Remove or keep up?
A bottom-up assessment of online hate speech

Angeliki Monnier (Metz)

Abstract

Multi-component methods of analyzing hate speech seem to open new horizons in training human annotators and algorithms to facilitate online hate detection and moderation. One of the limitations of these endeavors lies on the fact that the components proposed are either treated as if they were of equal significance or hierarchized within scales that are based on scholars' understandings and assessments – informed and relevant as these might be – with no feedback from common laypeople. They constitute "top-down" understandings of the performative power of hate speech. The present contribution seeks to address this gap by analyzing how ordinary people, especially young people in our case, assess different forms of hate speech. Our aim is to identify what they estimate to be acceptable or tolerable, and the actions they would choose to undertake, if they were to monitor online posts (i. e., take them down or keep them up). We consider hate speech as an "ordinary concept" reflecting "ordinary people's norms". The research reveals an overall tendency to content removal, especially in the case of calls to lethal aggressions and swearwords which seem to operate as offense intensifiers. Incitements to non-lethal actions and emotional expressions may be tolerated or even accepted by some, being perceived as non-hateful". A distinction between hate action, hate speech, and hate expression revealed to be significant. At the same time, punctuation seems to have no significant effect on the overall judgments. Finally, combinations of hate features increase the perceived hatefulness and, even more substantially, the willingness to take down the content in question.

1 Introduction

Academic works on hate speech, coming from different disciplines, have exponentially increased over the last years, drawing on multiple epistemic and methodological approaches. Researchers in computer science or computational linguistics who look into the implementation of automatic detection tools (for a summary cf. Fortuna/Nunes 2018) often embrace a lexical framing, based on pre-established lists of "hate" words, sometimes combined with analysis of their frequency and syntactic configurations. Despite its merits, this stance, which builds upon binary categorizations (hate/no hate), fails to capture the fact that hate speech hinges on patterns of thought, even fallacious ones. Against this backdrop, feature-based approaches (cf. Reiners/Schemer 2020) analyze hate speech in online user comments focusing on some "manifest features": group-related labels and swearwords, negative trait and action attributions, treatment recommendations and calls to collective action, verbal and/or pictorial expression of

emotions. Modular methods (cf. Strippel et al. 2020) divide negative comments in categories like insults, generalizations, violent implications and/or dehumanization to assess hate speech. Semantic-based stances coupled with analysis of speech acts (cf. Monnier/Seoane/Gardenier 2020) insist on the discursive, "pre-discursive" and "post-discursive" components that characterize this type of content, between diagnosis and prognosis: linguistic marks, existing representations, action-oriented incitements, respectively. Narrative approaches (cf. Monnier/ Boursier 2022) highlight the underlying stories that forge hate speech directed at specific targets, while designating actors and roles (opponents, helpers, etc.). Finally, other multifactorial protocols (cf. Poletto et al. 2017) establish and distinguish intensities of hate, based on the presumed intensity, intent and impact of the contents: offensive messages with potential hurtful effects towards the target population, aggressive calls inciting to more or less violent, even lethal, actions.

Multi-component methods (feature-based, semantic-based, speech-act-based, etc.) seem to open new horizons in quantitative content analyses, and can also be used to train human annotators, or even algorithms (cf. Reiners/Schemer 2020) with a view to facilitating online hate content detection and moderation. One of the limits of these endeavors lies though on the fact that the components proposed are either treated as if they were of equal significance (i. e., all hate features have the same performative power) or hierarchized within scales that are based on scholars' understandings and assessments – informed and relevant as the latter might be – with no feedback from ordinary people. They constitute "top-down" understandings of the performative power of hate speech. They build on the assumption that the meaning of hate lies basically with the language as such, and that well-trained annotators would be able to identify it. This is why these studies tend to multiply reliance tests to ensure that annotators have understood and correctly applied guidelines, categorizations and definitions. The problem is that, beyond its legal definitions, hate speech is an "ordinary concept" (Brown 2015), i. e., its meaning is not only a matter of language, but also a matter of perception and assessment.

The present contribution[1] seeks to address this gap by analyzing how ordinary people, especially young people in our case, assess what researchers commonly consider as being distinctive hate speech features. Our aim is to identify what they estimate to be acceptable or tolerable, and the actions they would choose to undertake in relation to online user-generated hate contents (i. e., take them down or keep them up). Our assumption is that hate speech lies with the language but also with the people. Annotator disagreements are not necessarily the symptom of poor method and research design, but also a mark of problematic contents, which are difficult to appraise, and are subject to personal – and inevitably sociocultural – readings. That is the reason why our approach calls for survey-feedbacks instead of a restricted number of annotations. Our contribution mainly addresses the following research questions: Are all hate speech forms (i. e., features) equally perceived in terms of hatefulness? Are all hate speech forms equally perceived in terms of attitude to adopt (remove, tolerate, etc.)? Finally, does the accumulation of two hate forms in the same comment modify initial single perceptions?

[1] The author would like to thank the anonymous reviewer for his/her valuable comments.

2 Hate speech as an "ordinary concept"

Defining hate speech has been an epistemological and empirical challenge for researchers from all disciplines since the 1980s, when the term was first coined (cf. Matsuda 1989). The dare seems to have become more demanding in the era of the participative internet (cf. Benesch et al. 2021), which has opened new spaces for the expression of affects (cf. Papacharissi 2014). However, as Alex Brown (2015) highlights, hate speech designates a heterogeneous collection of expressive phenomena that are not strictly attached to emotions, feelings, or attitudes of hatred. Besides its legal standing, hate speech is an "ordinary concept", reflecting ordinary people's sense of what is acceptable and unacceptable, tolerable and intolerable, based on ordinary people's moral values and principles (cf. Brown 2015: 427). In this sense, hate speech needs to be understood in association with its related concept, freedom of speech.

Freedom of expression was first recognized as a human right under the 1948 Universal Declaration of Human Rights (cf. UN 1948). The 1966 International Covenant on Civil and Political Rights, adopted by the United Nations General Assembly on 16 December 1966 and entered into force on 23 March 1976, declares that "everyone shall have the right to hold opinions without interference" and "everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice". The exercise of these rights carries though "special duties and responsibilities" and may "therefore be subject to certain restrictions", when necessary, "[f]or respect of the rights or reputation of others" or "[f]or the protection of national security or of public order, or of public health or morals" (UN 1967: article 19). In this sense, freedom of speech can be regulatable.

The hassle is that freedom of speech can be framed in two rather opposed ways: combative or tolerant (cf. Ramond 2011, 2013). A combative interpretation presumes that any word or image, even offensive, feeds the public debate and, therefore, serves democracy; it benefits everyone, including the offended minority. Within this frame, hate speech is acceptable, even useful. The second interpretation advocates a pluralist stance in line with Philosopher Paul Ricœur's viewpoint (1996), which insists on the necessity to preserve personal convictions while practicing a tolerant openness to otherness. Ricœur's tolerant pluralism is not synonymous with relativism, but a fundamental value of democracy: offending the other does not indicate braveness, whilst respecting the other shouldn't signify cowardice or compromise (cf. Héran 2020). In light of this stance, freedom of speech also implies accepting that all people do not embrace the same values and norms. Consequently, hate speech is problematic because it expresses personal convictions but lacks tolerance towards the other (cf. Sunstein 1999). Visions of hate speech are thus related to visions of the other.

Online hate speech regulation has been a major topic of discussion over the past few years.[2] In France, the law of 24 June 2020 aimed at combating hateful content on the Internet, known as

[2] In fact, several studies have shown that the communicational affordances of contemporary technical devices (availability, portability, etc., cf. Schrock 2015; Murthy et al. 2015), combined with those of participatory platforms (unpredictability, politicisation, etc., cf. Veikou/Siapera 2015), accentuate the role of emotions in the construction of affective publics (cf. Papacharissi 2014).

the "Avia law" (JO 2020; it was named after Laetitia Avia, the Deputy who proposed the bill), intended to remove terrorist and child pornography content from any site, as well as hateful and pornographic content from the main social networks, collaborative platforms and search engines, within 24 hours. Political figures, a large number of organisations and legal experts criticised the law, which they presented as a danger to freedom of expression, in particular because decisions to remove content would be taken by private operators without the intervention of a judicial judge, who is the constitutional guarantor of individual freedoms. The bill was adopted by the National Assembly on 13 May 2020. Seized by opposition senators, the Constitutional Council ruled that the text was largely contrary to the Constitution, in particular because it disproportionately infringed on freedom of expression. On 24 June, President Emmanuel Macron promulgated the law purged of its provisions, deemed unconstitutional (cf. EDRi 2020). Following this decision, only minor provisions remained in the law: the creation of a public prosecutor's office specialising in online hate messages; the simplification of flagging content; the creation of an "online hate observatory", attached to the CSA (*Conseil Supérieur de l'Audiovisuel, '*Supreme Audio-visual Council', which was the French audiovisual regulatory authority that regulated radio stations and television channels in France between 1989 and 2021; cf. JO 2020).

Besides national regulations, most social media platforms – at least until recently – have prohibited hate speech, this being a politically "safe position […] that does not require hedging, balance, or exceptions" (Gillespie 2018: 59). In reality though, hate moderation reveals to be a paramount challenge for platforms, media and States all over the world. In France, over the past years, many major mainstream media have shut down their online comment sections and forums, being unable to respond to the huge moderation efforts and investments needed.

Generally speaking, platform guidelines and State regulations echo U. S. legal language and tradition, particularly in terms of protected categories. The latter refer to groups of people that need to be legally shielded, because of their specific characteristics, which are often summarized as follows: race, ethnicity, nationality, religion, sexual orientation. Within this frame, not all targets of hate speech are protectable, and there are even semi-protected categories, e. g., migrants. According to the Facebook moderator training documents published by *The Guardian* in May 2017, "although dehumanizing statements about them [migrants] should be removed, cursing at them, calling them thieves, and urging them to leave the country do not amount to hate speech" (Gillespie 2018: 112).[3]

[3] Among the non-protected categories are concepts, institutions, beliefs, countries, etc., in other words abstract entities – but not the people who compose them, who are, for their part, protected, because they are considered vulnerable. In addition, groups potentially targeted on the basis of their social class, their appearance or their political ideology are not considered protected. On the other hand, refugees, immigrants, asylum seekers and migrants – who interest us in this study – constitute a "semi-protected" category; in other words, they are groups against which certain hate speech can be tolerated, because they are supposed to participate in the public debate about migration. For example, when an online comment relates to a characteristic related to the appearance of people (e. g., "these migrants are dirty"), it is to be authorized since appearance is not part of the criteria judged discriminatory; but if the comment dehumanizes and insults a group targeting its own identity (e. g., "migrants are dirt"), it is to be removed.

3 Online hate speech against migrants in France

In France, issues of immigration and national identity have occupied public arenas and have played an important role in election campaigns, since the 1970s. Press analysis between 1974 and 1984 (cf. Bonnafous 1991) has shown that mainstream media gradually stopped reporting on the living and working conditions of immigrants, while putting emphasis on difficulties in cohabitation and assimilation. Economic concerns (1974 crisis) – but also a certain failure of the social-class model to address anxieties about the future – seem to have enhanced "the link between the emptiness thus produced and the fear of the Other" (Bonnafous 1991: 273).

The individual-universalist humanist stance having presumably failed to take a clear, positive and well-argued position on the question of immigration, the extreme right has succeeded in imposing, if not its ideas, at least its agenda. Since then, hate speech against migrants has been mostly formulated in nationalist terms or through racist expressions. Foreigners are described as a threat to national identity, culture and economic prosperity (cf. Hargreaves 2012, 2016; Mahfud et al. 2016; Monnier/Boursier/Seoane 2022).

Within this frame, francophone online hate speech against migrants on social media reproduces negative representations and offenses, employs swearwords, calls out for action, expresses – and sometimes graphically illustrates – emotion, etc. (cf. Monnier/Seoane/Gardenier 2020; Gardenier/Monnier 2020). Various components of hate speech, among those described earlier in this paper, can be observed (slur, offense, aggression, dehumanizing, etc.). They usually overlap within same discussion threads or comments, forging contents of different lengths, various emotional intensities and combined meanings. Can all these components be assessed as equivalent in terms of hatefulness? Is there a cumulative effect when these features are matched together? Beyond academia categorizations and scales, what do ordinary people think about these comments? What are the actions they would undertake when faced with them? In order to be able to provide some answers to these questions, we need to first disentangle hate speech in some of its major forms, then submit the latter for assessment to ordinary people. This is what our research aspires to achieve.

Our endeavor was part of the M-PHASIS project (Migration and patterns of hate speech in social media), funded by the French National Research Agency (ANR) and the Deutsche Forschungsgemeinschaft (DFG) from 2019 to 2022 (grant n° ANR-18-FRAL-0005, and conducted jointly by the University of Lorraine in France (Center for Research on Mediations and Lorraine Laboratory for Research in Computer Science and its Applications), University of Mainz and University of Saarland, in Germany. The project aimed to study hate speech against migrants in social media, in order to better understand its emergence and prevalence within user-generated online contents. Its premise was reports indicating an increase of online hate speech against migrants and minorities in several European countries, in the frame of recent migration flows and subsequent polarizations in public debates (cf. Bricks 2016; SELMA 2019). Even though the recent Ukrainian refugee movement seemed to have triggered a widespread wave of empathy within Europe, hate speech and its toll were still observed (cf. Wypych/Bilewicz 2022[4]).

[4] An erratum for this article was reported online on 03.08.2023. However, corrections do not alter the study's general conclusions cited in the present paper.

4 Research design and methodology

Our research builds upon Brown's (2015) aforementioned stance on hate speech as an "ordinary concept" reflecting "ordinary people's norms". Resorting to ordinary people as a form of crowdsourcing is already part of the protocols implemented by several platforms in their effort to assess hatefulness, offensiveness, harmfulness, content violence, and dangerousness, in general. Social media companies outsource content moderation to subcontracted "screeners" (Roberts 2019), who look into flagged contents. Google hires "quality raters" (Google 2025) charged with content appraisal and rating, reporting to the company's computer scientists with the aim to improve its algorithms. Interestingly, raters are asked to rate Google results based on their location, as the company now understands that evaluation norms cannot be totally generic but need to be associated with the specific location of a user. However, beyond these practices in the professional realm, resorting to ordinary people's perceptions in assessing hate speech seems to be less common in academic research. This is the gap we try to fill in.

Our study is based on a survey conducted over 230 undergraduate university students in different fields of Social Sciences (Geography, History, Information and Communication Sciences, Language Sciences, Philosophy, Psychology, Sociology, Theology). More than 2,000 students were sent a link with an invitation to participate in a survey. They were presented with a list of expressions falling under the general umbrella of hate speech against migrants, as defined in some of the aforementioned studies (offense, slur, aggression, etc.). They were told that they were participating in a survey on online hate speech (M-PHASIS project) and that they would have to evaluate a set of "negative" contents regarding migrants, by filling in an online questionnaire. We took the precaution not to label contents as hateful from the outset, in an effort to avoid reception bias in the evaluation process – even though students were aware that the theme of the setting was hate speech. The survey was not mandatory and was not conducted during classes. Exactly 230 respondents filled in the questionnaire voluntarily. They were not asked to provide any personal information (gender, age, class, level, background, etc.), and no track of these elements can be made by the author.

Our approach aimed at simultaneously addressing different forms (i. e., features and levels) of hate speech: existing negative representations regarding migrants, "hate words", punctuation elements that often characterize online affective writing style (capitals), aggressive contents of various scales and calls for action (lethal and non-lethal), as well as verbal emotional expressions. It built upon recent works (cf. Poletto et al. 2017) which analyze hate speech as a complex and multilayer concept, related, among other things, to the illocutionary force of an utterance. According to this stance, offensiveness and aggressiveness can be categorized as forms of hate, with different levels of intensity depending on the speech act involved. Similar stands tend to demarcate "soft" hate (cf. Assimakopoulos/Baider/Millar 2017) or "ambient" hate (cf. Siapera 2019), as to signify the proliferation of hostile, though not necessarily illegal, discourses, particularly on social media and participatory platforms.

Drawing on these schemas, our research distinguished offensiveness and aggressiveness. Offensiveness focuses on the potential hurtful effect of a comment and is similar to insults or judgements based on stereotypes. In 1st grade offense (or "weak" offense, according to Poletto et al. 2017), the target is associated with typical human flaws (e. g., laziness), its status of

disadvantaged or discriminated minority is questioned, etc. In 2nd grade offense (strong offense), an overtly insulting language is used, and the target is addressed to by means of outrageous or degrading expressions (e. g., slurs). Similarly, in 1st grade aggression (weak), a comment explicitly calls for action against the targeted group – all types of action with the exception of lethal ones. In 2nd grade aggression (strong), the comment comprises an incitement to lethal action.

Beyond offensiveness and aggressiveness, our analysis grid included emotional manifestation as part of hate speech's features (cf. Reiners/Schemer 2020), as well as an item pertaining to formal issues (the use of capitals). Indeed, works have shown that the expression of emotions in user-generated content often goes hand in hand with punctuation markers such as ellipsis, exclamation marks, uppercase letters, etc. Though these features were not at the center of our research, we still thought it useful to examine the performative effect of at least one of them. We thus opted to compare the acceptability of a swearword written in lowercase letters with that of the same word written in capitals. Finally, for the sake of methodological consistence and efficiency, the protocol of this research excluded other targets of hate commonly observed in antimigrant discourse (politicians, NGOs, etc.), as well as veiled forms of hate expression (irony, sarcasm, etc.).

Consequently, the experimental corpus of the survey was implemented by the author on the basis of manipulated contents, meant to be comparable and structured as follows:

- An offensive content attacking the dignity and honor (the ethos) of the targeted population, formulated as a phrase (code: OFF);
- An offensive content consisting only of a swearword (slur) (code: SW);
- The same offensive content (swearword) written in capital letters (code: SWCAP);
- An aggressive content inciting to a negative action (e. g. expulsion) (code: ACTN);
- An aggressive content inciting to a lethal action (killing) (code: AGRL);
- An emotional expression of hate, formulated as a phrase (code: EM).

The above 6 categories were first presented separately to participants (Table 1), then matched by two in all possible combinations (14), resulting in a total of 20 statements. A second dataset, comprising the same generic categories, was added as a control pool, bringing up the number of utterances to 40[5] (see Appendix 1). For the sake of methodological consistency, we chose to finish all comments with an exclamation mark.

[5] We decided to avoid further combinations of utterances (by three, four, etc.). Despite its obvious utility for the research, this option would have significantly increased the length of the (already long) questionnaire putting at stake the respondents' commitment.

Existing categorizations	**Intensity**	**Hate features**	**Code**	**1st group of comments**	**2nd group of comments**
				Single utterances **(originals in French, English translations)**	***Single utterances*** **(originals in French, English translations)**
Offensiveness	1	Offense	OFF	*Ce sont des lâches qui fuient leur pays pour venir nous emmerder!* 'They are cowards who flee their country to come and piss us off!'	*S'il n'y aurait pas toutes ces aides sociaux, ma main à couper que ils auraient choisi un autre pays à gratter!* 'If it wasn't for all this welfare, I'd bet that they would have chosen another country to scratch!'
	2	Swear-word	SW	*Des parasites!* 'Parasites!'	*Des merdes!* 'Shits!'
Aggressiveness	1	Incitement to a negative action	ACTN	*Dehors!* 'Get out!'	*Rentrez chez vous!* 'Go home!'
	2	Incitement to a lethal action	AGRL	*À la guillotine!* 'To the guillotine!'	*Foutez les donc en l'air quand vs les voyez sur l'autoroute et vs dites que vs les avez pas vu! Accident ça arrive!* 'So screw them up when you come across them on the highway, and say that you didn't see them! Accidents happen!'
Emotion	-	Emotional expression	EM	*Je les hais!* 'I hate them!	*Ils me dégoutent!* 'They disgust me!
Formal aspects	-	Swear-word in capital letters	SWCAP	*DES PARASITES!* 'PARASITES!'	*DES MERDES!* 'SHITS!'

Table 1: The experimental dataset

For each of these comments, students were asked to decide whether they would: (A) remove the comment because it was hateful; (B) not remove it because it was not hateful; (C) not remove it even though the comment was hateful because everyone has the right to express oneself. These answers encapsulate two kinds of assessments, the first one referring to the hatefulness of the content, the second one concerning the eventual moderation action to undertake. More specifically, contrary to B, answers A and C share a common appraisal of a comment as non-

hateful. Contrary to A, answers B and C share a common action-orientation, advocating a comment's removal (Figure 1).

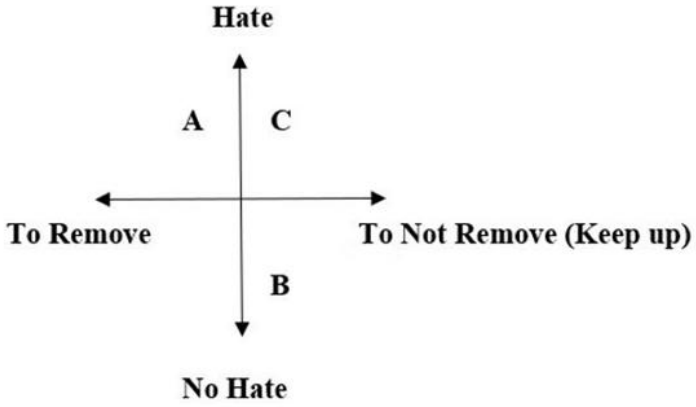

Figure 1: The spatial configuration of appraisals (Hate/No Hate axis) and action-orientations (To remove/To Not Remove) in the questions asked to evaluators

These postures deduce three types of acceptability perceptions, between the intolerable, the tolerable and the "endorsed" – this last category describing contents that were not disapproved of because they were not perceived as hateful (Table 2). As it will be shown later in this study, these distinctions tend to correspond to specific "hate features".

Types of answer	Appraisal of the hatefulness of the comment, as declared by evaluators	Moderation action to undertake, as suggested by evaluators	Deduced evaluators' acceptability perception of the comment
A	Hate	To remove	Intolerable
B	No Hate	To not remove	Endorsed
C	Hate	To not remove	Tolerable

Table 2: Categories of appraisal, moderation and acceptability

5 Research results and discussion

The analysis of the hate/no hate appraisal axis (A+C vs B) shows that most of the comments presented to the evaluators were perceived as hateful (A+C), though differences of ratings do appear between distinctive features (Figure 2).

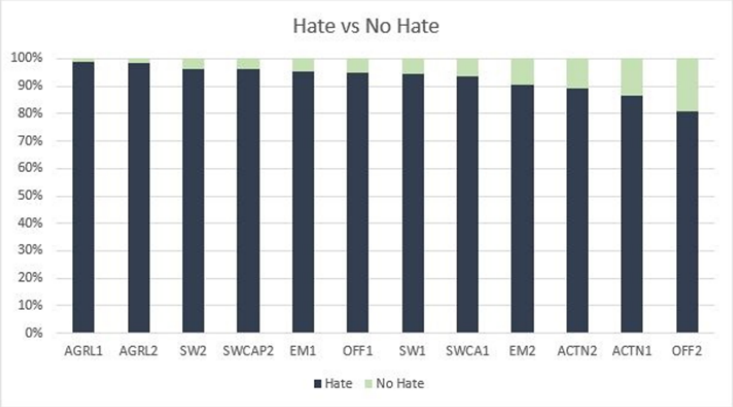

Figure 2: Appraisals: Hate vs No Hate

Indeed, calls to lethal behaviors (AGRL) are logically deemed to be the most hateful. They are followed by swearwords (SW, SWCAP), which confirm Poletto's et al. (2017) assumption of their strong offensive nature; calls for non-lethal actions (ACTN) are generally considered to be the less hateful.

Results about hatefulness are consistent between groups 1 and 2, with the major exception of the "offense" category (OFF). Offense in group 1 (OFF1) appears to be more hateful than offense in group 2 (OFF2). In order to understand this discrepancy, the perception of swearwords needs probably to be taken into consideration. The survey shows that swearwords are strongly objectionable elements, generally appraised as hateful, no matter if in lowercase or uppercase letters. Swearwords have many different functions in many different social contexts (cf. Holmes 2013), but have traditionally been met with skepticism even aversion, and are prone to censorship, especially in written texts (cf. Marsh/Hodsdon 2010). In our case, they seem to operate as offense intensifiers. Within this frame, we might assume that OFF1 might be assessed as more hateful than OFF2 because it contains compelling predicative adjectives (*cowards*) that resemble to swearwords. The same argument may be applied to explain the discrepancy between emotional expressions (EM): as EM1 directly uses the verb *hate*, it may be seen as more hateful.

In general, these results are also consistent with the methodological design of the survey. Since the experiment was designed using only features falling under the umbrella of hate speech, it comes as no surprise to observe that respondents have also flagged most contents as hate. However, the constant presence of a fraction "No Hate" (B), even though limited, attracts our attention. It reveals that in all categories (offense, slurs, aggression, etc.) a more or less small portion of evaluators believed that the comments were not hateful (even in case of calls for lethal behaviors). This could of course be interpreted as a sign of some respondents' seriousness and/or commitment when answering the survey (desire for provocation, etc.). It could also be a sign of a certain banalization of hate speech in contemporary societies, especially among the young. This process is not exclusive to the digital environment, but the latter may constitute a factor that tones down content (cf. Pasta 2022). Finally, the persistence of the "No Hate" fraction among the appraisals of what we have considered to be hate content in the first place could be also reflecting anti-migrant dispositions among the reviewers who embrace the positions expressed. In this sense, it indicates both an emblematic divide regarding ordinary perceptions of hatefulness, and – perhaps – a hiatus on the topic of migration.

These findings bring us to put in perspective legal understandings of hate speech as verbal attacks addressed to "protected" categories (Fortuna/Nunes 2018), in the sense explained earlier in this paper, used by platforms such as Facebook. The heterogeneity of the results in terms of hate evaluation (i. e., the coexistence of those who see hate along with those who do not see it, within same comments) suggests that lay users do not take into consideration – or even are aware of – protected, not-protected or semi-protected targets, when assessing hate speech. Findings also minimize the role of formal components such as the use of capital letters. Punctuation certainly accentuates the expression of emotion, but, contrary to what linguistic approaches have to some point assumed (cf. Monnier/Seoane/Gardenier 2020), it does not seem to provide a relevant indicator when it comes to hate evaluation.

The analysis of the To remove/To not remove (tolerate or endorse) orientation axis (A vs C+B) displays a clear dominance of evaluators' preference for content removal (A) (Figure 3).

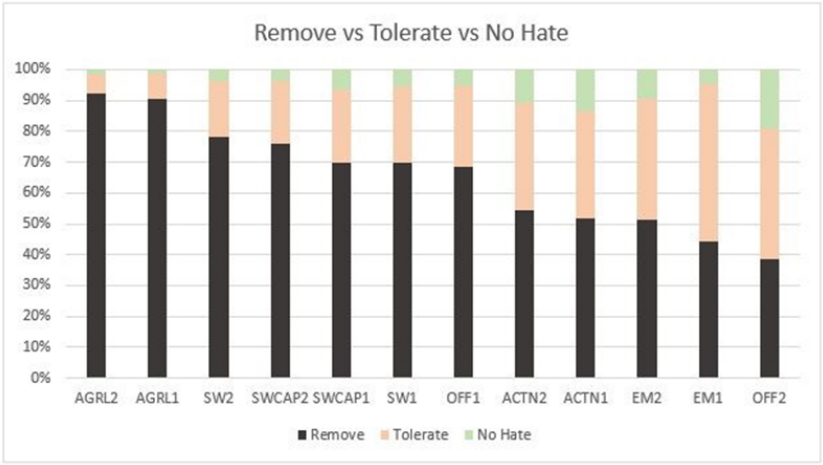

Figure 3: Moderation and acceptability: To remove, to tolerate, to endorse (No Hate)

However, results are less consistent between categories and two major "realms" seem to emerge (Table 3). The first one assembles comments that are considered as hateful and need to be taken down according to a large majority of evaluators (70% and more). These are swearwords, in both lowercase and uppercase letters (SW, SWCAP), and incitements to lethal actions (AGRL). They correspond to the 2nd grades of offensiveness and aggressiveness respectively, as explained above. We will call them "remove features". The second realm comprises all the other forms of single utterances: offense (OFF), call for (non-lethal) action (ACTN), and expression of emotion (EM). For them, "not remove features", opinions diverge: approximately half of the referees are prone to tolerating these statements even though they judge them as hateful ("tolerable" features), or, less often, appraise them as being no hateful ("endorsed" features). These results tend to be quite consistent for both groups, with the exception again of the offense, probably for the same reasons as described above: we might assume that OFF1 might be assessed as more objectionable and condemnable than OFF2, because it contains compelling predicative adjectives (*cowards*) that resemble to swearwords.

Remove features (more than 70% consensus)	**Not remove features (tolerate or endorse)**
Swearwords (SW)	Offence (OFF)
Swearwords in capital letters (SWCAP)	Calls to negative but non-lethal action (ACTN)
Calls to lethal action (AGRL)	Expression of emotion (EM)

Table 3: Remove vs Not remove features

These results indicate that common binary evaluations often used in research (e. g., violating/delete vs non-violating/ignore, cf. Fortuna/Nunes 2018) might not fully reflect the range of publics' possible reactions to hate speech. Acknowledging a hate expression does not necessarily translate in the willingness to remove the element in question. In some cases – probably falling under what Assimakopoulos/Baider/Millar (2017) described as "soft hate" – tolerance holds sway over removal.

Is there a cumulative effect when two hate features are combined? It seems so. When a "tolerate" or "endorse" class is combined with a "remove feature", the result is clearly "to remove" (more than 70% consensus) (Table 4). Less surprisingly, when two "remove features" match together, they produce a "remove feature" as well, marked though by a higher overall consensus concerning its removal (the initial average 79.4% "To remove" score rises to 91.9%). Finally, interestingly, when two "Not remove features" ("tolerate" or "endorse" features) match together, they remain in the "Not remove features realm" (i. e., below 70% remove consensus) but their overall "to remove" recommendation rate increases (the initial 51.6% score goes up to 64%). This is also true for the overall assessment of hatefulness, though the increase is modest: when features are separately considered, their average hate score is 92.8%; when their compounds configurations are taken into consideration, the average hate score rises to 96.8%. It has to be noted, however, that the increase in hatefulness assessment is not proportionate to the increase in the "to remove" recommendation rate. In other words, when two hate features are combined within an utterance, the willingness to remove the comment augments more than its perceived hatefulness.

Feature(s)	**Average "To remove" score**
Not remove feature (OFF)	53,6 %
Not remove feature (OFF)+Remove feature	82,1 %
Not remove feature (ACTN)	53,2 %
Not remove feature (ACTN)+Remove feature	79,6 %
Not remove feature (EM)	47,9 %
Not remove feature (EM)+Remove feature	79,3 %
Remove feature (all categories, average)	79,4 %
Remove feature+ Remove feature (all categories, average)	91,9 %
Not remove feature (all categories, average)	51,6 %
Not remove feature+ Not remove feature (all categories, average)	64,0 %
Average hate score for single features (all categories, average)	92,8 %
Average hate score for compound features (all categories, average)	96,8 %

Table 4: The cumulative effect of hate features: compound hate features increase the willingness to take down hate content.

In this sense, our results seem to be in line with and corroborate existing research which also underlines that cumulating different features of verbal hate expressions impacts hate evaluation. Precisely, analysing annotators' agreement during coding processes, Poletto et al. (2017) found that the highest consensus occurs when a hate speech feature is followed by irony and stereotype. The effect of hate features accumulation on hate assessment certainly needs to be further investigated within multifactorial protocols.

Finally, the overall predominance of the "remove" stance suggests that most evaluators give importance to the harmfulness potential of hate content to the detriment of its necessity or usefulness. They seem to espouse the common approach of hate speech as a threat to society (cf. Calvès 2015; Brown 2017) and endorse regulation practices for contents deemed to be dangerous. However, results also reveal a certain resistance to content banishment, perhaps seen as censorship and as a form of political correctness.

Current debates on political correctness and its limits – within academia and beyond – stress its potential counterproductive effects. Originally grounded in respect for difference and sensitivity to suffering, political correctness has often become a distraction or a silencer. Instead of redefining morality or eliminating prejudice, it might end up accentuating feelings of offense (cf. Hughes 2010). Others decry "safetyism", as it develops especially among the young, i. e., a philosophy that encourages youth to think of themselves as fragile and to demand "that institutions provide them with spaces of emotional and physical comfort" (French 2020: 104). They condemn "a rising hostility to free speech" (French 2020: 104) and deplore that "censorship is now a grassroots effort", "undirected and driven largely by students, to scrub campuses clean of words, ideas, and subjects that might cause discomfort of give offense" (Lukianoff/Haidt 2015: 1).

Beyond those who express justified concerns on the stakes of political over-correctness, all kinds of zealots also denounce such practices, bringing confusion in the debate as to the other that needs to be protected: those expressing extremist views or those targeted by them. Both censorship and hate speech can potentially be experienced as forms of violence. Choosing to defend one to the detriment of the other is a major political decision linked again to a society's visions of the other, but also – and perhaps most importantly – to visions of the harm done on them. The findings of the present study suggest that lay users do condemn "wrongs" understood in the strong sense, as defined by Joel Feinberg (1984), i. e., calls to physical attacks (aggressions) and "serious offences", especially those conveyed through incivility (swearwords). Aside from these cases, they tend to defend a rather "liberal" (Mill 1859), "content-neutral", acceptance of the freedom of expression, according to which it is less the content of an opinion that should justify its prohibition, but the way in which it is expressed (cf. Ramond 2013: 125). In other words, hate action and hate speech (in a narrow sense, i. e., incivility) are not to be tolerated, whereas hate expression (i. e., opinions even offensive) is (Figure 4). Most probably, the necessity to protect freedom of expression as manifested by respondents in our study is to be attributed to a general assumption that characterizes democracies – thoroughly developed by Thomas Scanlon (1972) – according to which citizens can after all be critical and autonomous in their assessments when confronted with hateful content.

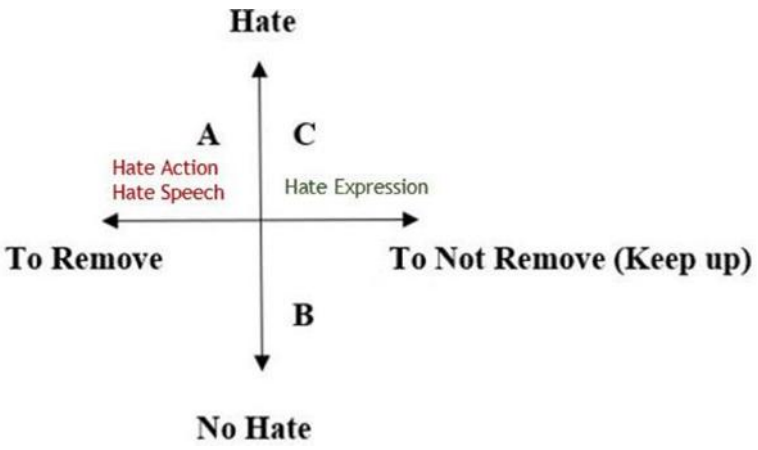

Figure 4: According to the evaluators, hate action (calls to aggression) and hate speech (in a narrow sense, i. e., swearwords) are not to be tolerated, whereas hate expression (opinions even offensive) is

6 Conclusion

In this research, we have explored hatefulness assessment by young people in France. Our aim was to study how distinctive hate forms are evaluated by ordinary people. Inevitably, results apply to France's specific context, as well as to the specific population of the 230 respondents of the survey. However, some significant key elements seem to emerge. First, research confirms that most of the content presented to participants was appraised as hateful: offensiveness, aggressiveness and expression of emotions were understood as hate. The study also highlighted that in most cases, participants opted for content removal. With no surprise, calls to lethal aggressions were the most objectionable comments, along with, interestingly enough, swearwords, which seem to operate as offense intensifiers. Incitements to non-lethal actions and expressions of emotions engender some rates of tolerance or even acceptance and endorsement, judged by some as non-hateful. On the contrary, punctuation seems to have no significant effect on the overall judgments. Though incivility (swearwords) is less tolerated, emotions and non-violent-oriented ideas are deemed to be acceptable even when they are not endorsed. Finally, combinations of hate speech features increase the perceived hatefulness and, even more substantially, the willingness to take down the content in question. A distinction between hate action, hate speech, and hate expression revealed to be significant.

The extension of the experimental survey to larger audiences – in terms of age as well as in terms of geographical and cultural settings – is certainly needed in order to be able to comment on the capacity to generalize the above results. However, these findings could open new horizons for the development of automated assessments of hate, which could consider, for instance, the accumulation of hate forms within a same comment. More research needs certainly to be conducted, exploring further combinations of hate forms, something that the experimental frame of our study was not able to address.

Of course, hate speech moderation its only part of the problem and inevitably part of the solution. Despite word filters and human moderation, users constantly find ways through the system, using inventive methods to post punishable hate. They veil abuse through subtler, more sophisticated and even unexpected guises, such as, for example, the invocation of biblical passages (cf. Roberts 2019: 160–161). Moreover, by solely attending to content, governments, technology companies, academia and civil society fail to understand and address the connective infrastructure that brings "hate cultures" together (cf. Ganesh 2018: 44) as well as the overall patterns of antisocial behavior (cf. Cheng/Danescu-Niculescu-Mizil/Leskovec 2015). They also neglect to measure the impact of other aspects that appear to be significant, such as the sparsity of pro-minority comments in shaping community perceptions and the need for supportive voices. Research on this field seems promising, though results are also nuanced (Buntain et al. 2020; Chandrasekharan et al. 2017, 2021; Palakodety/KhudaBukhsh/Carbonell 2020; Ribeiro et al. 2020). As polarizations increase, tending to dominate contemporary public spheres, it is also the overall network structures and interrelations of hate that need to be addressed, besides its content.

References

Assimakopoulos, Stavros/Baider, Fabienne/Millar, Sharon (2017): *Online Hate Speech in the European Union. A Discourse-Analytic Perspective*. New York: Springer. link.springer.com/book/10.1007/978-3-319-72604-5 [16.06.2025].

Benesch, Susan et al. (2021): "Dangerous Speech: A Practical Guide. By the Dangerous Speech Project". dangerousspeech.org/libraries/guide [16.06.2025].

Bonnafous, Simonne (1991) : *L'immigration prise aux mots*. Paris: Éd. Kimé.

Bricks 2016: Maneri, Marcello (ed.): *#Silence Hate. Study on hate Speech Online in Belgium, Czech Republic, Germany and Italy. BRICKS Building Respect on the Internet by Combating hate Speech.* University of Milano Biccoca. bricks-project.eu/wp/wp-content/uploads/2016/10/relazione_bricks_eng2-1.pdf [16.06.2025].

Brown, Alex (2015): *Hate Speech Law: A philosophical Examination*. New York: Routledge.

Brown, Alex (2017): "What is Hate Speech? Part 1: The Myth of Hate". *Law and Philosophy* 36/4: 419–468. link.springer.com/article/10.1007/s10982-017-9297-1 [16.06.2025].

Buntain, Cody et al. (2020): "YouTube Recommendations and Effects on Sharing Across Online Social Platforms". arxiv.org/abs/2003.00970 [16.06.2025].

Calvès, Gwénaële (2015) : "Les discours de haine et les normes internationales". *Esprit* 10 : 56–66.

Chandrasekharan, Eshwar et al. (2017): "September 6. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech". *Proceedings of the ACM on Human-Computer Interaction* 1 (CSCW). doi.org/10.1145/3134666.

Chandrasekharan, Eshwar et al. (2021): "Quarantined! Examining the Effects of a Community-Wide Moderation Intervention on Reddit". arxiv.org/abs/2009.11483 [16.06.2025].

Cheng, Justin/Danescu-Niculescu-Mizil, Cristian/Leskovec, Jure (2015): "Antisocial Behavior in Online Discussion Communities". ICWSM Conference. arxiv.org/abs/1504.00680 [16.06.2025].

EDRi (2020): *French Avia law declared unconstitutional: what does this teach us at EU level?*. Association European Digital Rights, June 24. https://edri.org/our-work/french-avia-law-declared-unconstitutional-what-does-this-teach-us-at-eu-level/ [04.11.2025].

Feinberg, Joel (1984): *Offense to Others*. Oxford: Oxford University Press.

Fortuna, Paula/Nunes, Sérgio (2018): "July 31. A Survey on Automatic Detection of Hate Speech in Text". *ACM Computing Surveys* 51/4:. 1–30.

French, David (2020): *Divided we Fall, America's Secession Threat and How to Restore our Nation*. New York: St. Martin's Press.

Ganesh, Bharath (2018): "The ungovernability of digital hate culture". *Journal of International Affairs* 71/2: 30–46. jia.sipa.columbia.edu/news/ungovernability-digital-hate-culture [16.06.2025].

Gardenier, Matthijs/Monnier Angeliki (2020) : "Atténuer la radicalité. Stratégies de communication de groupes vigilantistes anti-migrants". *Mots. Les langages du politique* 123 : 63–78. doi.org/10.4000/mots.26737.

Gillespie, Tarleton (2018): *Custodians of the Internet. Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven: Yale University Press.

Google (2025): General Guidelines, September 11, https://static.googleusercontent.com/media/guidelines.raterhub.com/fr//searchqualityevaluatorguidelines.pdf [04.11.2025].

Hargreaves, Alec G. (2012) : "De la victoire de la gauche à la percée de l'extrême droite : l'ethnicisation du jeu électoral français". *Histoire@Politique* 16 : 154–165. shs.cairn.info/revue-histoire-politique-2012-1-page-154?lang=fr [16.06.2025].

Hargreaves Alec G. (2016) : "La percée du Front national". *Hommes & Migrations* 1313 : 29–35. doi.org/10.4000/hommesmigrations.3555.

Héran, François (2020) : "October 30. Lettre aux professeurs d'histoire-géographie. Ou comment réfléchir en toute liberté sur la liberté d'expression". *La Vie des idées.* laviedesidees.fr/Lettre-aux-professeurs-d-histoire-geo-Heran.html [16.06.2025].

Holmes, Janet (2013): *An Introduction to Sociolinguistics*. London/New York: Routledge.

Hughes, Geoffrey (2010): *Political Correctness. A History of Semantics and Culture*. Malden, MA/Oxford: Wiley-Blackwell.

JO (2020) : *LOI n° 2020-766 du 24 juin 2020 visant à lutter contre les contenus haineux sur internet*, Journal officiel électronique authentifié n° 0156 du 25/06/2020. legifrance.gouv.fr/jorf/id/JORFTEXT000042031970 [04.11.2025].

Lukianoff, Greg/Haidt, Jonathan (2015): "September. The Coddling of the American Mind". *The Atlantic*. theatlantic.com/magazine/archive/2015/09/the-coddling-of-the-american-mind/399356/ [16.06.2025].

Mahfud, Yara et al. (2016) : "Distance culturelle, perception du multiculturalisme et préjugés envers les immigrés en France". *L'Année psychologique* 116/2 : 203–225. shs.cairn.info/revue-l-annee-psychologique1-2016-2-page-203?lang=fr [16.06.2025].

Matsuda, Mari J. (1989): "Public Response to Racist Speech: Considering the Victim's Story". *Michigan Law Review* 87/8: 2320–2381. repository.law.umich.edu/mlr/vol87/iss8/8/ [16.06.2025].

Marsh, David/Hodsdon, Amelia (2010): *Guardian Style*. 3rd ed. London: Guardian Books.

Mill, John Stuart (1859): *On Liberty*. London: John W. Parker and Son.

Monnier, Angeliki/Seoane Annabelle/Gardenier, Matthijs (2020) : "Analyser le discours de haine en ligne: réflexions méthodologiques". In : Vergely, Pascale/Carbou, Guillaume (eds.) : *Médias et émotions. Catégories d'analyse, problématiques, concepts*. Rome, Roma TrePress : 65–80. romatrepress.uniroma3.it/wp-content/uploads/2020/10/Prismes-n.-2-2020.pdf [16.06.2025].

Monnier, Angeliki/Boursier, Axel (2022) : "La structure actantielle des discours de haine dans les plateformes participatives en ligne". *Cahiers de narratologie* 42. journals.openedition.org/narratologie/13848 [16.06.2025].

Monnier, Angeliki/Boursier, Axel/Seoane, Annabelle (eds.) (2022): *Cyberhate in the Context of Migrations*. London/New York/Shanghai: Palgrave Macmillan/Springer.

Murthy, Dhiraj et al. (2015): "August 25. Do We Tweet Differently from Our Mobile Devices? A Study of Language Differences on Mobile and Web-Based Twitter Platforms". *Journal of Communication* 65/5: 816–837.

Palakodety, Shriphani/Khuda Bukhsh, Ashiqur R./Carbonell, Jaime G. (2020): "The Refugee Experience Online: Surfacing Positivity Amidst Hate". *Frontiers in Artificial Intelligence and Applications* 325 (ECAI 2020): 2925–2926. doi.org/10.3233/FAIA200456.

Papacharissi, Zizi (2014): *Affective Publics. Sentiment, Technology, and Politics*. New York: Oxford University Press.

Pasta, Stefano (2022): "Social network conversations with young authors of online hate speech against migrants". In: Monnier, Angeliki/Boursier, Axel/Seoane, Annabelle (eds.): *Cyberhate in the Context of Migrations*. Cham, Palgrave Macmillan/Springer: 187–214.

Poletto, Fabio et al. (2017): "Hate Speech Annotation: Analysis of an Italian Twitter Corpus". In: Basili, Roberto/Nissim Malvina/Satta, Giorgio (eds.): *Proceedings of the 4th Italian Conference on Computational Linguistics* (CLIC-IT), 11–12 December 2017, Rome. Torino, Accademia University Press: 263–268. books.openedition.org/aaccademia/2448?lang=fr [16.06.2025].

Ramond, Denis (2011) : "Liberté d'expression : De quoi parle-t-on?". *Raisons politiques* 44 : 97–116. shs.cairn.info/revue-raisons-politiques-2011-4-page-97?lang=fr [16.06.2025].

Ramond, Denis (2013) : "L'ironie de la liberté d'expression". *Raisons politiques* 52 : 123–141. shs.cairn.info/revue-raisons-politiques-2013-4-page-123?lang=fr [16.06.2025].

Reiners, Liane/Schemer, Christian (2020): "A Feature-based Approach to Assess Hate Speech in User Comments". *Questions de communication* 38: 529–548. doi.org/10.4000/questionsdecommunication.24808.

Ribeiro, Manoel Horta et al. (2020): "Do Platform Migrations Compromise Content Moderatio? Evidence from r/The_Donald and r/Incels". doi.org/10.48550/arXiv.2010.10397.

Ricœur, Paul (1996) : "L'usure de la tolérance et la résistance de l'intolérable". *Diogène* 176: 166–176.

Schrock, Andrew Richard (2015): "Communicative Affordances of Mobile Media. Portability, Availability, Locatability, and Multimediality". *International Journal of Communication* 9: 1229–1246. ijoc.org/index.php/ijoc/article/viewFile/3288/1363 [16.06.2025].

Roberts, Sarah T. (2019): *Behind the Screen. Content Moderation in the Shadow of Social Media*. New Haven/London: Yale University Press.

Scanlon, Thomas (1972). "A Theory of Free Expression". *Philosophy and Public Affairs* 1/2: 204–226.

SELMA: Social and Emotional Learning for Mutual Awareness (2019): "Hacking Online Hate: Building an Evidence Base for Educators". hackinghate.eu/news/hacking-online-hate-building-an-evidence-base-for-educators/ [16.06.2025].

Siapera, Eugenia (2019): "Organised and Ambient Digital Racism: Multidirectional Flows in the Irish Digital Sphere". *Open Library of Humanities* 5/1. doi.org/10.16995/olh.405.

Strippel, Christian et al. (2020): "Modularized Hate Speech Annotation: A Human-Labeled Dataset of German User Comments on Flight and Migration". Poster presented at the *6th International Conference on Computational Social Science* (IC2S2), 17–20 July 2020, Amherst St. Cambridge (MA).

Sunstein, Cass R. (1999): "The Law of Group Polarization". *The Chicago Working Paper Series, Working Paper Nº 91*. https://ssrn.com/abstract=199668 [04.11.2025].

UN 1948: UN General Assembly, Resolution 217 A: *Universal Declaration of Human Rights*. A/RES/217(III) (December 10, 1948). un.org/en/about-us/universal-declaration-of-human-rights [14.11.2025].

UN 1967: "International Covenant on Civil and Political Rights". *Treaty Series* 999. https://treaties.un.org/doc/treaties/1976/03/19760323%2006-17%20am/ch_iv_04.pdf [04.11.2025].

Veikou, Mariangela/Siapera, Eugenia (2015): "Rethinking Belonging in the Era of Social Media: Migration and Presence". In: de Been, Wouter/Arora, Payal/Hildebrandt, Mireille (eds.): *Crossroads in New Media, Identity and Law*. London, Palgrave Macmillan: 119–137.

Wypych, Michał/Bilewicz, Michał (2022): "Psychological toll of hate speech: The role of acculturation stress in the effects of exposure to ethnic slurs on mental health among Ukrainian immigrants in Poland". *Cultural Diversity and Ethnic Minority Psychology* 30/1: 35–44.

APPENDIX 1: The **compound hate features of the data set**

N°	**Hate features**	**Code**	**1st group of comments**	**2nd group of comments (control dataset)**
			Compound utterances	**Compound utterances**
			(originals in French and English translations)	**(originals in French and English translations)**
1	Offense and swearword	OFF/SW	Ce sont des lâches qui fuient leur pays pour venir nous emmerder! Des parasites!	S'il n'y aurait pas toutes ces aides sociaux, ma main à couper que ils auraient choisi un autre pays à gratter! Des merdes!
			'They are cowards who flee their country to come and piss us off! Parasites!'	'If it wasn't for all this welfare, I'd bet that they would have chosen another country to scratch! Shits!'
2	Offense and call to negative (not lethal) action	OFF/ACTN	Ce sont des lâches qui fuient leur pays pour venir nous emmerder! Dehors!	S'il n'y aurait pas toutes ces aides sociaux, ma main à couper que ils auraient choisi un autre pays à gratter! Rentrez chez vous!
			'They are cowards who flee their country to come and piss us off! Get out!'	'If it wasn't for all this welfare, I'd bet that they would have chosen another country to scratch! Go home!'
3	Offense and call to aggressive lethal action	OFF/AGRL	Ce sont des lâches qui fuient leur pays pour venir nous emmerder! À la guillotine!	S'il n'y aurait pas toutes ces aides sociaux, ma main à couper que ils auraient choisi un autre pays à gratter! Foutez les donc en l'air quand vs les voyez sur l'autoroute et vs dites que vs les avez pas vu ! Accident ça arrive!

			‘They are cowards who flee their country to come and piss us off! To the guillotine!’	‘If it wasn’t for all this welfare, I’d bet that they would have chosen another country to scratch! So screw them up when you come across them on the highway, and say that you didn’t see them! Accidents happen!’
4	Offense and expression of emotion	OFF/EM	Ce sont des lâches qui fuient leur pays pour venir nous emmerder! Je les hais!	S’il n’y aurait pas toutes ces aides sociaux, ma main à couper que ils auraient choisi un autre pays à gratter! Ils me dégoutent!
			‘They are cowards who flee their country to come and piss us off! I hate them!’	‘If it wasn’t for all this welfare, I’d bet that they would have chosen another country to scratch! They disgust me!’
5	Offense and swearword in capital letters	OFF/SWCAP	Ce sont des lâches qui fuient leur pays pour venir nous emmerder! DES PARASITES!	S’il n’y aurait pas toutes ces aides sociaux, ma main à couper que ils auraient choisi un autre pays à gratter! DES MERDES!
			‘They are cowards who flee their country to come and piss us off! PARASITES!’	‘If it wasn’t for all this welfare, I’d bet that they would have chosen another country to scratch! SHITS!’
6	Swearword and call to negative (not lethal) action	SW/ACTN	Des parasites! Dehors!	Des merdes! Rentrez chez vous!
			‘Parasites! Get out!’	‘Shits! Go home!’
7	Swearword and call to aggressive lethal action	**SW/AGRL**	Des parasites! À la guillotine!	Des merdes! Foutez les donc en l’air quand vs les voyez sur l’autoroute et vs dites que vs les avez pas vu! Accident ça arrive!
			‘Parasites! To the guillotine!’	‘Shits! So screw them up when you come across them on the highway, and say that you didn’t see them! Accidents happen!’
8	Swearword and expression of emotion	SW/EM	Des parasites! Je les hais!	Des merdes! Ils me dégoutent!
			‘Parasites! I hate them!’	‘Shits! They disgust me!’

9	Call to negative (not lethal) action and call to aggressive lethal action	ACTN/AGRL	Dehors! À la guillotine! 'Get out! To the guillotine!'	Rentrez chez vous! Foutez les donc en l'air quand vs les voyez sur l'autoroute et vs dites que vs les avez pas vu! Accident ça arrive! 'Go home! So screw them up when you come across them on the highway, and say that you didn't see them! Accidents happen!'
10	Call to negative (not lethal) action and expression of emotion	ACTN/EM	Dehors! Je les hais! 'Get out! I hate them!'	Rentrez chez vous! Ils me dégoutent! 'Go home! They disgust me!'
11	Call to negative (not lethal) action and swearword in capital letters	ACTN/SWCAP	Dehors! DES PARASITES! 'Get out! PARASITES!'	Rentrez chez vous! DES MERDES! 'Go home! SHITS!'
12	Call to aggressive lethal action and expression of emotion	AGRL/EM	À la guillotine! Je les hais! 'To the guillotine! I hate them!'	Foutez les donc en l'air quand vs les voyez sur l'autoroute et vs dites que vs les avez pas vu ! Accident ça arrive! Ils me dégoutent! 'So screw them up when you come across them on the highway, and say that you didn't see them! Accidents happen! They disgust me!'
13	Call to aggressive lethal action and swearword in capital letters	AGRL/SWCAP	À la guillotine! DES PARASITES! 'To the guillotine! PARASITES!'	Foutez les donc en l'air quand vs les voyez sur l'autoroute et vs dites que vs les avez pas vu ! Accident ça arrive! DES MERDES! 'So screw them up when you come across them on the highway, and say that you didn't see them! Accidents happen! Accidents happen! SHITS!'
14	Expression of emotion and swearword in capital letters	EM/SWCAP	Je les hais! DES PARASITES! 'I hate them! PARASITES!'	Ils me dégoutent! DES MERDES! 'They disgust me! SHITS!'