An Italian Twitter Corpus of Hate Speech against Immigrants

Manuela Sanguinetti[⋆], **Fabio Poletto**[⋆], **Cristina Bosco**[⋆], **Viviana Patti**[⋆], **Marco Stranisci**[⋄]
[⋆]Dipartimento di Informatica, University of Turin, Italy
[⋄]ACMOS, Turin, Italy
fabio.poletto@edu.unito.it, {msanguin, bosco, patti}@di.unito.it,marco.stranisci@acmos.net

Abstract

The paper describes a recently-created Twitter corpus of about 6,000 tweets, annotated for hate speech against immigrants, and developed to be a reference dataset for an automatic system of hate speech monitoring. The annotation scheme was therefore specifically designed to account for the multiplicity of factors that can contribute to the definition of a hate speech notion, and to offer a broader tagset capable of better representing all those factors, which may increase, or rather mitigate, the impact of the message. This resulted in a scheme that includes, besides hate speech, the following categories: aggressiveness, offensiveness, irony, stereotype, and (on an experimental basis) intensity. The paper hereby presented namely focuses on how this annotation scheme was designed and applied to the corpus. In particular, also comparing the annotation produced by CrowdFlower contributors and by expert annotators, we make some remarks about the value of the novel resource as gold standard, which stems from a preliminary qualitative analysis of the annotated data and on future corpus development.

Keywords: hate speech, social media, immigrants, Italian

1. Introduction

The global spread of the so-called “web 2.0” and of social network sites allows users to find, create and share knowledge more easily than ever before, with scarcely any skill and cost required. This enormously increased the amount of user-generated content, within a process that some call ”democratization” of the web (Silva et al., 2016). Yet, this freedom also allows for the publication of content which is abusive and harmful both towards the principles of democracy and the rights of some groups of people - namely *hate speech* (henceforth, HS). HS can be defined as any expression “*that is abusive, insulting, intimidating, harassing, and/or incites to violence, hatred, or discrimination. It is directed against people on the basis of their race, ethnic origin, religion, gender, age, physical condition, disability, sexual orientation, political conviction, and so forth*” (Erjavec and Kovačič, 2012).

Although definitions and approaches to HS are varied and depend on the juridical tradition of the country, many agree that what is identified as such can not fall under the protection granted by the right to freedom of expression, and must be prohibited. Online platforms like Twitter or Youtube discourage hateful content, but its removal mainly relies on users reports and lacks a systematic control. In this regard, a promising direction of research is the training of automated classifiers based on manually annotated corpora.

Our Contribution The work hereby presented deals with the creation of a Twitter corpus aimed at obtaining a large and richly-annotated dataset for the development of an automatic system of HS identification[1]. Moreover, given the multiplicity of factors that can contribute to the definition of the HS notion, the annotation scheme was specifically designed to account for this complexity and to offer a broader tagset capable of better representing the possible nuances of the message.

The annotation process will be described in the paper also comparing the contribution given by CrowdFlower users and that of expert annotators; this is in order to make a point on the complexity involved in the development of this kind of resource, where a balance among truth and subjectivity must be achieved.

The complete resource is going to be made freely available and accessible for non-commercial use by the end of 2018[2], along with the annotation guidelines.

The paper is organized as follows. Section 2. briefly overviews some cornerstone researches in the field of HS detection, while in Section 3. we present the criteria we used to collect our corpus and design the annotation scheme, and we describe the parallel annotations carried out on two different sub-sets. The annotation scheme used for both tasks is described in details in Section 4., and we discuss the annotation results in Section 5.. Finally, in Section 6. we present some conclusions on the present work and ideas for future developments.

2. Related Work

Hate speech is a complex notion, especially in a computational perspective. Attempts to define annotation labels that can account for such complexity are found in Ross et al. (2017), where data are labeled with regard to HS (*yes - no*) and to offensiveness (on a scale of 1 through 6), but also in Del Vigna et al. (2017), where the labels used are *no hate - weak hate - strong hate*, and in Kwong and Wang (2013), where tweets are classified for their offensiveness rated on a scale of 1 through 5. Despite the fact that offensiveness is often used interchangeably to refer to HS – which is not necessarily the case (Waseem, 2016) – all these works suggest that a simple binary label does not meet the required level of complexity for analyzing HS (although this may

[1] The work forms part of the wider Hate Speech Monitoring program coordinated by the Computer Science Department of the University of Turin. See `http://hatespeech.di.unito.it/`

[2] `https://github.com/msang/hate-speech-corpus`

come at the expense of reliability). We thus opted for such an approach, though separating the key notions of HS and offensiveness. Moreover, we also included multiple annotation categories, thus building a multi-faceted scheme, as described in section 4..

Contrary to many works where data are collected through a set of typically hateful words, Waseem and Hovy (2016) combine it with a set of neutral words, which are found to frequently occur with hateful content, without directly conveying hate. This allows the identification of a broader range of HS expressions that are not necessarily conveyed by offensive words. Although a much simpler scheme is used, compared to the one hereby described, a set of neutral keywords is a choice that proved consistent also with our findings.

While most of the available works are based on English language, there are a few that collect and analyze an Italian corpus. Del Vigna et al. (2017) identify six categories of HS (*Religion*, *Disability*, *Social status*, *Politics*, *Race*, *Sex and gender issues*, plus *Others*), and test a three-fold label for the annotation process. Musto et al. (2016) give an important contribution to the understanding of HS in relation to other social phenomena: by collecting geo-tagged data from Twitter, the authors create a Hate Map that locates the breeding grounds of five different types of hateful content (*Homophobic*, *Racist*, *Sexist*, *Anti-semitic* and *Against disability*). However, contrary to the approach we have followed in this work, the keywords used for filtering the data consisted of swear words frequently used against the five HS targets.

Eventually, stereotypes are as well among the crucial elements of prejudice and hatred against minority groups (Brown, 2011). Their relevance in analyzing HS, also highlighted in Warner and Hirschberg (2012), led us to introduce in our annotation scheme a novel orthogonal layer specifically devoted to mark the presence of stereotypes in the corpus. However, while in Warner and Hirschberg's contribution the use of stereotypes implicitly presupposes the presence of hateful content (although the words used to convey it may not be hateful themselves), in our study, stereotype alone is not sufficient to define hate speech.

To conclude, although inspired by the related work mentioned in this section, we followed the idea of developing a novel, finer-grained scheme where several facets of the phenomena involved can be represented. As further described in the remainder of the paper, this proved a very challenging direction.

3. Corpus Creation and Description

The corpus development forms part of the Hate Speech Monitoring program[3], 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 (Bosco et al., 2017).

Considering that among the minority groups targeted by 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.

For the data filtering, we opted for a common keyword-based approach, selecting a small set of neutral keywords associated with each target. We obtained a dataset of 236,193 tweets, from which we randomly selected a subset to be annotated. The detailed description of the entire pipeline of the data collection and annotation can be found in Poletto et al. (2017).

Given the higher degree of complexity that applying such scheme entailed, we first annotated 1,827 tweets, then we performed another data filtering starting from neutral words that more frequently occur in texts annotated as HS in this first dataset: *invadere* (invade), *invasione* (invasion), *basta* (enough), *fuori* (out), *comunist** (communist*), *african** (African), *barcon** (migrants boat*). After a further removal of duplicates and off-topic tweets, this resulted in a new portion of 4,182 tweets to be annotated. The final version of the corpus thus consists of 6,009 tweets, annotated according to the scheme and guidelines described in the next section, and by two different groups of annotators. The first section of the corpus, i.e.the tweets of the preliminary dataset and 1,327 tweets of the newly retrieved data, were annotated by a team of expert annotators. The annotation task was carried out by four independent annotators working in pairs, with one half of the corpus assigned to each pair. A fifth independent annotator was finally involved in order to solve the cases where at least one category was labeled differently by the previous two annotators. Furthermore, with the twofold aim of enlarging our annotated corpus and of comparing the accuracy of our team against that of a different group of judges, we had a new set of 2,855 tweet annotated on CrowdFlower. Here we carefully describe the settings we used for collecting this annotation.

CrowdFlower[4] is a crowdsourcing platform that allows researchers to have their data evaluated or annotated by contributors, who can be selected or discarded according to their accuracy. For our task, we uploaded a novel dataset to be annotated and provided a subset of 600 tweets from our gold standard corpus, used as test questions to monitor the contributors' reliability throughout their job. The annotation scheme we asked contributors to apply is exactly the same we used for our annotation, and is described in details in Section 4..

To compute contributors' accuracy, CrowdFlower simply checks if their answers to a given test question match the gold standard exactly. If not, the whole question is marked as failed. If a user fails too many test questions, his reliability gets below a threshold: he is then discarded and his judgments are marked as tainted. In our case though, the task was extremely complex and presented a huge array of possible combinations: for each tweet, contributors had to answer 5 or 6 (intensity being dependent on the presence of hate speech) multiple choice questions, with up to 4 an-

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

[4] https://www.crowdflower.com/

swers each. Due to this reason, in order to avoid discarding too many contributors, we chose to assess users' reliability considering only their judgments on hate speech, and we required to keep a minimum reliability of 65% throughout the job.

Since the corpus language is Italian and we believe that only native speakers can fully grasp even the subtlest linguistic cues, at first we made our experiment available only to those users who claimed to be Italian speakers residing in Italy, assuming them to be native speakers[5]. Yet, due to the poor number of participants, we then opened the task also to Italian-speaking users residing abroad - who, with due exceptions, are likely to be second language speakers. Anyway, this measure only slightly increased the number of contributors.

Furthermore, contributors on CrowdFlower can give feedbacks on a test question when they miss it: this is sometimes helpful, as some test questions can be unclear or unfair and thus undermine accuracy of otherwise reliable judges. We removed a few after observing that contributors would repeatedly fail and contest them - this was not in order to artificially increase their accuracy score, but to make sure it was only tested against fair questions.

The annotators results on both sub-sets are reported and discussed in Section 5.. Next section briefly introduces our annotation scheme along with the main guiding principles for the annotation task.

4. Annotation Scheme: Tagset Design and Issues

HS identification is a challenging task that can be subject to individual biases (Waseem, 2016; Ross et al., 2017). In Weber (2009) these challenges are discussed by illustrating the European Court of Human Rights *modus operandi*, and in particular stressing the fact that there is no single distinctive factor in drawing the line between lawful and illicit, but a set of variables that the Court must consider 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.

After the first annotation phase, we measured the Inter-Annotator Agreement (also described in Poletto et al. (2017)) and the results showed a high disagreement in all annotation categories (with a coefficient ranging from k=0.37 for offensiveness to k=0.54 for hate speech). In light of these results, we discussed the possible sources of disagreement, and revised the guidelines accordingly. Nevertheless, considering the complexity of this annotation task also for humans, we also discussed the inherent complexity of the task and the possibility of finding a single ground truth, given the topic addressed. What emerged from such discussion is described in Section 5.

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 (Searle, 1969): 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. Furthermore, a few more aspects are not considered HS in our study: offensiveness (either weak or strong) alone, blasphemy, historical denialism, overt incitement to terrorism, offense towards public servants and police officers, and defamation.

Below we provide a brief description of the remaining categories:

aggressiveness: 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*. 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 Taliban and radical Muslims it's all useless)

offensiveness: 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:

[5] This is, of course, an approximation. Unfortunately, to our knowledge, CrowdFlower does not allow stricter rules on users' language.

Italiani sfrattati e immigrati viziati
(Italians [are] evicted and 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)

irony: similar to Bosco et al. (2013), 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)

stereotype: 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 (as also mentioned in Section 2.). 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 don't work themselves to death. they have everything paid for.)

The features and tags conceived for the last category, that of intensity, are discussed more in detail in the next section.

4.1. Going Deeper in the Annotation Task: the Incitement Degree

What emerged from a more detailed observation of the annotated data, especially regarding tweets that were considered as HS, is that these data consistently differed from one another, spanning over a broad range of intensity and harm. We thus proceeded to a further step of the research, developing an annotation framework which could account for different types of HS on the basis of what we namely defined as its "intensity": i.e. the degree to which incitement (to hate, and even violent acts) is present in the tweet.

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 terroristici. Io sono pronto alla guerra.
  (They're pissing me off with all these terrorist attacks. I'm ready for war.)

To sum up, the complete annotation scheme is composed of the following categories and tags:

- **hate speech**: *no - yes*
- **aggressiveness**: *no - weak - strong*

- **offensiveness**:*no - weak - strong*
- **irony**: *no - yes*
- **stereotype**: *no - yes*
- **intensity**: *0 - 1 - 2 - 3 - 4*

4.2. Annotation Examples

Table 1 shows few examples of how such categories and their tags are applied in our corpus. As stated above, the only annotation constraint posed by our scheme is related to the annotation of intensity, which depends on that of HS: if the latter is not present, its intensity degree will be equal to 0, otherwise the degree will range from 1 through 4. Except for this case, all the other labels are mutually independent, in that the presence of a given category does not imply nor exclude any of the others. It is therefore possible, among other things, that a tweet contains HS, but not other phenomena represented by the other categories (see tweet number 1 in Table 1), that other phenomena are encountered along with HS (tweet number 2), or even that all the possible phenomena but HS are encountered (tweet number 3).

In the example tweet number 1, the message expresses a feeling of strong aversion towards migrants and their presence on the Italian soil, and implies a subtle encouragement to act in order to ban those who are in Italy or prevent others from coming. Hence the choice to annotate it as HS, with intensity equal to 2 (because of the implicit incitement to take action).

The tweet number 2 reports a news headline about a young Somali arrested in Italy for crimes committed in a refugee camp. This would be considered a tweet with a neutral content, if not for the comment that precedes the headline (*Risorse da accogliere...*, ”resources to be welcomed...”) and that completely reverses the annotators judgment. In fact, the comment reflects not only, once again, a strong aversion towards immigrants, but also an implicit incitement to see immigrants as a whole as criminals and a potential threat to the country and the safety of its citizens. This attitude is considered not only loaded with hate and stereotypes, but also as weakly aggressive and offensive. In addition, irony has also been detected in this tweet, in particular in the sarcastic use of the term *risorse* (”resources”), referred to immigrants, and in the use of the expression *da accogliere* (”to be welcomed”), which clearly intends exactly the opposite.

Finally, tweet number 3 is not considered an example of HS because it neither contains incitement to hate or violent actions, nor is targeted to any of the minority groups selected in our study. In fact, the tweet is presumably addressed to politicians, who reportedly tend to give a higher priority to migrants' needs compared to those of their compatriots. Such assumption is considered as strongly influenced by stereotypes, as well as weakly aggressive and offensive towards migrants (implicitly considered as people not worthy of help). The whole message is finally expressed in sarcastic tones.

The examples just described mainly serve the purpose of making clear to the reader some of the annotation choices adopted in the corpus creation; however, they also highlight a critical point of our study, which is related to the definition of precise and unambiguous linguistic criteria for the selection of proper labels.

Although there are recurring expressions that can be easily associated with HS, especially in reference to immigrants (e.g. *stop invasione!*, ”stop invasion!”), the multiple ways in which it can be conveyed, as well as our choice to use only neutral keywords - rather than more explicit terms - to filter the corpus, somehow prevent the selection of precise lexical patterns in the identification of HS, as well as of the other categories. As a result, also recalling what stated at the beginning of Section 4., the selection of the tags to be associated with each tweet is determined case by case, based on its very content, on the context it refers to (whenever such information can be extrapolated from the text), and on the general principles indicated in the guidelines. Needless to say that this kind of approach has several drawbacks, being a strong disagreement one of those. On the other hand, in this work we rather look at the latter point as a *signal* (in Aroyo and Welty's (2015) words) of the inherent complexity of the task, given in particular by the potential ambiguity of the data at hand, as well as of the possibile solving strategies that can be put forward.

Section 5.1. is namely devoted to a wider discussion of such disagreement, its distribution in the two sub-corpora (i.e the one produced using CrowdFlower and the one annotated by field experts) and its possible causes.

5. Results and Discussion

In this section we extend the preliminary qualitative analysis of the data presented in a previous study on the tag distribution (Poletto et al., 2017). Figure 1 sums up such distribution over the final version of our corpus. However, bearing in mind that the main goal of our work is studying HS and the possible factors contributing to its automatic identification, we hereby provide an analysis of the annotated data centered on HS and, in particular, on its intensity, rather than on every single categories conceived in our scheme.

The categories that co-occur more frequently with HS are, expectedly enough, stereotype (72% of cases), aggressiveness (66%) and offensiveness (51%)[6]. Therefore, in the analysis of intensity degrees and their distribution, we thus focused on these aspects, so as to better understand whether an interdependence among all these categories actually exists and, ultimately, to come up with a ”data-driven” definition of HS based on these findings. For this reason, we did not include in the frequency count the tweets annotated with a 0 degree, as they do not contain HS.

What emerged from the distribution of the intensity degrees (in Figure 1) preliminarily confirms what we discussed in the previous section, i.e. that HS and incitement are often mitigated and conveyed in subtler ways. In fact, most of the hateful tweets contain an implicit incitement (intensity degrees equal to 1 and 2), while a far smaller number of users explicitly incite to engage in violent or discriminatory actions: we thus observe a general trend by Twitter users to

[6] Irony is present in only 11% of hateful tweets.

tweet	hs	aggr.	off.	iro.	ster.	intens.
(1) *basta migranti in Italia, basta!* (no more migrants in Italy, I've had enough!)	yes	no	no	no	no	2
(2) *Risorse da accogliere... Omicidi e stupri nel campo profughi in Libia: arrestato 22enne somalo a Milano.* (Resources to be welcomed ... Murders and rapes in the refugee camp in Libya: 22-year-old Somali arrested in Milan.)	yes	weak	weak	yes	yes	1
(3) *hai la mia solidarietà ma se lasciamo fare a questi ti porteranno via anche l'auto per metterci qualche migrante* (you have my sympathy but if we leave it up to them they might as well get your car to put migrants in it)	no	weak	weak	yes	yes	0

Table 1: Annotation examples of three different tweets having immigrants as a target, one containing hate speech only (tweet number 1), along with its intensity, one containing HS as well as other categories (tweet number 2), and one where all categories are present except for HS and its intensity (tweet number 3).

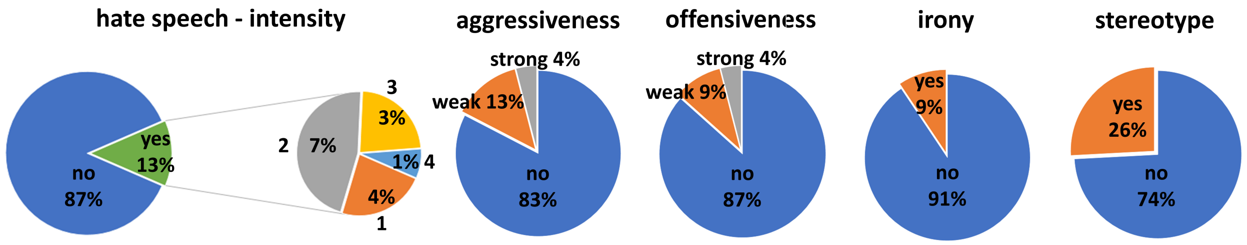

Figure 1: Distribution of all tags in the final version of the corpus.

limit the exposure and the risks arising from reprehensible, or even dangerous, claims.

We then investigated the possible interconnections between the intensity degree and the presence of stereotype, aggressiveness and offensiveness attributed in the previous phase. Results in Figure 2, 3 and 4 show the distribution of these tags across the 1 through 4 intensity degrees.

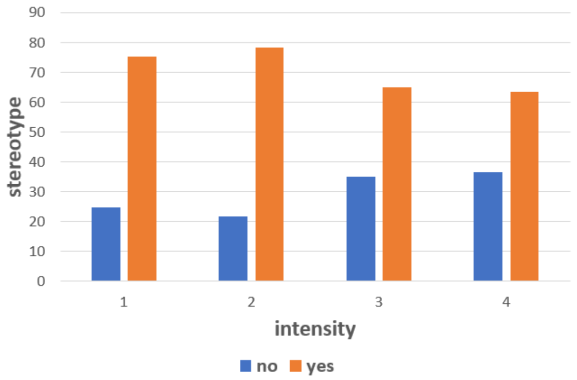

Figure 2: Distribution of stereotype tags (expressed in percentage) across the 1 through 4 intensity degrees.

The presence of stereotype is more frequent in all intensity degrees, though mostly in the lower ones (1 and 2). Such findings are quite consistent with our interpretation and definition of implicit incitement as typically based on, and promoting, prejudices, discrimination and hatred against a given target group (see Section 4.1.). On the other hand, stereotype largely co-occurs also with higher degrees, which suggests us that it might be considered a fundamental factor in the definition of HS.

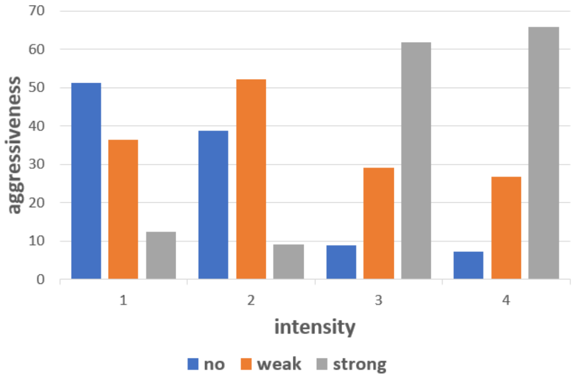

Figure 3: Distribution of aggressiveness tags (expressed in percentage) across the 1 through 4 intensity degrees.

As regards aggressiveness, Figure 3 shows that almost all cases where aggressiveness is absent are concentrated in tweets annotated with a lower intensity degree (1 and 2); to a partially similar extent, most of the tweets considered as weakly aggressive constitute an example of implicit incitement (therefore with an intensity degree equal to 1 or 2). Conversely, the tweets expressing explicit incitement, in its different degrees (namely 3 and 4), are for most part strongly aggressive. Aggressiveness as well can thus be taken into account while providing a definition of HS and incitement.

This does not seem to be the case with offensiveness, whose

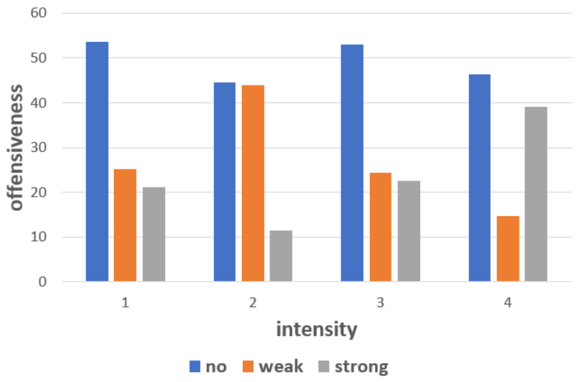

Figure 4: Distribution of offensiveness tags (expressed in percentage) across the 1 through 4 intensity degrees.

distribution shows a less coherent pattern, with all its possible tags spanned over all intensity degrees. The strongly offensive tweets, as expected, are mostly concentrated in the highest degree cluster, and weakly offensive tweets are more frequent in the 2-degree instances; on the other hand, the majority of tweets in all the intensity degrees were not considered as offensive at all. This partially confirms the idea that offensive language does not necessarily involve forms of hate or violence (Waseem, 2016) and supports our choice to select only neutral keywords while filtering the data for the corpus (see Section 3.).

All such patterns will become useful when we will exploit the corpus for automatic HS detection, in a machine learning perspective.

5.1. Agreement Discussion

Given the complexity of the scheme described in the previous sections and the subjectivity involved in the topic, we expected the annotation to pose a number of challenges and problems. That is why, after a first stage where only expert annotators were involved, we also carried out the annotation experiment using CrowdFlower.

As the following analysis will show, a large number of cases appeared to be particularly tricky, also resulting in a very poor annotation agreement (see Table 2).

Besides, we observed peculiar patterns in the behavior of CrowdFlower annotators. First of all, conversely to our expectations, the number of contributors remained low (five annotators carried out more than 90% of the job). Secondly, as said before, the guidelines we used for our annotation are the same we provided CrowdFlower users with; nonetheless, some of their replies and feedbacks seem to suggest that they have not read or taken into due account our definitions and examples. Thus, although their accuracy score remained above the threshold, we should keep in mind that their judgments can not be compared to those by experts annotators, and that their fluctuation is probably due not only to shortcomings in the annotation scheme but also to a certain negligence among the judges.

Considered that the annotation process was carried out in different stages and with different methods, as described in Section 3., agreement as well was computed separately and with different coefficients, based on the number of judments available for each tweet. In the first sub-corpus, two expert annotators worked on the same set of tweets, while for the CrowdFlower experiment each tweet was expected to have at least three judgments; therefore the inter-annotator agreement was assessed by using the Cohen's κ (Carletta, 1996) for the former and the Krippendorff's α (Krippendorff, 2007) for the latter.

	hs	aggr.	off.	iro.	ster.	intens.
experts	0.45	0.48	0.45	0.32	0.41	0.21
CF	0.38	0.25	0.30	0.12	0.20	0.31

Table 2: Agreement for each annotation category in both sub-sets, i.e. the one annotated by our expert team (first row in the table) and the one by CrowdFlower contributors (second row).

The results for the two groups, however low in both cases, show the greater reliability of the corpus annotated by the experts, hence its (relatively) better quality overall. On the other hand, the strongest disagreement between the expert annotators is found for intensity, which, conversely, is the second category after HS where CrowdFlower users seem to reach a higher number of consistent annotations.

This highlights that intensity is the most controversial point of our scheme. While we believe it is crucial to acknowledge that not all hate speech is the same and that there are indeed different shades of intensity, our results show that much work is still to be done before these shades can be effectively defined and detected. The low agreement suggests the presence of shortcomings in the guidelines, which are still ambiguous and not always helpful in settling doubtful cases. Furthermore, distinctions between the four levels are often based on pragmatic rather than semantic features: this results in annotators giving more weight to the attitude of the author than to the actual content of its tweet.

Thus, according to our guidelines, tweet (4) below is to be considered more intense - and therefore more dangerous - than tweet (5), only because the former's author uses a first-person construction which entails individual responsibilities, while the latter's uses a more detached and impersonal form.

(4) *Milva e la ”sua” Goro: ”Se vivessi ancora lì, i migranti li avrei ospitati io” ///dacci l'indirizzo...te li porto io...almeno una dozzina*
(Milva and ”her” Goro: ”If I still lived there, I'd have hosted those migrants myself” ///give us your address...I'll bring you some...a dozen at least)[7]

(5) *Sarebbe da VIETARE il culto dell'islam, bisognerebbe DISTRUGGERE le moschee, DEPORTARE tutti gli islamici e dichiarare l'islam FUORI LEGGE!*
(Islamic faith should be BANNED, mosques should be DESTROYED, all Muslims should be DEPORTED and Islam should

[7] In October 2016, some residents in the little town of Goro and Gorino, Italy, erected barricades to prevent 12 asylum-seekers from entering the town and being hosted in a tourism facility, as determined by legal authorities. Milva is a popular Italian singer, born in Goro, who spoke out in favor of the asylum-seekers.

be OUTLAWED!)

Cases such as this suggest that the present scheme is not always suitable for understanding intensity and dangerousness of HS. Therefore, future work will necessarily have to focus on a thorough rethink of how intensity is conceived and annotated. The scheme will have to be simpler, featuring maybe only two levels - for example ”weak” and ”strong”, as proposed in (Del Vigna et al., 2017) for hate speech and in this paper for aggressiveness and offensiveness; and it will have to be clearer with regard to distinctive features.

6. Conclusion

In this paper we describe an Italian Twitter corpus of HS against immigrants and propose a novel multi-layered annotation scheme to account for different aspects of this multifaceted and complex phenomenon. Besides the presence of HS, we annotated its intensity, as well as the presence of aggressiveness, offensiveness, irony and stereotypes. A preliminary analysis of annotation results is proposed, that opens new perspectives for the exploitation of our data set for the development of HS detection systems.

The choice of such a rich and fine-grained scheme is not flawless, nor without drawbacks, all highlighted and discussed in this paper. On the other hand, namely due to its greater complexity, the corpus lends itself to more detailed and systematic analyses of the possible linguistic patterns associated not only with HS itself, but also to all the other categories included in our scheme.

Acknowledgments

C. Bosco, F. Poletto, M. Sanguinetti and V. Patti were partially funded by Progetto di Ateneo/CSP 2016 (*Immigrants, Hate and Prejudice in Social Media*, S1618_L2_BOSC_01) and by Fondazione CRT (*Hate Speech and Social Media*, 2016.0688), which instead totally funded the work of F. Poletto.

7. References

Aroyo, L. and Welty, C. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. *Artificial Intelligence Magazine*, pages 15–24.

Bosco, C., Patti, V., and Bolioli, A. (2013). Developing corpora for sentiment analysis: The case of irony and SENTI-TUT. *IEEE Intelligent Systems*, 28(2):55–63.

Bosco, C., Viviana, P., Bogetti, M., Conoscenti, M., Ruffo, G., Schifanella, R., and Stranisci, M. (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.

Brown, R. (2011). *Prejudice: Its Social Psychology*. Wiley.

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

Del Vigna, F., Cimino, A., Dell’Orletta, F., Petrocchi, M., and Tesconi, M. (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.

Erjavec, K. and Kovačič, M. P. (2012). ”You don’t understand, this is a new war!” Analysis of hate speech in news web sites’ comments. *Mass Communication and Society*, 15(6):899–920.

Krippendorff, K. (2007). Computing Krippendorff’s alpha reliability. Departmental papers (ASC) 43.

Kwok, I. and Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, pages 1621–1622.

Musto, C., Semeraro, G., de Gemmis, M., and Lops, P. (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.

Poletto, F., Stranisci, M., Sanguinetti, M., Patti, V., and Bosco, C. (2017). Hate speech annotation: Analysis of an italian Twitter corpus. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*, volume 2006 of *CEUR Workshop Proceedings*, Rome, Italy. CEUR-WS.org.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., and Wojatzki, M. (2017). Measuring the reliability of hate speech annotations: The case of the European refugee crisis. *CoRR*, abs/1701.08118.

Searle, J. R. (1969). *Speech acts: an essay in the philosophy of language*, volume 626. Cambridge University Press.

Silva, L., Mondal, M., Correa, D., Benevenuto, F., and Weber, I. (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.

Warner, W. and Hirschberg, J. (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.

Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California, June. Association for Computational Linguistics.

Waseem, Z. (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.

Weber, A. (2009). *Manual on hate speech*. Council of Europe.