Codebook

Instructions:

Thank you for helping to label these responses. Your job is to look at a given message and decide if it is toxic or not, using only the comment text of that particular message to decide. Labeling is up to your discretion, although examples have been provided for reference.

Here are the categories of different types of toxicity that we want you to identify and focus on:

- obscene
- threat
- insult
- identity_attack
- Sexual_explicit
- emotes (that are used in a toxic way)

If you see a comment that demonstrates one or more of the above categories, this is likely a toxic message.

For some context, these messages are taken from a Twitch stream and may contain emotes in them. Emotes can appear as strange words like "HYPERPOGGER" or "NODDERS". As a general rule of thumb, any word that you don't recognize from the English language is likely an emote, and they will usually be in all caps, though not always.

What to do if you encounter an emote:

If you encounter something that you think is an emote, Google search something like "twitch emote <EMOTE_NAME> image", and it will show you an image of what that emote actually is. You should then look at that image, and decide whether or not the message including that emote is toxic - given now you know what the emote looks like. You should do this every time you encounter an emote. You will also fill out appropriate columns based on this information and decision (see below).

How to label:

- Under the "Human Label" column in the Excel file, write "yes" or "no" depending on whether you think the message is toxic or not (yes being toxic).
- Then, under the column "Human Explanation", include your category of why something is toxic (ie "obscene").
- If you encounter an emote that you thought was toxic, but after looking it up, it did not seem toxic, add this to the (ie "Human Thought Emote Was Toxic But Was Not"). Do the same for the opposite, where you thought it was not toxic, and after looking it up, it was toxic (ie Human Thought Emote Was Not Toxic But Was").

- Make sure to also fill out the Human Emote Explanation column with either what type of toxicity the emote demonstrates after looking it up, or what kind of toxicity you thought it was before looking it up.

Do I need to label non-toxic comments?
You do NOT need to add categories for “no” (non-toxic) comments unless you changed your mind about the message at some point after looking up the emote (see the “If you encounter an emote you thought was toxic” section above).

Codebook Structure

Raw Data

user
- **Definition**: The username of the individual who posted the comment.
- **Data Type**: String
- **Notes**: Usernames are anonymized to protect privacy and align with ethical data usage guidelines.

comment_text
- **Definition**: The full text of the comment posted by the user, including any emotes.
- **Data Type**: String
- **Notes**: Comments may contain Twitch emotes, symbols, and text, which should be considered in context when labeling.

Human Annotation Fields

Human 1 Label, Human 2 Label, Human 3 Label
- **Definition**: The individual label assigned by each annotator regarding the toxicity of the comment.
- **Data Type**: Categorical (e.g., toxic, non-toxic, etc.)
- **Options**: `Toxic`, `Non-Toxic`, etc.

Human Explanation:
- **Definition**: Any additional explanation the human annotator provides to justify their labeling.
- **Data Type**: String

Human Thought Emote Was Toxic Before Visual But Not After:
- **Definition**: Indicates whether the annotator initially thought the emote was toxic before seeing its visual representation but changed their mind after.
- **Data Type**: Boolean (True/False)

Human Didn't Think Emote Was Toxic Before Visual But Did After:

- **Definition**: Indicates whether the annotator initially thought the emote was non-toxic but considered it toxic after visual representation.
- **Data Type**: Boolean (True/False)

Human Thought Emote Was Toxic Before and After Visual:

- **Definition**: Whether the annotator thought the emote was toxic both before and after seeing the visual.
- **Data Type**: Boolean (True/False)

Human Emote Toxic Label:

- **Definition**: Whether the emote was ultimately labeled as toxic by the annotators.
- **Data Type**: Categorical (e.g., toxic, non-toxic, etc.)

Human Majority Label:

- **Definition**: The final label is determined by a majority vote from the three annotators.
- **Data Type**: Categorical (e.g., toxic, non-toxic, etc.)

Emote-Related Fields

Contains Global Emote

- **Definition**: Indicates whether the comment contains a global Twitch emote (i.e., one available to all users of the platform).
- **Data Type**: Boolean
- **Options**: `True`, `False`
- **Notes**: Global emotes are typically well-known to the annotators and the wider Twitch community, and their toxicity levels may be interpreted based on community norms.

Contains Channel Emote

- **Definition**: Indicates whether the comment contains a channel-specific emote (i.e., an emote exclusive to the streamer's channel).
- **Data Type**: Boolean
- **Options**: `True`, `False`
- **Notes**: Channel emotes are often specific to a streamer's community, and prior knowledge of the streamer's culture may influence annotators' judgments of these emotes.

Channel Emote Map

- **Definition**: Contains mapping information regarding the specific channel emotes used in the comment.
- **Data Type**: String
- **Notes**: Provides additional context by identifying the emotes specific to a particular streamer's channel, which may be necessary for understanding in-group and community-specific usage.