GBV Counterspeech Evaluation Guideline

[Content Warning] This research involves exposure to offensive language, which may cause mental or emotional distress. Please consider this carefully before deciding to participate. Participation is entirely voluntary, and we appreciate your consideration, even if you choose not to proceed. Remember, you are free to withdraw from the study at any point.

Task Overview

This evaluation study aims to assess the quality and effectiveness of counterspeech texts designed to respond to gender-based violence (GBV) content online. Your feedback will contribute to a deeper understanding of what constitutes a good counterspeech response.

Some key concepts are defined below:

- **Gender-Based Violence (GBV)**: A complex and multifaceted issue that includes hybrid behaviours of physical, digital, verbal, psychological, and sexual violence. It can take both implicit and explicit forms and often occurs across multiple spaces and contexts. GBV contains various forms of abuse and specialist focuses, such as coercive control, domestic violence, intimate partner violence, sexual harassment, and stalking.
- **Counterspeech (CS)**: A type of response that actively challenges and pushes back against gender-based violence (GBV), aiming to reduce harm, eliminate hate speech, and promote respectful dialogue.

The evaluation is conducted in **three rounds**:

- **Round 1 & Round 2**: Given a list of counterspeech for one GBV text, answer three binary questions for each counterspeech**.**
- **Round 3:** Make a realistic choice between different ways of responding to a GBV text, including counterspeech, a moderation message, or a disengagement option.

Your feedback is welcome after the evaluation to see if any issues you meet during the counterspeech evaluation process.

Round 1 & Round 2

In this round of evaluation task, you will see a **single GBV text** and **several counterspeech responses**. You need to read each counterspeech response, then **answer three questions**:

- **Q1:** Does the response directly and appropriately address the harmful content?

Instruction: For this question, think about whether the response is relevant. E.g. does it seem off topic or have details that seem unnecessary?

- **Q2:** Does the response feel persuasive or effective?

  Instruction: For this question, think about how the tone feels, i.e. does the tone feel off? Does the response seem convincing?

- **Q3:** Do you think the response could promote positive and educational dialogue?

  Instruction: Remember, the point is not to fight fire with fire (e.g. responding to something hateful with something equally hateful), but fire with water. Here, think about whether the response feels constructive and could potentially build awareness for others that may see the response.

You need to give **yes/no** feedback for each counterspeech response:

- If you think it is **yes**, click **thumb up** button.
- If you think it is **no**, click **thumb down** button.

NOTE: If the response is **“No response”** or **“I can't engage with content that promotes HS.”**, click **thumb down** button for all questions.

Round 3: Choose Your Preferred Response

In this task, remember that you are **placed in a more realistic scenario** similar to what users may encounter on social media platforms.

In each example, you will see a **single GBV post** and **4 responses**. These include two counterspeech responses and two non-counterspeech alternatives. Read the explanations of the different responses below, and **select one or more responses** you would prefer for this GBV post.

- **Response 1 & 2: Counterspeech Responses**
  Counterspeech aims to challenge the harmful message, promote constructive dialogue and reduce harm. However, some responses might unintentionally worsen the situation or make bystanders feel uncomfortable, depending on the tone and context.

- **Response 3: content moderation-style message**
  This kind of warning is commonly used on social platforms like Reddit, YouTube, and other community-guided platforms. It flags the harmful post that breaks community rules and discourages further engagement. This approach may reduce harm, but it avoids directly confronting the content.

  E.g. “This post has been flagged for violating community guidelines on gender-based violence. Continued engagement in this thread may lead to content being filtered or restricted.”

- Option **4: disengagement option**
  This option allows you to disengage from the conversation entirely. It is often a valid and self-protective decision, particularly in triggering or unsafe contexts. However, disengagement may leave the original harm to persist unaddressed, or may miss the opportunity to support others affected by the content.

  E.g. “I do not wish to engage in this conversation.”