Aposteriori Unimodality

Dimitris

March 2025

1 Methodology

1.1 Problem Formulation

Let $\{c(d, 1), c(d, 2), \ldots\}$ be the comments[1] of a discussion d and $A(d, i)$ be the set of annotations for comment $c(d, i)$. 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 annotations from different annotators:

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

where $a(d, i, \theta)$ is a single annotation for comment $c(d, i)$ and $\theta \in \Theta$ is the annotator's SDB.

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

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

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 Section 1.1 allows us to isolate the effects of each SDB dimension, but we still lack a mechanism with which to analyze that effect. In this section, we present the "pol-statistic" (polarization statistic) as a tool to not only attribute polarization to a dimension ξ, but also to specific groups within that dimension $g \in G_\xi$.

Intuitively, ξ influences polarization within an annotation set when the annotations, divided according to each group $g \in G_\xi$, show greater polarization

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

than the full set of annotations. For instance, consider an annotation task for bigoted speech and a misogynistic comment. In this case, the annotations may be generally polarized, although annotations from female annotators might show low polarization (most may agree the comment is bigoted), and so may annotations from male annotators (but most may agree that it is *not* bigoted). This suggests that the overall polarization is driven by disagreements between male and female annotators.

However, using all annotations in a discussion is not ideal because we can not control for factors like different comment content. For example, if women find one comment toxic more than men, and another much less toxic than men, the opposing effects might balance each other out, leading to a false negative. To avoid this, we apply our statistic only within the same comment's annotations instead. 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 set of annotations A for a comment $c(d,\, i)$ for which its annotators belong to the group g of 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 adjustment, since it is widely used, generally conservative, and especially so towards correlated hypotheses. 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 largest weaknesses (large numbers-hundreds-of hypotheses) is unlikely to be met in annotation groups.

This procedure is referred to 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