Aposteriori Unimodality

Dimitris Tsirmpas, John Pavlopoulos

March 2025

1 Methodology

1.1 Problem Formulation

Let $\{c(d, 1), c(d, 2), \ldots\}$ be the comments[1] of a discussion d. We assume that annotating a comment depends on three variables: its contents, the annotator's SocioDemographic Background (SDB), and uncontrolled factors such as mood and personal experiences. Assuming that each comment is assigned multiple annotators, we can define the annotation set $A(d, i)$ for comment $c(d, i)$ as:[2]

$$A(c) = \{a(c, \theta) \mid i = 1, 2, \ldots, |d|, \theta \in \Theta\} \tag{1}$$

where $a(c, \theta)$ is a single annotation for comment $c(d, i)$ and Θ is the set of annotator SDBs.

Since our goal is to pinpoint which specific characteristics contribute to polarization, we need a way to isolate individual attributes within a SDB. Θ is usually composed of multiple "dimensions" (e.g., age, sex, educational level), each of which is split between various groups. We can thus model $\theta \in \Theta$ as:

$$\theta = \{(dim, gr) \mid dim \in Dims, gr \in Groups(dim)\} \tag{2}$$

where $Dims = \{dim_1, dim_2, \ldots, dim_k\}$ is the set of SDB dimensions, and $Groups$ is the set of possible groups for dimension dim (e.g., $Groups(dim_{gender}) = \{male, female, \ldots\}$).

1.2 Quantifying changes in polarization

The mechanism defined in §1.1 allows us to isolate the effects of each SDB dimension *dim*, but we still lack a mechanism with which to analyze that effect. In this section, we present the "pol-statistic" (polarization statistic) as a comment-level tool which not only attributes polarization to a dimension *dim*, but also to specific groups within that dimension ($gr \in Groups(dim)$).

Intuitively, dimension *dim* partially explains polarization when the annotations divided according to each group $gr \in Groups(dim)$, show less polarization

[1] Also referred to as "dialogue turns" in some publications.

[2] For the sake of simplicity, we use c to mean any comment $c(d, i)$.

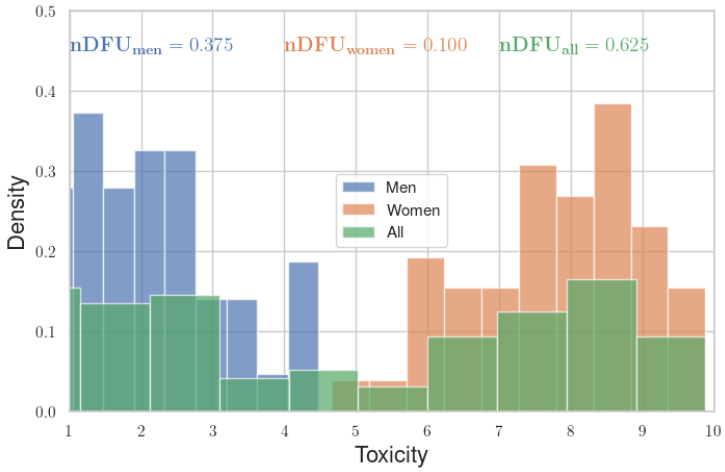

Figure 1: Hypothetical example of a polarizing comment, where male and female annotators agree between themselves, but disagree with the opposite gender. Overall polarization ($nDFU_{all} = 0.625$) is much greater than the polarization exhibited by the annotations grouped by gender ($nDFU_{men} = 0.3725$, $nDFU_{women} = 0.100$). In this example, the annotation set A was generated as $A_{men} \sim \mathcal{N}(2, 1.3)$, $A_{women} \sim \mathcal{N}(8, 1.3)$, $|A_{men}| = |A_{women}| = 50$.

between each group 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. The annotations are generally polarized ($nDFU_{all} = 0.625$), although the set of annotations from female annotators might exhibit low polarization between themselves ($nDFU_{women} = 0.100$—most agree the comment is toxic). The set of male annotations on the other hand, 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.

Given this observation, we would be tempted to aggregate all annotations for each comment in a discussion, and test whether intra-group polarization is greater than global polarization. However, this formulation would not work well, as illustrated in Figure 2. It features a hypothetical discussion with two comments, both of which are toxic, but where male and female annotators disagree on *which* comment is the toxic one. If we aggregate the annotations for the two comments, the opposing polarization effects might balance each other out, leading to a false negative. 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. To avoid this, we apply our statistic only on annotations that reference the same comment.

1.3 The pol-statistic

The comment-level polarization statistic for group $gr \in dim$ is obtained by:

$$pol_{actual}(c,\ dim,\ gr) = nDFU(P(A(c),\ dim,\ gr)) \tag{3}$$

where $P(A(c),\ dim,\ gr) = \{a(c,\ \theta) \in A | (dim,\ gr) \in \theta\}$ is the partition of A for a comment c, for which its annotators belong to the group gr of SDB dimension dim.

If the polarization in a comment c is driven by group g, we would expect polarization inside the groups to be higher than expected. We can estimate the expected polarization in comment c between groups $gr \in Groups(dim))$ by randomly splitting the comment's annotations in $|Groups(dim)|$ groups, with each random group having size equal to that of the observed groups. For instance, 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. We can then sample the randomly partitioned annotations t times. Formally, we define it as:

$$pol_{expected}(c,\ dim,\ gr) = \frac{1}{t}\sum_{i=1}^{t} nDFU(\tilde{P}_i(A(c))) \tag{4}$$

where each $\tilde{P}_i(A(c),\ dim,\ gr)$ is a random partition of the annotation set $A(c)$ into $|Groups(dim)|$ subsets with sizes matching the original, individual partitions $(|\tilde{P}_i(A,\ dim,\ gr)| = |P(A,\ dim,\ gr)|,\ \forall gr \in Groups(dim),\ i = 1,\ \ldots,\ t)$.

The difference between the observed and expected polarization for a comment c given group $gr \in Groups(dim)$ is then given by the "pol-statistic", defined as:

$$pol(c,\ dim,\ gr) = pol_{actual}(c,\ dim,\ gr)) - pol_{expected}(c,\ dim,\ gr) \tag{5}$$

1.4 The Aposteriori Unimodality Test

By obtaining the pol statistics for all comments in a discussion d we can apply a mean test with the null hypothesis :

$$H_0 : \forall gr \in Groups(dim),\ \frac{1}{|d|}\sum_{c \in d} pol(c,\ dim,\ gr) = 0 \tag{6}$$

versus the (multiple) alternative hypotheses:

$$H_{gr} : \frac{1}{|d|}\sum_{c \in d} pol(c,\ dim,\ gr) > 0 \tag{7}$$

We refer to this test as the *"Aposteriori Unimodality Test"*, where a small p-value suggests that we can not rule out that annotators of the gr group make a significant contribution to the overall annotator polarization. The scope of the test and its relationship with the previously presented statistics are demonstrated in Figure 3.

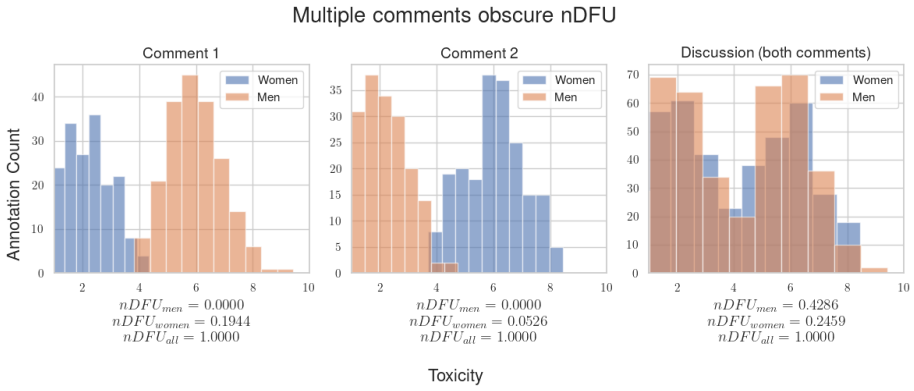

Figure 2: Hypothetical example of a polarizing discussion with two comments, for which the annotators disagree which is the toxic one. If we aggregate the two comments, the polarization scores for both men and women significantly rise, obscuring whether the exhibited polarization can be partially attributed to gender. In this example, the annotation set A was generated as $A_{men} \sim \mathcal{N}(6, 1)$, $A_{women} \sim \mathcal{N}(2, 1)$, $|A_{men}| = |A_{women}| = 200$ for the first comment, and $A_{men} \sim \mathcal{N}(2, 1)$, $A_{women} \sim \mathcal{N}(6, 1)$, $|A_{men}| = |A_{women}| = 200$ for the second.

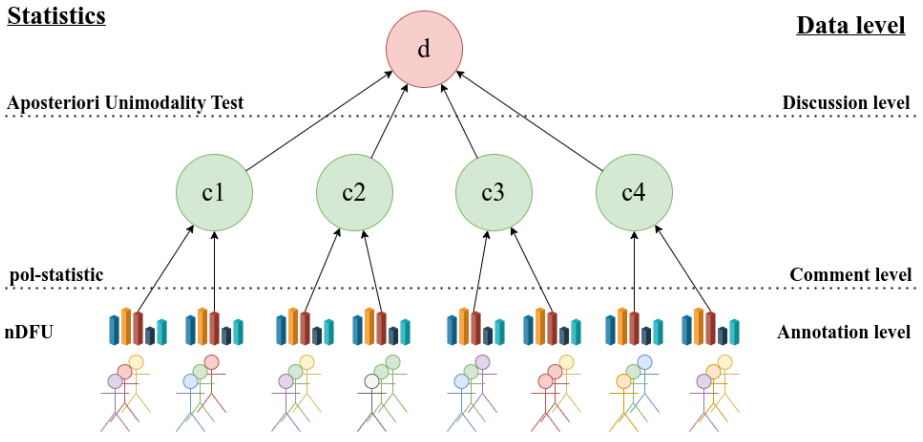

Figure 3: An overview of the Aposteriori Unimodality Test. We gather statistical information from the annotations (different SDBs are denoted by different colors) via the nDFU measure, which is aggregated on the comment-level by the pol-statistic. The Aposteriori Unimodality Test is applied on the discussion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

1.5 Technical Details

The means test performed on the pol-statistics of each comment (Equations 6, 7) can be theoretically performed by most well-known statistical mean tests. In our case, we use the one one-sample Student t-test. Additionally, since we are simultaneously considering $|Groups(dim)|$ hypotheses, we apply a multiple comparison correction to the resulting p-values. We choose the Bonferroni method [1], since it is widely used, generally conservative, and especially so towards correlated hypotheses [2]. The last point is important, since many annotation groups are likely to be inter-correlated (e.g., age categories such as $50-60$ and $60-70$ years). Furthermore, one of its biggest weaknesses (the presence of very large numbers of hypotheses [2]) is unlikely to be met in annotation tasks.

The Aposteriori Unimodality Test is actually parameterized by two more parameters: the t parameter, which determines how many times we sample random partitions (Equation 4), and the Family-Wise Error Rate (FWER), which is used to tune the strength of the multiple comparison correction mentioned above. We can increase the t parameter to get a better estimate of the expected comment polarization, while increasing the computational cost of the method. We can also increase the FWER to make our test more conservative towards multiple hypotheses [2]). 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 $p < 0.05$)

2 Acronyms

SDB SocioDemographic Background

nDFU normalized Distance From Unimodality

FWER Family-Wise Error Rate

References

[1] J Martin Bland and Douglas G Altman. "Multiple significance tests: the Bonferroni method". In: *BMJ* 310.6973 (1995), p. 170. ISSN: 0959-8138. DOI: 10.1136/bmj.310.6973.170. eprint: https://www.bmj.com/content/310/6973/170.full.pdf. URL: https://www.bmj.com/content/310/6973/170.

[2] 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.