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.

We don't have a specific tag to mark a tweet as hateful: the other annotations will make that clear. It is not always an easy decision, though.

The `Group` Tag

```
<!ELEMENT Group (#PCDATA)>
<!ATTLIST Group spans #IMPLIED>
<!ATTLIST Group hate (no | yes) #IMPLIED "no">
<!ATTLIST Group id ID #REQUIRED>
<!ATTLIST Group ref IDREF>
<!ATTLIST Group slur (hateful | reclamation | sarcasm | outside_ref)>
<!ATTLIST Group race (Black | White | Asian | Latino |
Native_American | other)>
<!ATTLIST Group religion (Muslims | Jews | Protestants | Catholics |
other)>
<!ATTLIST Group class (Rich | Poor | other)>
<!ATTLIST Group gender (Women | Men | Nonbinary | other)>
<!ATTLIST Group genderID (Trans | other)>
<!ATTLIST Group orientation (Straight | Gay | Lesbian | Bi | Pan |
other)>
<!ATTLIST Group immigrant_status (Immigrants | Refugees | other)>
<!ATTLIST Group nationality CDATA>
<!ATTLIST Group disability (Mental | Physical | other)>
<!ATTLIST Group age (Old | Young | other)>
<!ATTLIST Group other CDATA>
```

The `Group` tag marks mentions of groups of people that may be the target of hate speech. 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. *Any* group that fits into one of the provided categories, regardless of whether the group itself is specifically listed, should be tagged. Details on each of the attributes follow:

`spans`

The `Groups` tag should consume all and only the words that refer to the group in question. This is usually not a problem:

— "Chilling: GOP Establishment message to America: u will live by **Muslim rapists & terrorists** whether u like it or not!"
— "#LibLabCon or #UKIP time to put differences and vote #Brexit and #LeaveEU before overun with **#Rapeugees** and destroyed #UK"

If uncertain, use the following algorithm to determine the extent of the tag:

1. Select the entire noun phrase containing the reference to the group.
2. Exclude all articles (*the*, *an*, *a*) at the beginning. If the reference to the group clearly extends beyond the current selection, tag the selection from step 1. Otherwise, proceed to Step 3.
3. Exclude all prepositional phrases at the end. Once again, if the reference to the group clearly extends beyond the current selection, tag the selection from step 2.

Sometimes, a group is mentioned implicitly:

— "There's nothing exotic about the color of shit."

If this is the case, the `Group` tag is still necessary, but should be nonconsuming and placed exactly at the beginning of the tweet.

`hate`

Specify the `hate` attribute as `yes` if, based on the guidelines in the first section, you determine that the tweet is hateful toward the group referenced by this `Group` tag. Otherwise, make it `no`.

`id`

Every `Group` tag must have an `id`, because other tags (`Sentiment` and `Stereotype`) will refer back to it. To conform to XML standards, the id must begin with a letter, a colon or an underscore, subsequently contain those symbols or digits. We recommend short ids, but it really does not matter.

`ref`

Specify the `ref` attribute if the group in question has been mentioned before in the same tweet (and thus already has `Group` tag). In cases of multiple mentions, the first is considered the anchor, and subsequent coreferent `Group` tags need not have any specified attributes (except `id`, which is

required by the D.T.D.). Any attributes of these coreferent tags that are specified (`race`, `hate`, *etc.*) will be ignored and replaced with their values in the anchor tag.

`slur`

Use the `slur` attribute to say whether the text enclosed in the Group tag is inherently derisive or offensive (*sand nigger*, *mudshark*, *retard*, *trannies*). Choose the value of the tag according to the guidelines below:

- `none` if not a slur
- `hateful` if the slur is used in its derisive sense
- `reclamation` if the slur is used in a non-offensive sense, especially by a member of the community it refers to as part of an effort to remove its negative connotations. (*Queer* and *nigga* are common examples of reclaimed slurs.)
- `sarcasm` if the slur is used in its derisive sense, but the statement as a whole is meant to be sarcastic—or otherwise implies that the author does not endorse the idea as literally stated.
- `outside_ref` if the slur is used in a quotation or other reference to another individual's derisive statement

Categories

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.

If one group mention is targeted toward a group that is defined by multiple categories—for example, Hispanic immigrants or Muslim refugees—then the `Group` tag should tag each of them.

If a group mention is targeted toward a group that is not specified within one of the categories (Kurds, perhaps), choose the `other` option in the category that most closely fits the group (in this case, `race`) and then type the name of the group in the `other` attribute. Details follow:

`race`

We interpret "race" to refer to any ethnic heritage. Some people draw a distinction between "race" and "ethnicity"; we do not.

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.