Hate Speech Corpus
Annotation Guidelines

October 4, 2017

Contents

1 Introduction

Hate speech (HS) identification is a challenging task that can be subject to individual biases, especially considering the fact that there is no single distinctive factor in drawing the line between HS and not-HS, but a set of variables that the should be considered case by case.
Bearing this in mind, we attempted to annotate each tweet not only based on the presence or absence of HS, but also on other parameters that may even increase, or rather mitigate, the impact of the message.

As a result, we came up with a set of annotation categories and guidelines that attempt to encompass all those variables in a single coherent framework. Such categories include, besides HS, aggressiveness, offensiveness, irony and stereotype, and intensity (of HS).

2 Hate Speech

Labels: No – Yes

As regards HS category alone, we decided to consider two aspects for its identification:

- the **target**, which must be a group identified as one of the three categories included in the search, or even an individual considered for its membership in that category (and not for its individual characteristics);
- the **action**, or more precisely the illocutionary force of the utterance [**?**]: this means that we must deal with a message that spreads, incites, promotes or justifies hatred or violence towards the given target, or a message that aims at dehumanizing, delegitimizing, hurting or intimidating the target.

The joint presence of both elements in a tweet was considered essential to determine whether the tweet contained HS, as in the example below:

la prossima resistenza la dovremo fare subito contro gli invasori islamici!
"our next resistance movement should be right against Muslim invaders!"

In case even just one of these conditions was not detected, HS was assumed not to occur.

3 Aggressiveness

Labels: No – Weak – Strong

It focuses on the user intention to be aggressive, harmful, or even to incite, in various forms, to violent acts against a given target; if present, it can be distiguished between *weak* and *strong*. For example, a message that implies or legitimates discriminating attitudes or policies is considered weakly aggressive:

Gli Italiani prima di tutto!
"Italians first!"

while the reference – whether explicit or just implied – to violent actions is considered strongly aggressive:

tutto tempo danaro e sacrificio umano sprecato
senza eliminazione fisica dei talebani e dei radicali musulmani è tutto inutile
"it's all a waste of time, money and human lives
without the extermination of talebans and radical Muslims it's all useless"

4 Offensiveness

Labels: No – Weak – Strong

Conversely to aggressiveness, it rather focuses on the potentially hurtful effect of the tweet content on a given target; offensiveness also, if present, can be distinguished between *weak* and *strong*, based on the extent of the offense. If, for example, the given target is associated with typical human flaws, this is considered weakly offensive:

Italiani sfrattati e immigrati viziati
"Italians [are] evicted while immigrants [are] spoiled"

while if the target is addressed to by means of outrageous or degrading expressions, the tweet is annotated as strongly offensive:

Barletta, sgomberato mega-campo rom... #raccoltadifferenziata
"Barletta, big Roma camp evacuated ... #recycling"

5 Irony

Labels: No – Yes

This has been used as a general term to cover other nuances such as sarcasm, humor, and satire. In the corpus, irony has a binary value (*no* or *yes*). The introduction of this category in the scheme was led by preliminary observations of the data, which highlighted how it was a fairly common linguistic expedient used to mitigate or indirectly convey a hateful content, as in the example below:

Toh, che caso: clandestino, islamico radicale e terrorista
"Uh, what a coincidence: clandestine, radical Muslim and terrorist"

6 Stereotype

Labels: No – Yes

It determines whether the tweet contains any implicit or explicit reference to (mostly untrue) beliefs about a given target. Even in this case, the inclusion of this category in the scheme is motivated by some considerations on the fact that hatred against minority groups is often characterized by the presence of prejudices. In the scheme, stereotype as well has a binary value (*yes* or *no*); here an example:

gli immigrati non muoiono di fatica . sono spesati di tutto.
"immigrants aren't going to die from exhaustion. they have everything paid for."

7 Intensity

Labels: 0 – 1 – 2 – 3 – 4

In a pragmatical perspective, we noticed that some mitigation devices seemed to play a role in determining the intensity of hateful discourse. In our corpus, we observed that such forms of mitigation seem to interact in determining different degrees of HS. The framework describes five degrees of intensity modulated by mitigation strategies, with a 1-4 value scale for HS tweets, and 0 for the other ones:

- **degree 0**: there is no incitement at all. The message at issue, despite being annotated as aggressive, offensive or other, does not contain HS:

  Come sempre #Italia rifugio sicuro per terroristi!"
  "As usual #Italy [is] a safe haven for terrorists!"

- **degree 1**: there is no explicit incitement, but the acts ascribe a negative feature or quality to a targeted group. These cases are more similar to insults or judgements based on stereotypes; sometimes they suggest that the negative feature may pose a threat to the reader;

  Anche il PD se ne accorge: "I migranti sanno solo ostentare l'ozio. La gente è stufa."
  "Even the Democratic Party realized it: Migrants can only show off their laziness. People are fed up."

- **degree 2**: there is no explicit incitement, but the acts aim at dehumanizing or delegitimizing the targeted group, or claim that the granting of its basic rights and needs is instead an unjust privilege, or that it damages the reader, and should therefore no longer be granted. These acts are not calls to violence, but they raise aversion or hate towards the targeted group;

  La polizia i controllori fermano solo italiani rom e immigrati non li avvicina nemmeno rischiano la vita.
  "Policemen [and] conductors only inspect Italians they don't even get close to Roma or immigrants they risk their lives."

- **degree 3**: there is explicit incitement to violent or discriminatory actions, but the speaker refrains from assuming responsibilities for those actions and only justifies them or express his/her wish that they may happen;

  Quella schifosa rom prende anche in giro, speriamo che cn i loro fuochi tossici si brucino e crepino tutti alla svelta, TOLLERANZA 0.
  "That filthy Roma woman is even mocking, [I hope] they are all burned down by their toxic fires and croak quickly, NO TOLERANCE."

- **degree 4**: there is explicit incitement to violent or discriminatory actions; the speaker overtly suggests or calls for these actions, and declares him/herself ready to carry them out, or take part in their realization.

Hanno rotto il cazzo con tutti questi atti terroristi. Io sono pronto alla guerra.
"They're pissing me off with all these terrorist attacks. I'm ready for war."