Guidelines for HateMirage Data Annotation

Faux Hate refers to hate speech that arises indirectly from fake or misleading narratives, where misinformation fuels hostility toward specific groups or individuals. Such content is subtle, context-dependent, and often difficult for automated systems to interpret correctly.

Initially, the explanations for the Faux Hate dataset were generated using the GPT-4 model due to the complex and nuanced nature of this data. However, synthetic explanations may not fully capture human reasoning, emotion, or contextual understanding. Therefore, we are now manually labeling the explanations to evaluate the reliability and quality of GPT-generated outputs.

Each explanation is structured into three components: Target, Intent, and Implication, which together capture who is attacked, why the comment was written, and what effects it could have.

Annotation Components

Each explanation must be structured into three parts: **Target, Intent, and Implications**.

1. **Target**
    - Identify *who or what* is being attacked in the comment.
    - Capture both explicit and implicit references.
    - Use words or short phrases.
    - If multiple targets exist, list them separated by commas.
    - *Example*: *"immigrants, government policies"*

2. **Intent**
    - Describe the agenda or motive of the author when writing the comment.
    - Highlight what might have influenced them (e.g., misinformation, stereotypes, political bias, personal grievance).
    - Focus on the underlying reason for posting the content.
    - *Example*: *"To blame immigrants for unemployment and fuel resentment against them."*

3. **Implications**
    - Explain the effect this comment could have on the online environment and its audience.
    - Describe the themes or hidden meanings that readers may capture.
    - Indicate what beliefs or negative attitudes it may reinforce.
    - *Example*: *"This may strengthen anti-immigrant sentiments, normalize blaming minorities, and increase polarization."*

Additional Notes

- If any context is unclear, annotators are encouraged to refer to fact-checking websites or other reliable background sources to better understand the narrative behind the comment.
- To ensure high-quality annotation and minimize cognitive fatigue, each annotator is limited to 50 records per day and encouraged to take regular breaks. Annotators may also step away from the task if they find it mentally or emotionally challenging. All participants provide informed consent, and ethical standards are maintained throughout the process.
- AI tools may be used for rephrasing or grammar correction only, not for generating full explanations. This ensures the work remains manual, credible, and human-grounded while maintaining clarity and readability.
- Annotators should also maintain consistency across all records so that the dataset remains coherent, reliable, and useful for downstream research and evaluation.