DALC Annotation Guidelines - Offensive Language

DATA ANNOTATION

OFFENSIVE LANGUAGE: any message that contain any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct. This includes insults, threats, and posts containing profane language or swear words. (Zampieri et al. 2019)

Offensive language is a different phenomenon than abusive language. A message can be **perceived** as offensive just because it is a bit rude or impolite. A message is abusive when there is an intent from the speaker/producer of the message to result in a debasement, harassment, threat, or aggression of an individual or a (social) group (but not necessarily of an entity, an institution, an organisation, or a concept).

Offensive message may or may not be addressed to a target. Abusive messages always have a target.

Our annotation will be conducted using two attribute sets:
- EXPLICITNESS: EXPLICIT | IMPLICIT | NOT
- TARGET: INDIVIDUAL | GROUP | OTHER

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 and the effect of the receivers. Explicitness is measures by referring to the presence of profanities, slurs, offensive terms.

Target: the focus of this attribute is on the (potential) receiver of the message. It makes explicit to whom the message is "addressed to".
We distinguish between two major types of targets:
- INDIVIDUAL : this value is used when the target of a message is a specific individual
- GROUP: this value is used when the target of the message is a (social) group of people; and
- OTHER: this value is used when the target of the message are concepts, institutions and organisations, or non-living entities.

ANNOTATION GUIDELINES

Tweets to be considered for the annotation:
- The message is not a retweet:
- The whole message is not a quotation of someone else
- The message is not a meme or simply a link to an external website

The first attribute to annotate concern EXPLICITNESS:

- A message is marked as EXPLICIT if it is interpreted as potentially offensive if it contains a profanity or a slur.

A. *Liberalen zijn idioten :* EXPLICIT
B. *Islam is onzin.* EXPLICIT

- A message is marked as IMPLICIT if it is interpreted as potentially offensive (intension of the speaker to debase/offend; effect on the message on the receiver) if it DOES NOT contain a profanity or a slur.

C. *Minder minder Marokkanen* : IMPLICIT
D. *Europa heeft een plan om blanke mensen te vervangen door mensen uit Afrika.* IMPLICIT

- A message is marked as NOT if it is interpreted as not being offensive or rude (no intension of the speaker to debase/offend; no effect on the message on the receiver). We do consider as OFFENSIVE messages that debase or offend the author of the message (e.g. messages at the first singular or plural person)

E. *Ik ben een idioot* : EXPLICIT
F. W*ij experts zijn soms zo dom*: EXPLICIT

OTHER EXAMPLES:

G. *Godverdomme:* EXPLICIT
H. *Wat een kutstreek! :* EXPLICIT
I. *Wat een shitdag!:* EXPLICIT
J. *Wat een klotezooi!:* EXPLICIT
K. *Wat een zeikerige voetbalwedstrijd.:* EXPLICIT
L. *Dat vind ik verdomd vervelend!:* EXPLICIT
M. *Ik ben het spuugzat!:* EXPLICIT

Key questions:

- Q1: does the message contain any word or combination of words that explicitly make the message offensive? YES —> mark the message as EXPLICIT
- Q2: can the be perceived as offensive by the receiver although there is no explicit profanity, slur, or offensive terms? YES —> mark the message as IMPLICIT
- Q3: does the message contain a profanity?YES —> mark the message as EXPLICIT

The second attribute to annotate concern the TARGET.

NOTE: a message can be OFFENSIVE and NOT have a target

N. *Wat een shitdag!:* EXPLICIT

Annotate this attribute AFTER you have annotated the EXPLICITNESS attribute.

- A target is marked as INDIVIDUAL if the message is OFFENSIVE and the target of the offense is an individual (e.g. a specific person)

O. *hij is een idioot :* INDIVIDUAL

- A target is marked as GROUP if the message is OFFENSIVE and the target of the offense is a (social) group (e.g. an ethnic or religious minority)

P. *Liberalen zijn idioten. :* GROUP
Q. *Conservatieven zijn corrupt. :* GROUP
R. *Minder minder Marokkanen* : GROUP

- A target is marked as OTHER if the message is OFFENSIVE and the target of the abuse is an organisation, an institution, or a concept

S. *Europa is een deken van fascisten. :* NOT
T. *Europa heeft een plan om blanke mensen te vervangen door mensen uit Afrika.* OTHER
U. *Religie is onzin. :* OTHER
V. *Islam is onzin.* OTHER
W. *Het katholicisme is onzin* OTHER
X. *De VVD is corrupt.*: OTHER (NOTE: unless there are evidence that this is true, then members/ voters of VVD may feel offended by such a statement)