Title: Datasheet codebook for dataset annotation appendix for “The unappreciated role of intent in algorithmic moderation of abusive content on social media”
Authors: Xinyu Wang (1), Sai Koneru (1), Pranav Narayanan Venkit (1), Brett Frischmann (2), Sarah Rajtmajer (1)
Date: July 29th, 2025
Note: The material contained herein is supplementary to the article named in the title and published in the Harvard Kennedy School (HKS) Misinformation Review.

Appendix A: Datasheet codebook for dataset annotation

Regarding information provided to annotators:

- What are the definitions and scope of online abuse presented to the annotator?
- What underlying taxonomy is provided, and how should it be applied (e.g., modified, integrated) during annotation?
- What contextual information is provided to annotators to assist in the annotation process?
- What is the platform's content moderation policy, and does the annotation rubric align with this policy?
- What is the platform's moderation policy for abusive content and does the annotation rubric adhere to the policy?

Regarding information solicited from annotators:

- Is abusive or offensive language present?
- Is there identifiable intent behind the dissemination of the content, if there is abusive language present?
- Who are the initiators and the targets?

Regarding assessment and reporting:

- When was the data collected and when was it labeled?
- Who are the annotators (demographics, etc.)?
- What is the agreement score amongst annotators?
- What are the data points with low agreement? What are potential reasons for disagreement?