# Hate Speech Annotation: Analysis of an Italian Twitter Corpus

**Fabio Poletto**
Dipartimento di StudiUm
University of Turin
f.poletto91@gmail.com

**Marco Stranisci**
Acmos
marco.stranisci@acmos.net

**Manuela Sanguinetti,**
**Viviana Patti,**
**Cristina Bosco**
Dipartimento di Informatica
University of Turin
{msanguin,patti,bosco}@di.unito.it

## Abstract

**English.** The paper describes the development of a corpus from social media built with the aim of representing and analysing hate speech against some minority groups in Italy. The issues related to data collection and annotation are introduced, focusing on the challenges we addressed in designing a multifaceted set of labels where the main features of verbal hate expressions may be modelled. Moreover, an analysis of the disagreement among the annotators is presented in order to carry out a preliminary evaluation of the data set and the scheme.

**Italiano.** *L'articolo descrive un corpus di testi estratti da social media costruito con il principale obiettivo di rappresentare ed analizzare il fenomeno dell'hate speech rivolto contro i migranti in Italia. Vengono introdotti gli aspetti significativi della raccolta ed annotazione dei dati, richiamando l'attenzione sulle sfide affrontate per progettare un insieme di etichette che rifletta le molte sfaccettature necessarie a cogliere e modellare le caratteristiche delle espressioni di odio. Inoltre viene presentata un'analisi del disagreement tra gli annotatori allo scopo di tentare una preliminare valutazione del corpus e dello schema di annotazione stesso.*

## 1 Introduction

Hate is all but a new phenomenon, yet the global spread of Internet and social network services has provided it with new means and forms of dissemination. Online hateful content, or Hate Speech (HS), is characterised by some key aspects (such as virality, or presumed anonymity) which distinguish it from offline communication and make it potentially more dangerous and hurtful (Ziccardi, 2016). What is more, HS is featured as a complex and multifaceted phenomenon, also because of the variety of approaches employed in attempting to draw the line between HS and free speech (Yong, 2011). Therefore, despite the multiple efforts, there is yet no universally accepted definition of HS.

From a juridical perspective, two contrasting approaches can be recognised: while US law is oriented, quite uniquely, towards *granting freedom of speech* above all, even when potentially hurtful or threatening, legislation in Europe and the rest of the world tends to *protect the dignity and rights of minority groups* against any form of expression that might violate or endanger them.

Several European treaties and conventions ban HS: to mention but one, the Council of European Union condemns publicly inciting violence or hatred towards persons or groups defined by reference to race, colour, religion, descent or national or ethnic origin. The *No Hate Speech Movement*[1], promoted by the Council of Europe, is also worth-mentioning for its efforts in endorsing responsible behaviours and preventing HS among European citizens.

The main aim of this paper is at introducing a novel resource which can be useful for the investigation of HS in a sentiment analysis perspective (Schmidt and Wiegand, 2017). Providing that among the minority groups targeted by HS, the present socio-political context shows that some of them are especially vulnerable and garner constant attention - often negative - from the public opinion, i.e. immigrants (Bosco et al., 2017), Roma and Muslims, we decided to focus our work on HS against such groups. Furthermore, providing the spread of HS in social media together with their

[1] https://www.nohatespeechmovement.org

current relevance in communication, we focused on texts from Twitter, whose peculiar structure and conventions make it particularly suitable for data gathering and analysis.

## 2 Related Work

One of the earlier attempts to develop a corpus-based model for automated detection of HS on the Web is found in Warner and Hirschberg (2012): the authors collect and label a set of sentences from various websites, and test a classifier for detecting anti-Semitic hatred. They observe that HS against different groups is characterised by a small set of high frequency stereotypical words, also stressing the importance of distinguishing HS from simply offensive content.

The same distinction is at the core of Davidson et al. (2017), where a classifier is trained to recognise whether a tweet is hateful or just offensive, observing that for some categories this difference is less clear than for others.

An exhaustive list of the targets of online hate is found in Silva et al. (2016), where HS on two social networks (Twitter and Whisper) is detected through a sentence structure-based model.

One of the core issues of manually labelling HS is the reliability of annotations and the inter-annotator agreement. The issue is confronted by Waseem (2016) and Ross et al. (2017), who find that more precise results are obtained by relying on expert rather than amateur annotations, and that the overall reliability remains low. The authors suggest that HS should not be considered as a binary "yes/no" value and that finer-grained labels may help increase the agreement rate.

An alternative to lexicon-based approaches is suggested in Saleem (2016), where limits and biases of manual annotation and keyword-based techniques are pointed out, and a method based on the language used within self-defined hateful web communities is presented. The method, suitable for various platforms, bypasses the need to define HS and the inevitable poor reliability of manual annotation.

While most of the available works are based on English language, Del Vigna et al. (2017) is the first work on a manually annotated Italian HS corpus: here the authors apply a traditional procedure on a corpus crawled from Facebook, developing two classifiers for automated detection of HS.

## 3 Dataset Collection

The dataset creation phase was divided into three main stages.

We first collected all the tweets written in Italian and posted from 1st October 2016 to 25th April 2017.

Then we discussed in order to establish *a*) which minority groups should be identified as possible HS targets, and *b*) the set of keywords associated with each target, in order to filter the data collected in the previous step. As for the first aspect, we identified three targets that we deemed particularly relevant in the present Italian scenario; based also on the terminology used in European Union reports[2], the targets selected for our corpus were immigrants (class: ethnic origin), Muslims (class: religion), and Roma. As regards the second aspect mentioned above, we are aware of the limits of a keyword-based method in HS identification (Saleem et al., 2016), especially regarding the amount of noisy data (e.g. off-topic tweets) that may result from such method; on the other hand, the choice to adopt a list of explicitly hateful words[3] may prevent us from finding subtler forms of HS, or even just tweets where a hateful message is expressed without using a hate-related lexicon. With this in mind, we then filtered the data by retaining a small set of neutral keywords associated with each target. The keywords selected are summarised below:

| **ethnic group** | **religion** | **Roma** |
|---|---|---|
| *immigrat** | *terrorismo* | *rom* |
| *(immigrant*)* | *(terrorism)* | *(roma)* |
| *immigrazione* | *terrorist** | *nomad** |
| *(immigration)* | *(terrorist*)* | *(nomad*)* |
| *migrant** | *islam* | |
| *stranier** | *mussulman** | |
| *(foreign)* | *(muslim*)* | |
| *profug** | *corano* | |
| *(refugee*)* | *(koran)* | |

The dataset thus retrieved consisted of 370,252 tweets about ethnic origins, 176,290 about religion

[2] See the 2015 Eurobarometer Survey on discrimination in the EU: http://ec.europa.eu/justice/fundamental-rights/files/factsheet_eurobarometer_fundamental_rights_2015.pdf

[3] Such as the ones extracted for the Italian HS map (Musto et al., 2016): http://www.voxdiritti.it/ecco-la-nuova-edizione-della-mappa-dellintolleranza/

and 31,990 about Roma.

The last stage consisted in the creation of the corpus to be annotated. In order to obtain a balanced resource, we randomly selected from the previous dataset 700 tweets for each target, with a total amount of 2,100 tweets.
However, a large number of tweets were further removed from the corpus, during the annotation stage (because of duplicates and off-topic content). Despite the size reduction, though, the distribution of the targets in the corpus remained quite unchanged, resulting in a balanced resource in this respect.

At present, the amount of annotated data consists of 1,828 tweets. In the next section, we describe the whole annotation process and the scheme adopted for this purpose.

## 4 Data Annotation: Designing and Applying the Schema

Being HS a complex and multi-layered concept, and being the task of its annotation quite difficult and prone to subjectivity, we undertook some preliminary steps in order to make sure that all annotators share a common ground of basic concepts, starting from the very definition of HS.
When determining what can, or cannot, be considered HS (thus in a *yes-no* fashion), and based on the juridical literature and observations reported above in Section 1, we considered two different factors:

- the **target** involved, i.e. the tweet should be addressed, or just refer to, one of the minority groups identified as HS targets in the previous stage (see Section 3), or even to 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, in that it is capable of spreading, inciting, promoting or justifying violence against a target.

Whenever both factors happen to co-occur in the same tweet, we consider it as a HS case, as in the example below:

| **target** | **tweet** |
|---|---|
| religion | *Ci vuole la guerra per salvare l'Italia dai criminali filo islamici* ("We need a war to save Italy from pro-Islamic criminals") |

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

In line with this definition, we also attempted to extend the scheme to other annotation categories that seemed to significantly co-occur with HS; this in order to better represent the (perceived) meaning of the tweet, and to help the annotator in the task, by providing a richer and finer-grained tagset[4]. The newly-introduced categories are described below.

**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 the message reflects an overtly hostile attitude, or whenever the target group is portrayed as a threat to social stability, the tweet is considered *weakly* aggressive, while if there is the reference – whether explicit or just implied – to violent actions of any kind, the tweet is *strongly* aggressive.

| **tweet** | **aggressiveness** |
|---|---|
| *nuova invasione di migranti in Europa* (A new migrant invasion in Europe) | weak |
| *Cacciamo i rom dall'Italia* (Let's kick Roma out of Italy) | strong |

**Offensiveness** (labels *no - weak - strong*): conversely to aggressiveness, it rather focuses on the potentially hurtful effect of the tweet content on a given target. A tweet is considered *weakly* offensive in a large number of cases, among these: the given target is associated with typical human flaws (laziness in particular), the status of disadvantaged or discriminated minority is questioned, or when the members of the target group are described as unpleasant people; on the other hand, if an overtly insulting language is used, or the target is addressed to by means of outrageous or degrading expressions, the tweet is expected to be considered as *strongly* offensive.

[4]The whole scheme description along with the detailed guidelines are available at `https://github.com/msang/hate-speech-corpus`

| tweet | offensiveness |
|---|---|
| *I migranti sanno solo ostentare l'ozio* (Migrants can only show off their laziness) | weak |
| *Zingari di merda* (You fucking Roma) | strong |

**Irony** (labels *no* - *yes*): it determines whether the tweet is ironic or sarcastic rather than based on the literal meaning of words. 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.

| tweet | irony |
|---|---|
| *ora tutti questi falsi profughi li mandiamo a casa di Renzi ??!* (shall we send all these fake refugees to Renzi's house??!) | yes |

**Stereotype** (labels *no* - *yes*): it determines whether the tweet contains any implicit or explicit reference to (mostly untrue) beliefs about a given target. There is a whole host of stereotypes and prejudices associated with the target groups selected for our research; from an exploratory observation of the data in the corpus, the following cases were identified: the members of a given target are referred to as invaders, freeloaders, criminals, filthy (or having filthy habits), sexist/mysoginist, undemocratic, violent people.
Furthermore, we also take into account the role that conventional media may have in spreading stereotypes and prejudices while reporting news on refugees, migrants, and minorities in general. Based on what suggested in the Italian journalists' Code of Conduct, known as "Carta di Roma"[5], in order to ensure a correct and responsible reporting about these topics, we also applied this criterion to any tweet containing a news headline that implicitly endorses, or contributes to the spread of, such stereotypical portrayals (see the example below).

| tweet | stereotype |
|---|---|
| *Roma in bancarotta ma regala 12 milioni ai rom* (Rome is bankrupt but still gives 12 millions to Roma) | yes |

[5]See https://www.cartadiroma.org/

**Annotation process** The annotation task consisted in a multiple-step process, and it was carried out by four independent annotators after a preliminary step where the guidelines were discussed and partially revised.
The corpus was split in two, and each part was annotated by two annotators. The annotator pairs then switched to the other part, in order to provide a third (possibly solving) annotation to all those tweets where at least one category was labelled differently by the previous two annotators. A further subset of around 130 tweets still received different labels by the different annotators (namely for aggressiveness and offensiveness). In order to solve these remaining cases, a fifth independent annotator was finally involved. As a result, the final corpus only contains tweets that were fully revised.

Regarding the results of the annotation in terms of label distribution, we found that 16% of all tweets have been considered containing HS, of which 23% against immigrants, 38% against Muslims and 39% against Roma. When considered alone, aggressiveness occurs in 14% , offensiveness in 10%, irony in 11% and stereotype in 29% of tweets. However, the labels that co-occur more frequently with hate speech are those indicating the presence of aggressiveness (81%), stereotypes (81%), and offensiveness (56%), and, overall, they co-occur altogether 52% of the times; irony is labelled in 11% of HS tweets. While, within the whole corpus, 57% of cases are just tweets with a "neutral" content, which means that no one of the categories were annotated as such.

### 4.1 Agreement Analysis

The development phase related to the inter-annotator agreement (IAA) is not only a necessary step for validating the corpus and evaluating the schema adopted, but also a tool that provides more details about the trends and biases of individual annotators with respect to specific annotation categories.

In this study, we measured the IAA right after the first annotation step was completed, i.e. the one where just two annotators were involved (see Section 4). In line with related cases[6], our data showed a very low agreement: in 47% of cases, the annotator pair annotated at least one of the five

[6]See (Del Vigna et al., 2017), (Gitari et al., 2015), (Kwok and Wang, 2013), (Ross et al., 2017), (Waseem, 2016), to mention a few.

categories using different labels. In fact, the disagreement took place mostly in one (40%) or two (16%) categories, while just 4 tweets received a completely different annotation by the annotator pairs. More specifically, we measured the agreement coefficient, using Cohen's kappa (Carletta, 1996), for each individual category. Results – also reported in Table 1 – show that the category with the highest agreement is namely the one related to the presence of hate speech (abbreviated to 'hs' in the table), followed by irony ('iro.') and stereotype ('ster.').

| | hs | aggr. | off. | iro. | ster. |
|---|---|---|---|---|---|
| before merge | 0,54 | 0,18 | 0,32 | 0,44 | 0,43 |
| after merge | 0,54 | **0,43** | **0,37** | 0,44 | 0,43 |

Table 1: Agreement (Cohen's $k$) for each annotation category before and after merging labels for aggressiveness and offensiveness.

Considering that the lowest agreement was found in aggressiveness ('aggr.') and offensiveness ('off.') – the only categories where three labels were used, instead of two – the agreement was recalculated by merging the *weak-strong* labels; it thus increased considerably (especially in aggressiveness), though still remaining far below an acceptable threshold.
The low agreement with regard to the degree of offensiveness can be attributed to the absence of clear indications within the annotation guidelines in this respect.
Finally, among the annotation criteria established in the preliminary stage, one in particular proved to be quite misleading, i.e. whenever a clearly hateful tweet did not actually refer to the target identified by one of the selected keywords, HS and stereotype were assumed not to occur. On the other hand, the remaining categories should be annotated accordingly. This principle was conceived in order to provide annotated data that could be considered a true reflection of HS towards the targets we identified in our study, though still "preserving" the meaning and the intent of the tweet in itself, regardless of the target involved. This, together with other points of the guidelines, will be further discussed and clarified in the next project phase.

## 5 Conclusion and Future Work

We introduced in this paper the collection and annotation of an Italian Twitter corpus representing HS towards some selected target. Our main aim is at producing a corpus to be used for training and testing sentiment analysis systems, but some effort must still be applied to achieve this goal. The current contribute is mainly in designing and trying a novel schema for HS, but the relatively low agreement shows that modelling this phenomenon is a very challenging task and a further refinement of the guidelines and of the scheme must be applied, together with the application to larger data sets.

## Acknowledgments

The work of Cristina Bosco, Viviana Patti and Manuela Sanguinetti was partially funded by Progetto di Ateneo/CSP 2016 (*Immigrants, Hate and Prejudice in Social Media*, project S1618_L2_BOSC_01) and partially funded by Fondazione CRT (*Hate Speech and Social Media*, project n. 2016.0688).

## References

Cristina Bosco, Patti Viviana, Marcello Bogetti, Michelangelo Conoscenti, Giancarlo Ruffo, Rossano Schifanella, and Marco Stranisci. 2017. Tools and resources for detecting hate and prejudice against immigrants in social media. In *Proceedings of First Symposium on Social Interactions in Complex Intelligent Systems (SICIS), AISB Convention 2017, AI and Society*, Bath, UK.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254, June.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *International AAAI Conference on Web and Social Media*.

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on Facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), Venice, Italy, January 17-20, 2017.*, pages 86–95.

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In Marie desJardins and Michael L. Littman, editors, *AAAI*. AAAI Press.

Cataldo Musto, Giovanni Semeraro, Marco de Gemmis, and Pasquale Lops. 2016. Modeling community behavior through semantic analysis of social data: The italian hate map experience. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization, UMAP 2016, Halifax, NS, Canada, July 13 - 17, 2016*, pages 307–308.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the European refugee crisis. *CoRR*, abs/1701.08118.

Haji Mohammad Saleem, Kelly P. Dillon, Susan Benesch, and Derek Ruths. 2016. A web of hate: Tackling hateful speech in online social spaces. In *Proceedings of the First Workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS 2016)s*, Portoro, Slovenia.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain, April. Association for Computational Linguistics.

Leandro Silva, Mainack Mondal, Denzil Correa, Fabrcio Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. In *Proceedings of the 10th International Conference on Web and Social Media, ICWSM 2016*, pages 687–690. AAAI Press.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, LSM '12, pages 19–26, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas, November. Association for Computational Linguistics.

Caleb Yong. 2011. Does freedom of speech include hate speech? *Res Publica*, 17(4):385, Jul.

Giovanni Ziccardi. 2016. *L'odio online. Violenza verbale e ossessioni in rete*. Saggi / [Cortina]. Cortina Raffaello.