Aposteriori Unimodality: Attributing polarization to sociodemographic groups

Dimitris Tsirmpas, John Pavlopoulos

March 2025

1 Introduction

Annotations are essential for a wide range of tasks, especially in fields such as content moderation and toxicity and racism detection. Annotations for these tasks are especially essential for training systems that protect vulnerable and minority groups in online spaces, where they are often targeted [22, 7]. However, systems trained on such data frequently fail. One of the reasons for this failure is that if a comment targets marginalized groups, there can be cases where most annotators (belonging by definition in the majority groups) overlook it, while the few marginalized annotators vehemently disagree in vain. This issue exists even in representative samples. An example of such cases would be racist comments that are presented with coded language (usually referred to as a "dog-whistle" [17]), which are not picked up by most annotators, but only by ones belonging in the specific, targeted minority.

Inter-annotator disagreement is often used to detect such cases [12]. Polarization [15, 16] is a better instrument in this regard, since it detects multi-modal annotation distributions, which are typically present in such cases. However, polarization can be caused by a number of factors, and there is currently no work attributing it to specific annotator subgroups. Pavlopoulos and Likas [16] propose such a mechanism which only tests for the extreme edge-case where total polarization is zero, which is not the case in real-world data. There are also approaches such as tampering with the sample [8], creating subgroup-specific classification problems [1], or parameterizing the distribution of annotations [4]. These approaches however are only useful for downstream tasks, and can not be used to explain differences in annotation between groups.

In this paper, we propose the "Aposteriori Unimodality Measure", a statistic that attributes polarization to individual annotator sociodemographic groups. We provide a formulation which avoids inherent statistical limitations, as well as issues not yet acknowledged in literature, such as the presence of inherent polarization and conflicting polarization directions. We apply our test to four datasets; two with human comments and annotations [19, 13] as well as two generated and annotated by Large Language Model (LLM) agents [21]. We find interesting patterns in polarization in the human datasets, and verify current findings in literature w.r.t. LLM SocioDemographic Background (SDB) prompts.

2 Methodology

In this section, we provide a formal mathematical formulation for the problem of attributing polarization to specific annotator characteristics (§2.1), and offer an intuitive rationale for how established polarization metrics can be leveraged (§2.2). We then introduce a data-point-level statistic that attributes polarization to individual SDB groups (§2.3), and subsequently develop a statistic that generalizes this mechanism to an entire dataset.

2.1 Problem Formulation

SocioDemographic Backgrounds Since our goal is to pinpoint which specific characteristics contribute to polarization, we need to isolate individual groups within a SDB. Let Θ be one of multiple "dimensions" the set of all annotator SDB groups for a single dimension (e.g. "male", "female" and "non-binary" if we are investigating annotator gender).

Annotations Now let $d = \{c_1, c_2, \ldots\}$ be a dataset composed of multiple annotated data-points. We assume that annotating a data-point depends on three variables: its contents, the annotator's SDB, and uncontrolled factors such as mood and personal experiences. Assuming that each data-point c is assigned multiple annotators, we can define the annotation multi-set $A(c) = \{(a, \theta)\}$, where a is a single annotation. As an example, the annotations for a comment in a sentiment analysis task would be formulated as:

$$\begin{gathered} A(\textit{"Could be better, could be wose"}) = \\ \{(\textit{positive}, female), (\textit{positive}, \textit{male}), (\textit{negative}, female), \ldots\} \end{gathered} \tag{1}$$

Polarization Each set of annotations features a certain degree of *polarization.* Unpolarized annotations are usually unimodal, which makes intuitive sense; there is a difference between the annotators disagreeing on the details (which would be shown roughly as a bell curve around the median annotation), and them fundamentally disagreeing with each other (which would likely be shown as a multimodal distribution). Pavlopoulos and Likas [16] create a polarization metric, the "normalized Distance From Unimodality (nDFU)", which measures whether an annotation set shows no polarization (unimodal distribution — $nDFU \to 0$), up to complete polarization (multimodal distribution — $nDFU \to 1$). Thus, we will define a test that, given that annotation polarization exists, tests whether the nDFU of a data-point's annotations can be partially explained by θ.

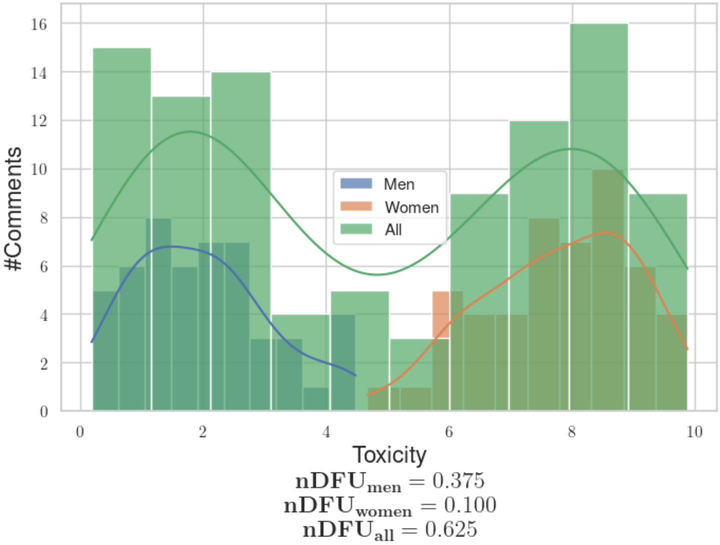

Figure 1: Hypothetical example of a polarizing comment, where male and female annotators agree between themselves, but disagree with the opposite gender. Overall polarization ($nDFU_{all} = 0.625$) is much greater than the polarization exhibited by the annotations grouped by gender ($nDFU_{men} = 0.3725$, $nDFU_{women} = 0.1$).

2.2 Quantifying changes in polarization

Data-point polarization Intuitively, θ partially explains the polarization of a data-point c when the annotations grouped by θ are more polarized compared to the full set of annotations. Figure 1 exhibits a hypothetical example where a misogynistic comment is annotated for toxicity by male and female annotators.[1] The annotations are generally polarized ($nDFU_{all} = 0.625$). The annotations by female annotators exhibit low polarization between themselves ($nDFU_{women} = 0.1$), since most agree the data-point is toxic. The set of male annotations, also shows low polarization ($nDFU_{men} = 0.3725$), but for the opposite reason—most men agree that it is *not* toxic. This suggests that the overall polarization is driven by disagreements between male and female annotators.

Dataset polarization Given this observation, we would be tempted to aggregate all annotations in the dataset and compare the polarization of each θ with the full set of annotations $A(c)$. However, this formulation would not work well. Figure 2 illustrates a hypothetical discussion with two comments, both of which are toxic, but where male and female annotators disagree on *which* com-

[1] The use of only two predominant genders is made only for the purposes of demonstration.

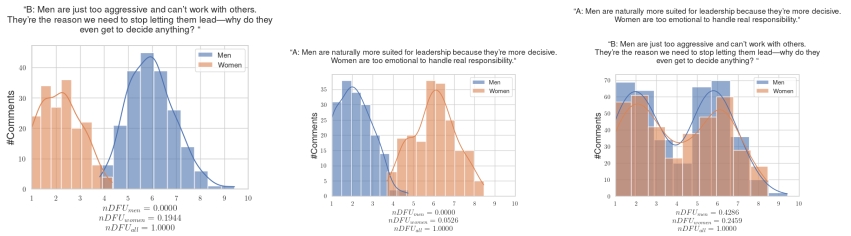

Figure 2: Hypothetical example of a polarizing discussion with two comments, for which the annotators disagree on which is the toxic one. If we aggregate the two comments, the polarization scores for both men and women significantly rise, obscuring whether the exhibited polarization can be partially attributed to gender.

ment is the toxic one. If we aggregate the annotations to a single discussion, the opposing polarization effects might balance each other out, leading to a false negative. In simpler terms, polarization has a *direction*, and care must be taken to not combine polarization effects of opposite directions.

In our example, while it is obvious that gender partly explains the polarization found in each of the individual comments ($nDFU_{all} \gg nDFU_{men}$, $nDFU_{all} \gg nDFU_{women}$), this observation is much harder to make when aggregating the two comments. This is graphically shown as the presence of three bimodal distributions instead of two unimodal distributions, and one bimodal (aggregated) distribution. To avoid this, we apply our statistic only on annotations that reference the same data-point.

2.3 The pol-statistic

As demonstrated above, it is necessary to define a statistic for each individual data-point. We thus introduce the *"pol-statistic"*, which quantifies the change in nDFU when annotations are grouped by $\theta \in \Theta$.

In order to define this statistic, we need to measure (1) the polarization of a data-point when grouped by θ, and (2) the "inherent" polarization present in the annotations, which is driven by either the contents of the data-point, or by the uncontrolled factors mentioned in §2.1. Note that we can not directly compare the θ-subset of annotations with the full set of annotations, since any statistical comparison between them violates the assumption of independent samples.

The observed polarization of a data-point on the group of annotations θ can be directly computed as:

$$pol(c, \theta) = nDFU(P(A(c), \theta) \tag{2}$$

where $P(A(c), \theta_i) = \{a(c;\theta) \in A(c) | \theta = \theta_i\}$ is the partition of A for a data-point c, for which its annotators belong to the SDB group θ_i.

We can estimate the inherent polarization in data-point c by bootstrapping. We randomly partition the data-point's annotations in $|\Theta|$ groups, with matching group sizes (e.g., if we have 100 annotations, 80 of which are made by male annotators and 20 by female annotators, we will create random partitions of sizes 80 and 20). The randomly partitioned annotations are then sampled t times. Formally, the expected inherent polarization will be given by:

$$\frac{1}{t}\sum_{i=1}^{t} nDFU(\tilde{P}_i(A(c))) \tag{3}$$

where $\tilde{P}_i$ is the random partition operator with matching group sizes.

2.4 Aposteriori Unimodality

By obtaining the pol-statistics for all data-points in a dataset d for SDB θ, we can get the mean observed polarization for the dataset:

$$pol_O(d,\,\theta) = \frac{1}{|d|}\sum_{c\in d}(pol(c,\,\theta) - \frac{1}{t}\sum_{i=1}^{t} nDFU(\tilde{P}_i(A(c)))) \tag{4}$$

and similarly the mean apriori dataset polarization:

$$pol_E(d) = \frac{1}{|d|}\sum_{c\in d}\sum_{i=1}^{t} nDFU(\tilde{P}_i(A(c))) \tag{5}$$

We can then define the Aposteriori Unimodality Statistic (κ) with a formulation inspired by Cohen's Kappa [6], as:

$$\kappa = apunim(d,\,\theta) = \frac{pol_O(d,\,\theta) - pol_E(d)}{1 - pol_E(d)} \tag{6}$$

The statistic is technically unbounded, but typically resides in the $[-1,\,1]$ range. If $\kappa \approx 0$, the exhibited polarization among annotators of group θ does not surpass what would be expected by chance. If $\kappa > 0$, then the group partially explains a rise in polarization. Unlike Cohen's Kappa, the case of $\kappa < 0$ does have meaning; the group actually exhibits lower polarization compared to the whole. The scope of the statistic and its relationship with the previously presented statistics are demonstrated in Figure 3.

Like most metrics, we also include a p-value alongside the κ value. This can be computed either non-parametrically, by repeatedly shuffling annotations and computing a null distribution of *apunim*, or parametrically, by assuming a normal distribution (which is generally a safe assumption if $t \geq 30$). In the latter case, we compute:

$$z = \frac{\hat{\kappa} - \mathbb{E}[\kappa]}{SE(\hat{\kappa})},\, p = 1 - \Phi(z) \tag{7}$$

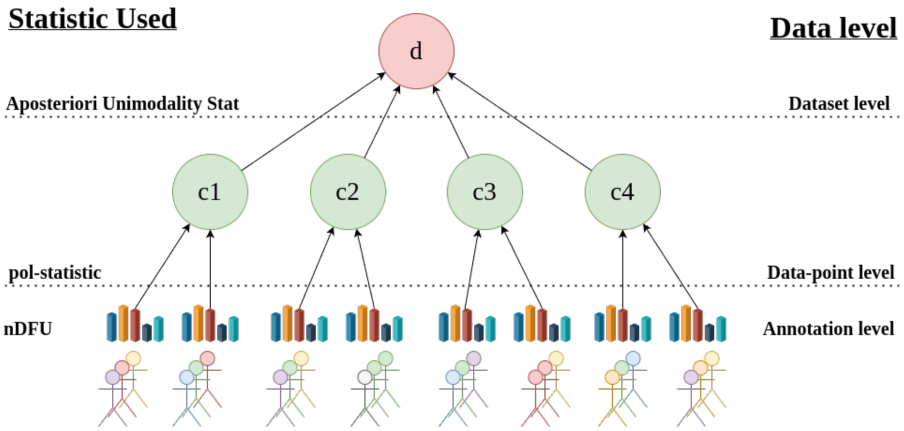

Figure 3: An overview of the Aposteriori Unimodality Test. We gather statistical information from the annotations (different SDBs are denoted by different colors) via the nDFU measure, which is aggregated on the data-point-level by the pol-statistic. The Aposteriori Unimodality Test is applied on the discussion level, operating on the pol-statistics of the individual data-points.

2.5 Technical Details

Since we are simultaneously considering $|\Theta|$ hypotheses, we apply a multiple comparison correction to the resulting p-values. We choose the Holms method [10]. The test is parameterized by the Family-Wise Error Rate (FWER), which is used to tune the strength of the correction; we can increase this value to make our test more conservative towards multiple hypotheses [5]). In general, it is safe to set FWER equal to the significance level of our test (e.g., $FWER = 0.95$ if we are looking for $p < 0.05$).

3 Results

We apply our test on four datasets; two human-annotated datasets, and two with comments and annotations generated by LLMs.

3.1 Human Datasets

We use the datasets provided by Kumar et al. [13] and Sap et al. [19]. The datasets feature various online comments, each annotated by a number of human annotators (5 and $4-6$ annotators per comment respectively). The Sap et al. [19] dataset includes racism annotations for 626 Twitter/X comments, and standard SDB information (age, race, education, gender). The Kumar et al. [13] dataset measures toxicity on various online comments, and the provided SDBs include mostly annotator experiences (e.g., whether they have been personally targeted online) as well as standard SDB information such as sexual orientation, age and education. Since the dataset is extensive ($XXXX$ tweets), we only select a sample of 5000 comments. This is due both to computational constraints, as well as statistical tests being unreliable on large enough samples [20]. Assuming a confidence level of $\alpha = 0.05$ we test whether any of the provided SDB dimensions can partially explain polarization in the toxicity/racism annotations.

With respect to the Kumar et al. [13] dataset, we find statistically significant differences for Conservatives ($p = 0.0486$), parents ($p = 0.000039$), religious people ("religion: very important" — $p = 0.00073$), young people ("ages: 25-35" — $p = 0.001681$), as well as people that have not encountered toxic content ($p = 0.007374$), and people who have been targeted in the past ($p = 0.004186$). We can not claim that transgender annotators, annotators with differing stances on whether toxic comments are problematic, or annotators with different education or sexual orientation, contribute to the polarization within the dataset.

Concerning the Sap et al. [19] dataset, we find statistically significant results for white ($p = 6.531785e - 08$), and male ($p = 0.01289$) annotators, but not for age or educational level.

3.2 Synthetic Datasets

We use two synthetic datasets; one is the Virtual Moderation Dataset (VMD) presented in Tsirmpas, Androutsopoulos, and Pavlopoulos [21]. This dataset features 140 discussions, each having 14 comments (and usually up to 14 facilitator comments), each comment annotated by 10 LLM annotators, each supplied with a different SDB. The second dataset, is an extension of VMD, where we select four random unmoderated discussions, and employ 100 distinct LLM annotators.

We find no statistically significant results in any of the SDB dimensions, in any of the datasets (annotator age, gender, sexual orientation, education, employment and political alignment). While polarization does exist in the dataset, it can not be attributed to any of the synthetic SDB groups. This suggests that SDB prompting can not be used to simulate annotations for human groups, which is consistent with relevant literature on the topic [2, 9, 18, 11, 3, 14].

3.3 Effect of number of annotators

In §2.4 we mentioned that the test relies on enough annotations for each SDB group and data-point. This can be an issue, given the cost required to utilize multiple human annotators for each data-point [18]. Figure 4 demonstrates the effect of the number of annotators to the *pol-statistic* estimation. We use the 100-annotator synthetic dataset and sample progressively $3 - 100$ annotators with replacement, then calculate the standard error with the mean *pol-statistic*. As expected, the standard error is inversely proportional to the number of annotators, although the difference is not great, even for a low number of annotators.

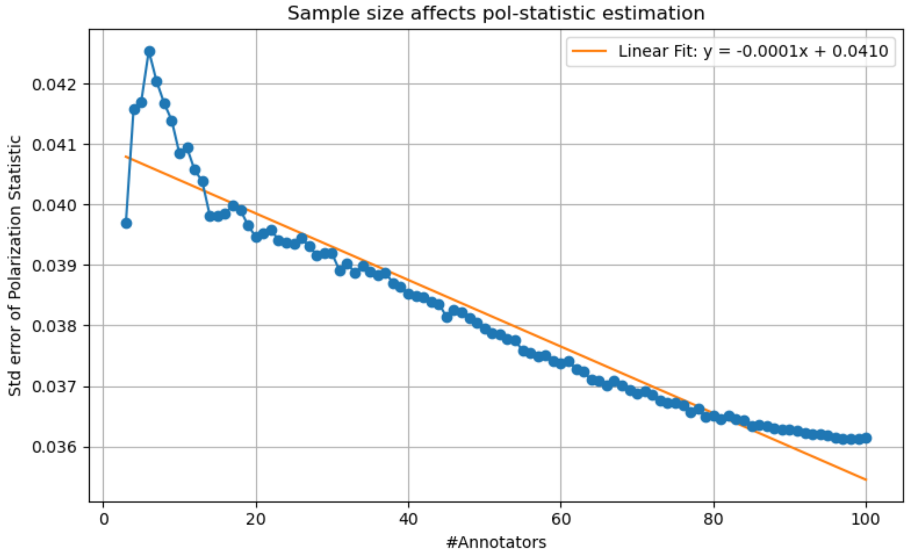

Figure 4: Standard error of the *pol-statistic* when sampled with replacement from the 100-annotator synthetic dataset for various number of annotators.

4 Conclusion

We introduced the “Aposteriori Unimodality Test”, a statistical test that detects whether certain sociodemographic groups partly cause polarization in annotations. Our test is resistant to low samples of annotations, bypasses common statistical issues such as sample dependence, and avoids newly-discovered issues such as the presence of inherent polarization and polarization direction. We apply our test two human-generated and two LLM-generated datasets and find interesting patterns on the former, and verify current literature on LLM SDB prompting on the latter.

5 Limitations

Since there are no other established tests that attribute polarization to sociodemographic characteristics, there is no way to verify that the results presented in §3 are accurate. Similarly, there is no qualitative way of evaluating our test, other than observations on the (non-) efficacy of LLM SDB prompting and our own intuition. Lastly, while our test seems to work sufficiently well with a small number of annotators per data-point, we still encourage researchers using this test to have at least a few annotators of each sociodemographic group that is being tested, in their dataset.

6 Ethical Considerations

While our works aims to help protect marginalized and disadvantaged groups by attributing polarization to certain subgroups, it can be taken advantage of by malicious actors. Given a dataset with fine-grained SDB information, these actors can use our test to target specific vulnerable groups. We thus urge researchers to keep datasets including such information protected, and provided to others only under explicit permission.

References

[1] Sohail Akhtar, Valerio Basile, and Viviana Patti. "Modeling Annotator Perspective and Polarized Opinions to Improve Hate Speech Detection". In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 8 (Oct. 2020), pp. 151–154. DOI: 10.1609/hcomp.v8i1.7473.

[2] Jacy Reese Anthis et al. *LLM Social Simulations Are a Promising Research Method*. 2025. arXiv: 2504.02234 [cs.HC]. URL: https://arxiv.org/abs/2504.02234.

[3] James Bisbee et al. "Synthetic Replacements for Human Survey Data? The Perils of Large Language Models". In: *Political Analysis* 32.4 (2024), 401–416. DOI: 10.1017/pan.2024.5.

[4] Silvia Casola et al. "Confidence-based Ensembling of Perspective-aware Models". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 3496–3507. DOI: 10.18653/v1/2023.emnlp-main.212. URL: https://aclanthology.org/2023.emnlp-main.212/.

[5] S.Y. Chen, Z. Feng, and X. Yi. "A general introduction to adjustment for multiple comparisons". In: *Journal of Thoracic Disease* 9.6 (2017). doi: 10.21037/jtd.2017.05.34; PMID: 28740688; PMCID: PMC5506159, pp. 1725–1729.

[6] Jacob Cohen. "A Coefficient of Agreement for Nominal Scales". In: *Educational and Psychological Measurement* 20.1 (1960), pp. 37–46. DOI: 10.1177/001316446002000104. eprint: https://doi.org/10.1177/001316446002000104. URL: https://doi.org/10.1177/001316446002000104.

[7] Shelby Dioum. *What Explains the Increase in Online Hate Speech?* Accessed: 2025-09-22. University of California, Davis, Mar. 2025. URL: https://www.ucdavis.edu/magazine/what-explains-increase-online-hate-speech.

[8] Stephanie Eckman et al. *Aligning NLP Models with Target Population Perspectives using PAIR: Population-Aligned Instance Replication*. 2025. arXiv: 2501.06826 [stat.ME]. URL: https://arxiv.org/abs/2501.06826.

[9] Luke Hewitt et al. "Predicting Results of Social Science Experiments Using Large Language Models". Equal contribution, order randomized. Aug. 2024.

[10] Sture Holm. "A Simple Sequentially Rejective Multiple Test Procedure". In: *Scandinavian Journal of Statistics* 6.2 (1979), pp. 65–70. ISSN: 03036898, 14679469. URL: http://www.jstor.org/stable/4615733 (visited on 07/07/2025).

[11] Bernard J. Jansen, Soon gyo Jung, and Joni Salminen. "Employing large language models in survey research". In: *Natural Language Processing Journal* 4 (2023), p. 100020. ISSN: 2949-7191. DOI: https://doi.org/10.1016/j.nlp.2023.100020. URL: https://www.sciencedirect.com/science/article/pii/S2949719123000171.

[12] Petra Kralj Novak et al. "Handling Disagreement in Hate Speech Modelling". In: *Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Ed. by Davide Ciucci et al. Cham: Springer International Publishing, 2022, pp. 681–695. ISBN: 978-3-031-08974-9.

[13] Deepak Kumar et al. "Designing toxic content classification for a diversity of perspectives". In: *Proceedings of the Seventeenth USENIX Conference on Usable Privacy and Security*. SOUPS'21. USA: USENIX Association, 2021. ISBN: 978-1-939133-25-0.

[14] Terrence Neumann, Maria De-Arteaga, and Sina Fazelpour. *Should you use LLMs to simulate opinions? Quality checks for early-stage deliberation*. 2025. arXiv: 2504.08954 [cs.CY]. URL: https://arxiv.org/abs/2504.08954.

[15] John Pavlopoulos and Aristidis Likas. "Distance from Unimodality for the Assessment of Opinion Polarization". In: *Cognitive Computation* 15.2 (2023), pp. 731–738. ISSN: 1866-9964. DOI: 10.1007/s12559-022-10088-2. URL: https://doi.org/10.1007/s12559-022-10088-2.

[16] John Pavlopoulos and Aristidis Likas. "Polarized Opinion Detection Improves the Detection of Toxic Language". In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Yvette Graham and Matthew Purver. St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 1946–1958. URL: https://aclanthology.org/2024.eacl-long.117/.

[17] Anne Quaranto. "Dog whistles, covertly coded speech, and the practices that enable them". In: *Synthese* 200.4 (2022), p. 330.

[18] Luca Rossi, Katherine Harrison, and Irina Shklovski. “The Problems of LLM-generated Data in Social Science Research”. In: *Sociologica* 18.2 (2024), 145–168. DOI: `10.6092/issn.1971-8853/19576`. URL: `https://sociologica.unibo.it/article/view/19576`.

[19] Maarten Sap et al. “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*. Ed. by Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 5884–5906. DOI: `10.18653/v1/2022.naacl-main.431`. URL: `https://aclanthology.org/2022.naacl-main.431/`.

[20] David Trafimow et al. “Manipulating the Alpha Level Cannot Cure Significance Testing”. In: *Frontiers in Psychology* 9 (2018), p. 699. DOI: `10.3389/fpsyg.2018.00699`.

[21] Dimitris Tsirmpas, Ion Androutsopoulos, and John Pavlopoulos. *Scalable Evaluation of Online Facilitation Strategies via Synthetic Simulation of Discussions*. 2025. arXiv: `2503.16505 [cs.HC]`. URL: `https://arxiv.org/abs/2503.16505`.

[22] United Nations. *Targets of Hate*. Accessed: 2025-09-22. Sept. 2025. URL: `https://www.un.org/en/hate-speech/impact-and-prevention/targets-of-hate`.

A Appendix

A.1 Acronyms

VMD	Virtual Moderation Dataset
LLM	Large Language Model
SDB	SocioDemographic Background
nDFU	normalized Distance From Unimodality
FWER	Family-Wise Error Rate

A.2 Demonstrative examples

Our paper uses synthetic examples to demonstrate the intuition behind the Aposteriori Unimodality test. These examples use random sampling of manually selected distributions. Table 1 summarizes the synthetic annotation distributions used to demonstrate the Aposteriori Unimodality test. In each scenario, annotations for men and women were sampled independently from Gaussian distributions with distinct parameters.

Figure	**Group**	**Distribution**	**Size**
Figure 1	A_{men}	$\mathcal{N}(2,\,1.3)$	50
	A_{women}	$\mathcal{N}(8,\,1.3)$	50
Figure 2 (1st comment)	A_{men}	$\mathcal{N}(6,\,1)$	200
	A_{women}	$\mathcal{N}(2,\,1)$	200
Figure 2 (2nd comment)	A_{men}	$\mathcal{N}(2,\,1)$	200
	A_{women}	$\mathcal{N}(6,\,1)$	200

Table 1: Synthetic annotation sets used to illustrate the Aposteriori Unimodality test. Each group corresponds to samples drawn from a Gaussian distribution with specified mean and standard deviation.