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: the content of the comment, the annotator 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 and analyze individual attributes within a SDB. We define θ as a structured composition of multiple attributes, where each attribute corresponds to a specific demographic or personal trait, and each attribute has an associated value. More formally, we can express θ as a set of feature-factor pairs:

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

where $\Xi = \{\xi_1, \xi_2, \ldots, \xi_k\}$ is the set of features (e.g., age, gender, education level, political affiliation), and M_i is the set of possible factors for attribute ξ_i (e.g., if ξ_1 corresponds to gender, then $M_1 = \{\textit{male}, \textit{female}, \ldots\}$).

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

1.3 The Aposteriori Unimodality Test

Define the partition $P(A, \xi, \mu)$ as the set of annotations A for a comment $c(d, i)$ for which the SDB θ includes the feature-factor pair (ξ, μ). In other words,

$$P(A, \xi, \mu) = \{a(d, i, u) \in A | (\xi, \mu) \in \theta\} \tag{4}$$

Focusing on the annotations for a single comment, we introduce the "r-statistic" to measure the distance from unimodality. Suppose the feature of interest is ξ_j. Then, for comment $c(d, i)$, the distance from unimodality is defined as:

$$r(d, i, \xi_j) = \max_{\mu \in M_j} \{nDFU(A) - nDFU(P(A, \xi_j, \mu)\} \tag{5}$$

where A denotes the complete set of annotations associated with comment $c(d, i)$. Intuitively, r quantifies how much the observed polarization can be attributed to the feature ξ_j, with higher values of r indicating that a significant part of the polarization is due to variations in ξ_j. It is obvious that $r \to [0, 1]$.

The reason r is computed on the comment-level instead over the whole discussion is because each comment depends on a number of confounding variables, including previous comments, its content, its intended recipients, etc. However, annotations referring to each individual comment control for most of these variables, allowing us to isolate the effect of ξ on annotator polarization.

If the polarization in a discussion$d$ is not driven by the feature ξ_j, we would expect $r(d, i, \xi_j) \approx 0, \forall i = 1, 2, \ldots$. Consequently, even in the face of limited samples and random noise, we can apply a non-parametric t-test with the null hypothesis $H_0 : \sum_{i=1}^{k} r(d, i, \xi_j) = 0$, versus the alternative hypothesis $H_a : \sum_{i=1}^{k} r(d, i, \xi_j) > 0$. 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_j makes a significant contribution to the overall annotator polarization.

2 Acronyms

SDB SocioDemographic Background