Annotation Guide · 8 March 2016 · Anna Astori, Jake Freyer, José Molina, Annie Thorburn

Ugh... ML

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 or not 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 or condescension toward a particular demographic.
- There is a difference between insensitivity and hate. Tweets that show ignorance or coarse humor are not necessarily hate.
- Similarly, legitimate political opinions ("Marriage is between a man and a woman") are not hate, though many may disagree and find them insensitive.
- All this said, there will be many fuzzy cases and it is best to err on the side of marking something as hate. Some people are more sensitive than others, and one annotator cannot presume to anticipate everyone's opinion.

We don't have a specific tag to mark a tweet as hateful: there is an attribute of the **`Group`** tag that indicates hate. It is not always an easy decision, though.

Consider the following examples as a guide:

- *"If my son started showing signs of being gay.. Don't get me wrong I'll love him regardless but I'm correcting it asap."* — Hateful because the writer implies a belief that being gay is a quality that needs "correcting," and is therefore somehow inferior.
- *"PURE MARRIAGE IS A MAN N A WOMAN,NOT A MARRIED MAN N A MISTRESS OR 2 MEN HEBREW 13-4"* — **Not hateful** because the writer states an opinion about marriage, not

about gay people. Moreover, opposition to gay marriage is acceptable in contemporary public discourse, whereas homophobia is not.

- *"If #LoveWins then why don't we make incest legal. And beastiality. Who are you to say that my dog and I aren't romantically involved?"* — **Hateful** because the writer draws a parallel between homosexuality, incest and bestiality, the latter two of which are generally considered immoral.
- *"Come on walking dead bringing in gay dudes, we don't need to hit EVERY demographic possible."* — **Not hateful** because the writer is not expressing negative opinions about gay men, but rather noting that the demographic spread of a TV show appears forced. (Note that some may find this insensitive, but we do not consider it to be hateful.)
- *"I say live and let live but in Christianity homosexuals are not welcome in the kingdom of heaven."* — **Hateful** because this writer goes beyond just quoting the Bible to conclude that one's sexual orientation is a sin strong enough to deny entry into heaven.

The Group Tag

```
<!ELEMENT Group (#PCDATA)>
<!ATTLIST Group spans #IMPLIED>
<!ATTLIST Group hate (no | yes) #REQUIRED>
<!ATTLIST Group id ID #REQUIRED>
<!ATTLIST Group ref IDREF>
<!ATTLIST Group slur (none | hateful | reclamation | sarcasm | outside_ref) #RE-
   QUIRED>
<!ATTLIST Group race (Black | White | Asian | Latino | Native_American | Middle_
   Eastern | other)>
<!ATTLIST Group religion (Muslim | Jewish | Protestant | Catholic | Christian |
   other)>
<!ATTLIST Group class (Rich | Middle | Poor | other)>
<!ATTLIST Group gender (Female | Male | Nonbinary | other)>
<!ATTLIST Group genderID (Trans | Cis | other)>
<!ATTLIST Group orientation (Straight | Gay | Lesbian | Bi | Pan | other)>
<!ATTLIST Group immigrant_status (Immigrant | Refugee | 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 DTD 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:

Non-Category Attributes

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, and will have all of the same attributes, but is nonconsuming.

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. MAE auto-generates an **id** for every **Group** tag. See the **Ref** tag at the end of this file for details about linking.

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*). This also includes insulting terms that may not be common slurs (such as **rapeugees**). 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 stated.
- **`outside_ref`** if the slur is used in a quotation or other reference to another individual's derisive statement

Categories

Every hateful statement targets some group of people, and the groups of people consistently fall into a few recurring categories, characterized as different brands of bigotry: racism (race), xenophobia (nationality), sexism (gender), etc. The D.T.D. lists nine such categories, and in each category gives as value options several common groups within that category, particularly those often the subject of hate speech. Each category is described in detail later in this section.

To approach this section, identify the category that the tagged group fits into, and specify that attribute, leaving the others unspecified. If one group mention is targeted toward a group that is defined by multiple categories—for example, Italian Catholics 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.

If you see clear hate speech against a group that cannot even be categorized according to the convention defined here, then do not specify any category, but name the group in the **`other`** attribute.

Details on phrasing of an **`other`** entry are included for most categories. As a general rule, almost all specified values in each category are adjectives, and therefore user-specified values should be as well. Where it is not clear which name should be used for a group, decide based on the usage in the Wikipedia article about the group. Always capitalize the first letter of each word, and replace spaces with underscores.

Note: Sometimes people making prejudiced comments mix up their demographics (transgendered women with gay men, Sikhs with Muslims, etc.). There may be no way to annotate this well; for the sake of consistency, we say take the writer at their word and pretend the mix-up hasn't happened.

`race`

We interpret "race" to refer to any ethnic heritage. Some standards draw a distinction between "race" and "ethnicity"; for the purposes of this annotation project, we do not.

We should, however, address the issue of race overlapping with other categories included here, namely religion and nationality. Many groups of people, like Jews or Finns, may be described in terms of ethnic heritage and also religious or national identity. For convenience and consistency, we adopt the standard of resolving ambiguous cases in favor of the non-**`race`** option.

References to Jews, for example, should always be classified as **`religion`**, not **`race`**, despite significant shared heritage among Jews and stereotypes referencing physical features. References to Finnish people should likewise always be classified as **`nationality`** for the same reason. This standard will inevitably result in some nonideal classifications, but should nevertheless be applied evenly and without overthinking.

The options given are:

- **`Black`** for generalizations about people of Black African descent, whether specifically African, African American or the African diaspora
- **`White`** for people of European descent, not including people from the Middle East or the Indian subcontinent
- **`Asian`** for people of East Asian or Pacific islander descent, inclusive of Southeast Asia but not the Middle East or the Indian subcontinent
- **`Latino`** for people with heritage in Mexico, Central America, South America and the Caribbean.
- **`Native_American`** for native peoples of the Americas, especially of the United States and Canada, variously referred to as "American Indians," "Native Americans" and "First Nations."
- **`Middle_Eastern`** for peoples of the Middle East, North Africa and the Indian subcontinent, often superficially grouped together as "brown"
- **`mixed`** for people of multiracial ancestry, often persecuted on the belief that their mixed heritage makes them impure
- **`other`** for ethnic groups more specific than the above.

When you need to specify an unlisted group, choose **`other`** as the value for this attribute and then specify the **`other`** attribute. This will not always be straightforward, as many groups have multiple accepted names. To disambiguate, find the Wikipedia article about the group, and give the value of the other attribute as the adjective form of the group's name as cited there: the form most appropriate in the context *the ____ people*. Some examples: *Romani*, *Kurdish*, *Hutu*, *San*, *Tamil*, *Xinjiang*, *Sioux*.

`religion`

The category for mentions and references to any religious groups. The options given are:

- **`Muslim`**
- **`Jewish`**

- **Protestant**
- **Catholic**
- **Christian** for Christians generally, not specific references to Protestants and Catholics
- **other** for specific sects or denominations, or religions not listed here

Follow the guidelines in the race section for defining unlisted groups. If there is no convenient adjective form of, for example, the name of a religious denomination, just use the basic noun form. Examples: *Buddhist*, *Baptist*, *Mahayana_Buddhist*, *Dutch_Reformed_Church*, *Wiccan*.

class

The category for mentions of groups defined by their income or socioeconomic status. Prejudice against these groups would be called "classism." The options given are:

- **Rich**
- **Middle**
- **Poor**
- **other**

class is a category that is likely to be combined with other categories. For example, references to rednecks or white trash may be considered (for the purposes of this annotation) to target poor white people. Thus, that tag would have attributes **race**="White" **class**="Poor".

gender

Groups defined by sex or gender fall into this category. Prejudice against these groups is usually called "sexism." Note the subtle difference between **gender** and **genderID**, which specifically references groups on people based on whether they have changed their gender identity.

Thus, mentions of various "third gender" groups, like Hijras in South Asia, should be in this category, not **genderID**, because we deem that they are most strongly united by their shared gender, rather than their shift from one identity to another. (The line is admittedly fuzzier in real life, but we adopt this standard for consistency.)

- **Female**
- **Male**
- **Nonbinary** as a general term for those who do not identify as male or female
- **other** for subgroups within the above—in this case, likely culture-specific nonbinary genders

genderID

This is the category for groups defined by their shift (or lack thereof) from one sex or gender to another. By far the most common form of hate speech in this category is transphobia.

Also in this category would be groups who have no gender, multiple genders or fluid gender.

- **Trans** for transgendered or transsexual individuals
- **Cis** for those whose gender identity has not changed (sometimes called "cisgendered")
- **other**

orientation

This is the category for groups defined by sexual orientation. For simplicity, indicate general homophobia and mentions/hatred of LGBTQ individuals here.

- **Straight**
- **Gay** for gay men, or for various generalizations about people who are not straight
- **Lesbian** for lesbian women specifically
- **Bi** for bisexuals
- **Pan** for pansexuals
- **other** for other sexual orientations

immigration_status

Distinct from **nationality**, there are many instances of Internet bigotry against immigrants and refugees, wherever they may be.

- **Immigrant** for people who have come to the country from abroad by choice
- **Refugee** for people who have been driven from their homes
- **other**

nationality

Because there are almost 200 sovereign states in the world, our D.T.D. does not list explicit options for this. To determine the tagging for a given national group, refer to Wikipedia and use the title of the article about the country in question.

This is the only category where the user-input value should be a noun (the name of the country) rather than an adjective. Remember to capitalize the first letter of each word, and replace spaces with underscores.

Examples: *United_States*, *China*, *Syria*, *South_Africa*, *Federated_States_Of_Micronesia*,

`disability`

The options given are:

- **`Mental`** for remarks about people with mental disabilities such as Autism or ADD. This is the corresponding option for slurs like retard.
- **`Physical`** for remarks about people with physical disabilities such as blindness, deafness or having missing limbs. This is the corresponding option for slurs like lame.
- **`other`** (most, if not all, should fit in the two above)

`age`

The options given are:

- **`Old`**
- **`Young`**
- **`other`**

Ageism is much more common in hiring than on Twitter, but we include it for completeness.

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.

Sentiments

```
<!ELEMENT Sentiment (#PCDATA)>
<!ATTLIST Sentiment spans #IMPLIED>
<!ATTLIST Sentiment type (judgement | incitement | warning | reaction | other)
   #REQUIRED>
<!ATTLIST Sentiment other CDATA>
<!ATTLIST Sentiment nonhate (sarcasm | outside_ref | other)>
<!ATTLIST Sentiment id ID #REQUIRED>
```

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

- **`judgement`**: A personal feeling about members of a group. **`judgement`** will probably be the most common kind of sentiment:
  - *"I'm sorry but based on their actions, I think Jews are the lowest of the low."*
  - *"If #LoveWins then why don't we make incest legal. And beastiality. Who are you to say that my dog and I aren't romantically involved"*

— *"It's sad that in today's society we advertise and glorify gay marriage and sex change but never anything truly heroic"*

- **`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"*
- **`reaction`**: a reaction to an experience related to a group—*"If my son started showing signs of being gay.. **Don't get me wrong I'll love him regardless but I'm correcting it asap**."*
  - — Note: Tag the reaction itself as a sentiment, not the condition, hence the bold text above.

Stereotypes

```
<!ELEMENT Stereotype (#PCDATA)>
<!ATTLIST Stereotype spans #IMPLIED>
<!ATTLIST Stereotype type (physical | mental | behavioral | cultural)  #REQUIRED>
<!ATTLIST Stereotype nonhate (sarcasm | outside_ref | other)>
<!ATTLIST Stereotype id ID #REQUIRED>
```

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 **`Group`** tag and a **`Stereotype`** tag may be coterminous.

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

Common Attributes of Sentiments and Stereotypes

What to tag: The same standards for tag spans apply to sentiments and stereotypes:

- A single tag does not extend beyond one sentence. Related sentiments/stereotypes in adjacent sentiments should be separate tags.
- Tags should extend over as much of the sentence as is logical; sentiments will usually

be whole sentences, but stereotypes may be less.

— If tagging an entire sentence is appropriate, do so. Include punctuation.

— If a sentence of the form "*x* and *y*" contains two distinct assertions, give *x* and *y* separate tags. Still include punctuation (like the period at the end), but omit the conjunction between them.

— Sometimes sentiments and stereotypes are necessarily more restricted in scope. Still give it the largest extent possible: if not a sentence then a verb phrase, or a noun phrase, or a single word if necessary.

What not to tag: Do not tag non-hateful remarks, unless they are hateful comments made sarcastically or referencing other people. See the **nonhate** attribute below.

nonhate

Sometimes people quote outsiders or make statements sarcastically. (See also the **Group** tag's **slur** attribute.):

- *RT @equaldex: Ben Carson: Prisons prove being gay is a choice http://t.co/LZ8P7CGHqb #LGBT*

The options for this tag are **sarcasm**, **outside_ref** and **other**.

id

An **id** is specified automatically by MAE, for linking purposes. See below.

The Ref Tag

```
<!ELEMENT Ref EMPTY>
```

Because we are tagging target groups separately from sentiments and stereotypes about them, it is necessary to connect those with a link tag. So we use the **Ref** tag:

- Every **Sentiment** and **Stereotype** should be linked to the **Group** it targets by a **Ref** tag.
- Subsequent mentions of an already-tagged **Group** should be linked back to the original **Group** by a **Ref** tag.

Do not forget to make **Ref** tags!

Though our D.T.D. does not include them, MAE auto-generates **from** and **to** attributes:

from

from is the **id** of the beginning of a link. Remember we link to referenced **Group**s. Therefore, in a **Ref** tag, the **from** field may contain the **id** of:

- a **`Sentiment`**
- a **`Stereotype`**
- a **`Group`**, only if this is a subsequent of a **`Group`** being linked to the original mention

`to`

`to` is the end of the link. The **`to`** field may contain the id of:

- a **`Group`**

Don't put anything else there!