sciendo

Research on Education and Media
Vol. 15, N. 1, Year 2023
ISSN: 2037-0830 – DOI: 10.2478/rem-2023-0017

Hate Speech Research: Algorithmic and Qualitative Evaluations. A Case Study of Anti-Gypsy Hate on Twitter

Stefano Pasta[a]

[a] *Catholic University of the Sacred Heart, Milan, Italy, stefano.pasta@unicatt.it, https://orcid.org/0000-0002-7756-5427*

Abstract

Hate speech may be the research focus of the interdisciplinary field of hate studies, but it is also a difficult phenomenon to define. Internationally, there are several detection studies on automatically detecting hate speech. They can be grouped according to two approaches: the first includes searching using only machine learning methods, while the second includes studies that combine automatic searching with human classification. The case study on anti-Gypsy hate in Italian on Twitter in the second half of 2020 falls into the second category, and its methods are outlined here. Based on the results (annotation as 'hate'/'non-hate', identification of forms of rhetoric and anti-Gypsyism), the researchers propose classifying online content according to seven indicators called the 'spectrum of online hate'.

Keywords: *racism; artificial intelligence; humanities; Internet.*

1. Hate Speech Detection: Machine Learning and Human Interpretation

According to scientific research and European and international data, the phenomenon of hate speech is becoming more established. It is not easy to find a consensus on the definition of 'hate', which manifests as a spectrum of emotions, attitudes and behaviours. It is an ambiguous concept in many respects, but it is useful to broadly outline the plethora of 'anti-', '-ism' and '-phobia' instances, which take specific forms depending on the target group.

The Council of Europe's Committee of Experts on Combating Hate Speech has proposed this definition of hate speech:

> Hate speech is a complex and multidimensional phenomenon that has far-reaching consequences in contemporary democratic societies, in particular for human dignity, equality, participation and inclusion in society […]. Hate speech is understood as all kinds of expressions, which spread, incite to, promote or justify violence, hatred, discrimination or prejudice against a person, or a group of persons, that is based on presumed or real personal characteristics or status including ['race'/race], colour, language, religion, citizenship, national or ethnic origin, age, disability, sex, gender, gender identity and sexual orientation […].
>
> Hate speech that does not entail criminal, civil or administrative liability, but nevertheless causes prejudice and hate and raises concerns in terms of tolerance, civility, inclusion and respect for the rights of others, should be addressed through other, non-legal, means (Council of Europe, 2021, p. 5).

Research into contemporary forms of hate (Siegel, 2020; Santerini, 2021; Faloppa, 2020) and particularly studies on changes in the social web (Pasta, 2018a; 2022) agree that this phenomenon requires a multidisciplinary approach. In the psychological literature, hate is thought to be a combination to two components, one cognitive and one emotional (Sternberg & Sternberg, 2008; Heinzelmann, Höltgen & Tran, 2021) and five belief domains – superiority, injustice, vulnerability, distrust, and helplessness – are identified as particularly noteworthy (Eidelson & Eidelson, 2003). A study (Kommattam, Fischer, & Jonas, 2019) shows that individuals perceive less intense emotions in ethnic outgroup members than in ethnic ingroup members; this intensity bias in interethnic emotion perception points to a systematic downplaying of the intensity of outgroup emotions and suggests an empathy gap towards members from other ethnic groups. From a pedagogical perspective, citizenship education is questioned by hostile narratives, by a binary us/them, friend/enemy, inside/

outside view of the world and by the establishment of widespread and everyday processes of selecting target groups, not only through words but images and memes as well. In light of this, researchers in the field of education need to study the phenomenon, its types and rhetorical forms, in greater depth in order to prevent and combat the various forms expressed online through language and images.

The field of hate studies[1] combines the legal and digital fields with the humanities (sociology, pedagogy, anthropology, philosophy, linguistics and semiotics) and the interests of scholars, researchers, politicians, communication experts, human-rights activists and non-governmental organisation (NGO) managers. Numerous studies in various countries have focused on automating detection processes and creating an algorithm for identifying online hatred. The corpus analysed is almost always extracted from Twitter, as this is the only one of the main social networks that enables automatic access to data through application programming interfaces (APIs).

This field of research harbours a tension between human and non-human action and between technology and human action, with a tendency to limit interventions to artificial intelligence at the expense of more interpretive approaches. At a macro level, we can identify two groups among the international studies: studies using only machine learning methods and studies that combine automated searching with human classification.

The former category conducts searches using artificial intelligence alone, for example the study by the US scholars (Zannettou et al., 2020) who combined two classifiers (Hatesonar and Perspective API) based respectively on logistic regression to categorise comments as hateful, offensive or neither; the exploitation of crowdsourced annotations of text to create machine learning models for the algorithm. Other international models based solely on machine learning include the 'deep learning based fusion approach' (Zhou et al., 2020), 'multi-faceted text representations' (Cao, Lee, Hoang, 2020), the various 'deep neural network' models (DNN) (Amrutha & Bindu, 2019) and the 'random forest' method with handcrafted psycholinguistic features to enhance the interpretability and scalability tested in studies on hate speech during the COVID-19 pandemic in the Philippines and the US (Ueng & Carley, 2021).

Uppsala University's dictionary-based method (Isbister et al., 2018) also belongs in this group. Using this algorithm, researchers categorised user-generated content (comments in this case) in Swedish about 23 politicians, taken from sites critical of migration flows; the content was divided into six different categories (anger, naughtiness, swearwords, general threats, and death threats), each of which was represented by a dictionary of hate terms chosen by experts (psychologists and data scientists). The proximity, within the comment, of these terms to terms indicating the target group was then detected.[2] The model by Indian researchers Paul and Bora (2021) uses a simple sampling method to balance the data and implements deep learning models such as long short-term memory (LSTM) and bidirectional long short-term memory (Bi-LSTM) to improve accuracy in detecting hate speech on social networking sites.

The second group includes the studies that combine automated searching with human classification: The Hate Lab at Cardiff University uses support-vector machine methods combined with a bag-of-words approach and then analysing the output of UK-based content creators only, identifies anti-Semitic tweets using keywords (Ozalp et al., 2020). Again, to analyse hate tweets during the early months of the pandemic, a team from King Saud University worked on the Ar-COV-19 dataset, an ongoing collection of Arabic tweets related to COVID-19, using a pretrained convolutional neural network (CNN) model; 10,000 tweets were given a score between 0 and 1, with 1 being the most hateful text; they also used non-negative matrix factorisation to discover the main issues and topics discussed in hate tweets (Alshalan et al., 2020). Lexicon-based approaches were developed for the Sinhalese language by researchers at the University of Ruhuna (Sandaruwan, Lorensuhewa & Kalyani, 2019), which combined text mining and machine learning and used a corpus of 3,000 comments as a resource to identify words that are specific to Sinhala hate speech and offensive speech.

This group also includes studies based on automatic sentiment analysis and topic-modelling techniques and on manual annotation by previously trained coders, such as the study of tweets in Spanish about the rescue of refugees by the Aquarius ship (Arcila-Calderón et al., 2021). The model used by the Indian researchers from Thadomal Shahani Engineering College in Mumbai also combines manual tagging with text mining and the use of deep learning models built using LSTM neural networks, tested on a corpus of 56,745 tweets and comments (Dubey et al., 2020); the HateTrack system developed by University College Dublin (Siapera, Moreo & Zhou, 2018), on the other hand, has developed a method that combines LSTM networks with the manual annotation of racist content. Finally, we can also mention the studies on anti-Semitism on Twitter by Indiana University (Jikeli, Cavar & Miehling, 2019).

2. Hate Studies in Italy

Various approaches to hate-speech analysis are also being conducted in the Italian language. Some studies emphasise mapping based on geography and insults and offensive words (Vox Diritti's Intolerance Map, 2021[3]); others attempt a general quantification using only overtly offensive terms, as in the study by DataMediaHub and KPI6 (2020)[4]. A similar

[1] See the International Network for Hate Studies, Sussex University: https://internationalhatestudies.com/.

[2] As embedding, they use Gensim's (Řehůřek & Sojka, 2010) implementation of the Continuous Bag of Words (CBOW) model (Mikolov et al., 2013), which builds word vectors by training a two-layer neural network to predict a target word based on a set of context words.

[3] The 6th Intolerance Map is from 2021, produced by the University of Mian, the Sapienza University, the University Aldo Moro of Bari and Catholic University of Milan. See D'Amico and Siccardi (2021) and, for sexist hate speech, Brena (2021).

[4] http://www.datamediahub.it/2020/06/22/rapporto-sullhate-speech-in-italia/#axzz6uNY1PRQr.

approach to identifying anti-Muslim hatred was taken by the Hatemeter developed by the University of Trento, although it also combined the trialling of a computer-assisted persuasion tool with the use of an application similar to a chatbot, which, given a short input text, provides five suggestions that could be used to counteract the hate speech or reduce the intensity of the discussion (Di Nicola, Andreatta & Martini, 2020).

However, linguistic analyses based on semiotics and the philosophy of language have taken several steps forward in identifying the lexical components of hate speech, moving beyond the idea that detection can coincide with offensive terms alone (Femia, 2019; Ferrini & Paris, 2019).

In terms of method, too, various techniques have also been trialled for automatic speech processing (sentiment analysis, text mining etc.) and for data collection and annotation to capture the different facets of hate speech (Sanguinetti et al., 2018); the Evalita 2020 initiative promoted by the Associazione Italiana di Linguistica Computazionale made it possible to test natural language processing (NLP) tools for the automatic processing of hate speech. However, as Lazzardi, Patti and Rosso (2021) note, algorithmic categorisation alone is insufficient: for instance, in 2018 the Evalita and IberEval evaluation campaigns proposed a shared automatic misogyny identification (AMI) task, based on Italian and English tweets and on Spanish tweets respectively, but obtained poor results in categorising misogynistic behaviour[5]; a study with a similar objective and dataset obtained similar results (Fersini, Nozza & Rosso, 2020).

At the other end of the spectrum from analyses based solely on machine learning is Amnesty International Italia's Hate Barometer (2021), which was developed using manual annotation by activists and only uses automated tools to test the level of significance.[6]

Major detection projects combining algorithmic logic and manual annotation include: 'Contro l'odio' developed by the Acmos Association with the University of Turin and the University of Bari (Capozzi et al., 2020; Poletto et al., 2021)[7]; 'REASON – REAct in the Struggle against ONline hate speech' by the National Office Against Racial Discriminations (UNAR), the Research Centre on Intercultural Relations of the Catholic University of Milan, the IRS and the Carta di Roma Association[8]; the Innovative Monitoring Systems and Prevention Policies (IMSyPP) of Online Hate Speech run jointly by the Institut Jozef Stefan, University of Ca' Foscari, University of Cyprus, Agcom and Textgain Bvba, which monitor YouTube comments in English and Italian, tweets in Slovenian and Facebook and Twitter posts in Dutch (Novak et al., 2020)[9].

The Mediavox Observatory of the Catholic University, whose case study on anti-Gypsy hatred in Italian-language Twitter will now be presented, also combines a socio-educational approach and automatic computer processing. This methodology is applied to different target groups such as Muslims (Pasta, 2021) and Jews (Pasta et al., 2021) and, besides simply detecting hate speech, aims to study its characteristics more deeply in order to design coherent educational interventions.

3. The Case Study: Forms of Anti-Gypsyism

This study addresses the classification of anti-Gypsy hate speech, that is, specific racism towards Roma and Sinti people (Pasta, 2020a; Piasere, 2015) in Italian-language tweets in the second half of 2020. Through temporal analyses on sample classifications conducted manually by experts, the researchers wanted to find out whether monthly peaks in hate occur. The researchers then specified which forms of rhetoric and hate were prevalent.

The forms of rhetoric to choose from were insult, derision/irony, exclusion/separation, prejudice, dehumanisation, humiliation/disdain, fear, competition and incitement/violence. These emerged from a psycho-social and literary-historical analysis, conducted by researchers with expertise in hate speech, of linguistic forms of hostility towards the other.

In the absence of an agreed definition of anti-Gypsyism, the forms of hate were borrowed from the categorisation proposed by Mediavox researchers based on a comparison with the literature (Pasta, 2018b; 2019; Arrigoni & Vitale, 2008; Pasta & Vitale, 2017).

One form is *differentialist* anti-Gypsy speech, based on the exasperation of those with a differentialist-culturalist view (Balibar & Wallerstein, 1991) in response to the maximum degree of otherness represented by the Roma and Sinti, whereby cultural differences (with a distorted interpretation of culture as an immutable and static hereditary trait) are so strong and inalienable that they do not allow Roma and non-Roma to coexist. This form of anti-Gypsyism uses the category of the 'nomad', that is, people who are 'different', who reject integration and links with the local community to exacerbate social distance.

A second form is *anomie*-based anti-Gypsyism, which relates to a moral objection to Roma and Sinti behaviours seen as pestilent and having the potential to corrupt the majority society, that is, a fear of bad apples linked to the theme of contagion (Zimbardo, 2007). This form of anti-Gypsyism targets Roma and Sinti as beggars exhibiting morally disreputable behaviour (lazy, cheating, unwilling to work, superstitious, pushy), poor (blaming the poor for their poverty),

[5] The best machine learning results: Linear SVM for Italian and Spanish, and tf-idf combined with SGD, which differs from standard SGD in its parameter settings, for English.
[6] Monitoring is updated annually. See https://www.amnesty.it/barometro-dellodio-intolleranza-pandemica/ (2021).
[7] https://controlodio.it/.
[8] https://reasonproject.eu/il-progetto/.
[9] http://imsypp.ijs.si/.

deserving of whatever they have to endure, antisocial and dirty. It draws on prejudices linked to pseudo-scientific theories that scientists such as the criminologist Cesare Lombroso posited in their publications over a century ago (Bravi, 2009). One particular form of anomie-based anti-Gypsyism is linked to the accusation of theft. The association of all Roma and Sinti with theft is in fact one of the most historically rooted prejudices. It has a powerful impact and features frequently in hate speech.

Competition-based anti-Gypsyism accuses Roma and Sinti of exploiting 'Italian' welfare systems,[10] that is, resorting to the classic xenophobic theme of the depletion of welfare resources – and therefore the potential well-being of the 'natives' – by people who are 'different'. In addition to the classic racist argument concerning competition for resources (Wieviorka, 1995), in this case we also see the negative moral judgement of Roma and Sinti who, based partly on arguments mentioned above in relation to other forms, are accused of being cunning profiteers unwilling to work and strive.

Finally, *elimination*-based anti-Gypsyism advocates the removal of these groups from the community to the point of justifying the killing of Roma and Sinti, or causing damage to their property (places they live, personal belongings); the motivations underlying the other forms of anti-Gypsyism described above may be present, but in this case the hatred is such that the incitement to commit violence or discriminate against Roma and Sinti prevails. Incitement to exterminate Roma and Sinti people, referring consciously or otherwise to the Nazi-Fascist genocide of these people, meets no condemnation in the cultural climate (Pasta, 2020b).

4. Methodology

The methodology is based on social network analysis (SNA) techniques. Data were collected using the open-source Python library GetOldTweets3,[11] which allows tweets to be obtained by query search. Each downloaded tweet comes with several pieces of information as well as the tweet text, such as the author's username, mentions, hashtags, number of retweets and likes, date, time and geolocation.

Tweets published between 1 July 2020 and 31 Decembe 2020 were extracted using a search string that combined a lemma identifying the target group (zingar_ OR zingher_ OR rom OR nomad_, all Italian terms for 'gypsy') with (AND) a reference to typical elements of anti-Gypsyism according to the literature (feccia OR ladr_ OR rapit_ OR rapisc_ OR rolex OR ruspa OR stupr_ OR violent_ OR mendica_ OR elemosin, which can be translated respectively as scum, thief, kidnapped, kidnap_, Rolex, bulldozer, rape_, violent, begg_ and panhandl_)[12]. Then, using the technique of simple random sampling without repetition, a sample consisting of 100 tweets per month was selected, resulting in a sample data set of 900 total posts (James et al., 2017). The latter was manually classified by experts in the field ('annotators') who determined whether the tweet contained hate or not. If it was hate speech, they assigned it the corresponding forms of rhetoric and anti-Gypsyism.

In this case, concerning anti-Gypsyism specifically, it was not possible to conduct the last step usually taken by the Mediavox Observatory due to the absence of research on this form of hatred using only machine learning with Italian lemmas. However, we are including it in this outline of the methodology because its use in other case studies on Islamophobia (Pasta, 2021) and anti-Semitism (Pasta et al., 2021) contributed to the reflections reported in this article's conclusion. This last step in processing the results consists in applying a *confusion matrix*, that is, a tool for analysing errors made by a machine learning model (James et al., 2017). All the texts classified by the annotators are therefore also evaluated by an algorithm that can determine whether the tweet contains hate, after applying a series of typical natural language processing (NLP) procedures to 'clean' the texts, such as removing superfluous characters, converting the text to lower case and removing stopwords (Bird, Klein & Loper, 2009). The 'alternative' classification is made using a dictionary of negative words taken from another scientific study in the Italian language and therefore absent in the case of anti-Gypsyism. Other data set cases submitted to the confusion matrices revealed that the algorithm that only detects word roots does not perform well in identifying hate content, with evaluations of the same tweets by the manual annotation process (also known as labelling, which requires collaboration with experts) differing greatly from evaluations by the algorithm created using the dictionary of negative words from another methodology.

5. Results

Figure 1 shows the number of tweets downloaded each month; as can be seen, there are no peaks in the six months under consideration, since there are no events of national importance concerning Roma and Sinti in that period that would significantly shift the monthly average, as was the case, for example, in May 2020 with the Islamophobic tweets coinciding with the release and conversion to Islam of NGO worker Silvia Romano, who had been kidnapped by Al Shabab terrorists (Pasta, 2021).

[10] Even Italian Roma, who make up about half of the Roma population in Italy, are also considered 'foreigners'.
[11] https://pypi.org/project/GetOldTweets3/.
[12] Number of tweets extracted: 1,191.

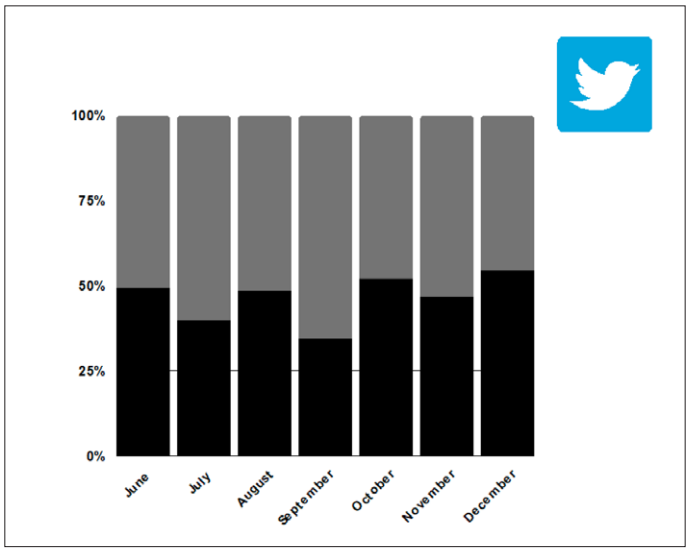

Fig. 1. Number of tweets downloaded each month.

The following figures show a more in-depth analysis following manual classification by the annotators. The first step is to define whether the text of the tweet contains hatred or not: the results in Figure 2 show an average of 45.6%, with a stable monthly distribution confirming that the phenomenon of anti-Gypsyism, although liable to flare up during particular events, is deep-rooted and introjected in Italian society (Jo Cox Committee, 2017).

Table 1. Frequency distribution of hate tweets.

Month	Number of tweets	Hate tweets	No hate tweets	% of hate
June	176	87	89	49,4%
July	179	71	108	39,7%
August	134	65	69	48,5%
September	145	50	95	34,5%
October	94	49	45	52,1%
November	75	35	40	46,7%
December	97	53	44	54,5%

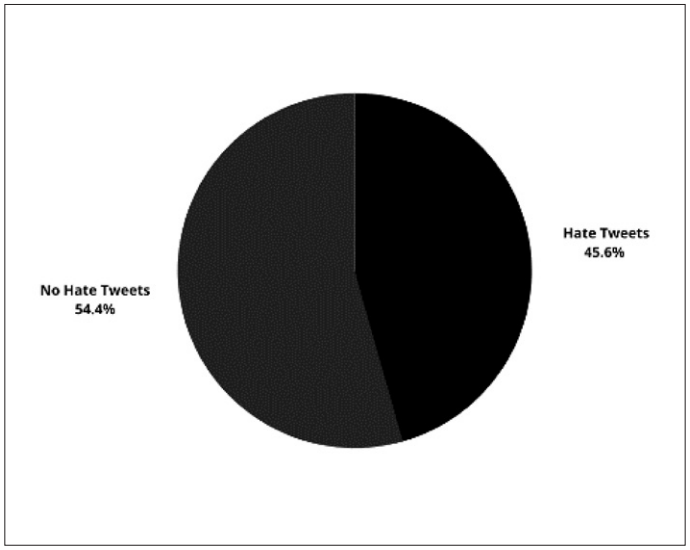

Fig. 2. Percentage of hate tweets.

Once it has been established that the tweet contains hate, the second step is to classify the rhetoric and anti-Gypsyism forms in the text. Accordingly, the results refer to the 410 hate tweets identified in the data set. In terms of rhetoric, 46.5% of the tweets were classified as prejudice and 34.5% as insults, with other forms in the minority (incitement/violence: 5%; derision/irony: 4%; competition: 3%; humiliation/disdain: 3%; exclusion/separation: 2%; dehumanisation: 1.5%; fear: 0.5%) (Figure 3).

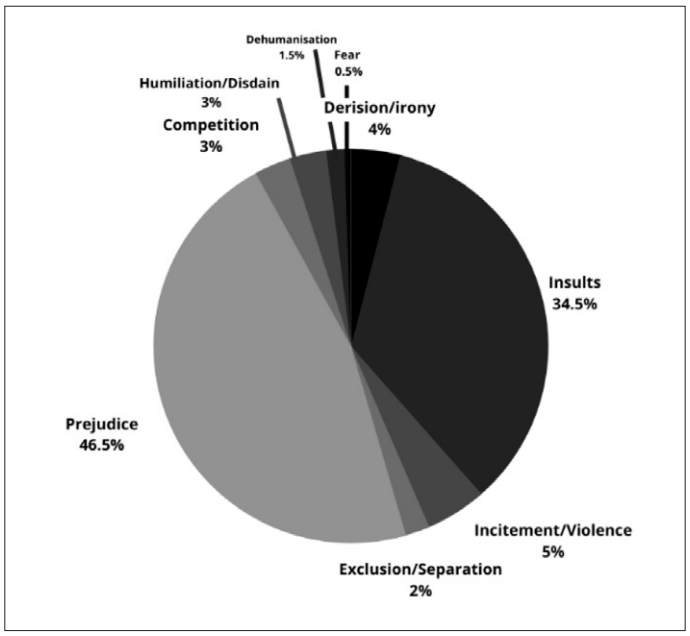

Fig. 3 Rhetoric of hate.

However, as for the forms of anti-Gypsy hate, 75.7% were anomie-based, 6.4% were differential, 9.3% were competition based and 8.6% were elimination based[13] (Figure 4).

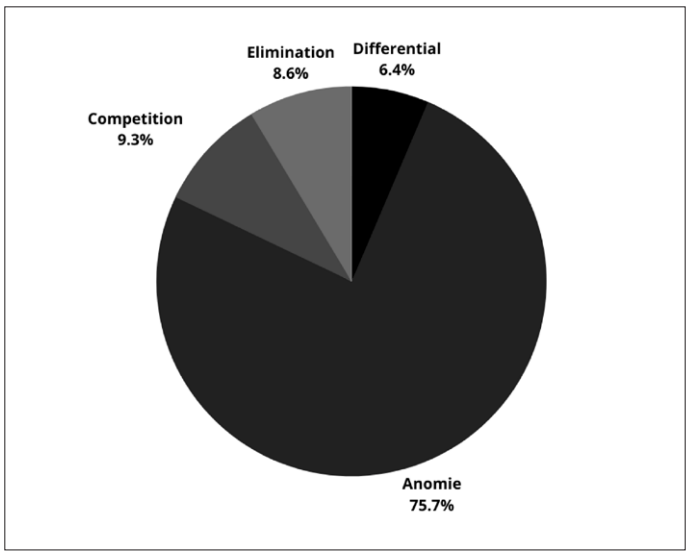

Fig. 4. Forms of anti-Gypsy hate.

In this article, we do not focus on the discussion of anti-Gypsy hate content online, rather the detection methods with regard to the potential for automatic search methods to detect hate.

6. Towards the 'Spectrum of Online Hate'

The analysis conducted reveals the need to contextualise signs (words, images, memes and so on). Indeed, in this study, a corpus of tweets in Italian addressed to a specific target was manually annotated. Other case studies have mentioned a strong discrepancy between the manual annotation process, also known as tagging, which requires the collaboration of

[13] 311 tweets were anomie-based, 26 tweets were differential, 38 were competition-based and 35 were elimination-based.

experts, and annotation using only the algorithm created with the negative word dictionary in other studies.[14] It is believed that semantic analysis alone is not sufficient to correctly automate such a complex process. It is necessary to have knowledge of the reality and therefore of the context in which it is found, a factor which annotators can verify.

On the one hand, as Stanley (1994) points out, one property of slurs is that they retain their derogatory meaning regardless of the form of the utterance in which they appear (Wodak & Meyer, 2011). On the other hand, as De Mauro (2016) has shown, even non-offensive terms can be used in a hostile way, especially since online platforms enable the use of other means, such as images, fonts, metaphors, stereotypes and prejudices, which acquire particular relevance when supported by the mechanism of 'othering', that is, a set of dynamics – processes, structures, including linguistic – that dialectically group subjects into an 'us' and a 'them', presented as groups that are homogeneous and alternative to each other (Powell & Menendian, 2018) – and 'perspectivation' strategies for us-versus-them polarisation (Graumann & Kallmeyer, 2002).

One of the characteristics that makes hate speech so elusive is the challenge of grasping meaning when an utterance is isolated from context. As Bruner has shown (1997), algorithmic logic deals with information that is already encoded, the meaning of which is established in advance. Computational logic deals with stimuli and responses, not with the meaning to be attributed to things; it processes information, while those who create culture and education interpret and produce meaning, an operation laden with ambiguity and above all sensitive to context. Therefore, in the detection of hatred, algorithms alone will not suffice, given that discourses and narratives are concealed, masked under an ordinary lexicon and able to employ different repertoires and registers. For this reason, studies that limit themselves to detecting insults, or that only detect the presence of hate words, are partial; it is therefore appropriate to experiment with search techniques that integrate the two phases of human and automatic classification, applying interdisciplinary approaches and drawing on in-depth knowledge of the manifestations of the phenomena under investigation.

This study by the Mediavox Observatory on detection mechanisms that combine human annotation with automatic computerised procedures reveals the need to go beyond the single hate/non-hate definition (although it is difficult to determine how this should be done), in line with other international groups. For example, this is the direction taken by researchers from the Institute of Linguistics and Language Technology at the University of Malta who, for hate speech about migration and LGBTIQ+ issues, use Cortese's Scale, annotating in a hierarchical manner: (1) whether the post communicates a positive, negative or neutral attitude; (2) if it is negative, to whom this attitude is directed at (individual/group); (3a) if it is addressed to an individual, whether this is because of his/her belonging to a group and if so to which one; (3b) if it is addressed to a group, the name of the group; (4a) how the attitude is expressed in relation to the target group (derogatory term / generalisation / insult / sarcasm (including jokes and trolling) / stereotyping / suggestion / threat); (4b) if the post involves a suggestion, whether it is a suggestion that incites violence against the target group (Assimakopoulos et al., 2020).

The Institute for Media and Communication Studies at the Freie Universität Berlin proposes focusing on the grey area between hate speech and offensive language based on a text-annotation study of 5,031 user comments in German on the topic of immigration and refugees on news sites, Facebook pages, YouTube channels and one right-wing blog (Paasch-Colberg et al., 2021).[15] It contains three attributes to specify different forms of offensive language, namely, insults and slurs, derogatory metaphors and comparisons and derogatory wordplays.

In light of these considerations, the Mediavox Observatory researchers (Milena Santerini, Stefano Pasta) propose classifying online content according to seven indicators called the 'spectrum of online hate'.

In Figure 5, manual annotation using the seven indicators was applied to the data set of tweets identified as containing anti-Gypsy hatred.

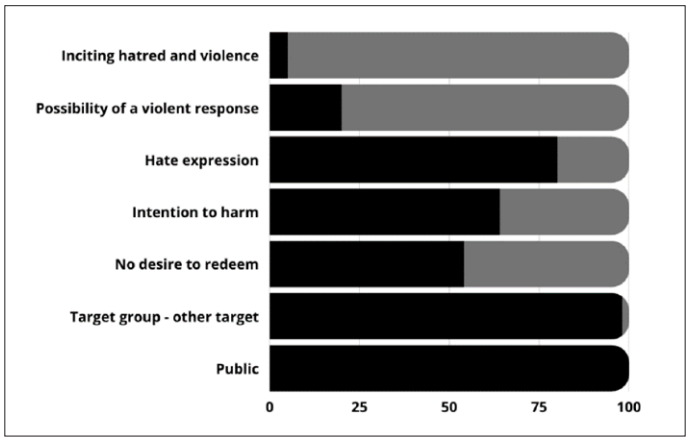

Fig. 5 - The 'spectrum of online hate'.

[14] In two case studies on anti-Semitism (Pasta et al., 2021), out of 235 tweets classified as hate speech by the annotators, the confusion matrix algorithm only identifies 14 correctly, while in the second study, investigating a link between anti-Semitism and COVID-19, of the 147 tweets classified as hate speech by the annotators, the algorithm only identifies 10 correctly.
In the case study on Islamophobia (Pasta, 2021), of the 165 tweets classified as hate speech by the annotators, the algorithm only correctly identified 13, and the percentage of hate speech identified by the algorithm was 6.9% of the tweets analysed compared to 18.6% by the annotators.

[15] The study uses both machine learning and manual annotation; the software used is BRAT, a browser-based tool for rapid annotation of structured text.

The seven indicators are as follows.

- *Public*: The content can be seen without restriction by users, so it is not a private message in a closed circle.
- *Target group – other target*: Hate content affects a target group, or an individual linked to that group, or an individual because of what he or she represents (a vaccine campaign testimonial, for example); often the targets are minorities or their members; this selection process distinguishes this content from cyberbullying.
- *No desire to redeem*: The hate-speech author is not interested in changing the victim's mind, only in insulting and hurting them. It should also be considered that very often the target of the hate speech is not the interlocutor but the subject of the speech. Therefore, the aim of the hate speech author is rarely to 'redeem' the target, not least because the most structured forms of hate do not allow for redemption since the target is hated 'for what they are' (Jewish, foreign, Black, Roma...) regardless of their individual behaviour.
- *Intention to harm*: The author's intent to harm the victim (individual or target) is clear in the content. Again, the target of the hate speech is frequently the object of the discourse rather than the interlocutor, so the speech often takes the form of an attempt to provoke hostility towards the target.
- *Hate expression*: This contains hate speech in an explicit verbal form, that is, hate words, verbally explicit insults, or speech in which one denies the other person as a person, considering them inferior or attributing negative qualities to them, insulting them or humiliating them.
- *Possibility of a violent response*: The tone of the speech is marked by such violence or intensity that, in line with the typical mechanisms of toxic online disinhibition, they could easily and quickly lead to the inciting of hatred and violence.
- *Inciting hatred and violence:* This is hate speech that explicitly and directly incites hatred and violence, through which the author aims to increase the numbers of co-producers of hate speech. In these cases, online hate is a particularly potent breeding ground for offline hate actions and hate crimes.

Digital content can be hate speech even when not all seven indicators apply (on the contrary, it is definitely hate speech if all seven indicators apply), but this scale can help identify the intensity of the hate speech and certain characteristics relevant to combating it.

References

Alshalan, R., & Al-Khalifa, H. (2020). A Deep Learning Approach for Automatic Hate Speech Detection in the Saudi Twittersphere", *Applied Sciences, 10*(23), 8614.

Amnesty International Italia (2021). *Barometro dell'odio. Intolleranza pandemica*. Roma.

Amrutha, B.R., & Bindu, K.R. (2019). Detecting hate speech in tweets using different deep neural network architectures. *International Conference on Intelligent Computing and Control Systems (ICCS)*, IEEE, pp. 923-926.

Arcila-Calderón, C., Blanco-Herrero, D., Frías-Vázquez, M., & Seoane-Pérez, F. (2021). Refugees Welcome? Online Hate Speech and Sentiments in Twitter in Spain during the Reception of the Boat Aquarius. *Sustainability*, *13*(5), 2728.

Arrigoni, P., & Vitale, T. (2008). Quale legalità? Rom e gagi a confronto. *Aggiornamenti Sociali*. 3, pp. 182-194.

Assimakopoulos, S., Vella Muskat, R., van der Plas, L., & Gatt, A. (2020). Annotating for Hate Speech: The MaNeCo Corpus and Some Input from Critical Discourse Analysis. *Proceedings of the 12th Conference on Language Resources and Evaluation*, Marseille, LREC, 5088–5097.

Balibar, É., & Wallerstein, I. (1991). *Race, nation, class: ambiguous identities.* London-New York: Verso.

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python.* Sebastopol: O'Reilly.

Bravi, L. (2009). *Tra inclusione ed esclusione. Una storia sociale dell'educazione dei rom e dei sinti in Itali*a. Milano: Unicopli.

Brena, S. (2021). Mappa dell'Intolleranza, così le donne sono da sempre nel mirino. In S. Pasta, & M. Santerini (Eds.), *Nemmeno con un click. Ragazze e odio online* (pp. 68-79). Milano : FrancoAngeli.

Bruner, J.S. (1997). *The Culture of Education*. Cambridge, MA: Harvard University Press.

Cao, R., Lee, R.K.W., & Hoang, T.A. (2020). DeepHate: Hate speech detection via multi-faceted text representations. *12th ACM Conference on Web Science*. New York: AMC, pp. 11-20.

Capozzi, A.T.E., Lai, M., Basile, V., Poletto, F., Sanguinetti, M., Bosco, C., Patti, V., Ruffo, G., Musto, C., Polignano, M., Semeraro, G., & Stranisci, M. (2020). "Contro L'Odio": A Platform for Detecting, Monitoring and Visualizing Hate Speech against Immigrants in Italian Social Media. *Italian Journal of Computational Linguistics*, *6*(1), 77-97.

Council of Europe (2021). *Draft Recommendation of the Committee of Ministers to member States on combating hate speech*. Strasbourg: Committee of Expert on Combating Hate Speech.

D'Amico, M., & Siccardi, C. (Eds.) (2021). *La Costituzione non odia. Conoscere, prevenire e contrastare l'hate speech on line*. Torino: Giappichelli.

De Mauro, T. (2016). Le parole per ferire. *Internazionale.*

Di Nicola, A., Andreatta, D., & Martini, E. (2020). *Hatemeter. Hate speech tool for monitoring, analysing and tackling Anti-Muslim hatred online*, eCrime, Trento.

Dubey, K., Nair, R., Khan, M.U., & Shaikh, P.S. (2020). Toxic Comment Detection using LSTM. *2020 Third International Conference on Advances in Electronics, Computers and Communications (ICAECC),* 1-8.

Eidelson, R., & Eidelson, J. (2003). Dangerous ideas. Five beliefs that propel groups toward conflict. *American Psychologist Journal, 58*(3), 182-192.

Faloppa, F. (2019). *#Odio. Manuale di resistenza alla violenza delle parole*. Milano: Utet.

Femia, D. (2019). Discorso dell'odio e risorse per il trattamento automatico delle lingue. In R. Petrilli (Ed.), *Hate speech* (pp. 147-164). Torino: Round Robin.

Ferrini, C., & Paris, O. (2019). *I discorsi dell'odio*. Roma: Carocci.

Fersini, E., Nozza, D., & Rosso, P. (2020). AMI @ EVALITA2020: Automatic misogyny identification. *7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian.* EVALITA. CEUR-WS.

Graumann, C.F., & Kallmeyer, W. (2002). *Perspective and Perspectivation in Discourse*. Amsterdam: Benjamins.

Heinzelmann, N., Höltgen, B.T.A., & Tran, V. (2021). Moral discourse boosts confidence in moral judgments. *Philosophical Psychology, 34* (8), 1192-1216.

Isbister, T., Sahlgren, M., Kaati, L., Obaidi, M. & Akrami, N. (2018). Monitoring targeted hate in online environments. *arXiv preprint:* 1803.04757.

James, G., Witten, D., Hastie, T., & Tibshirani, T. (2017). *An Introduction to Statistical Learning.* Berlin: Springer.

Jikeli, G., Cavar, D., & Miehling, D. (2019). Annotating antisemitic online content. towards an applicable definition of antisemitism. *arXiv:1910.01214*, 1-27.

Jo Cox Committee on hate, intolerance, xenophobia and racism (2017). *Final Relation*. Rome: Camera dei deputati.

Kommattam, O., Fischer, A., & Jonas, K. (2019). Perceived to feel less: intensity bias in interethnic emotion perception. *Journal of Experimental Social Psychology*, 84, 1-8.

Lazzardi, S., Patti, V., & Rosso, P. (2021). Categorizing Misogynistic Behaviours in Italian, English and Spanish Tweets. *Procesamiento del Lenguaje Natural, Revista*, 66, 65-76.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Proceedings of NIPS*, 3111–3119.

Novak, P.K., Mozetič, I., De Pauw, G., & Cinelli, M. (2020). *Hate speech detection and trends. Multilingual Hate Speech Database.* Ljubljana: IMSyPP.

Ozalp, S., Williams, M.L., Burnap, P., Liu, H., & Mostafa, M. (2020). Antisemitism on Twitter. *Social Media+ Society*, *6*(2), 1-20.

Paasch-Colberg, S., Strippel, C., Trebbe, J., & Emmer, M. (2021). From Insult to Hate Speech: Mapping Offensive Language in German User Comments on Immigration. *Media and Communication*, *9*(1), 171-180.

Pasta, S. (2018a). *Razzismi 2.0. Analisi socio-educativa dell'odio online*. Brescia: Morcelliana Scholé.

Pasta, S. (2018b). Addressing Antigypsyism. In D. De Vito, A. Ciniero, L. Bravi & S. Pasta, *Civil society monitoring report on implementation of the national Roma integration strategy in Italy. Focusing on structural and horizontal preconditions for successful implementation of the strategy* (pp. 2939). Luxembourg: Publications Office of the European Union.

Pasta, S. (2019). The media and the public perception of the Roma and the Sinti in Italy. *Trauma and Memory*, *7(*1), 46-52.

Pasta, S. (2021). Detection di odio antimusulmano tra machine learning e valutazione qualitativa. In S. Polenghi, F. Cereda, & P. Zini (Eds.), *La responsabilità della pedagogia nelle trasformazioni dei rapporti sociali. Storia, linee di ricerca e prospettive* (pp. 1169-1179). Lecce-Rovato (BS): Pensa Multimedia.

Pasta, S. (2020a). 'Theory of nomadism' in regional laws and national legislative vacuum. In D. De Vito, S. Pasta, A. Ciniero, & L. Bravi*, Civil society monitoring report on implementation of the National Roma Integration Strategy in Italy. Identifying blind spots in Roma inclusion policy* (pp. 11-20). Luxembourg: Publications Office of the European Union.

Pasta, S. (2020b). Didattica della memoria. Insegnare il Porrajmos, contrastare l'antiziganismo e prevenire l'elezione a bersaglio di rom e sinti. *Consultori Familiari Oggi, 28*(1), 54-68.

Pasta, S. (2022). Social network conversations with young authors of online hate speech against migrants. In A. Monnier, A. Boursier, & A. Seoane (Eds.), *Cyberhate in the Context of Migrations* (pp. 187-214). London: Palgrave MacMillan.

Pasta, S., & Vitale, T. (2017). 'Mi guardano male, ma io non guardo'. Come i rom e i sinti in Italia reagiscono allo stigma. In A. Alietti (Ed.), *Razzismi, discriminazioni e disuguaglianze. Analisi e ricerche sull'Italia contemporanea* (pp. 217-241). Milano: Mimesis.

Pasta, S., Santerini, M., Forzinetti, E., & Della Vedova, M. (2021). Antisemitism and Covid-19 on Twitter. The search for hatred online between automatisms and qualitative evaluation. *Form@re. Open Journal per formazione in rete*, *XXI*(3), 288-304.

Paul, C., & Bora, P. (2021). Detecting Hate Speech using Deep Learning Techniques. *International Journal of Advanced Computer Science and Applications (IJACSA*), *12*(2), 619-623.

Piasere, L. (2015). *L'antiziganismo*. Macerata: Quodlibet.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, *55*, 477–523.

Powell, J., & Menendian, S. (2018). The Problem of Othering. *Othering and Belonging*.

Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valletta, ELRA, 45-50.

Sandaruwan, H.M.S.T., Lorensuhewa, S.A.S., & Kalyani, M.A.L. (2019). Sinhala Hate Speech Detection in Social Media using Text Mining and Machine learning. *19th International Conference on Advances in ICT for Emerging Regions (ICTer)*, 1-8.

Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., & Stranisci, M. (2018). An Italian Twitter Corpus of Hate Speech against Immigrants. *Proceedings of the 11th International Conference on Language Resources and Evaluation*, LREC, Miyazaki, 2798-2805.

Santerini, M. (2021). *La mente ostile. Forme dell'odio contemporaneo.* Milano: Raffaello Cortina.

Siapera, E., Moreo, E., & Zhou, J. (2018). *Hate Track Tracking And Monitoring Racist Speech Online*. Dublin: Irish Human Rights and Equality Commission.

Siegel, A.A. (2020). Online hate speech. In N. Persily, & J.A. Tucker (Eds.), *Social Media and Democracy* (pp. 56-88). Cambridge: Cambridge University Press.

Stanley, J. (1994). *How Propaganda Works*. Princeton: Princeton University Press.

Sternberg, R. J., & Sternberg, K. (2008). *The nature of hate*. Cambridge: Cambridge University Press

Uyheng, J., & Carley, K.M. (2021), Characterizing network dynamics of online hate communities around the COVID-19 pandemic. *Applied Network Science*, 6(20), 1-21.

Vox-Osservatorio italiano sui diritti (2021). VI Mappa sull'Intolleranza in Italia. Retrieved from www.voxdiritti.it.

Wieviorka, M. (1995). *The arena of racism*. New York: Sage.

Wodak, R., & Meyer, M. (2011). *Methods of Critical Discourse Analysis* London: Sage.

Zannettou, S., Elsherief, M., Belding, E., Nilizadeh, S., & Stringhini, G. (2020). Measuring and Characterizing Hate Speech on News Websites. *12th ACM Conference on Web Science*, AMC, New York, 125-34.

Zhou, Y., Yang, Y., Liu, H., Liu, X., & Savage, N. (2020). Deep Learning Based Fusion Approach for Hate Speech Detection. *IEEE Access*, vol. 8, 128923-128929.

Zimbardo, P. (2007). *The Lucifer Effect: How Good People Turn Evil*. New York: Random House.