Hate Speech Corpus
Annotation Guidelines

October 6, 2017

Contents

1 Introduction

The corpus development forms part of the Hate Speech Monitoring program[1], coordinated by the Computer Science Department of the University of Turin (Italy), with the aim at detecting, analyzing and countering HS with an inter-disciplinary approach.
Providing that among the minority groups targeted by hate speech (HS), one is especially vulnerable and garners constant attention - often negative - from the public opinion, i.e. immigrants, we decided to work mainly on HS against immigrants. Nevertheless, considering that an operational definition of HS may be better extracted from data where a larger set of targets are considered and compared, we collected data where also other HS targets occur, namely Roma and Muslims.

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, stereotype, and intensity (of HS).

2 Targets

Religion (*Muslims*) HS against Muslims may include:

- insults, threats, denigratory or hateful expressions
- incitement to hatred, violence or violation of rights towards individuals or groups of faith other than their own

[1] http://hatespeech.di.unito.it/

- associations between Islamic faith and propensity to fundamentalism, terrorism, murder or a supposed plan of invasion or conquest of Europe

Ethnic group (*immigrants*) HS against immigrants may include:

- insults, threats, denigrating or hateful expressions
- incitement to hatred, violence or violation of rights to individuals or groups perceived as different for somatic traits (e.g. skin color), provenance, cultural traits, language, mode of arrival to Italy
- presumed association of origin/ethnicity with cognitive abilities, propensity to crime, laziness or other vices
- references to the alleged inferiority (or superiority) of some ethnic groups with respect to others
- delegitimation of social position or credibility based on origin/ethnicity
- references to certain backgrounds/ethnicities as a threat to Italian security or welfare or as competitors in the distribution of government resources
- dehumanization or association with animals or entities considered inferior

Roma HS against Roma may include:

- insults, threats, denigrating or hateful expressions; incitement to hatred, violence or violation of rights towards individuals or groups of Roma/Sinti or nomadic/semi-nomadic housing
- association of Roma/Sinti people with predisposition to delinquency, dirt and, in general, socially reprehensible acts or habits
- use, with offensive intent and towards other users, of epithets that denigrate Roma/Sinti people

3 Hate Speech

Labels: No – Yes

Two aspects are taken into account 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, delegitimating, hurting or intimidating the target.

Yes The joint presence of both elements in a tweet is considered essential to determine whether the tweet contains HS:

`la prossima resistenza la dovremo fare subito contro gli invasori islamici!`

No In case even just one of these conditions is not detected, HS is assumed not to occur.
Here a list of other aspects that are NOT considered hate speech in our study:

- offensiveness (either weak or strong) alone
- blasphemy
- historical negationism
- overt incitement to terrorism
- offense towards public servants and police officers
- difamation

4 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 distinguished between *weak* and *strong*.

Weak A message is considered weakly aggressive if:

- it **implies or legitimates discriminating attitudes or policies** :

  `Gli Italiani prima di tutto!`

- there is an *allusion* to a potential threat posed by the presence of the target, or its alleged outnumbering with respect to the Italian population:

  `Una nuova invasione di migranti in Europa, la minaccia fa tremare anche l'Italia https://t.co/smZv5K7T0O`

- there is a sense of dissatisfaction and frustration due to the (perceived) privileged treatment granted to the target group by the government:

  `Bisogna diffondere il fatto che il governo vuole requisire le case sfitte per darle ai migranti.  Atto antidemocrati...`

Strong A tweet is considered strongly aggressive if there is the reference – whether explicit or just implied – to **violent actions** of any kind:

`tutto tempo danaro e sacrificio umano sprecato`
`senza eliminazione fisica dei talebani e dei radicali musulmani è tutto inutile`

No If a tweet is considered as hateful or offensive, does not necessarily implies it is also aggressive:
`@TutteLeNotizie @MediasetTgcom24 e ce l'hai ancora visto che appoggi clandestini rom iussoli e stranieri criminali liberi impuniti a zonzo`

5 Offensiveness

Labels: No – Weak – Strong

Conversely to aggressiveness, it rather focuses on the potentially **hurtful effect** of the tweet content **on a given target**

Weak If, for example, the given target is associated with typical human flaws, this is considered weakly offensive:

`Italiani sfrattati e immigrati viziati`

Strong 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`

No A tweet may be aggressive, or it may portray a stereotypical image of the given target group, but it does not mean this is also offensive:

`Trovato in Croazia il tesoro della \regina" rom:  sequestro da 6 milioni di euro`

6 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.

Yes `Toh, che caso:  clandestino, islamico radicale e terrorista`

7 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.

Yes A tweet is considered as containing a stereotype in (at least) one of the following cases:

- the members of a given target are referred to as invaders:

  – `che diversità c'è tra loro e il pd, hanno firmato insieme per autirizzare l'invasoone di immigrati`
  – `Se gli italiani continuiamo a non fare figli ci ritroveremo presto sottomessi dai musulmani.  Fermiamoli prima sia troppo tardi`

  opportunists/freeloaders/good-for-nothing:

  – `gli immigrati non muoiono di fatica .  sono spesati di tutto.`

  criminals:

  `@TutteLeNotizie @MediasetTgcom24 e ce l'hai ancora visto che appoggi clandestini rom iussoli e stranieri criminali liberi impuniti a zonzo`

  filthy, or having filthy habits:

  `Mentre la #Raggi investe 12milioni di euro in favore dei #rom un migrante CAGA tranquillamente davanti l'altare del...`

- there is a news headline that implicitly endorses, or contributes to the spread of, stereotypes and prejudices:

`Roma è in bancarotta ma regala 12 milioni ai rom - il Giornale`

No Tweets that disclose or disprove a stereotype, or debunk fake news, should not be labeled as containing stereotypes

`Psicosi (e leggende) galoppano su Facebook: Attenzione al furgone pieno di rom accanto alla... https://t.co/FZO8FBl6tn`

****NOTE****

As also stated in Section **??**, whenever a clearly hateful tweet does not actually refer to the target selected in our corpus, HS is assumed not to occur. On the other hand, the remaining categories are expected to be annotated accordingly.

Example:

`#POLITICA la ue di m.:"l'italia discrimina i migranti"sono lontani da noi,SONO GLI ITALIANI DISCRIMINATI!!PEZZI DI M https://t.co/a7L0k447Vu`

Hate Speech: no; **Aggressiveness**: strong; **Offensiveness**: strong; **Ironyb**: no; **Stereotype**: yes

8 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!"`

- **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."`

- **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.`

- **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.`

- **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.`