation variation, graded human ratings, opposing stances, human disagreement

1. Introduction

Recent work has demonstrated that many natural language processing (NLP) datasets and tasks exhibit inherent annotation variation (Plank, 2022; Sorensen et al., 2024). This variation occurs across multiple linguistic domains and task types. In *syntax*, it is observed in annotation tasks such as part-of-speech tagging (Plank et al., 2014; Zeman, 2010); in *semantics*, such as semantic similarity (Wang et al., 2023) and natural language inference (Huang and Yang, 2023; Jiang and de Marneffe, 2022; Pavlick and Kwiatkowski, 2019; Liu et al., 2023; Zhang and de Marneffe, 2021); in *pragmatics*, including irony detection (Casola et al., 2024) and dialogue act annotation (Verdonik, 2023); and in other *socially-relevant NLP tasks*, such as hate speech detection (Sang and Stanton, 2022; MacAvaney et al., 2019), offensive language (Kocoń et al., 2021; Davani et al., 2024), and toxic content detection (Kumar et al., 2021).

It is increasingly acknowledged that human annotators do not always make identical decisions or hold the same opinions on many NLP tasks (Plank, 2022; Uma et al., 2021; Basile et al., 2021). However, the annotation pattern is not consistent across all items in the same task: some cases are clear-cut and have near-perfect agreement from multiple annotators, while others can be more ambiguous, resulting in variance in annotation patterns. Such annotation variation can arise from item ambiguity or vagueness (e.g., insufficient context), language complexity (e.g., use of slang or jargon), or annotators' personal beliefs, values, expertise, or personality (Sap et al., 2020, 2022).

Reflected in Likert rating distributions, annotations for some items show sharper peaked distributions, indicating strong consensus among annotators, while others may display flatter or multi-modal distributions, reflecting interpretation variability or the presence of opposing opinions among annotators.

Estimating which items are likely to elicit disagreement has important practical and theoretical implications. In *annotation practice*, identifying disagreement-prone items allows researchers to optimize annotation workflows by prioritizing difficult or perspective-divergent cases. For instance, socially controversial items such as potentially offensive or politically charged content often require a larger number of annotators to capture the diversity of latent perspectives, whereas less controversial items may require fewer annotations.

From a *linguistic perspective*, analyzing uncertainty patterns allows researchers to uncover the latent factors underlying annotation uncertainty, such as detecting cases that lack sufficient context (Sandri et al., 2023), and provide insights into human perception of complex language phenomena, such as irony, sarcasm, or figurative language. Studying perspective disagreement can further reveal culturally dependent interpretations (e.g., Western vs. non-Western perspectives (Sap et al., 2022; Huang and Yang, 2023; Larimore et al., 2021), liberal vs. conservative viewpoints (Luo et al., 2020)) or conflicts in judgment. In *other decision-making domains*, including legal, political, and medical decision-making, annotation variation may reflect conflicting interests and opposing perspectives between different parties (e.g., employers vs. employ-

Proceedings of NLPerspectives @ LREC 2026, pages 33–43
12 May 2026.

ees, producers vs. consumers) (Angouri, 2012). Automatically identifying disagreement-prone items can help flag conflicting cases, prioritize expert review, and improve decision fairness (Patel et al., 2018).

In this work, we investigate whether the item-level annotation patterns can be inferred solely from item features, and we mainly focus on inappropriate language detection tasks in this work, including hate speech, offensive, and toxic language classification. These tasks have been extensively studied and are known to exhibit substantial annotator disagreement (Sang and Stanton, 2022), and there is a lack of universally accepted standards or definitions. For example, definitions of hate speech vary across research objectives (Talat and Hovy, 2016), legal frameworks (European Commission, 2016), and platform policies. It is often impractical to specify every possible case in annotation guidelines, particularly in crowdsourced settings where annotators are not formally trained. Thus, hate speech annotation often relies on annotators' perceptions, linguistic intuition, and individual understanding of what constitutes hate speech. In the era of large language models (LLMs), these challenges become even more critical (Weidinger et al., 2021). Given the rapidly growing user base of LLM-powered systems, detecting toxic and inappropriate language is essential for mitigating risks, preventing the amplification of harmful content, and ensuring safer user interactions.

To more faithfully capture the nuances of human judgment and annotation distributions, we use datasets annotated with Likert-scale ratings instead of discrete binary labels in this work. We also tailor the training objective by employing loss functions designed for ordinal data rather than simple discrete classes, including the Earth Mover's Distance (and its variant) and cumulative cross-entropy. This work focuses on estimating two aspects of human annotation variation:

- **Annotation variance:** whether the degree of dispersion in annotator responses can be inferred from item features.
- **Opposing stances:** whether the conflicting stances among annotators can be effectively modeled and predicted.

By modeling item-level annotation variance and stance opposition, our work takes an initial step toward characterizing patterns or structures of human annotation variation. We hope this study will inspire further research on annotation uncertainty and perspective-aware NLP systems.

2. Related Work and Positioning

Existing literature on annotation variation can be broadly divided into two main streams. The first stream focuses on analyzing human annotation and human interpretation variation (Hong et al., 2025; Jiang and de Marneffe, 2022), investigating types or causes of variation across annotators (Xu et al., 2023), disentangling noise from genuine disagreement (Weber-Genzel et al., 2024), and analyzing cultural background influence (Huang and Yang, 2023). The second stream focuses on modeling human annotation variation with machine learning approaches (Uma et al., 2021; Mostafazadeh Davani et al., 2022; Zhou et al., 2022). The methods include soft-label training (Uma et al., 2021; Fornaciari et al., 2021), incorporating socio-demographic features for group perspective simulation (Gordon et al., 2022), multi-task learning (Mostafazadeh Davani et al., 2022) with each task corresponding to a specific annotator, and using an annotator embedding layer (Mokhberian et al., 2024) to learn annotator-specific labels.

Work that explicitly predicts or infers disagreement from text features remains relatively scarce. In this direction, Zhang and de Marneffe (2021) aim to tease apart agreed and disagreed items in natural language inference (NLI) and propose an ensemble approach by integrating three specialized models trained to predict three labels: entailment, neutral, or contradiction. In subjective tasks such as hate speech detection, Wan et al. (2023) model disagreement by directly predicting whether or not annotation variation exists for an item as a binary classification problem, and a regression problem by predicting annotation variation with the value $1 - P_{majority}$, the proportion of annotations that do not fall into the majority label. Baumler et al. (2023) and van der Meer et al. (2024) estimate the uncertainty of human annotations and incorporate it into an active learning framework.

Despite these advances, prior work has largely treated opinion divergence as a categorical prediction problem (e.g., three-way NLI labels or binary hate speech decisions) and has not examined the degree of annotation uncertainty and whether the full structure of fine-grained annotation distributions can be recovered from item-level signals (Zhang, 2025). To address this gap, we perform a detailed analysis to model the full distribution of Likert-scale ratings and examine its effectiveness in inferring annotators' opinion divergence with item textual features.

3. Rating Variation Prediction

To model opinion divergence, including both variance and opposing views, we use datasets with Likert-scale ratings. Unlike discrete labels, these graded ratings capture fine-grained human perception differences, reflect gradations in perceptions of inappropriate language, enabling analysis of multimodal and polarized patterns in distribution. We aim to test whether the structure of annotation variation across items can be captured from textual features. Specifically, we examine the magnitude of opinion divergence in Section 3.1 and the presence of opposing stances among annotators in Section 3.2.

3.1. Inferring Annotation Variance

For discrete classes, the entropy of annotator labels is commonly used to quantify uncertainty. For Likert ratings, where the distance between ratings is meaningful, we treat them as equally spaced values. For each item i, the degree of annotation variation is computed as the unbiased variance of the ratings from N_i annotators:

$$\sigma_i^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (r_{ij} - \bar{r}_i)^2, \quad (1)$$

where r_{ij} is the rating of annotator j on item i, and $\bar{r}_i$ is the mean rating for that item. Perfect agreement corresponds to $\sigma_i^2 = 0$.

To estimate annotation variation, we consider two approaches:

1. **Direct variance prediction**: we train a regression model to predict the unbiased variance σ_i^2 directly from item-level features.
2. **Distribution-based prediction**: we predict the full Likert rating distribution $\mathbf{p}_i = [p_{i1}, \ldots, p_{iK}]$ for item i, then compute the variance from the predicted distribution:

$$\hat{\sigma}_i^2 = \sum_{k=1}^{K} p_{ik}(k - \hat{\mu}_i)^2, \quad \hat{\mu}_i = \sum_{k=1}^{K} p_{ik} \cdot k. \quad (2)$$

To evaluate the performance of prediction, we mainly measure with the following metrics: (1) *Mean Squared Error (MSE)* between predicted and true variance, (2) *Spearman's rank correlation r* between predicted variance and true variance across items and (3) F_1 score of whether or not opinion divergence or rating difference is in the annotation of an item among annotators.

	Comments	Annotations
1	"Mr #Trump will be loving today. As it is the one day of the year when #FakeNews is acceptable. #aprilfools "	[0, 0, 0, 2, 2, 2]
2	"Nigga at da end of the day we all would be gone, or somewhere else. and speakin about it is not gonna fucking matter! "	[0, 0, 2, 2, 2]
3	"I hate when guys call their girls bitches and hoes. That's your girl. You respect her. "	[0, 0, 0, 2, 2]

Table 1: Examples of opposing stances in the offensive dataset (Sap et al., 2020). Three-ordinal ratings are used for labels, with scores from 0 to 2, with 0 as *not offensive* and 2 as *offensive*.

3.2. Identifying Opposing Opinions

Beyond variance, it is also crucial to assess distribution structures (Akhtar et al., 2021; Van der Eijk, 2001) and whether genuine opposing opinions exist for an item. Some cases with divergent judgments on offensive classification are shown in Table 1. The annotations suggest that disagreement can arise from different interpretations of what constitutes offensive content. In Example 1, disagreement emerges in a case that involves political satire. Some annotators interpret mocking a political figure as offensive, possibly due to perceived contempt or disrespect, whereas others regard it as legitimate political expression or humor rather than offensive content per se. In some cases, annotators label a comment as offensive due to the presence of offensive terms, even when the overall intent of the comment is not to insult. For example, in Example 3 of Table 1, a comment condemning derogatory terms toward women may still be marked offensive by some because it quotes them, while others focus on the critical intent and label it non-offensive. These examples show that annotation variation is often driven by differences in how annotators weigh lexical content, speaker intent, and contextual meaning about respect and harm.

Reflected in the Likert distribution, this manifests as bimodal patterns rather than a single Gaussian mode. We propose a metric to quantify opposing stances, referred to as **opposition index**.[1] Let the

[1]Traditional bimodality measures, such as the Bimodality Coefficient (BC), or mixture-model-based modality tests, are typically designed for continuous dis-

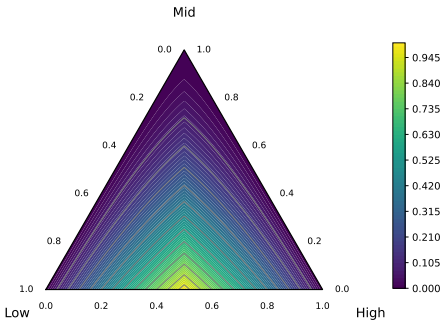

Figure 1: Opposition Index Illustration

predicted or observed distribution for an item be divided into three segments: the low, mid, and high ratings of the Likert scale, with probabilities denoted as $P_{\text{low}}, P_{\text{mid}},$ and P_{high}. P_{low} corresponds to the proportion of annotators giving the lowest ratings (e.g., not toxic). P_{high} corresponds to the proportion giving the highest ratings (e.g., extremely toxic); and P_{mid} corresponds to the proportion in the middle of the scale. We define the opposition stance index as:

$$\text{Index}_i = 2 \cdot \min(P_{\text{low}}, P_{\text{high}}) \cdot (1 - P_{\text{mid}}) \quad (3)$$

The final index value ranges from 0 to 1, with 1 indicating maximal polarization (half of the annotators select the low end and half the high end, with no intermediate ratings), with a value of 0 indicating consensus, reflected in a unimodal distribution centered in the middle or skewed toward either end of the scale, with no annotations spanning both extremes. Figure 1 illustrates how the index behaves: when $P_{\text{low}} = 0.5$ and $P_{\text{high}} = 0.5$ with negligible P_{mid}, the index reaches its maximum, reflecting clear opposing stances among annotators.

4. Objective Functions for Likert Distribution Prediction

To model both annotation variation and opposing opinions discussed in the previous section, we infer the full Likert rating distribution for each item, apart from predicting a single summary statistic of variance. In this section, we propose objectives specifically designed for Likert-scale ratings, leveraging their ordinal structure rather than treating them as categorical labels. For an item i with K Likert categories, the target distribution is represented as:

$$\mathbf{p}_i = [p_{i1}, p_{i2}, \ldots, p_{iK}], \text{ where } 1 < 2 < \cdots < K. \quad (4)$$

where p_{ik} denotes the proportion of annotators assigning rating k to item i, and $\sum_{k=1}^{K} p_{ik} = 1$. For example, in a 5-point Likert setting (ratings range from 0 to 4), if the annotations for an item are $\{1,1,0,1,2\}$ collected from 5 annotators, the empirical distribution is represented as a vector: $\mathbf{p}_i = [0.2, 0.6, 0.2, 0.0, 0.0]$.

To train the model to predict distributions, we experiment with loss functions listed below:

Earth Mover's Distance (EMD), also known as the Wasserstein distance (Rubner et al., 2000), explicitly accounts for the ordinal distance between rating categories. EMD penalizes prediction errors proportionally to the distance between categories.

EMD with Mean Regularization We further propose a multi-task objective that combines Earth Mover's Distance with an explicit constraint on the predicted mean rating with mean squared error.[2]

Ordinal Cumulative Cross Entropy To capture the ordinal structure of Likert-scale labels, we customize cross-entropy loss to measure the distributional difference of the Likert ratings. In our approach, a K-level Likert-scale problem is transformed into $K-1$ binary decisions with positive class as $y > k$. The total loss is then computed as the sum of the losses over all $K-1$ thresholds.

Kullback–Leibler Divergence We also consider the Kullback–Leibler (KL) divergence (Kullback and Leibler, 1951) as as a comparison to cumulative probability approaches and use it to quantify the dispersion from the predicted distribution to the target distribution (soft-label representation, which we refer to as KL-soft in this work).[3]

tributions with sufficiently large sample sizes. When the number of annotators per item is small (e.g., around five), these statistics become unstable and lack statistical power, making them unsuitable for reliably detecting bimodality in item-level annotation distributions.

[2]While EMD captures overall distributional shifts and respects the ordinal structure of the Likert scale, it does not directly constrain the expected rating (i.e., the mean of the distribution). Two distributions with similar cumulative shapes may still differ in central tendency. To address this, we introduce an additional mean-squared error term on the expected rating.

[3]While KL divergence is widely used for multi-class classification, the standard categorical formulation treats all class mismatches equally, ignoring the ordinal relationships among Likert ratings.

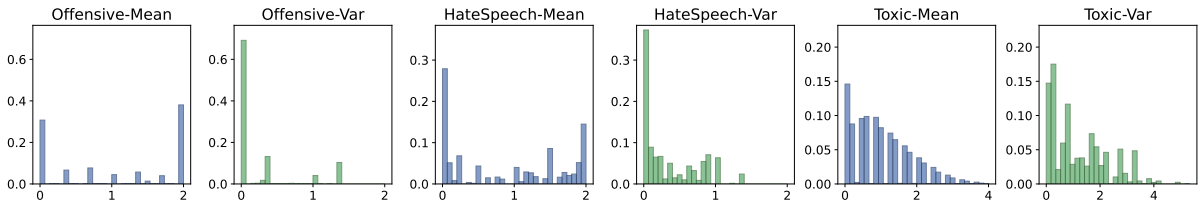

Figure 2: Summary of Dataset Statistics: Mean and Variance. The y-axis values indicate normalized density, which accounts for the relative proportion of data points in each bin.

5. Experiments and Implementation

This section introduces the datasets used for the experiments in this study and experiment implementation details.

5.1. Datasets

We conduct experiments on three subjective datasets annotated by multiple annotators. Figure 2 shows the summary of the statistics (annotation mean and variance) of three datasets.

Offensive Language. The Offensive Language dataset (Sap et al., 2020) is annotated with three categories of offensiveness: *no*, *maybe*, and *yes*. We map these categories to a three-point Likert scale (0 - 2). The dataset contains approximately 150k annotated items. To ensure reliable empirical annotation distributions, we filter out items annotated by fewer than three annotators. After filtering, the dataset comprises approximately 128.6k annotations over 35.8k unique items, with an average of 4 annotations per item.

Hate Speech. The hate speech dataset by Kennedy et al. (2020) provides graded annotations of hatefulness. Labels are provided on a three-level Likert scale, where 0 denotes non-hateful content and higher values indicate increasing hate speech severity. Items annotated by fewer than four annotators are removed, resulting in approximately 67k annotations over 5,990 unique items, and each item is roughly annotated by 11 annotators on average.

Toxicity. Annotations in the dataset (Kumar et al., 2021) follow a five-point Likert scale ranging from *not toxic* (0) to *extremely toxic* (4). The dataset contains approximately 107.6k text instances, most of which are annotated by 5 annotators. We retain items with at least five annotations and merge repetitive comments, resulting in approximately 106k items.

5.2. Implementation

Data Splits. Each dataset is randomly divided into training, validation, and test sets, following a 50%, 25%, 25% ratio. It is partitioned at the level of distinct text instances to prevent any items from appearing in multiple data splits.

Model Setting. Models are implemented in PyTorch, and text inputs are encoded using the pretrained Sentence-Transformer model `all-MiniLM-L6-v2`[4] (Reimers and Gurevych, 2019). To allow fair comparison across different prediction objectives, we keep the model architecture fixed, including input features, number of hidden layers, and layer dimensions, for all experiments. Models are trained with early stopping. Training is terminated if the validation performance does not improve for five consecutive epochs. The best-performing model on the validation set in the training history is selected for evaluation and reporting.

Baseline Setup. We use the aggregated binary distribution (where responses greater than $(K-1)/2$ for K-class Likert are treated as the positive class) as a baseline for each dataset, training with binary cross entropy loss, and compare its prediction with direct variance regression and full Likert distribution prediction.

Evaluation Protocol. For reliability, each experiment is repeated with five independent random splitting seeds, and the mean of the evaluation metrics is reported as the model performance.

Metric Computation. For the computation of the opposition index, we treat rating 0 as the low value and 2 as the high value for three-class Likert ratings, and treat ratings 0 and 1 as the low-value group and ratings 3 and 4 as the high-value group for five-class Likert ratings, representing two opposing stance camps.

[4] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

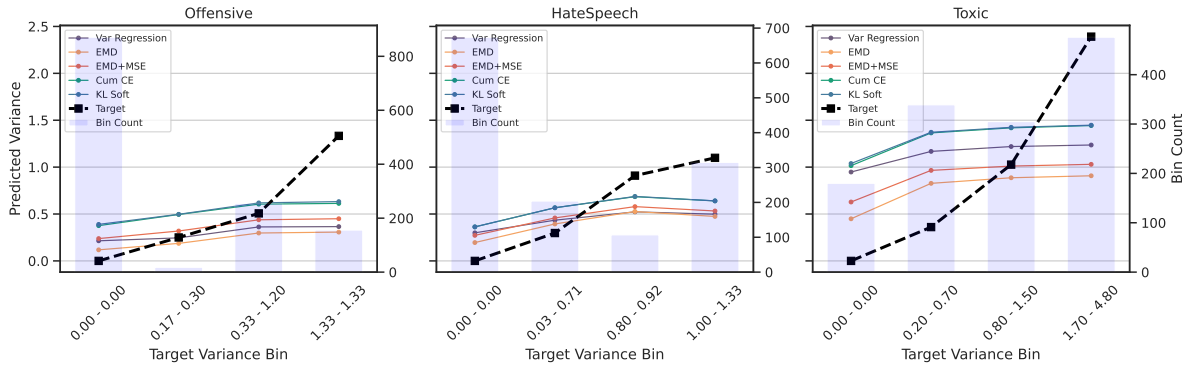

Figure 3: Item-level variance grouped by target variance bins: true versus predicted variance.

Model	Var_MSE ↓	Var_Corr ↑	Disagree_F1 ↑
Offensive			
Binary CE	0.218 ± 0.003	0.344	0.5917 ± 0.007
Var Reg	**0.195** ± 0.008	0.389	**0.6112** ± 0.018
EMD	0.227 ± 0.008	**0.393**	0.5957 ± 0.018
EMD+MSE	0.206 ± 0.003	0.386	0.6071 ± 0.018
Cum CE	0.248 ± 0.006	0.371	0.6076 ± 0.014
KL Soft	0.251 ± 0.014	0.372	0.6096 ± 0.022
Hate Speech			
Binary CE	0.275 ± 0.014	0.435	0.7275 ± 0.014
Var Reg	**0.197** ± 0.003	0.424	0.7233 ± 0.016
EMD	0.216 ± 0.019	**0.454**	0.7365 ± 0.013
EMD+MSE	**0.197** ± 0.011	0.445	0.7313 ± 0.017
Cum CE	0.204 ± 0.008	0.449	**0.7408** ± 0.016
KL Soft	0.206 ± 0.008	0.445	0.7371 ± 0.011
Toxic			
Binary CE	2.203 ± 0.021	0.290	0.9185 ± 0.001
Var Reg	**1.005** ± 0.027	**0.308**	0.9185 ± 0.007
EMD	1.179 ± 0.060	0.307	0.9186 ± 0.008
EMD+MSE	1.087 ± 0.049	0.303	0.9187 ± 0.007
Cum CE	1.056 ± 0.028	0.306	0.9191 ± 0.007
KL Soft	1.061 ± 0.012	0.298	0.9191 ± 0.007

Table 2: Comparison of models for estimating item-level annotation variance (mean ± std).

6. Results and Discussion

This section presents the experimental results for both annotation variance estimation (Section 6.1) and opposing stance prediction (Section 6.2).

6.1. Annotation Variance Estimation

Variance Prediction Models predict annotation variance values with reasonable accuracy. They achieve a variance MSE of around 0.2 on the 3-point Likert annotation tasks (Offensive and Hate Speech), and a variance MSE of approximately 1 for the 5-point Likert task (Toxic). Among all methods, directly predicting the unbiased variance using a regression model achieves the best performance. Among models that predict the Likert distribution and then compute variance, those trained with the EMD with mean regularization (EMD+MSE) achieve the second-best performance. They consistently outperform the EMD and KL-soft models.

Prediction across Annotation Variance Bins Apart from overall performance, we also analyze variance prediction across bins divided based on human-annotated variance levels (see Figure 3).

The bin-grouped analysis reveals a similar pattern across models. All variance predictions exhibit a monotonic trend: items with near-perfect agreement are assigned the lowest predicted variance scores, and items with higher empirical variance tend to receive higher predicted variance. However, the predicted variance values tend to concentrate around a middle range. The difference between low, medium, and high variance bins is attenuated. For instance, the magnitude differences are underestimated for the highest variance category. Models do not distinguish well between items with moderate and high variance. It may be due to the relatively smaller number of examples in these bins for the offensive and hate speech datasets. For the lowest variance bins, predicted variance values rarely reach exactly 0, particularly for distribution-based models. As a result, the lowest bin is not as low as the target variance.

Spearman Correlation Models show a moderate positive correlation with human annotation variance, ranging from approximately 0.3 to 0.45 across three datasets. Some models (e.g., EMD) trained to predict Likert distributions often achieve higher Spearman correlations with human annota-

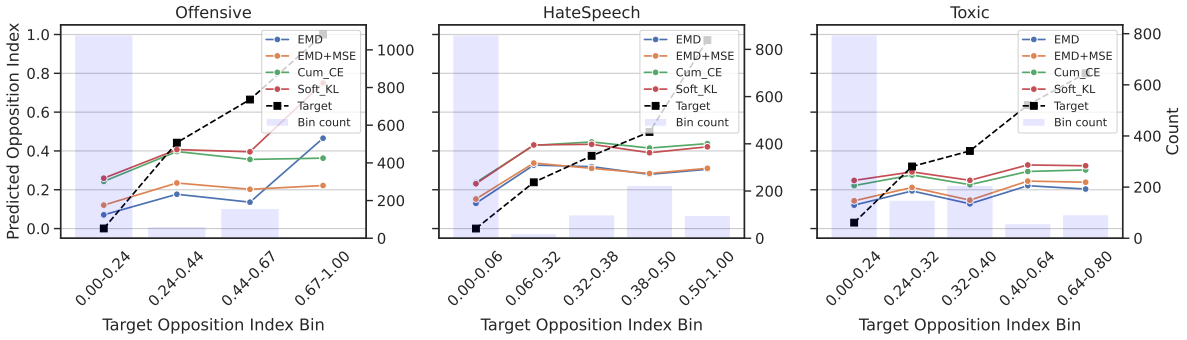

Figure 4: Opposition index values binned according to target scores.

tion variance compared to regression models that directly predict the variance. This suggests that even when variance is not explicitly supervised, distribution-based training objectives can recover the annotation variation structure across items. By contrast, directly regressing the variance can force the model to overfit the noise in the dataset. In contrast, predicting the full rating distribution allows the model to capture more structured patterns in annotator judgments. Rather than fitting a single summary statistic, the model learns the overall shape of the response distribution, providing a richer and more robust representation of opinion divergence. However, as expected, the loss function (KL-soft) that does not account for rating distances performs the worst, particularly for tasks with more Likert classes, such as the 5-rating Toxic dataset.

F1 Score F_1 score in this study measures whether an item's annotations exhibit rating differences (that is, whether the variance is greater than 0). Across all three datasets, the F1 scores from different models are generally very similar. On the Offensive dataset, the Variance Regression model achieves the highest F1 score (0.611 ± 0.018), slightly outperforming the other models. For hate speech, distribution prediction with cumulative cross-entropy loss achieves the best performance. On the Toxic dataset, all models achieve high (~0.92) and nearly identical F1 scores, likely because the dataset is annotated using a 5-point Likert scale, and most items have variance greater than 0. The imbalance tendency toward the class of presence of annotation variation (around 85% items) may lead to most results at class 1 (the presence of annotation divergence). Overall, the models achieve strong performance in detecting variation, with F_1 scores exceeding 0.6 for nearly all models across the three datasets.

6.2. Opposing Stance Prediction

For opposing stances measured by the polarization index, we group items into intervals based on their target opposition index values and analyze the results within each interval.

Ideally, items with higher true opposition values should also receive higher predicted scores from models. However, this pattern does not fully hold. While models successfully distinguish between items with no opposition (index close to 0) and items with moderate opposition, they struggle to capture extreme polarization. For items whose true opposition index approaches 1, predicted values tend to remain in a mid-range (approximately 0.4), indicating underestimation of highly polarized cases.

Several factors may explain this phenomenon. First, as shown in Figure 4, the number of items in the highest opposition index bin is relatively small, which limits the model's ability to learn stable patterns for extreme polarization when such cases are rare in the training set.

Secondly, items with a high polarization index are more prone to noise when a few annotators deviate significantly from the majority. When annotation noise inflates the opposition index, the resulting patterns may not reflect stable item features, causing the model to regress toward the mean.

Finally, extreme opposition may partly result from other factors beyond the text itself, such as annotators' ideological differences, personal experience, or differing interpretations of the guidelines, which cannot be inferred from textual features alone. Across the three datasets, items are labeled by annotators with diverse socio-demographic backgrounds, which are not evenly distributed across items. As a result, certain influences cannot be fully inferred from item-level features alone.

7. Limitations of Current Experiments

Firstly, although this paper examines the predictability of annotation variation from textual features, it cannot be assumed that the state-of-the-art encoder model, which converts texts into embeddings, perfectly captures all textual information (Lucy and Gauthier, 2017; Zhang et al., 2024; Zhang and Çöltekin, 2024). There can be information loss during the embedding process, and some linguistic cues may not be fully represented. Secondly, the number of annotations per item is limited, which limits the reliability of opinion divergence and distributional estimates. Additionally, the observed rating distributions may be sensitive to sampling noise. The datasets used in this work are crowd-sourced. Although crowd-sourced data increases annotator diversity, it also introduces additional noise, making human opinion modeling more challenging. Finally, Likert scales are restricted to three or five categories in the datasets we experiment with. With few annotators and coarse-grained scales, the space of possible variance or distributional values becomes highly discrete. For example, when only three annotators are available, certain distribution proportions (e.g., 0.33 or 0.66) occur frequently due to combinatorial constraints rather than meaningful underlying differences. This discreteness reduces the granularity of human opinion divergence and can affect the interpretability of predicted distributions.

8. Conclusion

This study investigates the extent to which annotation variation can be inferred from item-level features alone. We explore two aspects of annotation variation: human annotation variance estimation and opposing stances prediction. Our results show that variance derived from predicted distributions achieves performance comparable to direct variance regression when appropriate loss functions, such as Earth Mover's Distance and its variant with mean regularization, are used. Beyond predicting a single summary statistic like variance, distribution-based approaches can better capture disagreement structure, such as annotators' opposing stances. To quantify this, we propose the opposition index and demonstrate its use across three datasets. These findings have practical implications for future annotation design: resources can be allocated more efficiently by assigning more annotators to items likely to exhibit opinion divergence, while reducing effort on items with clear consensus.

9. Bibliographical References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2021. Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection. *arXiv preprint arXiv:2106.15896*.

Jo Angouri. 2012. Managing disagreement in problem solving meeting talk. *Journal of Pragmatics*, 44(12):1565–1579.

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.

Connor Baumler, Anna Sotnikova, and Hal Daumé III. 2023. Which examples should be multiply annotated? Active learning when annotators may disagree. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10352–10371, Toronto, Canada. Association for Computational Linguistics.

Silvia Casola, Simona Frenda, Soda Marem Lo, Erhan Sezerer, Antonio Uva, Valerio Basile, Cristina Bosco, Alessandro Pedrani, Chiara Rubagotti, Viviana Patti, and Davide Bernardi. 2024. MultiPICo: Multilingual perspectivist irony corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16008–16021, Bangkok, Thailand. Association for Computational Linguistics.

Aida Mostafazadeh Davani, Mark Díaz, Dylan Baker, and Vinodkumar Prabhakaran. 2024. D3CODE: Disentangling disagreements in data across cultures on offensiveness detection and evaluation. *arXiv preprint arXiv:2404.10857*.

European Commission. 2016. Code of conduct on countering illegal hate speech online. Accessed: 2026-03-25.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, Massimo Poesio, et al. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association*

for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics.

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

Pingjun Hong, Beiduo Chen, Siyao Peng, Marie-Catherine de Marneffe, Benjamin Roth, and Barbara Plank. 2025. Agree, disagree, explain: Decomposing human label variation in NLI through the lens of explanations. *arXiv preprint arXiv:2510.16458*.

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

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

Chris J Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*.

Jan Kocoń, Alicja Figas, Marcin Gruza, Daria Puchalska, Tomasz Kajdanowicz, and Przemysław Kazienko. 2021. Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Information Processing & Management*, 58(5):102643.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

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. Association for Computational Linguistics.

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2023. We're afraid language models aren't modeling ambiguity. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Li Lucy and Jon Gauthier. 2017. Are distributional representations ready for the real world? evaluating word vectors for grounded perceptual meaning. In *Proceedings of the First Workshop on Language Grounding for Robotics*, pages 76–85, Vancouver, Canada. Association for Computational Linguistics.

Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020. Detecting stance in media on global warming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3296–3315. Association for Computational Linguistics.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

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

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

Pinal Patel, Disha Davey, Vishal Panchal, and Parth Pathak. 2018. Annotation of a large clinical entity corpus. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2033–2042, Brussels, Belgium. Association for Computational Linguistics.

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

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

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 2000. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 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, Dubrovnik, Croatia. Association for Computational Linguistics.

Yisi Sang and Jeffrey Stanton. 2022. The origin and value of disagreement among data labelers: A case study of individual differences in hate speech annotation. In *International Conference on Information*, pages 425–444. Springer.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

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

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*.

Zeerak Talat and Dirk Hovy. 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

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.

Cees Van der Eijk. 2001. Measuring agreement in ordered rating scales. *Quality and Quantity*, 35(3):325–341.

Michiel van der Meer, Neele Falk, Pradeep K. Murukannaiah, and Enrico Liscio. 2024. Annotator-centric active learning for subjective NLP tasks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18537–18555, Miami, Florida, USA. Association for Computational Linguistics.

Darinka Verdonik. 2023. Annotating dialogue acts in speech data: Problematic issues and basic dialogue act categories. *International Journal of Corpus Linguistics*, 28(2):144–171.

Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's voice matters: Quantifying annotation disagreement using demographic information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14523–14530.

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

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.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

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

Daniel Zeman. 2010. Hard problems of tagset conversion. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources*, pages 181–185.

Leixin Zhang. 2025. Proposal: From one-fit-all to perspective aware modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1016–1025, Vienna, Austria. Association for Computational Linguistics.

Leixin Zhang, David Burian, Vojtěch John, and Ondřej Bojar. 2024. Unveiling semantic information in sentence embeddings. In *Proceedings of the Fifth International Workshop on Designing Meaning Representations @ LREC-COLING 2024*, pages 39–47, Torino, Italia. ELRA and ICCL.

Leixin Zhang and Çağrı Çöltekin. 2024. Tübingen-CL at SemEval-2024 task 1: Ensemble learning for semantic relatedness estimation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1019–1025, Mexico City, Mexico. Association for Computational Linguistics.

Xinliang Frederick Zhang and Marie-Catherine de Marneffe. 2021. Identifying inherent disagreement in natural language inference. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4908–4915, Online. Association for Computational Linguistics.

Xiang Zhou, Yixin Nie, and Mohit Bansal. 2022. Distributed NLI: Learning to predict human opinion distributions for language reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 972–987, Dublin, Ireland. Association for Computational Linguistics.