Modeling Multi-Rater Behavior with Bayesian Nonparametric MIRT: Inferring Latent Traits and Group Structure

Alex Cucco[1], Lara Fontanella[1], Pasquale Valentini[2], and Sara Fontanella[3]

[1] Department of Socio-Economic, Managerial, and Statistical Studies, G. d'Annunzio University Chieti-Pescara
[2] Department of Economics, G. d'Annunzio University Chieti-Pescara
[3] National Hearth and Lung Institute, Imperial College London, London

Abstract. Human annotation is a key step in data-driven modeling, yet traditional approaches seek consensus among raters, treating disagreement as error and failing to capture the complexity of human interpretation. This has given rise to the perspectivist approach, which explicitly models annotator variability and embraces multiple viewpoints. In this study, we apply a Bayesian Nonparametric Multidimensional Item Response Theory model to multi-rater annotation, adopting a formulation where annotated texts are treated as persons carrying latent traits and annotators function as items. This allows us to automatically identify groups of annotators and assign to each text a set of scores over latent dimensions whose number is inferred directly from the data. We demonstrate the approach through a case study involving social media comments on immigration, annotated independently by multiple raters for the presence of racist content. The model uncovers the structure of annotator heterogeneity, offering a model-based alternative to consensus-based labeling. We identified two distinct annotator clusters with systematically different perspectives, yielding group-specific severity scores. Disagreement was found to concentrate around politically charged language, where the boundary between opinion and hateful rhetoric emerged as contested.

Keywords: perspectivism; annotator heterogeneity; latent trait modeling; Bayesian MIRT; Dirichlet process

1 Introduction

Supervised learning pipelines commonly rely on aggregated annotations, implicitly treating the target label as a single, well-defined quantity. In practice, this target is typically obtained via majority voting, expert adjudication, or other aggregation procedures that collapse multiple judgments into a consensus label [11, 27]. Under this paradigm, inter-rater disagreement is largely operationalized as annotation noise and is therefore reduced or removed prior to model training. In Natural Language Processing (NLP), both traditional feature-based

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_6

approaches [8, 34] and more recent transformer-based architectures [3, 20] are commonly trained and evaluated against these aggregated labels as if they provided an unambiguous ground truth.

For socially embedded and value-laden phenomena, however, the assumption of a unique ground truth label is often conceptually and empirically fragile. Tasks such as hate speech and abusive language detection are not purely technical acts of classification: they require judgments about intent, context, implicit meaning, and the boundaries of social norms. As a consequence, annotator variability can reflect systematic differences in perception shaped by cultural background, personal experience, and social position [2,26]. In such settings, disagreement is not necessarily measurement error; rather, it may constitute informative structure about the phenomenon being studied and about the population of raters.

Motivated by this observation, perspectivist approaches to NLP [12] and data annotation advocate preserving individual judgments instead of enforcing consensus [19, 22]. Under this paradigm, disagreement is treated not as noise but as valuable signal, offering a richer representation of meaning. The analytical focus shifts from approximating a single truth to modeling the structure of rater heterogeneity and understanding patterns of disagreement [6, 31]. In measurement terms, this same shift can be expressed as moving away from estimating a single "true label" toward inferring the latent construct of interest jointly with rater-specific tendencies and the resulting patterns of disagreement. Standard aggregation methods, such as majority voting, risk suppressing minority perspectives by assuming a single gold standard for inherently subjective tasks. In social and sensitive domains such as hate speech detection [12], where the lived experience of targeted groups or the expertise of domain specialists may risk of being suppressed, failing to account for such divergence can result in models that perpetuate rather than mitigate harm. By contrast, a perspectivist approach preserves disagreement as a meaningful signal, ensuring that minority voices, if present, are not statistically erased, but rather explicitly modeled as distinguished and informative perspectives.

In this work, we adopt this measurement-theoretic perspective and cast racism annotation as a latent-trait inference problem. Specifically, we use Item Response Theory (IRT), a class of latent variable models that can be cast within the Item Factor Analysis framework, extending factor analysis to accommodate binary or categorical responses [36]. IRT models express the probability of a response to a given item, measured on a binary or categorical scale, as a function of person parameters (latent traits) and item parameters (e.g., discrimination and threshold). Departing from the canonical orientation in which annotators are treated as respondents, we reverse the mapping: text instances are treated as *persons* and annotators as *items*. This formulation provides a principled approach to model disaggregated judgments while explicitly accounting for rater heterogeneity. To accommodate the possibility that disagreement reflects multiple, qualitatively distinct interpretive axes, we consider multidimensional IRT (MIRT) models [23] and adopt an exploratory approach by fitting a hierarchical Dirichlet Process MIRT model. The model infers the number of latent dimen-

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_6

sions from the data and enforces a simple structure in which each item loads on at most one factor. This yields an automatic partition of annotators into clusters that share similar perspectives and, for each cluster, a latent trait score for every text, interpreted as its position along a racism-severity continuum as defined by that group. Within each cluster, annotator parameters capture both the strength with which each rater contributes to the definition of the continuum (*discrimination*) and the decision *thresholds* mapping latent severity to observed binary labels. Rather than collapsing annotations into a single consensus label, our approach leverages the full response patterns to expose systematic differences in subjective perception, producing interpretable estimates of both instance-level racism and annotator-level response behavior.

2 Related Work

In the absence of an external expert reference, annotation can be framed as a measurement problem in which observed labels are imperfect manifestations of latent structure in both the instances and annotators' response tendencies. IRT is a natural tool in this setting because it separates variability attributable to the instances from variability attributable to annotators. Importantly, IRT has been used in two distinct orientations in the annotation literature.

In the canonical formulation, annotators play the role of respondents (*persons*) with latent ability/expertise, while texts are the *items* characterized by parameters such as difficulty and discrimination. This orientation is typically used to calibrate test sets and to place respondents, human labelers or systems, on a common scale. [35] exemplified this approach with GLAD, which jointly estimates annotator expertise and item difficulty and used these quantities to aggregate labels more robustly than unweighted voting in the presence of heterogeneous annotator quality. [15, 16] likewise fitted 2PL/3PL (2/3 parameters logistic) models to large-scale textual entailment annotations to construct psychometrically calibrated evaluation sets: sentence pairs were treated as items, and item fit/discrimination were used to refine the benchmark and to derive scale-based scores for annotators or systems. [17, 18] extended the same logic beyond human annotation by treating ML classifiers as respondents and dataset instances as items within a 3PL framework. Across these canonical applications, disagreement is primarily used diagnostically, i.e., to identify ambiguous instances and heterogeneous annotator/system behavior, and the model output is often an aggregated item-level decision or a calibrated scale for evaluation.

A different line of work adopts the alternative orientation in which textual instances occupy the person position and receive latent trait scores, while annotators are parameterized as items whose characteristics describe how each rater maps the latent trait into observed labels. This orientation makes rater heterogeneity an explicit component of the measurement model rather than something to be absorbed by averaging. It has been used with different aims. [13] estimated a latent score for each text while characterizing annotators through item parameters, and then discretized the scores to produce a single label per instance, using

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_6

the richer parameterization mainly to improve label aggregation. [1], by contrast, used the same orientation with a graded response model to study and interpret annotator-specific response functions, leveraging rater characteristic curves as a complement to agreement coefficients and shifting the emphasis from "resolving" disagreement to explaining it. A closely related measurement perspective appeared in hate-speech annotation where [24] used many-facet Rasch measurement to place comment hatefulness, survey-item difficulty, and annotator strictness on a common scale, explicitly treating disagreement as perspective that can be quantified rather than eliminated. In a companion study, [25] analyzed differential rater functioning to quantify identity sensitivity, showing systematic differences in ratings depending on whether comments target an annotator's own identity group, illustrating how IRT-style models can support perspectivist analysis by making annotator-related variation directly interpretable. Related applications in other domains, such as lexical offensiveness ratings [30] and music annotation [21], similarly used this orientation to recover continuous latent scores for linguistic or musical objects while explicitly modeling rater-specific differences through discrimination and threshold structures.

Our work follows this alternative orientation within a Bayesian nonparametric MIRT framework that infers the number of annotator clusters directly from the data. This is a deliberate perspectivist choice: the goal is not merely to collapse judgments into a single binary label, but to uncover groups of annotators sharing similar perspectives and, for each group, to estimate where each text falls on a racism severity continuum as defined by that cluster. The resulting estimates provide both group-specific instance scores and directly interpretable annotator parameters within each cluster.

A model capable of recognizing and representing multiple perspectives is essential not only for capturing the inherent subjectivity of tasks such as hate speech detection, but also for building systems that are more fair, inclusive, and socially aware. Furthermore, training NLP models not on a single majority vote but on different latent dimensions is an interesting future direction that goes beyond promoting inclusion, as it opens the possibility of investigating and understanding different opinions; perspectives that, if left unrecognized and unidentified, risk introducing systematic bias into the model, ultimately causing it to perpetuate and reiterate harm.

3 Methods

Given a set of J raters, each providing a binary (e.g., presence/absence) annotation for the same I textual instances, we treat the annotations provided by each rater as a Boolean variable observed across the I instances. Before introducing the Dirichlet Process prior, the multidimensional IRT formulation can be viewed in direct analogy with classical exploratory factor analysis: the latent dimensions are extracted from the dependence structure of the observed response matrix. In other words, the latent annotation behaviors arise as factors that explain systematic patterns of agreement/disagreement across annotators. Within this

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_6

exploratory MIRT perspective, assuming the existence of M potential latent annotation behaviors, the two-parameter normal-ogive (2PNO) multidimensional model for dichotomous responses [28] can be written as

$$\Pr(Y_{ij} = 1 \mid \boldsymbol{\theta}_i, \boldsymbol{\lambda}_j, \gamma_j) = \Phi\Big(\boldsymbol{\lambda}_j^\top \boldsymbol{\theta}_i - \gamma_j\Big),$$

and consequently

$$\Pr(Y_{ij} = 0 \mid \boldsymbol{\theta}_i, \boldsymbol{\lambda}_j, \gamma_j) = 1 - \Phi\Big(\boldsymbol{\lambda}_j^\top \boldsymbol{\theta}_i - \gamma_j\Big),$$

for $i = 1, \ldots, I$ (texts) and $j = 1, \ldots, J$ (annotators), where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function.

In this reversed mapping, $\boldsymbol{\theta}_i \in \mathbb{R}^M$ denotes the latent trait vector for text i, locating the instance in a M-dimensional racism-related latent space. The vector $\boldsymbol{\lambda}_j \in \mathbb{R}^M$ collects the discrimination (loading) parameters for annotator j and quantifies how strongly that annotator's labeling behavior varies with each latent dimension, i.e., how informative the annotator is about the corresponding trait(s). The scalar γ_j is an annotator-specific threshold (often termed *difficulty* in IRT): it determines the location on the latent continuum at which annotator j is equally likely to assign label 1 versus 0, thereby capturing the annotator's response criterion (severity/leniency) for endorsing label 1. To address location and scale indeterminacy, latent traits are identified by fixing their marginal mean to zero and their variance to one. In a Bayesian setting, it is convenient to express the 2PNO MIRT model through a probit data-augmentation scheme, in which each observed binary annotation Y_{ij} is generated by an underlying continuous variable Z_{ij} satisfying

$$Z_{ij} = \boldsymbol{\lambda}_j^\top \boldsymbol{\theta}_i + \epsilon_{ij}, \qquad i = 1, \ldots, I, \;\; j = 1, \ldots, J,$$

where $\epsilon_{ij} \sim \mathcal{N}(0, 1)$ independently for all i and j.

The observed binary response is then obtained via a threshold mechanism:

$$Y_{ij} = \begin{cases} 1, & \text{if } Z_{ij} \geq \gamma_j, \\ 0, & \text{otherwise.} \end{cases}$$

To avoid fixing the number of latent dimensions, M, in advance and to induce an interpretable, sparse structure, we introduce a hierarchical Dirichlet Process prior. In particular, the prior supports an unknown number of latent dimensions (M^*) and, together with a simple-structure constraint, enforces that each annotator loads on at most one dimension (i.e., $\boldsymbol{\lambda}_j$ has at most one nonzero component). As a result, annotators are automatically organized into clusters associated with distinct latent annotation behaviors, while the corresponding text-specific scores, $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iM^*})^\top$, provide cluster-dependent positions of each instance along the inferred racism-severity continuum. The model reparameterizes the loading vector of each item as $\boldsymbol{\lambda}_j = \tilde{\lambda}_j \mathbf{h}_j$, where $\tilde{\lambda}_j > 0$ is a scalar discrimination parameter and $\mathbf{h}_j$ is a binary indicator vector that assigns

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_6

the item to exactly one latent dimension. The allocation vectors $\mathbf{h}_j$ are governed by mixing weights $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{M^*})'$ constructed via a stick-breaking representation of a Dirichlet Process, so that the number of active dimensions is learned from the data rather than specified a priori. A Gamma prior on the concentration parameter of the Dirichlet Process controls the tendency toward fewer or more clusters. Each latent dimension thus groups a subset of items (annotators) that share a common latent trait, and the corresponding person (text) scores are estimated conditionally on this grouping. In practice, by introducing a Dirichlet Process prior as specified, the model automatically explore and identify the number of latent behavior (and annotators groups) that better describe the data, allocating each annotator to groups. The fact that each annotator loads on at most one dimension, gives a hard partition of annotators avoiding a fuzzy situation in which an individual annotator can define different latent behavior.

4 Experimental Setup: Corpus Construction and Annotation Design

The objective of the data collection and corpus construction was to support the analysis of annotation behavior under heterogeneous and potentially contentious content. To this end, we curated a diverse set of social media comments on immigration, collected from Facebook, Instagram, and YouTube between 2014 and 2024. From an initial corpus of 185,734 comments, keyword filtering [10] yielded 39,570 potentially relevant instances. To maximize thematic and affective diversity, we applied Latent Dirichlet Allocation [5] on the relevant instances to infer 20 topics and construct a thematic graph linking comments that share at least one topic. Sentiment polarity and offensiveness of each textual content were then assessed using the Revised HurtLex lexicon [30] and three Italian sentiment lexicons [4,32,33]. This procedure defined six stratification classes obtained by crossing sentiment polarity $\{\text{positive}, \text{neutral}, \text{negative}\}$ with offensiveness $\{\text{offensive}, \text{non-offensive}\}$, yielding the 3×2 cells: {pos×off, pos×non-off, neu×off, neu×non-off, neg×off, neg×non-off}. We subsequently drew a network-based stratified sample [7] of 3,000 comments (500 per cell) via a space-filling design on the thematic graph [9], ensuring coverage of heterogeneous topics, sentiment profiles, and degrees of potentially offensive content. The annotation scheme for the selected comments was designed as a measurement instrument aimed at capturing individual differences in the perception of racism. For each comment, annotators provided judgments on multiple dimensions, including the presence of racism (binary). Ten annotators independently labeled the full set of 3,000 comments in Label Studio [29], with minimal training and no adjudication. Annotators were compensated for their work. This design preserves individual response patterns, which are essential for estimating annotator-specific behavior under our Bayesian Nonparametric MIRT framework.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_6

5 Results

In this section we will briefly describe the annotation outcome and then present the results obtained by applying the Bayesian Nonparametric MIRT model described.

5.1 Content annotation results

Figure 1 shows the distribution of comments according to the number of annotators (out of ten) who labeled them as racist. While 23.8% of comments were unanimously judged as non-racist, only 5.2% received unanimous agreement on the presence of racism. The majority of comments fall in intermediate regions where only a subset of annotators identifies racism, revealing a broad area of interpretative indeterminacy in which raters substantially differ in their perception of racist content. This pattern of widespread disagreement motivates the use of a modeling framework that treats annotator heterogeneity as informative structure rather than noise.

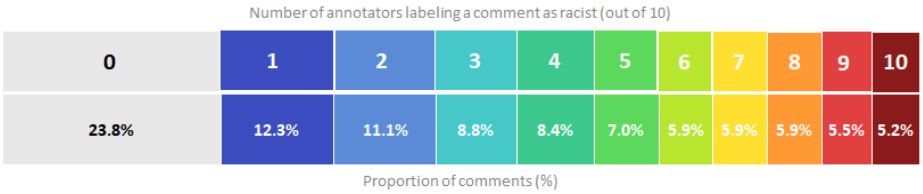

Fig. 1: Percentage distribution of comments by number of annotators indicating the presence of racism.

Considering individual annotator behavior, Figure 2 reveals a high level of heterogeneity in task performance. The number of comments labeled as racist ranges from 379 (Annotator 1 - A1) to 1,575 (Annotator 10 - A10), indicating substantial differences in annotators' propensity to identify racist content. This variability is consistent with the presence of distinct interpretive stances among raters, further supporting the need for a modeling approach that accounts for annotator-specific response tendencies. To quantify the degree of convergence among annotators, we computed Krippendorff's α [14], which is particularly suited to this setting because it accounts for chance agreement and supports any number of annotators. The overall agreement is $\alpha = 0.389$ (see Figure 3 for pairwise values). According to Krippendorff's commonly adopted reference values ($\alpha \geq 0.800$: reliable; $0.667 \leq \alpha < 0.800$: tentative; $\alpha < 0.667$: unreliable), the observed value falls well within the range conventionally interpreted as low agreement. Rather than indicating a failure of the annotation process, this result provides quantitative evidence that the perception of racist content is characterized by substantial and systematic inter-individual variability. In the

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_6

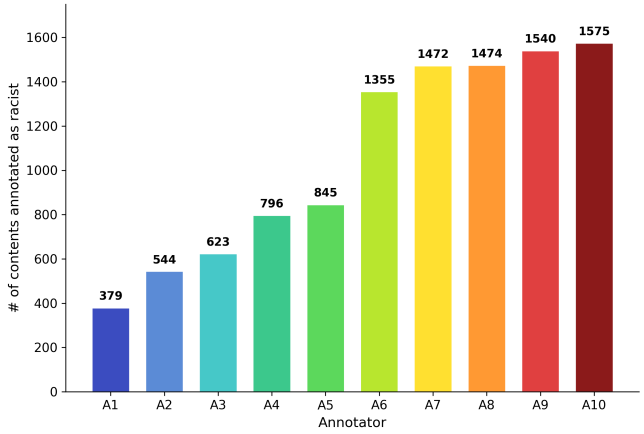

Fig. 2: Contents considered as racist by each annotator.

context of this experiment, low agreement operationalizes the empirical presence of multiple interpretive frames.

Fig. 3: Pairwise Krippendorff's alpha values between annotators.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_6

5.2 Bayesian Nonparametric MIRT model

To investigate the dimensionality and structure of the discrimination parameter matrix $\boldsymbol{\Lambda}$, we compare the performance of the proposed Bayesian Nonparametric model under two specifications: uncorrelated and correlated latent traits. Model fit is evaluated using the Log Pseudo Marginal Likelihood (LPML), a predictive criterion for which larger values indicate better performance. The correlated specification yields a higher LPML (−11,385) compared to the uncorrelated alternative (−14,038), providing clear evidence in favor of allowing dependence among the latent dimensions. The correlated specification allow to identify two latent traits, characterized by a positive correlation of 0.60. Table 1 displays the estimated annotator discrimination parameters $\hat{\boldsymbol{\lambda}}_1$ and $\hat{\boldsymbol{\lambda}}_2$ representing the contribution of each annotator in defining the corresponding trait distinguishing between different latent annotation behavior, along with the estimated threshold parameters $\hat{\boldsymbol{\gamma}}$ encoding the annotator-specific decision boundary, for label 1 endorsement.

Table 1: Annotator parameter estimates with 95% credible intervals.

	$\boldsymbol{\lambda}_1$		$\boldsymbol{\lambda}_2$		$\boldsymbol{\gamma}$	
Annotator	*Est.*	*95% CI*	*Est.*	*95% CI*	*Est.*	*95% CI*
A1	1.47	*[1.37, 1.57]*	0	–	2.03	*[1.85, 2.18]*
A2	0	–	1.53	*[1.45, 1.62]*	1.68	*[1.58, 1.75]*
A3	0	–	1.67	*[1.53, 1.78]*	1.58	*[1.47, 1.67]*
A4	2.00	*[1.88, 2.09]*	0	–	1.39	*[1.33, 1.47]*
A5	2.25	*[2.02, 2.38]*	0	–	1.41	*[1.25, 1.50]*
A6	1.20	*[1.13, 1.27]*	0	–	0.16	*[0.10, 0.21]*
A7	1.47	*[1.36, 1.61]*	0	–	0.03	*[-0.02, 0.07]*
A8	0	–	1.89	*[1.78, 2.01]*	0.03	*[-0.03, 0.10]*
A9	1.68	*[1.61, 1.78]*	0	–	-0.09	*[-0.15, -0.03]*
A10	0	–	1.42	*[1.34, 1.49]*	-0.12	*[-0.17, -0.07]*

The Dirichlet Process prior induces a simple structure in which each annotator loads on exactly one latent dimension, yielding two distinct clusters. Annotators A1, A4, A5, A6, A7, and A9 load on the first factor, while annotators A2, A3, A8, and A10 load on the second. This partition indicates that a single latent continuum is insufficient to capture the full complexity of annotation behavior: the two groups differ systematically in how they map latent racism severity into observed binary labels. Within each cluster, discrimination parameters vary considerably — ranging from 1.20 (A6) to 2.25 (A5) in the first group and from 1.42 (A10) to 1.89 (A8) in the second — reflecting heterogeneous discriminative power even among annotators who share the same interpretive orientation. Threshold parameters further differentiate raters: annotators with high thresholds (e.g., A1, $\hat{\gamma} = 2.03$) require substantially higher latent racist severity before

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_6

assigning a positive label, whereas those with thresholds near zero (e.g., A7 and A8, $\hat{\gamma} = 0.03$) or below (e.g., A9, $\hat{\gamma} = -0.09$; A10, $\hat{\gamma} = -0.12$,) are considerably more inclined to label content as racist.

Figure 4 presents a filled contour plot of the estimated latent trait scores $(\hat{\theta}_1, \hat{\theta}_2)$, interpolated over the two-dimensional score space, where the two axes correspond to the racism severity continua defined by the two annotator clusters. Regions are colored according to the total number of annotators (out of ten) who labeled the comment as racist. Comments in the first quadrant (high $\hat{\theta}_1$, high $\hat{\theta}_2$) are perceived as racist by both groups, while those in the third quadrant (low $\hat{\theta}_1$, low $\hat{\theta}_2$) are consistently judged as non-racist, representing areas of inter-group agreement. The second and fourth quadrants, by contrast, capture comments for which the two clusters diverge: a comment may score high on one dimension but low on the other, indicating that one group perceives racism where the other does not. The overall positive correlation between the two dimensions confirms that, for the majority of comments, the two groups produce broadly consistent assessments. However, the non-negligible spread into the off-diagonal quadrants reveals systematic disagreement on a subset of instances, underscoring the added value of a multidimensional representation over a single consensus score.

To further characterize the content associated with each region of the latent

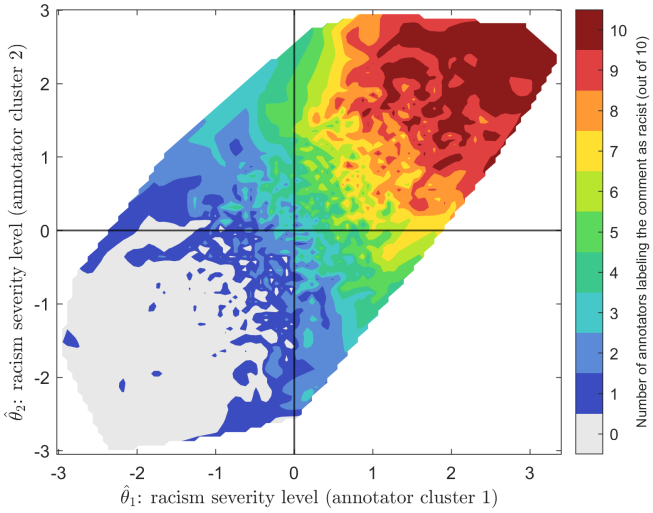

Fig. 4: Filled contour plot of the estimated latent trait scores $(\hat{\theta}_1, \hat{\theta}_2)$ for each comment, colored by the number of annotators (out of ten) who labeled the comment as racist.

space, Figure 5 displays word clouds of the most distinctive terms in each quadrant. The first quadrant, where both clusters agree on the presence of racism, is dominated by offensive language and stereotypical representations. The third

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_6

quadrant, where both clusters agree on the absence of racism, is characterized by a neutral tone associated with news reporting and factual content. The second and fourth quadrants — the areas of inter-group disagreement — show a higher prevalence of politically related terms. Unlike the neutral third quadrant, comments in these conflictual regions tend to exhibit an interpretive and opinionated style, reflecting personal commentary on news events rather than objective reporting. This suggests that annotator disagreement is most pronounced when content blends political discourse with implicit or ambiguous expressions of hostility, where the boundary between legitimate opinion and racist rhetoric is inherently contested.

Fig. 5: Word clouds of the most distinctive terms in each quadrant of the latent trait space, identified by comparison across quadrants.

6 Discussion, Limitations, and Future Work

This study has demonstrated that a Bayesian Nonparametric MIRT framework, applied with an IRT formulation in which texts are treated as persons and annotators as items, can uncover meaningful structure in multi-rater annotation data. By allowing the number of latent dimensions to be inferred from the data, the model identified two distinct clusters of annotators with systematically different

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_6

perspectives on racism, providing group-specific severity scores for each comment along with interpretable annotator parameters. The filled contour representation of the latent space further revealed that disagreement is not uniformly distributed across content: it concentrates in regions characterized by politically charged and interpretive language, where the boundary between legitimate opinion and racist rhetoric is inherently contested.

Some limitations should be acknowledged. First, the annotators involved in this study shared similar socio-demographic characteristics. While the observed disagreement was nonetheless substantial, indicating that perspectives are broad and not perfectly aligned even within a relatively homogeneous group, there is a clear need to develop sampling strategies that ensure adequate coverage of diverse annotator backgrounds and viewpoints.

Second, the current two-parameter normal ogive formulation assumes that all annotators provide genuine judgments for every instance. In practice, annotation tasks may be affected by careless or strategic response behavior, particularly when annotators are unselected or uncompensated. Extending the model to a three- or four-parameter normal ogive specification (3PNO/4PNO) would allow the explicit modeling of such behavior: a lower asymptote capturing random positive labeling regardless of latent severity, and an upper asymptote accounting for the possibility that annotators fail to identify racism even in clearly offensive content, for instance due to fatigue or disengagement. These extensions constitute a natural direction for future work.

More broadly, the framework presented here is not limited to racism detection and can be applied to any multi-rater annotation task where annotator heterogeneity is expected to carry substantive meaning. While the current application focuses on binary judgments, the underlying MIRT formulation naturally extends to ordinal rating scales, such as graded assessments of racism severity or abusiveness levels, through polytomous IRT models (e.g., the graded response model), enabling the same perspectivist analysis in settings where annotations are collected on multi-category scales. Future applications could explore domains such as misinformation detection, sentiment analysis, or other tasks involving socially situated and inherently subjective judgments, where perspectivist modeling may yield richer and more equitable representations of the phenomena under study.

Acknowledgments. This work is part of the research project PRIN-2022 PNRR "Identification and Critical Analysis of Online Racism and Xenophobia against (Im)migrants and Roma people" (Project Code: P2022APKJL), funded by the European Union – Next Generation EU.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Amidei, J., Piwek, P., Willis, A.: Identifying annotator bias: A new IRT-based method for bias identification. In: Scott, D., Bel, N., Zong, C. (eds.) Proceedings

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_6

of the 28th International Conference on Computational Linguistics. pp. 4787–4797. International Committee on Computational Linguistics, Barcelona, Spain (Online) (2020). https://doi.org/10.18653/v1/2020.coling-main.421
2. Aroyo, L., Welty, C.: Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. AI Magazine **36**(1), 15–24 (2015). https://doi.org/10.1609/aimag.v36i1.2564
3. Badjatiya, P., Gupta, S., Gupta, M., Varma, V.: Deep Learning for Hate Speech Detection in Tweets. In: Proceedings of the 26th International Conference on World Wide Web Companion. p. 759–760. WWW '17 Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (2017). https://doi.org/10.1145/3041021.3054223
4. Basile, V., Nissim, M.: Sentiment analysis on Italian tweets. In: Balahur, A., van der Goot, E., Montoyo, A. (eds.) Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. pp. 100–107. Association for Computational Linguistics, Atlanta, Georgia (2013), https://aclanthology.org/W13-1614/
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research **3**(null), 993–1022 (2003), https://dl.acm.org/doi/10.5555/944919.944937
6. Cabitza, F., Campagner, A., Basile, V.: Toward a Perspectivist Turn in Ground Truthing for Predictive Computing. Proceedings of the AAAI Conference on Artificial Intelligence **37**(6), 6860–6868 (2023). https://doi.org/10.1609/aaai.v37i6.25840
7. Cucco, A., del Gobbo, E., Fontanella, L., Fontanella, S., Ippoliti, L.: Covering the Online Spectrum of Opinion in Social Context: The Benefit of Network Node Sampling Through an Italian Case Study. In: Paszynski, M., Barnard, A.S., Zhang, Y.J. (eds.) Computational Science – ICCS 2025 Workshops. ICCS 2025. Lecture Notes in Computer Science. pp. 60–67. Springer Nature Switzerland, Cham (2025). https://doi.org/10.1007/978-3-031-97554-7_5
8. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated Hate Speech Detection and the Problem of Offensive Language. Proceedings of the International AAAI Conference on Web and Social Media **11**(1), 512–515 (2017). https://doi.org/10.1609/icwsm.v11i1.14955
9. del Gobbo, E., Fontanella, L., Ippoliti, L., Di Zio, S., Fontanella, S., Cucco, A.: A Space-Filling Sampling Approach for Collective Classification of Social Media Data. Advances in Data Analysis and Classification (2026). https://doi.org/10.1007/s11634-026-00670-z
10. Fontanella, L., Sarra, A., Del Gobbo, E., Cucco, A., Fontanella, S.: Exploring Anti-Migrant Rhetoric on Italian Social Media. In: Plaia, A., Egidi, L., Abbruzzo, A. (eds.) Proceedings of the Statistics and Data Science 2024 Conference: New Perspectives on Statistics and Data Science. pp. 108–113. Università degli Studi di Palermo, Palermo, Italy (2024)
11. Fortuna, P., Nunes, S.: A survey on automatic detection of hate speech in text. ACM Computing Surveys **51**(4), 1–30 (2018). https://doi.org/10.1145/3232676
12. Frenda, S., Abercrombie, G., Basile, V., Pedrani, A., Panizzon, R., Cignarella, A.T., Marco, C., Bernardi, D.: Perspectivist approaches to natural language processing: A survey. Language Resources and Evaluation **59**, 1719–1746 (2025). https://doi.org/10.1007/s10579-024-09766-4
13. Khattak, F.K., Salleb-Aouissi, A., Raja, A.: Accurate crowd-labeling using item response theory. In: Collective Intelligence (2016)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_6

14. Krippendorff, K.: Content Analysis: An Introduction to Its Methodology. Sage Publications, Thousand Oaks, CA, 4th edn. (2019). https://doi.org/10.4135/9781071878781
15. Lalor, J.P., Wu, H., Yu, H.: Building an Evaluation Scale using Item Response Theory. In: Su, J., Duh, K., Carreras, X. (eds.) Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 648–657. Association for Computational Linguistics, Austin, Texas (2016). https://doi.org/10.18653/v1/D16-1062
16. Lalor, J.P., Wu, H., Yu, H.: Learning Latent Parameters without Human Response Patterns: Item Response Theory with Artificial Crowds. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 4249–4259. Association for Computational Linguistics, Hong Kong, China (2019). https://doi.org/10.18653/v1/D19-1434
17. Martínez-Plumed, F., Prudêncio, R.B.C., Martínez-Usó, A., Hernández-Orallo, J.: Making sense of item response theory in machine learning. In: Proceedings of the Twenty-Second European Conference on Artificial Intelligence. p. 1140–1148. ECAI'16, IOS Press, NLD (2016). https://doi.org/10.3233/978-1-61499-672-9-1140
18. Martínez-Plumed, F., Prudencio, R.B.C., Martínez-Usó, A., Hernández-Orallo, J.: Item response theory in AI: Analysing machine learning classifiers at the instance level. Artificial Intelligence **271**, 18–42 (2019). https://doi.org/10.1016/j.artint.2018.09.004
19. Mostafazadeh Davani, A., Díaz, M., Prabhakaran, V.: Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. Transactions of the Association for Computational Linguistics **10**, 92–110 (2022). https://doi.org/10.1162/tacl_a_00449
20. Mozafari, M., Farahbakhsh, R., Crespi, N.: Hate speech detection and racial bias mitigation in social media based on BERT model. PLOS ONE **15**(8), e0237861 (2020). https://doi.org/10.1371/journal.pone.0237861
21. Nakano, T., Goto, M.: Using Item Response Theory to Aggregate Music Annotation Results of Multiple Annotators. In: Kaneshiro, B., Mysore, G.J., Nieto, O., Donahue, C., Huang, C.Z.A., Lee, J.H., McFee, B., McCallum, M.C. (eds.) Proceedings of the 25th International Society for Music Information Retrieval Conference, ISMIR 2024, San Francisco, California, USA and Online, November 10-14, 2024. pp. 1076–1084 (2024). https://doi.org/10.5281/zenodo.14877519
22. Plank, B.: The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation (2022), https://arxiv.org/abs/2211.02570
23. Reckase, M.D.: Multidimensional Item Response Theory. Statistics for Social and Behavioral Sciences, Springer, New York (2009). https://doi.org/10.1007/978-0-387-89976-3
24. Sachdeva, P., Barreto, R., Bacon, G., Sahn, A., von Vacano, C., Kennedy, C.: The Measuring Hate Speech Corpus: Leveraging Rasch Measurement Theory for Data Perspectivism. In: Abercrombie, G., Basile, V., Tonelli, S., Rieser, V., Uma, A. (eds.) Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022. pp. 83–94. European Language Resources Association, Marseille, France (2022), https://aclanthology.org/2022.nlperspectives-1.11/
25. Sachdeva, P.S., Barreto, R., von Vacano, C., Kennedy, C.J.: Assessing Annotator Identity Sensitivity via Item Response Theory: A Case Study in a Hate Speech

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_6

Corpus. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. p. 1585–1603. FAccT '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3531146.3533216
26. Sap, M., Swayamdipta, S., Vianna, L., Zhou, X., Choi, Y., Smith, N.A.: Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. In: Carpuat, M., de Marneffe, M.C., Meza Ruiz, I.V. (eds.) Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 5884–5906. Association for Computational Linguistics, Seattle, United States (2022). https://doi.org/10.18653/v1/2022.naacl-main.431
27. Schmidt, A., Wiegand, M.: A Survey on Hate Speech Detection using Natural Language Processing. In: Ku, L.W., Li, C.T. (eds.) Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. pp. 1–10. Valencia, Spain (2017). https://doi.org/10.18653/v1/W17-1101
28. Sheng, Y., Wikle, C.K.: Comparing Multiunidimensional and Unidimensional Item Response Theory Models. Educational and Psychological Measurement **67**(6), 899–919 (2007). https://doi.org/10.1177/0013164406296977
29. Tkachenko, M., Malyuk, M., Holmanyuk, A., Liubimov, N.: Label Studio: Data Labeling Software (2020–2025), https://github.com/heartexlabs/label-studio, open source software
30. Tontodimamma, A., Fontanella, L., Anzani, S., Basile, V.: An Italian lexical resource for incivility detection in online discourses. Quality & Quantity **57**, 3019–3037 (2023). https://doi.org/10.1007/s11135-022-01494-7
31. Uma, A., Fornaciari, T., Hovy, D., Paun, S., Plank, B., Poesio, M.: Learning from disagreement: A survey. Journal of Artificial Intelligence Research **72**, 1385–1470 (2021). https://doi.org/10.1613/jair.1.12752
32. Vassallo, M., Gabrieli, G., Basile, V., Bosco, C.: The Tenuousness of Lemmatization in Lexicon-based Sentiment Analysis. In: Bernardi, R., Navigli, R., Semeraro, G. (eds.) Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019). pp. 520–525. CEUR Workshop Proceedings, Bari, Italy (2019), https://aclanthology.org/2019.clicit-1.79/
33. Vassallo, M., Gabrieli, G., Basile, V., Bosco, C.: Polarity Imbalance in Lexicon-based Sentiment Analysis. In: Monti, J., Dell'Orletta, F., Tamburini, F. (eds.) Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020). pp. 334–340. CEUR Workshop Proceedings, Bologna, Italy (Mar 2020), https://aclanthology.org/2020.clicit-1.51/
34. Waseem, Z., Hovy, D.: Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In: Andreas, J., Choi, E., Lazaridou, A. (eds.) Proceedings of the NAACL Student Research Workshop. pp. 88–93. Association for Computational Linguistics, San Diego, California (2016). https://doi.org/10.18653/v1/N16-2013
35. Whitehill, J., Ruvolo, P., Wu, T., Bergsma, J., Movellan, J.: Whose vote should count more: optimal integration of labels from labelers of unknown expertise. In: Proceedings of the 23rd International Conference on Neural Information Processing Systems. p. 2035–2043. NIPS'09, Curran Associates Inc., Red Hook, NY, USA (2009), https://dl.acm.org/doi/10.5555/2984093.2984321
36. Wirth, R.J., Edwards, M.C.: Item factor analysis: Current approaches and future directions. Psychological Methods **12**(1), 58–79 (2007). https://doi.org/10.1037/1082-989X.12.1.58

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_6