DALC4Hate - Annotation Guidelines

Annotation guidelines are used to instruct users to conduct a task, in the case, the annotation of a specific language phenomenon.

Hate speech is a *subjective* task. Besides restrictions and rules, in the end, hate ultimately relates to the perception of the message by the reader. To keep track of subjectivity and possibly investigate its impact, we annotated the data in parallel: everyone annotates everything.

DEFINITION

Hate speech :

- **"**the unjustified assumption that a person or a group of persons are superior to others; it incites acts of violence or discrimination, thus undermining respect for minority groups and damaging social cohesion." (ECRI's General Policy Recommendation No. 15)[1]
- "Language that intends - through rhetorical devices and contextual reference - attack the dignity of a group of people, either through an incitement to violence, encouragement of the incitement to violence, or the incitement to hatred" (Kennedy at al., 2020; GAB Hate Corpus)

Hate speech is not offensive language. With this respect, hate speech is more similar to abusive language. The main **difference** is that hate speech does not simply dehumanise, debase, threat an individual. It does so because the individual is perceived as a representative of group of people that is assumed to be inferior and perceived as less-than-human.

DATA ANNOTATION

A message is hateful if - in general it has the following characteristics:
- there is a mention of an individual or a group of people (—> there is a target)
AND
1. the message is an incitement to violence against the target; OR
2. the message is an incitement to hatred against the target; OR
3. the message is an assault on human dignity against the target

Points 1., 2., and 3. can occur all at the same time, or in any combination. If any of these points occur, then the message is not- hateful (it can be offensive, abusive, or neither).

WHAT WE ANNOTATE

Explicitness: the focus is on the content of the message. It takes into account how the message is realised. The level of annotation focuses both on the surface form of the message. Explicitness is measures by referring to the presence of profanities, slurs, offensive terms, and similar.

Target: the focus of this annotation layer concerns which types of group are targeted by the hateful content. In a way, it can be understood as why the message is hateful.

---

[1] https://www.coe.int/en/web/european-commission-against-racism-and-intolerance/recommendation-no.15

HOW DO WE ANNOTATE

The annotation is conducted as follows:

- Identify if the message contains a target: a person, a group of people, an institutions
- Check if the content of the message can be considered hateful against the target
- Assign a value for EXPLICITNESS
- After the EXPLICITNESS layer has been assigned, fill in the values for TARGET

EXPLICITNESS: the possible values are:
- EXPLICIT: hateful messages that are explicit in their surface forms. There must be present a known slurs, offensive terms, swearword, ect
- IMPLICIT: hateful messages that do not present any explicit element but clearly represent a de-humanising attack on the target
- NOT: any message that is not hateful

TARGET: the possible values are:
- RACE: race or ethnicity; this includes anti-semitic, anti-crab, anti-asian; anti-turkish, anti-black, ect, messages. The target is de-humanised because of their belonging to an ethnic group
- NATIONALITY: nationality, regionalism; this includes xenophobia, attacks on immigrants; targets against specific countries. The target is de-humanised because of their nationality
- MISOGYNY: the target of hate is a woman or a group of women. The target is de-humanised because they is a woman
- RELIGION: religion and/or spiritual beliefs. This includes anti-muslim, anti-christian, anti-buddhist, ect, messages. The target is de-humanised because of their religious beliefs.
- SEX: the target is de-humanised for their sexual orientation. This includes anti-queer, anti-gay, anti-trans, anti-lesbo, anti-non-binary, ect, messages. The target is de-humanised because of their sexual orientation.
- IDEOLOGY: the target is de-humanised because of their political orientation, membership to party, representative of a political ideology.
- DISABILITY: the target is de-humanised because of their physical or mental disability.

Multiple labels may apply for the target. You are required to annotate the one that you perceived as the most relevant.

NOTE 1: If a message is simply a link to an external website, mark is as SKIP

The annotation of hate speech is best done when the full context of occurrence is available. For instance:

MESSAGE: *ahahah!*

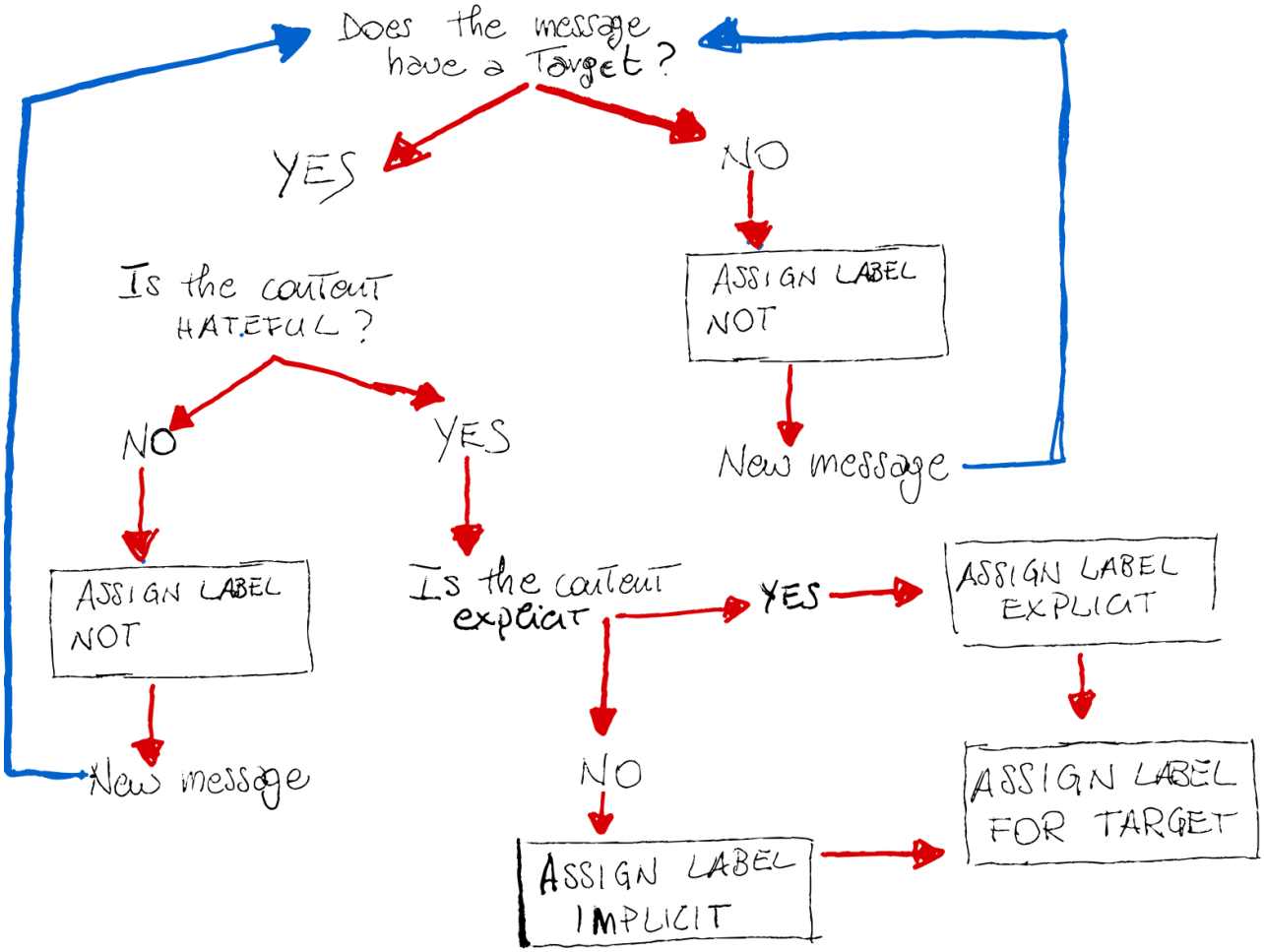
This message can be annotated as NOT - it is just a laugh. However, if the message is a reply, the annotations may change:

1.)
CONTEXT: Mijn kat at mijn eten!
MESSAGE: *ahahah!* —> NOT

2.)
CONTEXT: Die kut expats zuren altijd dat Nederlandse tomaten naar water smaken
MESSAGE: *ahahah!* —> EXPLICIT; NATIONALITY

In 2.) the message is actually endorsing a previous message that promotes exclusion based on nationality. The message is explicit because the slur “kut” is used.

To facilitate the annotation process, you are asked to follow the decision tree that you find below.
NOTE: if context is present, your answers must take the context into account.



TRAINING EXAMPLES

1) #Halsema in #thuisisolatie ? #aftreden en heel snel naar huis voor eigen veiligheid! Ook va 1.5 meter gooi ik raak met eieren en tomaten! Reden van demonstratie is absoluut belangrijk, dat staat buiten kijf, maar dan wel op afstand: #blijfthuis —>

2) @zeeger Maar hij houdt zich er wel aan! Lamlul! —>

3) @Percolator_HNJ @MinPres @VVD Een grote bom 💣.op iran —>

4) @hanwil56 volgens @JokeDrukker zijn dat geen echt moslims oid. Want islam is vrede ofzo #KNETTERGEK —>

5) Nederland is echt aan het verkankeren..\"Kick Out Zwarte Piet\" mag vrijwel in alle grotere steden tijdens de #sinterklaasintocht provoceren, demonstreren en uitlokken...#Leeuwarden, #Rotterdam, #DenHaag, #Nijmegen, #Apeldoorn, #DenHelder, #Eindhoven #Sinterklaasjournaal —>

6) @RobJetten Alle mensen met een mooi bruin huidje wegzetten als #slachtoffer. Allemaal in dat zelfde hokje. #polarisatie ten top #discriminatie ten top #slachtofferschap aanpraten ten top en allemaal voor je eigen agenda —>

7) Te veel over het zelfde kweek je Antipathie tegen homo's !!!!!! —>

8) nee, hou je kankerbek maarten —>

9) Beeldvorming over vluchtelingen begint walgelijk ongenuanceerd te worden. Alleen bij #pauw vind je nog wat nuance. Verder is het #NIMBY —>

10) Sluit me daar helemaal bij aan @RosdeLange protocollen hoef je zelfs niet rationeel te denken; Cruisecontrole slecht afgesteld @MinPres beter vanuit gevoel reilen/zeilen benaderen & Waar blijft toch die #woningnood noodwet @rivm bij stil gestaan #Nederland massaal tel uitzet?! —>