Aposteriori Unimodality

Dimitris

March 2025

1 Methodology

1.1 Intuition

Let's say

1.2 Problem Formulation

We define a conversation $d \in D$ as an ordered set of comments[1]:

$$d = \{c(d, 1), c(d, 2), \ldots, c(d, |d|)\} \tag{1}$$

We assume that annotating a comment depends on three variables: its contents, the annotator's SocioDemographic Background (SDB) (represented as $\theta \in \Theta$), and uncontrolled factors such as mood and personal experiences. Since each comment is likely to be annotated by a distinct subset of annotators, we define $S_{d,\,i} \subseteq \Theta$ as the set of annotators responsible for annotating the comment $c(d, i)$. We can then define the set of annotations as:

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

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). We can thus model θ as a set of attribute-value pairs, where each attribute corresponds to a specific socio-demographic trait, and is associated with at most one value. More formally:

$$\theta = \{(\xi_i, \mu) \mid i = 1, 2, \ldots, k, \mu \in M_i\} \tag{3}$$

where $\Xi = \{\xi_1, \xi_2, \ldots, \xi_k\}$ is the set of features or "dimensions" and M_i is the set of possible values for attribute ξ_i (e.g., if ξ_1 corresponds to gender, then $M_1 = \{male, female, \ldots\}$).

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

1.3 The Aposteriori Unimodality Test

The mechanism defined in Section 1.2 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 define the "Aposteriori Unimodality test" which will allows us to determine whether polarization in a comment can be attributed to a specific SDB attribute ξ.

Intuitively, ξ contributes to polarization in an annotation set when the annotations split by ξ are more polarized compared to the entire set of annotations. 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" (polarization statistic) as:

$$pol(d,\, i,\, \mu) = nDFU(A) - nDFU(P(A,\, \xi,\, \mu)) \tag{4}$$

where $P(A,\, \xi,\, \mu) = \{a(d,\, i,\, u) \in A | (\xi,\, \mu) \in \theta\}$ is the set of annotations A for a comment $c(d,\, i)$ for which the SDB θ includes the attribute-value pair $(\xi,\, \mu)$. Intuitively, $pol \rightarrow [0,\, 1]$ quantifies how much the observed polarization can be attributed to the value μ of attribute ξ.

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 $pol(d,\, i,\, \mu) \approx 0,\, \forall i = 1,\, 2,\, \ldots,\, |d|,\, \forall \mu \in M_\xi$. Consequently, even in the face of limited samples and random noise, we can apply a mean test with the null hypothesis $H_0 : \sum_{i=1}^{k} pol(d,\, i,\, \mu) = 0 \forall \mu \in M_\xi$, versus the alternative hypothesis $H_a : \exists \mu \in M_\xi : \sum_{i=1}^{k} pol(d,\, i,\, \mu) > 0$. Since we are considering $|M_\xi|$ tests, we apply a bonferroni correction to the resulting p-values.

This procedure is referred to as the *"Aposteriori Unimodality Test"*, where a small p-value suggests that we can not rule out that the feature ξ makes a significant contribution to the overall annotator polarization.

2 Acronyms

SDB SocioDemographic Background