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:

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

where $a(d, i, \theta)$ 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:

$$\theta = \{(\xi_i, g) \mid i = 1, 2, \ldots, k, g \in G_i\} \tag{2}$$

where $\Xi = \{\xi_1, \xi_2, \ldots, \xi_k\}$ is the set of SDB dimensions, and G_i is the set of possible groups for dimension ξ_i (e.g., if ξ_1 corresponds to gender, then $G_1 = \{male, female, \ldots\}$).

1.2 Quantifying changes in polarization

The mechanism defined in §1.1 allows us to isolate the effects of each SDB dimension ξ, 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 ξ, but also to specific groups within that dimension $g \in G_\xi$.

Intuitively, dimension ξ partially explains polarization when the annotations divided according to each group $g \in G_\xi$, show less polarization between each

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

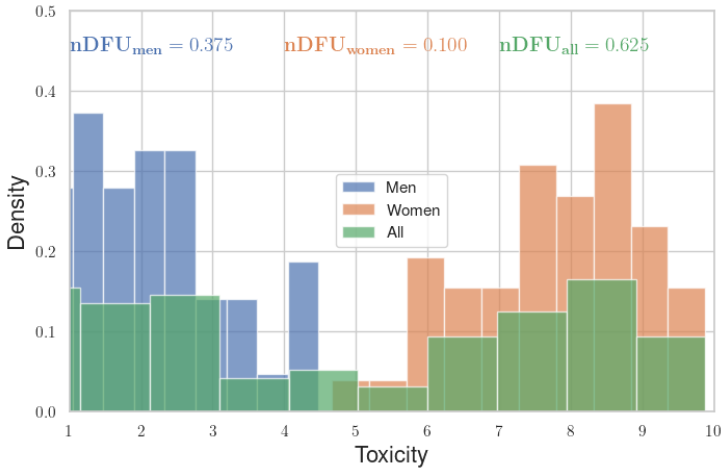

Figure 1: Hypothetical example of a polarizing comment, where male and female annotators agree between themselves, but disagree with the opposite gender. Overall polarization ($nDFU_{all} = 0.625$) is much greater than the polarization exhibited by the annotations grouped by gender ($nDFU_{men} = 0.3725$,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

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

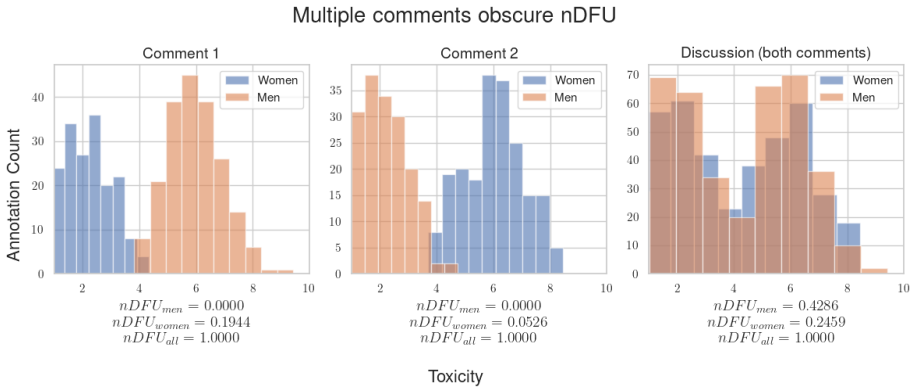

Figure 2: Hypothetical example of a polarizing discussion with two comments, for which the annotators disagree which is the toxic one. If we aggregate the two comments, the polarization scores for both men and women significantly rise, obscuring whether the exhibited polarization can be partially attributed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

same comment. Thus, we can define our "pol-statistic" as:

$$pol(c,\,\mu) = nDFU(A) - nDFU(P(A,\,\xi,\,g)) \tag{3}$$

where $P(A,\,\xi,\,g) = \{a(d,\,i,\,u) \in A | (\xi,\,\mu) \in \theta\}$ is the partition of A for a comment $c(d,\,i)$ for which its annotators belong to the group g of SDB dimension ξ.

1.3 The Aposteriori Unimodality Test

Although intuitive, the *pol statistic* can only be applied to individual comments, and is susceptible to inherent noise present in annotation tasks. If the polarization in a discussion d is not driven by the attribute ξ, we would expect $pol(c,\,g) \approx 0,\, \forall g \in G_\xi$. By obraining the pol statistics for all comments in a discussion d we can apply a mean test with the null hypothesis :

$$H_0 : \frac{1}{|d|}\sum_{c \in d} pol(c,\,g) = 0,\, \forall g \in G_\xi \tag{4}$$

versus the alternative hypotheses:

$$H_i : \frac{1}{|d|}\sum_{c \in d} pol(c,\,g_i) > 0,\, g_i \in G_\xi \tag{5}$$

Since we are considering $|G_\xi|$ tests, 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.

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 g group make a significant contribution to the overall annotator polarization.

2 Acronyms

SDB SocioDemographic Background

nDFU normalized Distance From Unimodality

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.