Identifying Hate

The first major step in annotating a tweet is to determine whether it is hateful. To make this determination, consider the following guidelines:

- Hate speech can be targeted at any group, regardless of whether it is commonly seen as a marginalized or disadvantaged group.
- Hate speech must contain hate: positive messages of empowerment are different from negative messages of derision.
- Hate need not be explicitly stated: the same words may be made hateful by their context, whether a previous part of the conversation, an associated photo or something else.
- Conversely, not everything that could be considered offensive is hate speech; there must be a component of hostility toward a particular demographic.

If you determine that a tweet is hateful, put a nonconsuming `<HateSpeech>` tag at the beginning of the tweet. In the attributes, indicate the categories of bigotry that pertain to the tweet: `race="yes"` for racism, `nationality="yes"` for xenophobia, etc.

If the tweet is not hateful, the algorithm will infer as much by the absence of a tag. There may still be other annotations to do, though!

Annotating Groups

Regardless of whether a tweet is hateful, scan through and apply a `<Group>` tag to every mention of a potential target group—so every mention of Jews, women, immigrants, *etc*. This is intended to help the algorithm identify common ways of referring to each group, and which terms are most likely to be offensive.

The D.T.D. lists nine categories of groups that are common targets of hate speech, and the most commonly cited groups within each category. For example, `religion` is one category, and `Muslims`, `Jews`, `Protestants` and `Catholics` are the provided examples.

Any group that fits into one of the provided categories, regardless of whether the group itself is specifically listed, should be tagged. For instance, Hindus are not listed as an example, but still tag all references to Hindus. Each category has an `other` option, and there is an `other` attribute that the user may fill with any word. So a tag for Hindus would look like this: `<Group religion="other" other="Hindus">`.

Future guidelines will give style recommendations for write-in groups.

Some additional points about tagging group names:

- If a term refers to multiple categories—for instance, a slur referring to black women—the annotator may invoke multiple categories in the same tag.

- There is a `target` attribute that should be employed when the group being tagged is the target of hate. So in a tweet encouraging white people to be violent against Asians, there should be tags for both groups—because the standard dictates that all groups mentioned should always be tagged—and the `target` attribute distinguishes the roles that each plays.
- If the term used to refer to a group is inherently derisive (not necessarily inappropriate), use the `slur` attribute. There are several options to accommodate slurs used non-hatefully: `reclamation` (e.g. "queer" as used in the L.G.B.T. community), `sarcasm` and `outside_ref` (e.g. "*x* says that ...")
- Targets are often implied, rather than explicitly stated. If no direct mention is made, place a `Group` tag at the beginning of the tweet, not encompassing any text, that characterizes the group in question. (Note: Only tag implicit groups if they are targets.)

Sentiments and Stereotypes

We consider *sentiments* and *stereotypes* to be the two other major elements of a hateful tweet. A sentiment is a feeling expressed toward a group, and a stereotype is a statement about the group. The distinction is subtle, but some examples may help to illustrate.

We recognize three types of sentiments (each an attribute in the tag):

- `judgement`: A personal feeling about members of a group—for example, "Jews are terrible."
- `incitement`: An exhortation to do something to/with/about a group—"Mexicans need to go home."
- `warning`: a prediction of some dire event relating to the group—"...before Muslims overrun our country"

Stereotypes also come in three broad varieties:

- `physical`: statements about a group's physical appearance (size, color, etc.)
- `mental`: statements about a group's mental ability or capacity (good at math, unintelligent, etc.)
- `behavioral`: statements about actions and behaviors associated with a group (violent, lazy, etc.)

In the case of a sentiment, the tag will usually encompass an entire sentence, though if multiple sentiments are expressed then it should be broken up. Stereotype tags may be entire sentences or single words/phrases, depending on how the stereotype is expressed.

Note that many derisive epithets are themselves stereotypes (references to skin color, for example). Thus, a `target` tag and a `stereotype` tag may be coterminous.

Do not tag sentiments or stereotypes that are not hateful or that do not appear in hateful tweets.