Mediascapes journal 20/2022

Platformization hate. Patterns and algorithmic bias of verbal violence on social media[*]

Miriam Di Lisio[**]
University of Naples Federico II

Rosa Sorrentino[***]
University of Naples Federico II

Domenico Trezza[****]
University of Naples Federico II

The paper presented is an analysis of the Hate Speech of tweets during the implementation of the EU's Digital Covid Certificate policy. The work starts from the assumption that Hate Speech is an often "submerged" phenomenon because it also includes some forms recognized as "incivility." Therefore, there are two research questions: the first asks what are the new categories of "hate" that emerge in the EU Digital Covid Certificate policy debate, while the second questions the methodological implications on the use of algorithms in detecting the phenomenon. The results we arrived at are, from a substantive point of view, of good interest because they show us how it is possible to witness a new kind of online hatred. However, the disagreements we encountered in constructing an unambiguous definition of HS for the supervised algorithm leave open many questions. Among them is the fact that the differences between HS, incivility, and even freedom of expression can be very small. In the context of large social platforms, where the criteria of the algorithm are not always explicit and are also the policies of the platform, this could be a problem.

Keywords: *Hate speech, Algorithms, Social Media, Digital Methods*

[*] Articolo proposto il 26/05/2022. Articolo accettato il 26/12/2022.
The work has been common. However, the following attribution is possible. Domenico Trezza: “Introduction. From Online Incivility to the platformization of hate” and “Research questions and methods for data building”; Rosa Sorrentino: “Hate speech. The normative evolution of an ambiguous concept”; Miriam Di Lisio: “Results. Description of HS Tweets”. Conclusions are common work.
[**] miriam.dilisio@unina.it
[***] rosa.sorrentino@unina.it
[****] domenico.trezza@unina.it

From online incivility to the platformization of hate

Digital environments from web 2.0 forward have offered the chance for users to generate their content and communicate anonymously. It is believed that the network can humanize communication and help reaching audiences in a more personalized and authentic way (Ward & Lusoli 2005; Jackson & Lilleker, 2011; Lilleker & Jackson 2014); but on the other side, social platforms have also been accused of facilitating the exacerbation of democratic debate and enabling the normalization of abuses (Amnesty International, 2018; Atlanta, 2018; Inter Parliamentary Union, 2016). This would have implied the abundance of phrases, incitements and sentiments of hate, especially towards public targets or vulnerable groups (Ziccardi, 2016). Such attitudes fall within the sphere of what is defined as “online incivility”, an umbrella term for offensive statements that violate the ideal type of democratic communication (Waisbord, 2018; Anderson et al., 2014; Papacharissi, 2004). Online incivility encompasses acts of online rudeness (Jamieson, 1997) and outrageous statements toward different actors or groups and implies that interlocutors are not treated with respect. It can, for example, take the form of name-calling, profanity, negative stereotyping, lack of interpersonal respect, (digital) shouting (Chen & Lu, 2017; Coe et al., 2014), quarrelsome and insolent conduct, harassment, and incitement, configurations that are finding their way among users of digital platforms (Antoci, Delfino, Paglieri, Panebianco, & Sabatini, 2016). These elements emerges especially in digital platforms, where - although the intervention of censoring (Boccia Artieri & Marinelli, 2018) - the proliferation of hate content is continuous, leading - similar to the “platformization of culture” (Duffy, Poell, & Nieborg, 2019) - to a “platformization of hate”, and making distinguishing the phenomenon extremely complicated.

Hate speech (HS) is commonly regarded as the most serious type of online incivility. The study of hate speech has been of particular interest to the social sciences as it takes on different physiognomies than media in general and social networks in particular (Nielsen, 2002). Given the complexity of the phenomenon, it is difficult to give a true definition of it. Therefore, - in line with this research objectives - it was preferred to define hate speech as any statement that expresses an attack, abuse, intimidation, and/or denigration of individuals and groups defined on the basis of an external group they are said to be a part of (Van Spanje & DeVreese, 2014; Walker, 1994; Warner & Hirschberg, 2012). These offensive discourses may be directed at different individuals based on ethnicity, religion, gender, or nationality and may contain threatening language or explicitly incite violence. Widespread on the Net, hate speech does not have yet a strict legal limitation because it clashes with another sensitive issue, that of freedom of expression. In some countries hate speech is indeed banned, although in many others there is no legal framework regulating verbal violence on the net. Despite this, in many European countries there have been many prosecutions for hate speech (Vrielink, 2016).

Although the classification of hate contents is part of the researcher's job, when referring to large portions of text the task is often handled automatically by algorithms. The tracking of violent content is now based on the use of supervised algorithms, that are obviously affected by imperceptible linguistic and meaning bias.

The results of tracking bias seem to be substantially related to over-representation and under-representation of the phenomenon. The first can have a considerable impact on freedom of expression, while the second can underestimate the problem and thus allow the circulation of inappropriate content. Also, as is well known, algorithmic classification not infrequently leads to distorted results, especially when dealing with the language and content of social media, for example with all the linguistic complexities arising from the use of abbreviations, slang, ironies, etc. (Aragona, 2021; Leavy, 2018).

However, before analyzing the phenomenon, it is necessary to define it. As we just illustrated, hate speech is very complex to define, disambiguate and delimitate. In fact, the experience in this work highlighted that it is not easy to unambiguously define and share the meaning, forms and effects of “hate”. We realized this since the structuring of research design, because - despite the effort made in theoretical reconstruction - many cases of disagreement were found in the relevant literature, but also among us authors. Incivility is understood, according to an orientation more focused on the style of the interaction, to the tone and words chosen by the communicator: more simply, incivility is circumscribed as an infraction of social rules and is represented by gratuitous insults, the use of sarcasm, and insults, all of which tell of a lack of respect for the other (Rega, & Marchetti, 2019). Starting from this generic definition, the question arises whether or not it is possible to extend the concept of hate speech to indirect, mild, or potential violent forms, looking not only at verbal expressions but also at the potential intentions behind them.This possibility inevitably leads to questioning the illustrated dividing line between hate speech and the broader concept of incivility, although it is often complex to analyze elements such as the real intentions of social actors and potentially hidden meanings in the text. For these reasons, the authors-although aware of the conceptual differences between HS and incivility-agreed that it was necessary to broaden the semantic scope of the term "hate speech" by deliberately and critically including forms assimilated to incivility. Such conscious encroachment is believed to provide a contribution to reflection on the topic, and to serve the objectives of the research. Therefore, it is necessary to point out that the construction of the algorithm was influenced by this common and broader representation we make of this social object.

Returning to the focus of the research, the Covid-19 pandemic has given impetus to the study of incivility and hate online in particular: in fact, during the emergency, the phenomenon has grown and the proliferation of hate content has been out of control (Caiani, Carlotti, & Padoan, 2021). Recent work by Druckmann et al. (2020) shows us that there was a strong association between the incivility of citizens and their attitudes toward public health emergencies.

In Italy, the lockouts, the implementation of the vaccination campaign, and the launch of the "EU Digital Covid Certificate" (DCC) for those vaccinated only, have inflamed the debate and in many cases exacerbated the violence of online verbal conflict (Uyeng, & Carley, 2021). The contribution aims to better understand the phenomenon under research, offering an analysis of the verbal violence of tweets in the time period starting August 9 and ending

September 25, which is the highlight of the introduction of the DCC, and thus when there was presemably a very hot debate online.

The paper is divided into three macro-paragraphs. The first introduces the issue of HS, which is closely related to incivility, trying to delimit the terminological field and making a literature review on work-related issues. The second section is methodological: it illustrates the phases of data construction, from the operational definition of HS to the preliminary choices for the automatic classifier construction. The third describes the analysis procedures and results, exploring the most significant contents of HS communication. Finally, the conclusions try to reason on the opportunity to monitor these areas of debate where new forms of hate seem to be emerging.

Hate speech. The normative evolution of an ambiguous concept

The affirmation of digital communication technologies has exponentially expanded the human capacity to share ideas, opinions, and moods, influencing the timing, methods, and contents of the communication. First with the development of websites and blogs, and then with the birth of social networks, people begun to produce and disseminate a large number of messages and contents of various kinds, rarely filtered and moderated (Nardi, 2019). However, the visibility offered by social networks has also led to an uncontrolled spread of violent and discriminatory content known as hate speech.

The concept of HS belongs to a category developed in the 1970s by US jurisprudence, which coined the term in relation to the countless cases of racism that occurred in university campuses in those years (Waldron, 2021). The term generally indicates all those discourses expressing hate and intolerance towards a person or a group, and which risk provoking violent reactions (Pino, 2008). A first and clear explanation of the concept is found in Recommendation No. 20 of the Committee of Ministers of the Council of Europe (1997), which states that:

> [...] the term "hate speech" shall be understood as covering all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin[1].

UNESCO[2] identifies four essential differences in HS that spreads through the web from “traditional” one. First and foremost, the *permanence* of hate, that is, the possibility of online hate manifestations to persist for long periods and in different formats, traveling through the web. The second is the possibility that in the net the manifestation of hatred could re-emerge again through another platform. Third, anonymity, which allows people behind the screen to feel protected and at ease in expressing hatred. Finally, the transnationality of online HS, an element that poses complications in identifying the legal mechanisms to contrast it (Ziccardi, 2021). These characteristics configure online hate speech as a phenomenon that – thanks to the possibilities offered by the Internet – becomes even more meaningful, extended, and transversal, and which, as in the past, continues to act against the most defenseless, discriminating on ethnicity, race, religion, gender, sexual orientation, socio-economic conditions, and appearance[3].

Because hate speech leverages values that are not conceived equally in all socio-political contexts, such as equality, human dignity, and freedom of expression (Cabo Isasi, & García Juanatey, 2016), it is extremely difficult to give a unanimous, unambiguous and complete definition, which would allow to identify and fight the phenomenon in a systematic way. Many are the difficulties that legislators have to face: from the identification of criteria through which to discern the "true" incitement to hatred from satire or expression of opinion to the definition of the role of platforms, arriving at the definition of the right boundaries between *freedom of expression* (or *freedom of speech*) and *hate speech*. Freedom of expression refers to the ability of an individual or group to express their beliefs, thoughts, ideas, and emotions on various issues without censorship. Nevertheless, it is not an absolute right. As anticipated, the speed with which it is possible to express oneself through mass media and the deceptive perception of "anonymity" and "disinhibition" - understood in both negative and positive senses - (Suler, 2004) about online making the experience of users on the web as seemingly free. The Internet is perceived as a space in which there is greater freedom of expression, which can also translate into the perception of being able to enact unprecedentedly strong verbal insults and violence without repercussions, highlighting how incivility is strongly present within the messages disseminated online (Rega, & Marchetti, 2019) and in the intentions that animate them.

While hate online is not a new phenomenon, its growth *trend* is surprising. In recent years, issues such as immigration, terrorism, politics and –last but not least– the Covid-19 pandemic, caused speculations which relies not only on the general feeling of fear and insecurity produced by the instability of the times, but also on the confusion and disinformation caused by the excess of news, mainly through social media. After all, the perception is that hate speech - overt or otherwise - has now become so normalized and fully entered the public debate that it is present in many communicative experiences - from print newspapers to WhatsApp groups - communication described as "uncivil", "aggressive," and "violent." (Waldron, 2021). As scholars Rega and Marchetti (2019) point out, "the problem is that public debate has been transformed, has become radicalized, and is increasingly characterized by the attack on the integrity of political opponents and the use of offensive language. Forms of incivility are not hindered, but in some cases promoted by political actors (top-down) who incite the active online public (bottom-up) to uncivil discourse". Therefore, with this in mind, it seemed appropriate to analyze what are the new target profiles of online violence that emerged during the pandemic, particularly during the period when DCC was introduced in Italy. For example, numerous studies have shown that in the lockdown period (that occurred in the early 2020s), due to the virus, new targeted categories were created (Bentivegna, & Rega, 2020). This is one of the motivations that prompted us to investigate the new HS target profiles in the green pass period. Also, in the period related to the Covid-19 pandemic it has been realized that the forms of resentment are not always the same, but change concerning motivations, historical moments, forms, and meanings, and turning out to be completely impossible to predict.

However, hateful expressions could be a manifestation of freedom of expression and, at the same time, are in contrast with the fundamental principles of protection of the person, respect for human dignity and the principle of non-discrimination. Consequently, by directly touching the roots of constitutionalism, HS needs to be framed within a legal framework (Pollicino, & De Gregorio, 2019), especially with the rise of Web 2.0. In 2001 it was introduced the Additional Protocol to the "Budapest Convention on Cybercrime"[4], signed by Italy in 2011 (although the Protocol has not yet been ratified). In 2019, the European Commission and major IT companies (Google, Facebook, Twitter and Microsoft) signed a "Code of Conduct" on HS online[5]: this code provides for a series of joint activities between public and private institutions, whose purpose is to make the verification and removal of hate comments that are reported on their platforms faster and more effective. With these measures European multilevel approach is trying to build a policy-network that could become a regulatory umbrella for national governments (Scamuzzi, Belluati, Caielli, Cepernich, Patti, Stecca, & Tipaldo, 2021). However, Codes of Conduct are fragile instruments, based on a voluntary commitment by the countries that sign them. Thanks to these codes, social companies have started to act against the phenomena of online hate. However, the question remains open linked to the linguistic and cultural dimension in which online hate take place. In addition, another issue is that of the reticence of States, which perceive European directives as non-binding and often interfering in their "sovereignty".

Currently in Italy there is no criminal law on HS, although the first regulation that stigmatized racial discrimination dates back to 1952, the "Scelba Law"[6]. Recently, following the media and political debate on the growing cases of harassment and violent verbal attitudes towards minors (cyberbullying), specific legislation was also adopted to protect minors and introduce some useful tools for removing harmful content from the network[7]. In this context should also be considered the recent Regulation proposed by the Communications Authority, adopted with Resolution No. 157/19/CONS ("AGCOM Regulation"), which aims to prevent conduct that incites hatred based on ethnicity, gender, religion or nationality in the context of audiovisual media services, and the creation of a task force against HS by Amnesty International Italy.

As already said, in Italy in recent times in addition to the cited *green pass* - "aimed at facilitating the free and safe movement of citizens in the European Union during the COVID-19 pandemic"[8] – also the Zan Bill (DDL Zan), a bill providing for an increase in penalties for crimes and discrimination against homosexual, transsexuals, women and disabled persons, inflamed the debate online and so the manifestations of hatred. Both measures have generated disputes on the part of those who see in them a restriction of freedom of expression (Pignatiello, 2021).

There is no legislation against hate speech to help counter the phenomenon as framed: reflection and regulation on hate speech and its digital manifestations are far from a definitive resolution. In this context, the Net, by facilitating and speeding up the spread of digital messages, on the one hand, intensifies freedom of expression and on the other offers itself as a container for an impressive flow of violent or potentially violent contents.

Research questions and methods for data building

The debate on COVID has been very hot, creating strong disparities between those against and those in favor of measures against the virus (lockdown, vaccines, green pass). Verbal violence has been very common, with frequent circulation of fake news (Di Lisio, & Trezza, 2021). The contribution presented in this paper is part of a broader work that, while presenting exploratory purposes, focuses on the last emergency period, related to the introduction of the DCC in Italy. DCC is the certificate that attests the vaccination and was introduced in Italy with the decree of June 17, 2021 and implemented from August 6, 2021. The introduction of this measure has been controversial and perceived by some people as a measure of 'control' by the institutions. This has exacerbated the debate, especially on social media where the spread of HS content has been very high, especially towards representatives of institutions but also towards ordinary users expressing agreement or disagreement towards this measure. Despite the exploratory purposes, there are some research questions to which we want to pay more attention because in our opinion they allow us to delimit the cognitive purposes regarding the HS phenomenon in the context of the COVID emergency and, specifically, in relation to the introduction of the certificate:

1# In this emergent phase in which the verbal conflict is very hot, what are major topics and new target emerging from this redefinition of HS?

2# How effective can the use of a supervised algorithm for detecting such a complex phenomenon return?

Data collection

The corpus consists of 64589 tweets about the DCC, posted between August 9 and September 25, at the height of the debate on the issue. The social platform Twitter is very relevant to our research goals because, following a ‘follow the medium’ approach (Rogers, 2009), we used the potential of this social to index open-access content via using semantic keys (hashtags) and extract tweets automatically via Twitter's Application Programming Interface (API). Twitter is therefore very useful for our purposes: it has the content easily identifiable by indexing with hashtags, and, most importantly, it is all public.

In addition, this platform has often been the context of scurrilous, hateful and offensive language. This was initiated based on a few simple inputs consistent with our needs related to the timeframe, numerosity, and other characteristics of the corpus. Extraction keys were selected based on topic trends in Italy over the previous 36 hours and then constantly updated. The base of hashtags was constituted by the textual keys: #greenpass, #passcovid, #certificatoverde, #certificatocovid, i.e., the most used tags in the discussions on green pass. We assumed, then, that during these days the social debate on the topic was very lively. The API extraction was done through the “Search Tweets” function of the R package “Rtweet” This function is associated with the academic version of Twitter's V2 APIs, which allows automatic extraction of tweets, with access to the platform's full archive, without limits. The main advantage of working with this function lies in the possibility of obtaining data already structured in matrix through 90 variables, most of which related to information of little interest for our purposes. The reduction of the matrix was therefore inevitable: we considered only the information related to the tweet and its source (account). Concerning the account, we have considered identity (id and nickname) and social engagement (number of followers, friends, listed, statutes and favorites) information. On the other hand, tweet information concerns general characteristics (date, text and text length) and engagement (number of retweets and favorites).

Preliminary operations on tweets

From database of tweets, a sample of 1500 tweets was extracted with the following objectives:

1. Perform an initial exploration of the tweets through manual tagging of the tweets;
2. Develop a classification plan for HS content;
3. Create a base of pre-labeled HS tweets for automatic supervised tweet classification.

To have a satisfactory criterion of representativeness of the main corpus, it was constructed to cover the entire time range, which we divided into 9 periods as shown in Fig.1

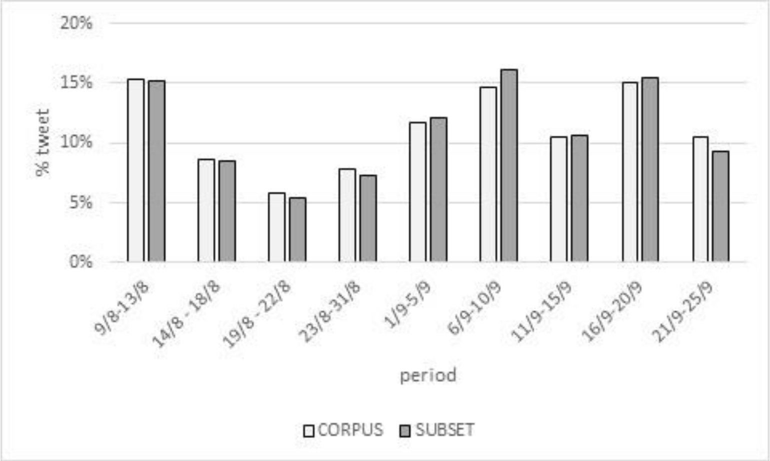


Figure 1 - % Tweet by period, corpus and subset

Classifying hs for machine learning goals

Collecting and annotating data to train automated classifiers to detect HS is challenging. In particular, identifying and agreeing on whether a specific text is HS is difficult and, as mentioned earlier, there is no universal definition of HS. Ross, et al. (2017) studied the reliability of HS annotations and suggest that annotators are unreliable. Therefore, it was crucial for our investigation to a) define the meaning of HS to have a classification criterion that is as shared and standard as possible; b) stipulate a scale of HS from 1 to 4, defining each level with a short description to better attribute the tweet.

In relation to the first point, we considered appropriate for our purposes the definition of HS in the work of Fortuna and Nunes (2018, 5):

> Hate speech is language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humor is used.

We considered this definition relevant because, compared to other work, it brings attention on language styles by including among the forms of HS irony, humor, and other styles that are not generally recognized as vectors of hate and verbal violence. In relation to the second point, i.e., the tweet classification scheme, we agreed on a scale of HS 1-4 as follows:

Level	*Definition*	*Tweet example*
1	Zero HS level. No incitement to hate and no bad language, verbal violence or intolerant content. These are generally news or tweets reporting statements, news events, etc.	*Israele: COVID19, eventi privati sono limitati a 100 persone all'aperto e 50 persone al chiuso; il GreenPass esteso ai bambini dai 3 anni in su*[9]
2	Almost zero HS level. They are tweets that express opinions or report experiences with a critical tone, but without the use of violent words or references to hate content.	*La narrazione secondo la quale i #Ristorantisti si rifiutano di controllare i documenti di identità, è falsa. L'altro giorno ho provato ad andare via senza pagare il conto e mi hanno chiesto tutto l'albero genealogico fino ai trisavoli. #greenpass*[10]
3	Moderate HS Level. In this class we have also included those tweets that are not obvious HS but more incivility, but which the extension of the concept adopted encourages us to consider HS	*#greenpass è la più delirante ipocrita idiozia realizzata dal governo @matteosalvinimi @LegaSalvini*[11]
4	High HS Level. Explicit use of violent language and hate speech against groups or individuals. Text is characterized by frequent use of foul words and capital letters	*FANCULO #greenpass....FANCULO sto GOVERNO di MERDACCE....FANCULO sto PARLAMENTO DI TRADITORI...FANCULO sti GIORNALAI SINISTRATI!*[12]

Table 2 - Levels of hs tweets, definition and example

This summary panel has been co-constructed and shared by the three authors of the paper. The classification process also included some discussions of tweets whose content was complex and did not allow for early attribution. In the analysis, doubts mainly involved intermediate categories 2 and 3. It has been important to define an argument that well discriminated these two levels, as they also distinguished no-HS content from HS content. In addition, the classification also included noting the type of HS where evident. For example, we have agreed to highlight all those cases where there was an explicit reference to a specific type of hate or intolerance (e.g., against migrants, or the LGBT community, etc.) or to new forms - which we have identified as intolerance towards what the institutions represent - or towards those who have expressed a position on the vaccine, the green pass, etc. To measure reliability, we tested a subset of 300 tweets from each analyst, calculating the coefficient according to the test-retest reability (Guttman, 1945). As we observe from the table 3, there is maximum agreement in 54% of the cases (the values of the diagonal) which grows up to 78% if we consider the neighboring values 1-2 or 3-4 (light gray cells)

The data to be monitored are indicated in dark gray. In fact, they reveal a significant discordance as they discriminate potentially HS content from content that is potentially not. We therefore have verified together 66 tweets placed in this 'grey' area, arriving at a single shared value for all of them. This control operation has been useful not only to set these tweets in a better position, but above all to recalibrate our attribution criteria.

		TEST				
		1	2	3	4	
RE-TEST	1	61	15	2		78
	2	37	66	21	1	125
	3	4	37	26	14	81
	4	1		7	8	16
	TOT.	**103**	**118**	**56**	**23**	**300**

max agreement (diagonal, white cells)	54%
agreement (gray cells)	78%

Table 3 – Levels of hs, test 1 and check (a.v.)

Based on these insights, we re-checked the tweets in the subset. The final distribution (tab.4) shows us that there is not a significant portion of tweets (almost 3 out of 10) express hate content or intolerance towards individuals and groups in an explicit or covert manner. Beyond the empirical evidence that we discuss below, this provided a good basis for training the supervised algorithm in automatic classification.

HS LEVELS	*NR OF TWEETS*	*TYPE*	*NR OF TW*
1	531	NO HS	**1062 (71%)**
2	531		
3	347	HS	**438 (29%)**
4	91		
	1500		***1500***

Table 4 – Frequency of tweets in dataset sample, by HS levels (a.v. and %)

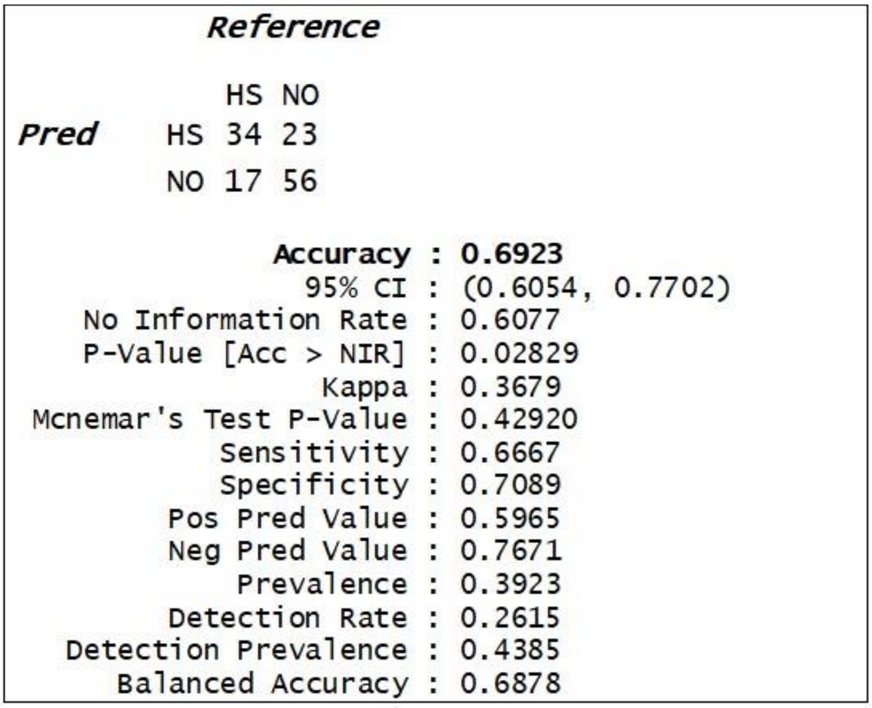
The automatic classifier for tracking hate

To increase the balance between HS and no HS tweets for classifier training, we tried to expand the subset with more HS level 4 tweets. These were manually extracted from the main sample by the most characteristic HS textual keys in the previous manual classification. They were aggregated according to three categories of HS lemmas:

- Usual insults: #asshole, #bitch, #idiot, #shit, #whore, #fag, #slut, #piece of, #lardball;
- Political/ideological/social opinion: #leftist, #piddini[13], #fascist, #communist, #covidiot, #zombie, #infect;
- Offensive definition: #you're just a, #you're a poor, #that's really, #sack of, #race of, #sleeve of, #son of,;
- Exhortation: #die!, #get the fuck!, #get the hell out!, #kill yourself! #Get lost!;

Using this method, an additional 180 tweets were detected for a total of 1680 tagged tweets.

The reduced database was useful for creating and training the automatic classifier. The approach used is “naïve bayes”. This method is a learning algorithm commonly applied to text classification[14], and it refers to an underlying probability model that makes the assumption of feature independence (it assumes that the presence or absence of a particular attribute in a textual document is not related to the presence or absence of other attributes). It was then necessary to divide the database into train (n=1550) to train the algorithm and test (n=130) to evaluate the accuracy of the classifier. Usually the recommended ratio between train and test is 80 /20, in this case a higher ratio was decided (93 / 7) because being a small subset, it was preferred to expand the number of tweets in the train database. The outputs of the confusion matrix (Fig.2) suggest that the classification model has a good accuracy (0.69) because it is significantly higher (p=.02) than the value of No Information Rate (0.60). The algorithm has a sensitivity (% HS tweets detected) of almost 67%.



```
            Reference

              HS NO
Pred      HS 34 23

          NO 17 56

               Accuracy : 0.6923
                 95% CI : (0.6054, 0.7702)
    No Information Rate : 0.6077
    P-Value [Acc > NIR] : 0.02829
                  Kappa : 0.3679
 Mcnemar's Test P-Value : 0.42920
            Sensitivity : 0.6667
            Specificity : 0.7089
         Pos Pred Value : 0.5965
         Neg Pred Value : 0.7671
             Prevalence : 0.3923
         Detection Rate : 0.2615
   Detection Prevalence : 0.4385
      Balanced Accuracy : 0.6878
```

Figure 2 – Confusion matrix and output values

Results. Description of hs tweets

The automatic classification with an accuracy of almost 70%, returns these results: our corpus contains more than 60% of tweets with HS content (tab.6). This is an objectively high and initially not expected share, also considering the lower percentage in the manually labeled subset. As noted in tab.7, tweets with HS present their own specific profile. First, they have a lower level of engagement (categorized on number of retweets and favorites) than no-HS content (67.8% and 64.4%): they therefore tend to circulate less (retweets) and also have less 'appeal' (favourites). Compared to no-HS content, in HS tweets we have found more 'quotes' (more than 12%), i.e. retweets with comments. Not surprising given that intolerant and hateful communication often targets people (and events), and is therefore carried by this form of social communication that allows you to express your own opinion on what is already circulating in the twittersphere.

TWEETS (N=64589)	*%*
HS CONTENT	60,82%
NO HSC	39,18%
TOT.	**100**

Table 5 – Frequency of tweets in dataset sample, by HS or NO HS CONTENT

TWEETS (N=64589)	*ENGAGEMENT*	***QUOTE***

	LOW	MEDIUM	HIGH	
HS CONTENT	67,80%	25,97%	6,24%	12,20%
NO HSC	64,37%	28,13%	7,50%	7,89%
COL TOT.	**66,45%**	**26,81%**	**6,73%**	**10,51%**

Table 6 – Frequency of type of tweets in dataset sample, by ENGAGEMENT and type (if Quote)

Lemmas and issues on the verbal violence about the covid certificate

We have observed how HS communication has been widespread in the twittersphere in relation to the greenpass debate. However, what is the most characteristic content? What were the words used by those who produced tendentially or totally hostile communication about the DCC? Table 8 shows the 30 most characteristic words in the HS-oriented and non-HS-oriented group of tweets, distributed in chi-square order. While in the second case most of the words present a moderate linguistic style more oriented to the procedures, users and places involved in the measure (school, extension, controls, cinema, worker, bar, staff...), the terms of the HS communication are marked not only, as expected, by strong and insulting expressions (cock, ass, shame) but also by a group of words that emphasize an attitude of strong contrast and discontent with the anti-Covid measures. One in particular seems to be the focal point of the controversial debate: dictatorship of health is the label used to define the anti-Covid policy system that is perceived as limiting individual freedom or even dangerous for health (nogreenpass, obbligovaccinale, novax, terzadose). In addition to this, the targets of polemics, and therefore of insults, are easily identified in the institutional and political establishment (Ward, & McLoughlin, 2020), in the media and in science, i.e. in all the actors that play a priority role in the communication of Covid risk and policies. In the first case, these are bipartisan political actors, governmental (Mattarella, Draghi, Salvini, Speranza) and non-governmental (Meloni). In the second case, a TV program is mentioned (Staseraitalia) that has repeatedly dealt with the DCC issue and has often been the object of verbal extremism on web platforms[15]. Finally, in the case of the scientific community ('science' lemma), there are many references to virologists as guests in TV programs, defined, in a disrespectful way, 'viro-star' for their high media exposure. The hostile communication on the implementation of the green pass therefore has its own very characteristic lexical connotation that seems to belong to an attitude of decided contrast to this type of policies. HS oriented tweets, before being a vehicle for free (and anonymous) insults, appear as a communicative tool to show one's dissent towards the governmental and scientific establishment.

HS				
LEMMAS	***SUB***	***TOT***	***CHI2***	***(p)***
DITTATURASANITARIA	601	624	389,4	0
DITTATURA	679	741	358,3	0
GREENPASS	37967	62420	357,7	0
CAZZO	650	710	341,6	0

NO HS				
LEMMAS	***SUB***	***TOT***	***CHI2***	***(p)***
PASS	2991	3953	1760	0
GREEN	2782	3643	1691	0
SCUOLA	1851	2457	1070	0
DECRETO	847	892	994,6	0

CAPIRE	1469	1954	257,3	0	PRIMA	1063	1227	967,2	0
CULO	370	391	223,3	0	OBBLIGO	2309	3546	728,5	0
DIRITTI	593	696	222,4	0	LAVORATORE	1287	1732	706,5	0
RICATTO	497	563	221,6	0	VIA	1068	1392	657,8	0
VERGOGNA	370	397	209,7	0	ESTENSIONE	743	874	638,6	0
GENTE	953	1235	200,5	0	BAR	574	620	629,6	0
NOVAX	2784	4117	182	0	ROMA	590	681	536,5	0
STASERAITALIA	402	455	179,9	0	CERTIFICATO	527	591	520,8	0
NOGREENPASS	1795	2553	178,5	0	CONTROLLI	590	704	485,4	0
PROPRIO	727	926	171,2	0	PERSONALE	517	594	476,5	0
OBBLIGOVACCINALE	955	1273	164,5	0	ACCEDERE	515	596	464,6	0
SINISTRA	354	399	161,6	0	ESTESO	394	432	414,7	0
MATTARELLA	334	374	157,1	0	VERDE	539	658	412,8	0
SALVINI	1190	1645	153,4	0	DOCUMENTO	465	545	403,9	0
POLITICO	828	1098	148,2	0	RISTORANTE	1062	1567	401,8	0
DRAGHI	2072	3045	145,6	0	STUDENTE	448	530	378,3	0
PAURA	403	479	141,5	0	OGGI	1374	2170	376,3	0
TERZADOSE	590	752	138,3	0	CINEMA	354	388	372,9	0
NOGREENPASSOBBLIGATORIO	577	734	137	0	CHIUSO	432	516	354,2	0
PARLAMENTARE	421	508	136,4	0	STORIA	464	570	348,3	0
SCIENZA	360	423	134,2	0	AGOSTO	379	448	320,8	0
LEGA	1352	1926	132,3	0	PUBBLICI	452	574	304,1	0
MELONI	430	530	123,6	0	CASO	372	445	303,6	0
CASA	721	971	114,8	0	DIPENDENTI	417	518	301,9	0
SPERANZA	721	976	110,3	0	ANNO	440	571	274,8	0

Table 7 – Hs and no-Hs lemmas by charateristics (chi-square)

The hs topics

We focused on HS communication by considering the most relevant topics (fig. 3). Topic analysis using Latent Dirichlet Allocation (LDA) (Blei *et al.*, 2003) has allowed us to extract clusters of meaning from the HS corpus. The extraction was done automatically by articulating the optimal partition into 8 clusters, whose meaning was constructed due to the most characteristic lemmas of the group, listed in parentheses.

1. “Vaccine effects" (vaccine, virus, infection, useful, effect, side effect, efficacy, third dose, severe, experimental, risk) is the most incisive cluster on the corpus (0.13) and collects all the words that are part of the HS discourse oriented on infection, vaccine and possible side effects.

2. “Restricted rights” (freedom, rights, Constitution, citizens, discrimination, regulation), in this topic, second in weight on the corpus (0.12), the verbally violent communication is oriented towards debating the supposed violations of the DCC on individual freedom. Most

of the content of this topic is therefore based on the intolerance of restrictive measures, seen as discrimination and as an attack on the Italian Constitution.

3. “Free alternatives” (swab, free, test, price, tax, pharmacy, increased, costs), another important semantic piece of HS-oriented content concerns the debate on alternatives to DCC, especially free swabs or a policy of limiting the costs of testing that would facilitate people who can not or are against vaccination.

4. “Political debate” (Salvini, Lega, Meloni, vote, Government, PD, Giorgetti, M5S, Draghi), this topic (weight 0.10) refers mainly to political actors who for different reasons are at the center of HS communication. In some cases because they are direct expressions of pro-DCC laws (e.g., M5s, Draghi, PD), in others because they are accused of lack of opposition (Salvini, Lega, Giorgetti), or, in the case of Meloni they are catalysts of a less than politically correct communication, likely due to her opposition role.

5. “Implementation critics” (work, obligation, school, salary, university, workers union, transports), this topic (0.09) groups together the HS communication based on the complex implementation of the DCC. This, in fact, has required a reorganization of the world of work and education, defining hard rules, the target of violent discussion, especially in terms of possible sanctions.

6. "No-vax in the world" (no-vax, Italy, France, demonstrations, strikes, nogreenpass, Spain, Denmark, flop, train station, England, Macron), in this cluster (0.09) the main focus of HS-oriented communication is related to the growing no-vax movement and to the angry debate that has affected protests and demonstrations even beyond Italy. For example, the numerous demonstrations in France against the DCC and against Macron's policies, the threat of station blocks, the protests in Spain and Denmark.

7. “Dictatorship of health” (noobligatorygreenpass, dictatorship of health, Bassetti, Burioni, Pregliasco, notcorrelation, Covidvaccine, Draghi, third dose), is the topic where the focus of HS debate is "dictatorship of health". According to the no-vax and other protest groups, the national and global government, with the support of the scientific community of virologists, are structuring a dictatorial power based on health control.

8. “Just vaccinate yourself” (no-vax, vaccinate yourself, just vaccinate yourself, fuck, asshole, mycousinnews, attention), the latter topic carries HS communication as a call for vaccination. This type of tweets are also characterized by the use of strong words towards the no-vax and their sources of information, considered unreliable

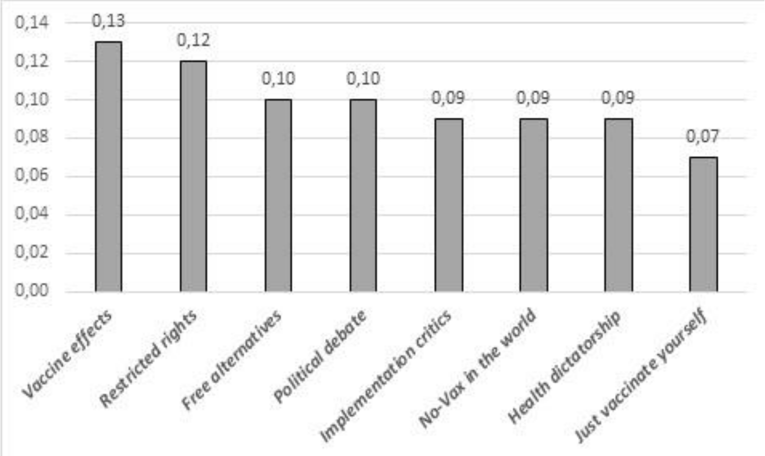


Figure 3 – Topic by weight on corpus

How are the emerging topics related? Investigating associations between topics has also helped us to measure the reliability of group labeling. To detect similarities, the 8 topics were projected on the Cartesian plane using Sammon method which computes associations between groups and words. As we can see in fig. 4, four aggregations are recognizable for each sector. In the first one there is space for the single “Political debate”. It is a very specific and “autonomous” topic for the almost exclusive presence of political actors. In the second one there are the topics “Implementation critics” and “Just vaccinate yourself”, meaning the continuous tension between not vaccinating and the implications that this choice has on the actual implementation of the DCC, not only for health but also for logistics. The third one includes the topics “Dictatorship of health” and “Restricted rights” where HS communication is against the current anti-Covid policies, considered inadequate and anti-constitutional. Finally, the fourth one recalls the 'no-vax' polemic semantics, joining the three topics “Free alternatives”, “Vaccine Effects” and “Novax in the world”.

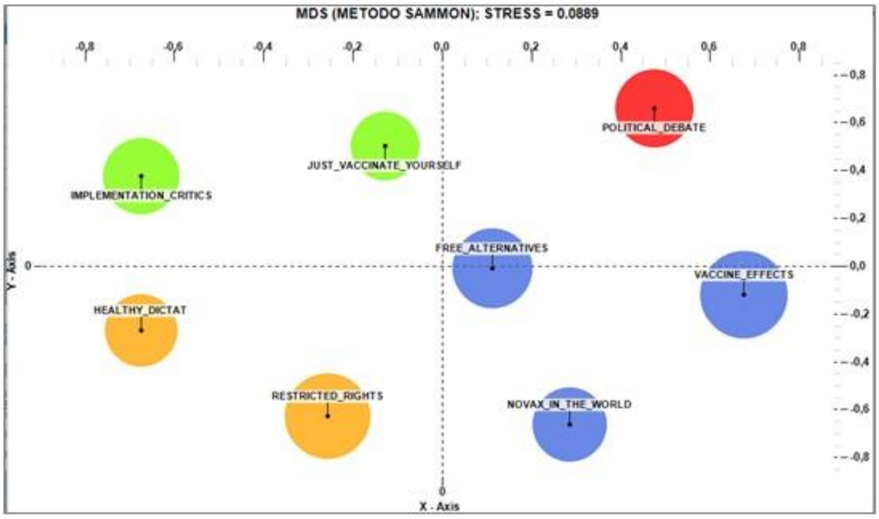


Figure 4 – Topic projected on Cartesian Plan via Sammon method

Conclusions. The ambiguity of digital hate

This work is currently under development: outcomes are still being defined and could lead to new empirical evidence. Nevertheless, what we have observed so far may be sufficient to develop an initial reflection on the evolution of the phenomenon of HS on social media in times of emergency and in response to a decisive measure to combat the pandemic, such as the DCC.

As we have discussed in the initial part of the study, despite having a greater awareness of HS today and especially of its social effects, there is still not a common agreement on its definition.

Studying HS therefore required us to well delimit the concept: in agreement with the definition of Fortuna and Nunes (2018), we considered HS not only the explicit and clear expressions of intolerance and hate, but also the less conspicuous and more hidden forms of hate such as irony or cold jokes and therefore apparently not recognizable as insulting phrases. Deliberately, therefore, we wanted to stretch the concept by including hidden forms that are closer to those of "incivility".

This 'broad' perception of the problem has obviously had implications for the research design and results. In fact, a supervised algorithm has been necessary to be built on the basis of a subset of tweets we labeled and fit to this terminological definition.

The results on about 65 thousand tweets suggest that about 3 out of 5 tweets are HS-oriented. This represents a wake-up call not so much about the degree of acceptability of the measure (which would still need to be investigated further) but instead about the easiness of spreading hate and incivility content on a popular platform like Twitter, creating a phenomenon geared toward the "platformization of hate". The content analysis allowed us to explore the thematic spaces in which verbally violent content most easily emerges.

This therefore allows us to provide an answer, even if partial, to the first of the two questions (In this emergent phase in which the verbal conflict is very hot, what are major topics and new target emerging from this redefinition of HS?),

The broadening of the concept of HS to include less obvious forms and thus closer to incivility has broken the banks of a phenomenon only seemingly "kept at bay" by censorship mechanisms or social media detect algorithms.

The debate over the implementation of the DCC, as expected, was characterized by very sharp language often resulting in slurring and verbally violent discussions. Although HS and, more generally, incivility, is often associated with usually content-poor phenomena such as “flaming” or “trolling” (O'Sullivan e Flanagin, 2003), in this case the unfriendly debate over the certificate was not limited to simple insults without argument. Moreover, the analysis of the topics has highlighted 8 issues that have evidently encouraged verbal conflict. By looking at a greater overview we are faced with two large semantic spaces: the first is attributable to the common no-vax discourse, characterized by a mixture of fake theories, refusal to vaccinate and anti-politics. The second HS-oriented space is instead more attentive to the concrete implementation of the DCC, both in terms of political governance (e.g., the polemic on the ambiguity of the Lega party) and in relation to the critical issues arising from the implementation of such an impressive measure (e.g., privacy question).

Relate to the second question (How effective can the use of a supervised algorithm for detecting such a complex phenomenon return?), the results seem to suggest that both under- and over-representation of verbal violence are possible outcomes on social media, depending on the interpretive reading of the text. Under-representation because some forms of incivility, which are precursors to expressions of hatred toward categories of users, are not "attended to" by social media control policies as well as overt expressions of HS.

However, disagreements that we have encountered in constructing an unambiguous definition of HS for the algorithm leave many questions open relative to the possible sovra-representation of HS. Among all, indeed the difference between HS and freedom of expression and HS and incivility can be very thin. Indeed, "uncivil" language can often conceal messages of hate. In the context of large social platforms, where algorithm criteria are not always explicit, and are platform policies-related, this could be a problematic issue.

Beyond the substantive results, the large size of the data was not an impediment to the use of even qualitative analysis strategies. Indeed, the construction of the supervised algorithm required manual text classification, which had a significant impact on the interpretation of the results and the meaning of this research. This could strengthen the idea that qualitative models would offer many possibilities for new data research (Bennato, 2021).

The ambiguity of the HS issue, as described in the opening of the paper, would make it complex to distinguish free expression from verbal discrimination of the individual. Therefore, online hate monitoring systems should refine their tools to keep tabs on hotly debated issues such as the implementation of the DCC: it is here that forms of verbal incivility and thus hate content might find new inspiration. These results might suggest strong caution to social media policy makers for two reasons: the first is precisely due to these new elements of hate that are not easily detectable because they are not explicit. The second is the complexity of HS, which could imply bias in the outputs of the detected HS algorithms.

Some limits and future work

Despite the many potentialities of this study, we have identified four limitations that are worth addressing in future developments of this work. The first limitation, as already mentioned, is related to the ambiguity of the research object. Indeed, we brought under the concept of HS also many forms of incivility because we started from the assumption that in some cases incivility can be a precursor to online hate. On the other hand, HS has no a specific definition and even among us authors there was not full agreement of the meaning. For this reason, attempting an automatic classification has been an ambitious task, which in some cases has inevitably returned an unreliable share of results. That would require two steps. The first, certainly an expansion of the labeled sample. The second is related to further methodological attempts to increase the reliability of the classification. For example, the combination with unsupervised or lexicon-based approaches could be a good solution. The other critical issue relates to the in-depth study that HS related issues would require. We have observed how the introduction of DCC has in some cases exasperated the debate and new forms of hostility and verbal violence are emerging. These new forms would be interesting to investigate from a qualitative point of view in order to better understand their meaning and underline possible sub-themes. The third limit is related to the search context, Twitter. Although Twitter is very useful for research because of the ease of scraping data (API) and indexing themes (hashtags), it has the obvious limitation that it is not representative of the digital universe and, indeed, is not the most used social media in Italy. Furthermore, some critical aspects of Twitter's API have been documented with regard to the actual representativeness of the content on the platform (Caliandro, 2021). That's why expanding the research context by investigating new platforms might be relevant. Finally, taking into account the images and not just the text could overcome one of the great limitations of this study. For instance, memes as the most popular communication tool of recent times are certainly also central to the spread of online hatred. The analysis of hate memes may therefore open up new avenues of understanding this phenomenon.

Biographical Note

Rosa Sorrentino is a Ph.D. candidate in Social Sciences and Statistics at the Department of Social Sciences, University of Naples "Federico II". She is a junior researcher of the S.F.O.R.A. and Norisc-19 projects of the same institution. Her research interests include: methodological issues, feminist and gender studies, and social and welfare policies from a territorial perspective.

Miriam Di Lisio (Naples, 1990) is a graduate of the master's program in "Public, Social and Political Communication" at the Department of Social Sciences at the Federico II University in Naples. She is a Ph.D. candidate in Social Sciences and Statistics and collaborates in the Master's degree program in "Direction, Management and Coordination of Health and Social Facilities" at the same institution. Her scientific interests mainly concern methodological issues and the study of sociology of science and social psychology.

Domenico Trezza is a Ph.D. in Statistical and Social Sciences at the Department of Social Sciences of the University of Naples Federico II. He is currently an IFEL consultant for the Education Policy Observatory of the Campania Region. He is engaged as a senior researcher of the S.f.o.r.a project of the Department of Social Sciences. His research interests include: methodological issues in the study of digital data, environment and risk perception, educational policy evaluation and social policy governance models

References

Amnesty International. (2018). Toxic Twitter: A Toxic Place for Women. Retrieved from https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-1/

Anderson, A. A., Brossard, D., Scheufele, D. A., Xenos, M. A., & Ladwig, P. (2014). The “nastyeffect:” online incivility and risk perceptions of emerging technologies, in

Journal of Computer-Mediated Communication, 19(3), 373–387, https://doi.org/10.1111/jcc4.12009

Antoci, A., Delfino, A., Paglieri, F., Panebianco, F., & Sabatini, F. (2016). Civilization versus incivility in online social interactions: an evolutionary approach, in *Plos One*, november 1, 2016, https://doi.org/10.1371/journal.pone.0164286

Aragona, B. (2021). *Algorithm Audit: Why, What, and How?* London: Routledge Focus.

Atlanta. (2018). (Anti)-social media: The benefits and pitfalls of digital for female politicians. London: Atlanta.

Bennato, D. (2021). The Digital Traces' Diamond. A Proposal to Put Together a Quantitative Approach, Interpretive Methods, and Computational Tools. *Italian Sociological Review*, *11, No.4S*. doi: http://dx.doi.org/10.13136/isr.v11i4S

Bentivegna, S., & Rega, R. (2020). Online hate speech in a communicative perspective: a research agenda, in *Mediascapes Journal*, E-ISSN 2282-2542, N. 16 (2020): SERIALITY IN THE POST-TELEVISION ERA.

Blei, D. M., Andrew, Y. N., & Jordan M. I., (2003). Latent dirichlet allocation, in the *Journal of machine Learning research* 3 pp. 993-1022. In: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Boccia Artieri, G., & Marinelli, A. (2018). Introduzione: piattaforme, algoritmi, formati. Come sta evolvendo l'informazione online, in *Problemi dell'informazione*, 43(3), 349-368. Doi: 10.1445/91657

Cabo Isasi, A., & García Juanatey, A. (2016). Hate speech in social media: a state-of-the-art review, in *Ajuntament de Barcelona*, december, 11 2016. In: https://ajuntament.barcelona.cat/bcnvsodi/wp-content/uploads/2017/01/Informe_discurso-del-odio_ENG.pdf

Caiani, M., Carlotti, B., & Padoan, E. (2021). Online hate speech and the radical right in times of pandemic: The Italian and English cases, in *Javnost-The Public*, 28(2), 202-218.

Caliandro, A. (2021). Repurposing Digital Methods in a Post-API Research Environment: Methodological and Ethical Implications, in *Italian Sociological Review*, *11*. Doi: http://dx.doi.org/10.13136/isr.v11i4S.433

Chen, G. M., & Lu, S. (2017). Online political discourse: Exploring differences in effects of civil and uncivil disagreement in news website comments, in *Journal of Broadcasting & Electronic Media*, 61(1), 108–125, https://doi.org/10.1080/08838151.2016.1273922

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of inci-vility in newspaper website comments, in *Journal of Communication,* 64(4), 658–679, https://doi.org/10.1111/jcom.12104

Di Lisio, M., & Trezza, D. (2021). Digital Methods to Study (and Reduce) the Impact of Disinformation, in *Culture e Studi del Sociale* (pp. 139-151). 6(1), Special issue. In: https://www.cussoc.it/index.php/journal/article/view/178/133

Druckmann, J. N., Clar, S., Krupnikov, J., Levendusky, M., & Barry Ryan G. (2021). Affective polarization, local contexts, and public opinion in America, in *nature human behavior*, November, 23, 2020, https://doi.org/10.1038/s41562-020-01012-5

Duffy, B. E., Poell, T., & Nieborg, D. B. (2019). Platform practices in the cultural industries: Creativity, labor, and citizenship, in *Social Media+ Society*, 5(4). doi:10.1177/2056305119879672

Fortuna, P., & Nunes, S. (2018). A Survey on Automatic Detection of Hate Speech in Text, in *ACM Comput Surv*. 2018;51(4):85:1–85:30. Doi: https://doi.org/10.1145/3232676

Guttman, L. (1945). A basis for analyzing test-retest reliability, in *Psychometrika 10*, 255-282, DOI: 10.1007/BF02288892

Inter Parliamentary Union. (2016). Sexism, harassment and violence against women Parliamentarians. Geneva: IPU.

Jackson, N., & Lilleker, D. (2011). Microblogging, constituency service and impression management: UK MPs and the use of Twitter, in *The Journal of Legislative Studies*, 17, 1.

Jamieson, K. (1997). Civility in the House of Representatives, in *APPC report 10*, Retrieved March 28, 2011, from http://democrats.rules.house.gov/archives/hear01.html

Leavy, S. (2018, May). Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning, in *Proceedings of the 1st international workshop on gender equality in software engineering* (pp. 14-16). doi: 10.1145/3195570.3195580

Lilleker, D., & Jackson, N. (2014). Interacting and representing: Can Web 2.0 enhance the roles of an MP?, in *ECPR Joint Sessions of Workshops*, 14–19 March 2009, Lisbon.

Nardi, V. (2019). I discorsi d'odio nell'era digitale: quale ruolo per l'internet service provider?, in *Diritto penale contemporaneo* (pp. 268-288), 2/2019. In: https://archiviodpc.dirittopenaleuomo.org/upload/4923-nardi2019a.pdf

Nielsen, L. B. (2002). Subtle, Pervasive, Harmful: Racist and Sexist Remarks in Public as Hate Speech, in *Social Issues*, Volume 58, Issue 2, Summer 2002, Pages 265-280, https://doi.org/10.1111/1540-4560.00260

O'Sullivan, P. B., & Flanagin A. J. (2003). Reconceptualizing 'flaming'and other problematic messages, in *New Media & Society* (pp. 69-94), 5.1. In: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.72.4856&rep=rep1&type=pdf

Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups, in *New Media & Society*, 6(2), 259–283.https://doi.org/10.1177/1461444804041444

Pignatiello, G. G. (2021). Countering anti-lgbti+ bias in the European Union. A comparative analysis of criminal policies and constitutional issues, in *Italian, Spanish and French legislation, Women's Studies International Forum*, 86. In: https://www.academia.edu/45674539/Countering_anti_lgbti_bias_in_the_European_Union_A_comparative_analysis_of_criminal_policies_and_constitutional_issues_in_Italian_Spanish_and_French_legislation

Pino, G. (2008). Discorso razzista e libertà di manifestazione del pensiero, in *Politica del Diritto* (pp. 287-305). XXXIX, 2. In: https://www.giorgiopino.net/uploads/1/3/1/5/131521883/pino_discorso_razzista.pdf

Pollicino, O. e De Gregorio, G. (2019). Hate speech: una prospettiva di diritto costituzionale comparato, in *Giornale Di Diritto Amministrativo*, n. 4/2019. In: https://www.academia.edu/40449867/Hate_speech_una_prospettiva_di_diritto_costituzionale_comparato

Rega, S., & Marchetti, R. (2019). Incivility in Politics 2018. End of public debate?, in *Comunicazione politica*, n. 1/2019, aprile. DOI: 10.3270/93027

Rogers, R. (2009). *The End of the Virtual: Digital Methods*, Amsterdam: Amsterdam University Press.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2017). Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis, in *User-Centred Social Media*, January, 27, 2017, arXiv:1701.08118v1

Scamuzzi, S., Belluati, M., Caielli, M., Cepernich, C., Patti, V., Stecca S., & Tipaldo, G. (2021). Fake news e hate speech, i nodi per un'azione di policy efficace, *Problemi dell'informazione* (pp.49-81), Fascicolo 1, aprile 2021. doi: 10.1445/100129

Suler, J. (2004). The Online Disinhibition Effect, in *CyberPsychology & Behavior,* Volume 7, Number 3, 2004.

Uyheng, J., & Carley, K. M. (2021), Characterizing network dynamics of online hate communities around the COVID-19 pandemic, in *Applied Network Science*, *6*(1), 1-21. doi: https://doi.org/10.1007/s41109-021-00362-x

Van Spanje, J., & De Vreese, C. (2014). The way democracy works: The impact of hate speech pro-secution of a politician on citizens' satisfaction with democratic performance, in *International Journal of Public Opinion Research*, 26(4), 501–516, https://doi.org/10.1093/ijpor/edt039

Vrielink, J. (2016). Do we want more or fewer prosecutions of opinions: The Geert Wilders trial 2.0, in *Netherlands Journal of Legal Philosophy*, 45(2), 3–11, https://doi.org/10.5553/NJLP/.000053

Waisbord, S. (2018). The elective affinity between post-truth communication and populist politics , in *Communication Research and Practice*, 4(1), 17–34, https://doi.org/10.1080/22041451.2018.1428928

Waldron, J. (2021). *The Harm in Hate Speech*, Cambridge: Cambridge University Press.

Walker, S. (1994). Hate speech: The History of an American Controversy. University of Nebraska Press.

Ward, S., & Lusoli, W. (2005). 'From weird to wired': MPs, the internet and representative politics in the UK', in *Journal of Legislative Studies*, 11(1), 57–81.

Ward, S., & McLoughlin, L. (2020). Turds, traitors and tossers: the abuse of UK MPs via Twitter, in *The Journal of Legislative Studies*, Volume 26, 2020 - Issue 1, 26(1), 47-73, https://doi.org/10.1080/13572334.2020.1730502

Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the World Wide Web, in *Workshop on Language in Social Media*, ACL, 19–26.

Ziccardi, G. (2016). *L'odio online. Violenza verbale e ossessioni in rete*. Milano: Raffaello Cortina Editore.

Ziccardi, G. (2021). Le espressioni d'odio sulle piattaforme digitali: alcune considerazioni informatico-giuridiche. In: https://air.unimi.it/retrieve/handle/2434/864733/1860884/ziccardidamico.pdf

Note

[1] Appendix to Recommendation No. R(97) 20, p. 107

[2] Source: https://unesdoc.unesco.org/ark:/48223/pf0000233231

[3] Source: https://www.amnesty.it/campagne/contrasto-allhate-speech-online/

[4] Convention on Cybercrime, Budapest, 23.XI.2001, European Treaty Series - No. 185

[5] https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

[6] L.645/1952

[7] www.amnesty.it/entra-in-azione/task-force-attivismo/

[8] https://www.dgc.gov.it/web/checose.html

[9] Trad: "Israel: COVID19, private events are limited to 100 people outdoors and 50 people indoors; GreenPass extended to children 3 and under"

[10] "The narrative that #Restaurants refuse to check IDs is false. The other day I tried to leave without paying my bill and they asked for my entire family tree down to my great-great-grandparents. #greenpass"

[11] "#greenpass is the most delusional hypocritical idiocy carried out by the State @matteosalvinimi @LegaSalvini"

[12] "FUCK YOU #greenpass....FUCK YOU'RE SHIT GOVERNMENT ....FUCK YOU'RE PARLIAMENT OF TRAITORS ... FUCK YOU'RE JOURNALAI LEFT! "

[13] "Piddini" is a derogatory term for the Democratic Party electorate

[14] To construct the text classifier, it was necessary to pre-process the text. The text analyses were done in the R environment by the 'tm' package. The text of the tweets was pre-processed through several normalization and textual content transformation functions. In fact, the 'tm_map' function of the tm package allowed us to 1.Reduce capitalization, 2. Remove stopwords, 3.Stripping white space, 4. Stemming, 5.Remove Punctuation

[15] https://www.open.online/2021/08/30/covid-19-no-vax-green-pass-squadrismo-digitale/