Humorous remarks in covert hate speech and counter-speech

Fabienne Baider
University of Cyprus

Christina Romain
University of Aix-Marseille

It has been 20 years since Susan Herring and her colleagues concluded that online communication, and especially various internet forums and social media, had become a new arena for the enactment of power inequalities, including those motivated by sexism, racism and heterosexism (Herring et al. 2002: 371). Indeed, and as Ruth Wodak noted, "The more anonymous the genre, the more *explicit exclusionary rhetoric* tends to be" (Wodak 2015: 207, our emphasis). The very same advantages of digitization-- such as connectivity, support of social relationships and access to new knowledge-- have led to this 'startling' and rapid rise in cyber-hate across the world (Citron 2014) to such a degree that the internet has been described as a prime location for the collection and analysis of (violent) discriminatory discourse. Such discourses manifest in various ways, including Twitter mobbing, trolling, cyberbullying and sexting—all of which fall under the umbrella term 'hate speech.' This phenomenon has attracted a vast and fast-growing field of inquiry in academic circles, and legal measures have been enacted to erase / delete such calls for violence and hatred. In the present paper we focus on covert hate speech and in particular on the way in which humor[1] is used in hate speech. We argue that it is important to investigate the use of humor in offensive messages because very often these humorous messages are, in fact, 'covert hate speech,' i.e., a discourse that manifests an illocutionary dimension similar to overt hate speech (the will to hurt), although this (hateful) intention is not obvious. Our data comprises 3000 Facebook posts that have been tagged as triggers of hate speech.[2] Definitions of covert hate speech, and decriptions/examples of humorous ways of conveying covert hate speech are the focus of the first section. In the second section we present our data, our methodology of analysis and some general results. The final section considers our results in light of other research specifically targeting racist humor, and

[1] Our use of the word *humor* includes irony and sarcasm, which will also be our focus.

[2] This article is based on data collected within IMsyPP, Innovative Monitoring Systems and Prevention Policies of Online Hate Speech (ImsyPP), cofunded by the European Union's Rights, Equality and Citizenship Programme (2014-2020). The project IMSyPP, grant no. 875263 is available at http://imsypp.ijs.si/. The European Commission's support for the production of this publication does not constitute an endorsement of the contents, which reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

explains why these ‘humorous’ ways of writing hateful or offensive messages should be labeled as ‘covert hate speech.’ In our concluding discussion, we draw on findings and data from two other European projects in which we have been involved, namely, the EU social justice projects C.O.N.T.A.C.T. and IMsyPP.[3]

1. Censorship and Covert Hate Speech

If we truly wish to derail hate speech, it is urgent to look closely at *covert* hate speech, and especially the way in which humor features in hate speech. To this end, we discuss the 2016 Code of Conduct, an important law that requires all hate speech messages on social media to be taken down. Yet we cannot help but note a very significant limitation of the Code--its reliance on automatic detection, which seems to unfairly affect specific communities. Moreover, the Code is unable to effectively ensure a ‘safe internet’ because of the pervasive presence of covert hate speech, which does not fall within the prescribed legal boundaries of hate speech.

1.1 The limitations of the 2016 Code of Conduct

The 2016 Code of Conduct,[4] as enacted by the European legislative bodies, mandates that social media companies remove or disable access to hate speech content within 24 hours after being reviewed; Facebook, for example, deleted almost 90% of all cases that were flagged. As such, *hate speech* is considered to be a performative speech act. These measures led to many questions including the vexing issue of freedom of expression, which has been debated elsewhere, mainly in legal publications (cf. Strossen 2018, inter alia). For that reason, and for the sake of safeguarding freedom of expression, the right to appeal any decision has been included to ensure a level of transparency. For instance, Facebook alone received 1.1 million appeals in a period of only three months (January 2019 - March 2019) and 130,000 pieces of content were restored after reassessment.

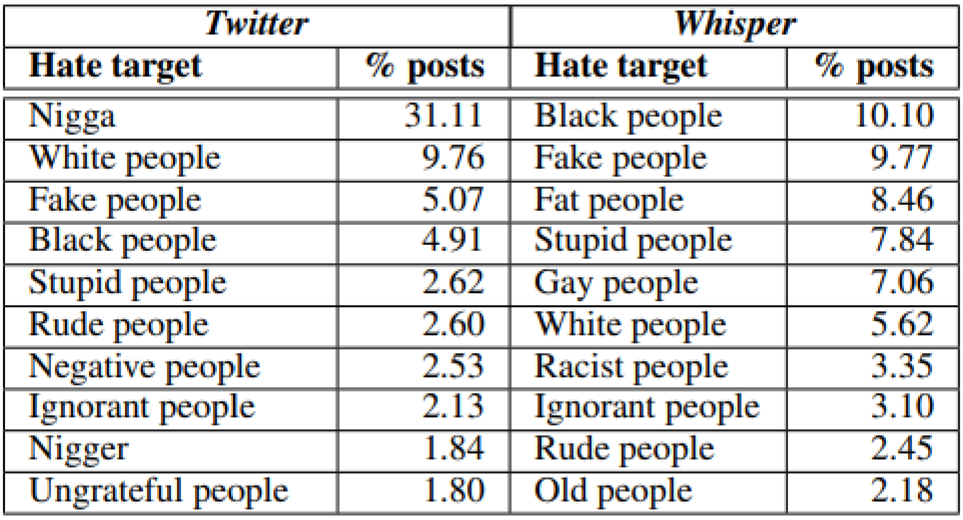
To apply the Code of Conduct, the EU has adopted its own definition, which is based on the European framework decision 2008/913 / JAI defining hate speech as publicly *inciting* to violence, or *hatred* directed against a group of persons or a member of such a group defined *by*

[3] The C.O.N.T.A.C.T. program (Creating an On-Line Network, Monitoring Team and Phone App to Counter Hate Crime Tactics) was cofunded by the European Union’s Rights, Equality and Citizenship Programme (2014-2020), with the grant no. 6706.

[4] https://ec.europa.eu/info/sites/info/files/code_of_conduct_on_countering_illegal_hate_speech_online_en.pdf

reference to race, colour, religion, descent or national or ethnic origin. In an earlier study (Baider 2020) we pointed out the reasons why this definition does not /cannot adequately address the issue-- to wit, many have been victimized for numerous reasons other than those listed above (Chakraborti 2015; Silva et al. 2016). This is clear in Silva et al.'s (2016) study, which analysed hate speech in a specific dataset (Table 1); the results of their survey show that while the highest percentage of hate victims were targeted because of race (including whites), there were many other categories of victims that do not appear in most definitions of hate speech.

Table 1. Ten top targets of hate speech on Twitter and Whisper (Silva *et al* 2016)



Twitter		*Whisper*	
Hate target	**% posts**	**Hate target**	**% posts**
Nigga	31.11	Black people	10.10
White people	9.76	Fake people	9.77
Fake people	5.07	Fat people	8.46
Black people	4.91	Stupid people	7.84
Stupid people	2.62	Gay people	7.06
Rude people	2.60	White people	5.62
Negative people	2.53	Racist people	3.35
Ignorant people	2.13	Ignorant people	3.10
Nigger	1.84	Rude people	2.45
Ungrateful people	1.80	Old people	2.18

The other identified categories, e.g., fat people, rude people, are not included in the EU hate speech law and therefore will be vulnerable to and unprotected from hateful messages. Moreover, although the EU definition targets racism in particular, one of the latest monitoring of the Code of Conduct (2021) revealed that the parameter of sexual orientation (18,2%) was as often flagged as xenophobia (18%)[5]. Therefore, the EU law does not seem to protect all EU citizens who are victims of hate speech in the same way.

1.2 Community of practice and context

[5] https://ec.europa.eu/info/sites/default/files/factsheet-6th-monitoring-round-of-the-code-of-conduct_october2021_en_1.pdf

Furthermore, understanding what qualifies a statement as hate speech is hard enough for human beings, never mind artificial intelligence systems. Indeed, monitoring online hate speech means recognizing –and taking down-- false positives as well. It has been shown that artificial intelligence models were 1.5 times more likely to flag tweets written by African Americans as *offensive* -- in other words, a false positive -- compared to other tweets (Sap et al. 2019). This likely happens because these communities use a particular vernacular with its own set of rules, which differ from those of other communities (Dixon et al. 2018). It is also reported that hate speech classifiers appear to be overly sensitive to group identifiers like *black, gay, transgender,* which are indicators of hate speech only when used in *specific settings* (Xia et al. 2020). We can therefore question the standards and registers used to evaluate hate speech as defined by the European Union; for example, the use of the words *nigger* and *bitch* is not always an insult—rather, these words can signify relational proximity in some communities of practice (cf. Culpeper 2011; Baider 2020). Hate speech detections / algorithms may be based on communication models that favor a certain idea/standard of correct, decent, authorized communication, and that, at the same time, may stigmatize other registers and communities of practice. The expression 'specific settings' used above is important. Indeed, in terms of understanding meaning, the current automatic detection models cannot detect something vital: the illocutionary force, its perlocutionary dimension, and the context of each utterance which "must be considered when analyzing its exclusionary force" (Wodak 2015: 207).

However, taking context into consideration is a very complex issue. In an earlier study, Hardacker (2010: 217) emphasized the vast number of contextual variables, and how this makes the task of being fair in taking down messages or in labelling messages as hate speech especially daunting.

1.3 Prevalence of covert hate speech

In any case, because overt racist, sexist, and homophobic statements have been labelled illegal in the EU cyberspace, and are supposedly systematically taken down, what is called covert hate speech has become prevalent as a way to –legally--express hatred on social media (Kumar et al. 2018). Covert hate speech offers a way to express hatred and disgust without being held responsible. At present there is no legal recourse to curtail covert hate speech, since legal action in this regard could ultimately impinge on freedom of expression.

Indeed, language does not have to contain insults to qualify as hate speech (Bartlett et al. 2020), and contemporary expressions of racism manifest lesser or more subtle forms of discrimination (Dovidio et al. 2010: 43). Some forms of speech do not reproduce racist stereotypes through conventional reference, such as explicit stereotypes; instead, they invite inferences (Hill 2008: 41). Among the most common discursive strategies used to demonize specific Others, as Wodak has observed, are "scapegoating, blaming the victim, victim–perpetrator reversal, trivialization and denial" (2015: 206).

Inference, or 'implicit meaning,' which Grice (1962) coined as *implicature*, is the basis for most of these indirect strategies, which are covert devices that allow one to express a very negative stance, especially by using humor (Chovanec 2018; Baider and Constantinou 2020). Indeed according to Attardo et al.'s theory (2003), irony and sarcasm can be defined by the co-presence of (at least) two distinct meanings, their antiphrastic nature (or at least their difference), and the contextual inappropriateness of the utterance. The statement/content contradicts the intended meaning or it contradicts the reality as we know it (Attardo 2000), and inferences are needed to understand the intended meaning.[6]

2. Theoretical Framework

Adopting Fairclough's tripartite model (Fairclough 1995), we consider texts placed within discursive and social practices and in specific contexts. The focus of the Facebook posts that comprise our data is on migration issues in the year 2015, a time when Europe saw thousands of new arrivals every month. This database is ideal for analyzing how ideological stances are constructed, whether for or against migration; and here we particularly examine the role of humor. Nevertheless, we must note that the same data would not be found today on the same social medium (Facebook), since the historical moment is different.

2.1 Hatred and humor: a potential sharing of hurtful stereotypes

Humor has a hegemonic dimension in as much as it 'puts in his/ her place' whoever appears to have violated norms and expectations, at least in the eye of the speaker (Billig 2005).

To understand the functions of humor, we adopted Charteris-Black's explanation of how metaphors function, and focused on how humor activates "emotions originating in pre-existent myths about classes, nations, and other social and ethnic groupings" (2006:204). We analyse

[6] In this paper we will focus more on sarcasm, the aim of which has been argued is to hurt the target of the utterance, and hurt is at the core of hate speech (Baider 2020; Culpeper 2021).

these ‘pre-existent myths’ as stereotypes and prejudices that “communicate emotionally potent and unambiguous evaluations on an ethical scale of right and wrong” (ibid). As far back as the earliest research targeting prejudice (Hamilton and Trolier 1986; Allport 1954; Lippmann 1922), it has been shown that creating /circulating negative stereotypes (i.e., constructing, reinforcing and perpetuating negative evaluations) has the power to incite hostility towards the target group. Stereotypes and prejudices, as well as their evaluative and emotional dimensions, are common to the emotion of hatred. Indeed, hatred has been defined as an intense emotion based on 'circulating' discourses and the institutionalized behaviors that legitimize it; the feeling of hatred, therefore, has a *cognitive* origin, which can be explained by an extremely negative *judgment* and negative *evaluation*, which are also characteristic of prejudice.

We hypothesize that sarcastic remarks especially --- since they are based on prejudices ‘disguised’ with a humorous tone—deliver a negative evaluation. They are more likely to be effective in terms of persuading the listener than overt discriminatory statements. Both hatred and this type of humor encourage identification of the targeted individual as belonging to a despised or feared community.

2.2 Humor and counter-speech

In this chapter we also address humor used in counter-narratives, i.e., arguments or strategies used to counter/respond to hateful comments. Counter-narratives have been defined in several ways (Benesch 2014; Mathew et al. 2018), and we have chosen to adopt Mathew et al.’s 2008 definition, whereby we consider a counter-narrative to be any response to hateful statements, even if that response is hostile. Gemmerli (2015: 4), in regard to extremist propaganda, defined a successful counter-narrative as one that wins the argument “by deconstructing and delegitimising extremist propaganda.” Counter-narratives also aim to foster critical thinking and spark reactions, i.e., in the best scenario, opening a dialogue.

To tag counter-narratives, as we have done in our research, we used Benesch’s taxonomy (2014), which lists eight criteria for identifying counter-speech: Presentation of facts; Warning of consequences of writing hateful comments; Pointing out hypocrisy; Humor; Denouncing the comment as such; Tone; Images; Affiliation. Other authors (Braddock and Horgan 2016) have noted additional argumentation strategies, including: Identification of contradictions in the argumentation: Identification of unequal elements being compared.

The majority of authors who work on counter-speech have found that humor is used as a strategy. For instance, Warrington (2017) noted that if direct counter-narratives “deconstruct, discredit or demystify violent extremist messaging through ideology, logic, fact: *humor can also be one strategy*" (our italics). Tuck and Silverman (2016: 20) have found instances where satire and humor are used to undermine the claims made online by haters; Hakoköngäs et al 2020 advised to use "bitter humor" to destabilize or even persusade haters; Gemmerli also suggested "*mak[ing] fun of* (…) the extremist ideology's claims'"(2015: 4, our italics). In fact, our data (cf. section 3.) reveal a widespread use of the strategy of 'making fun' to challenge offensive statements.

2.3 A socio-pragmatic perspective on humor

Jokes are based on--and at the same time they activate-- culturally based schemas; this means that they have an expressive potential for cognitive and emotional engagement. On the socio-cultural level, they serve not only to identify the Other in specific categories but also to evaluate/judge, either the Self or the Other(s). With regard to the socio-cultural use of humor in racist comments, we can refer to Weaver (2016, 2011), Chovanec (2018), Archakis and Tsakona (2019) and Baider and Constantinou (2020). Following Archakis and Tsakona's (2005) proposal, in our study we integrate a socio-pragmatic approach, which allows us to argue that the choice of humor should be interpreted with reference to the *purpose* of such use within particular discourse contexts.

3. Data and Methodology

As mentioned earlier, the data we analyzed was collected in 2015, i.e., prior to the 2016 Code of Conduct that mandates the removal of hateful content on and by social media. Therefore, in our data we have access to uncensored opinions of the 'general public,'[7] which can provide insight into the strategies used to express the values and beliefs that underlie the commentators' hateful messages. Moreover, we can also analyse the counter-stances, counter-narratives and their strategies.

3.1 Description of the database and tagging

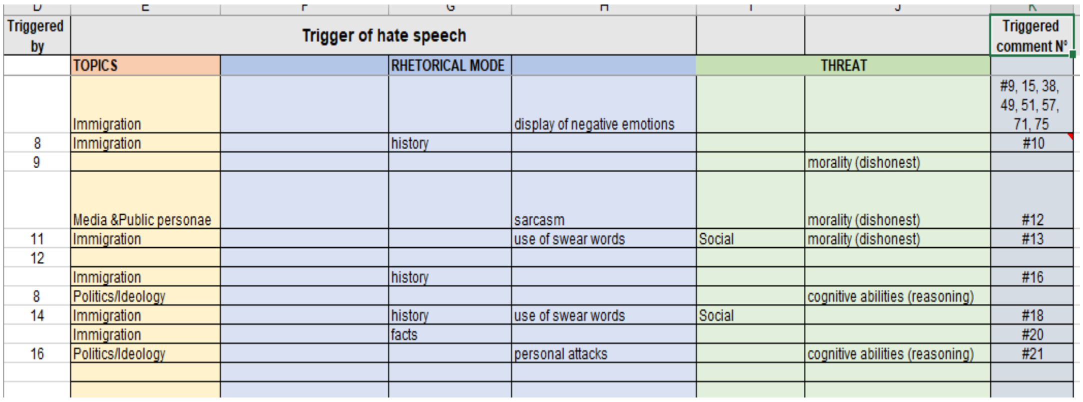
The dataset with which we worked is called FRENK,[8] and it has been extensively described in Ljubesic et al. (2019). It comprises Facebook posts and comments related to mainstream news

[7] Still to be researched, however, is whether the tweets are posted by the general public or only by an (unrepresentative) fringe of that general public.

[8] The acronym FRENK stands for "FRENK - Raziskave Elektronske Nespodobne Komunikacije" (Engl. "Research on Electronic Inappropriate Communication"). There is also a Slovenian dataset, but we worked only on the British dataset.

media from Great Britain and it covers two topics, migrants and the LGBTcommunity. We focus here on the migrant topic. In the first stage the Facebook comments were manually annotated for different types of socially unacceptable discourse (SUD) by the Slovenian team in the IMsyPP project. Over three months, our Cypriot team analysed 3000 comments and tagged what we considered triggers of hate speech (see Table 2), and what type of counter-speech was written to answer hate speech. We based our tagging on the topic or the main theme of the threads, the rhetorical mode used in hate speech and that used in counter-speech (inspired by Benesch's categories).

Table 2. Excerpt from our database tagged for triggers of hate speech



D	E	F	G	H	I	J	K
Triggered by	Trigger of hate speech						Triggered comment N°
	TOPICS		RHETORICAL MODE		THREAT		
	Immigration			display of negative emotions			#9, 15, 38, 49, 51, 57, 71, 75
8	Immigration		history				#10
9						morality (dishonest)	
	Media &Public personae			sarcasm		morality (dishonest)	#12
11	Immigration			use of swear words	Social	morality (dishonest)	#13
12							
	Immigration		history				#16
8	Politics/Ideology					cognitive abilities (reasoning)	
14	Immigration		history	use of swear words	Social		#18
	Immigration		facts				#20
16	Politics/Ideology			personal attacks		cognitive abilities (reasoning)	#21

For the counter-narratives, where we also tagged the rhetorical modes and the tone, we included the impact of such counter-narratives and the tone. It must be noted that the same counter-speech strategy can elicit a wide range of reactions, from a positive tone to a very aggressive one and all shades in between. The tagging of the counter-narratives is still ongoing and here we discuss only the preliminary results.

3.2 General results

Erjavec and Kovačič (2012: 900) distinguished several categories of social actors who post hate speech. In our data we found three main groups: the "producers of hate speech (…) mostly rearticulating the meaning of news items"; "the watchdogs [who] are motivated by drawing attention to social injustice"; and, finally, "the players who intervene to have fun." All categories of participants make use of humor, whether they are initiating or responding to discriminatory statements, whether they are 'watchdogs' or 'players.' In fact, the main strategies used to execute and to counter hate speech are basically the same, but in reverse:

. Both sides quote facts from more or less credible sources,[9] one side to build resentment, anger and fear, the other to deconstruct, discredit and demystify offensive messages;

• Both sides use emotional appeals either to highlight the impact of hate speech on the targets of such speech, or to explain the reasons why the newcomers should be sent back (for example, with terrorist stories);

. Both sides make use of narratives—whether to spread positive stories, testimonies and messages from influencers, or for the opposite reason, to spread fear and anger using horrific testimonies to legitimate negative feelings and offensive comments;

• Both sides use satire and humor to undermine the claims made by one or the other side.

Our results are discussed in more detail in the next section.

4. Parallelism between Hate Speech and Counter-Speech Strategies

In this section, we analyse the main humorous strategies used to articulate hate, as well as those used to counter offensive speech. In an earlier study we worked with Dynel's (2017, inter alia) theory on irony to analyze irony in another set of data (Baider and Constantinou 2020), which also focused on migration and dates back to 2015. We use our findings from this earlier work, which integrated Chovanec's (2018) results on racist discourse and migration.

4.1 Echoing discourses (external or internal to the threads)

When using the strategy of echoing discourse, commentators reiterate the pro-refugee discourse / arguments within the online thread, although often these arguments are not directly stated in the thread / article, but can be very easily inferred, as in quotation (1) below. Indeed, the pro-refugee arguments are well known and seem to represent what is considered the social doxa, i.e., a discourse that can be neither contested nor questioned (Bourdieu 1977).

> (1). [commenting on a shared video]
> And this is happening all over Europe. But of course if you are a concerned citizen and even DARE to voice your concerns, *you are automatically branded a racist.*

The way in which the commentator describes someone who rejects migration ('a racist'), contrasted with the way he/she feels the same person would be regarded by someone pro-migrant ('a concerned citizen') reveals oppositional identifications and therefore might point

[9] Indeed, we found questionable sources or statistics given as arguments on both sides.

to sarcasm. However, the use of the the verb ‘dare’ definitely shows that the comment is sarcastic, while writing the word in capital letters serves to heighten the sarcasm, as it metaphorically implies the oppressive force of the social doxa, hence the victimization of the commentator. This victimization can have as its aim, and can also result in , the creation of an in-group feeling among all other ‘victims’ of the social doxa.

> (2).
>
> a. Europeans who immigrated during WW2 weren't bringing Islam with them and their stupid demands to accommodate their medieval views.
>
> b. What demands? Where are these demands which are to accommodate their views? I could have missed an article or two.

In quotation (2), the counter-speech (2b) uses the strategy of identifying a weakness in the argumentation (*where are these demands*) contained in the insulting rant, and echoes the argument (*their stupid demands*), ridiculing it. Indeed, the counter-speech not only questions the very fact that they exist, but at the same time it ‘makes fun’ -- as Gemmerli puts it --- of this statement (*I could have missed an article or two*). This opened the way for a change in the thread, and indeed, the dialogue then ended.

In quotation (3), sharing a link also helps to change the direction of the online exchanges. In this ‘echoing of discourse’ we include sharing links and videos since they contain related discourse:

> (3) [responding to the sharing of link]
>
> okay thank you. I will look at it. I'm not denying anything. I just want to keep an open mind and if you say something I will consider it.

In the next section we examine how cultural stereotypical scripts are used to disparage the opponent, both expressing and reinforcing an ideological square.

4.2 The use of cultural scripts to build the ideological square

The ideological square developed by Van Dijk (1993), but already present in substance in Allport (1954), describes in a simple way the discursive strategies that are deployed in racist exchanges. These strategies aim to create a polarization between the features of positive *self*-presentation and the features of negative *other*-presentation. This polarization entails emphasizing the good aspects of ‘Us’ (who have the ‘right’ beliefs, behaviors, values, etc.) and toning down or erasing completely the positive ‘Them’ (i.e., the positive aspects of their beliefs, behaviors, values); in parallel it stresses the negative ‘Them’ (who have the ‘wrong’ behaviors, beliefs, values, etc.) and downplays or erases the negatives aspects of ‘Us.’ Such semantic macro-strategies are favored for creating in-group cohesion and out-group aversion.

> (4) And the Sharia! And ISIS! And many mosques! And jihadists! And explosions! And mutilated hands and legs! Progress is finally here!

In quotation (4) above, the sarcasm lies in the semantic opposition between the terrorist actions carried out by ISIS and its militants (*ISIS, jihadists, explosions, mutilated hands and legs*) and progress. The final ‘punch’ is expressed with the phrase ‘*is finally here,’* which draws a direct cause and effect relationship between the arrival of migrants and the above-mentioned terrorist actions. According to Dynel’s (2017) theory, the phrase *Progress is finally here* is meant to convey the opposite sentiment; the statement reads as optimistic, but in fact, it expects the reader to understand the opposite, negative belief, i.e., ‘we are regressing, going back to the medieval ages, etc.’ The exclamation marks infer and underline this false enthusiasm. The socio-pragmatic function of sarcasm here can be interpreted as twofold. On the one hand, the out-group is defined as a physical threat to the local inhabitants; therefore, the in-group is the victim of such communities, and this activates a survival / defensive mode among the readers. On the other hand, the statement allows one to make a connection between mosques (where Muslims practice their religion) and the Other’s religion, and terrorism. The comment thus echoes the right-wing discourse, drawing identical parallels.

> (5) You only see what you want to see. Pretty much like an ostrich.

In quotation (5) the counter-speech responds to (2a), and explains that the extremely negative description of Islam originates from a prejudice (*You only* see *what you want to see*). The humor comes later, with the comparison to an ostrich, an animal which is said -- wrongly -- to bury its head in the sand when feeling threatened. With this comparison the commenter implies that the person is afraid of the newcomers and is therefore prejudiced in his evaluation of the situation, anger and fear being at the heart of hatred (Baider 2013; Sternberg and Sternberg 2008). Interestingly, animal metaphors are often used in such exchanges when referring to migrants (Musolff 2015; O'Brien 2003), while here they refer to the 'other side,' the 'haters' and activate a strategy of reverse rhetoric.

4.3 Rhetoric statements describing fictions

The use of fiction--in other words, invented stories-- grounded on biased representations is very common in humorous, ironic and sarcastic statements. Such statements are most often expressed within rhetorical questions such those below, excerpted from our data:

> (6). Does Syria own the BBC?
>
> (7). How many Mexicans were involved in destroying the World Trade Center or bombing the Boston Marathon?
>
> (8). Yes all of the Americans in the Civil War in the 1800's fled America too didn't they? Oh wait......

Quotes (6) and (7) are rhetorical questions that can only be answered in the negative. Quote (6) is an indirect criticism of the BBC, which the commentator suggests focuses only on Syrian refugees and implies the social doxa, evoked in (4), as having 'invaded and conquered' (does Syria *own*) the national media in the UK. Quote (7) indirectly suggests the dichotomy often used in anti-migration discourse between good (Mexican) and bad (Muslim) migrants (Kuisma 2013; Wyszynski et al. 2020), a distinction based primarily on ethnic origin, color or religion, i.e., a racist distinction.

In quote (8) the speaker begins positively and appears to agree with the statement that defended Syrian refugees –after all, they were forced to flee because of civil war, as were some Americans during the American Civil War. However, this statement is totally and obviously false. Because the statement is untrue, the whole sentence is likewise untrue, leaving readers to understand that the commentator is actually challenging the statement. Moreover, the '*oh wait*,' and the omissions marks (dot, dot, dot) underline the irony of the statement, just in case readers do not understood that it was an intended falsehood. This expression *oh wait* has become a cliché in such sarcastic statements, and is usually seen as a marker of irony or sarcasm.

We also encountered a (few) examples of humorous counter-speech, which make fun of the offensive commentator without being unpleasant. This type of comment tends to put an end to hostility, although it doesn't directly challenge the assumptions made via hate speech comments:

(9)

a. The BBC is so pro-Islam that it makes me sick.

b. Bananas are filled with potassium. Well, that is a start.

In quote (9), the commentator (9b) interprets the metaphor ‘makes me sick’ literally and offers medical advice (to eat bananas since potassium can alleviate nausea).[10] This strategy of making fun without being hurtful is helpful for releasing tension in such polarized debates. However poking fun at someone does not really respond to their complaint, nor does it take it seriously. While our example is acceptable since it does not hurt a specific community, it can also be seen as an easy way out, insofar as it does not address the core of the hateful message (fear, anger, resentment, etc.), and therefore, does not advance the dialogue.

Nevertheless, we do believe that using this kind of humour, even though it does not address the grievance of the commentator, can at least help to defuse the debate, and as a consequence of reducing tension, could foster dialogue.

5. Discussion

At the outset of this study, we questioned the effectiveness of the current EU hate laws and the Code of Conduct, and also asked whether simply erasing hateful content can make the racist judgments/feelings disappear. The data we examined confirms that there are many hate messages that *covertly* express hate. These generally contain dehumanizing metaphors, prejudices and stereotypes that are no less powerful—in both meaning and effect-- than direct calls for hatred and violence, but which cannot be erased because they fall outside the EU definition of hate speech and the guidelines of the Code of Conduct. But we also want to emphasize that, in any case, simple erasure does not solve the problem: we must understand the reasons for such verbal violence, and this area demands much more research. We also examined and analysed the responses and counter-narratives to hate speech, and here we are in agreement with Chung et al. (2019), who argued that using humor might be a better way to respond to a hateful outburst, as it appears to change the dialogue into a more positive or neutral

[10] https://www.everydayhealth.com/digestive-health/diet/foods-that-help-relieve-nausea/#:~:text=If%20your%20nausea%20is%20accompanied,%2C%22%20says%20Palinski%2DWade.

tone / rhetoric.[11]

5.1 Discriminatory speech as counter-speech

What we found in our data, however, was that most counter-speech messages were delivered in an aggressive tone, generally in the form of sarcastic remarks such as the example below:

(10)

a. Can we shut up about refugees already?

b. Are you human? When your ancestral lot thrust themselves at the mercy of foreign lands, they were welcomed. How can you be such a sociopath to feel comfortable enough to suggest that people in need should be shut out? You are sick. Can we just shut up about you?

Questioning a person's humanness, and calling that person a 'sociopath'—never mind that the claim may be valid --will not help the dialogue. Criticizing someone who wants us 'to shut up' about refugees, and then demanding that he/she do the same, is lowering oneself to the same level as the racist. This only serves to polarize the debate.

(11)

The real solution to the refugee crisis is --if you don't like refugees, then stop making them refugees by bombing their homelands !!! *Dimwits* (our italics).

In quotation (11) the argument is quite valid, i.e., bombing the homelands of refugees does not help the refugee crisis. It identifies the cause of the arrivals,--the bombing of the Middle East, which is the fault of Europe, among other countries. However, calling someone a *dimwit* destroys the effectiveness of the argumentation.

It seems clear that sarcasm does not work as a counter-strategy: insulting people will not bring them over to your side. However, sarcasm is as pervasive in counter-narratives as it is in hate speech, a fact that is indicative of the polarization of the participants.

If *ad hominem* attacks such as those above serve to increase the rift between the two sides, they also polarize the debate between You (evil and immoral) and Me (good and moral)—Me taking the moral high ground and holding You in contempt. Another broad observation about such counter-narratives is that they work on logic and downplay any grievance.

By overlooking the social issues expressed in such outbursts, and ignoring of the roots of the fear, resentment or anger embedded in the discriminatory comments, the verbal violence is left

[11] The use of images may be helpful in some instances, but we did not observe any such sharing in our data, except sharing links to upsetting videos.

to spiral out of control. That process is neither stopped nor diverted—in fact, it may be reinforced. This is why Gemmerli (2015) concluded that counter-narratives are only a short-term solution: “This approach *attempts to affect the behaviour* of those who sympathise with or take part in violent extremism *in the short term*” (our italics).

5.2 Grievances should be acknowledged

Despite their shortcomings, we nevertheless argue that counter-narratives are important. Such narratives on the Internet can encourage the passive reader to condemn hateful comments, while they can also help to trigger positive feelings (such as empathy) for victims of hate speech.

Indeed, Braddock and Dillard (2016) performed a series of meta-analyses related to narratives and persuasion over a 30-year period, and concluded that “exposure to a narrative is *positively* related to the adoption of narrative-consistent viewpoints” (our italics).

Speckhard and Shajkovci (2018) have advised more research into strategies called *alternative narratives,* i.e., positive counter-speech, which aims to deflect the tension while also avoiding polarization. This kind of strategy has been described as more effective than the aggressive tone of counter-speech we found in our database. Alternative narratives would focus on an acknowledgment of the grievances and would listen to the feelings that motivate such comments. They would further acknowledge the feelings of the victims of hate speech and the consequences they suffer, rather than presenting direct counter-argumentation *per se* (Speckhard and Shajkovci 2018). We have seen that engaging with hateful narratives most often tends to reinforce or even validate them. Nevertheless, and although we believe that more research should look at the efficacy of such alternative strategies, we argue that such narratives work, even if only in the short term, since even limited effectiveness is important as it can help prevent an *escalation* of violence (Baider 2020; Baider and Borbori 2020).

Conclusion

In sum, our study found that most hateful comments do not call for violence explicitly, as does overt hate speech. They can be categorized as ‘polarizing’ strategies, disclosing an ‘us vs them’ mentality that might easily lead to hatred. Most counter-narratives tend to either reinforce this polarization, or create new a polarization between the two commentators (the good person vs the evil one). In our study of the role of humor in both (covert) hate speech and in counter-speech, we found that hate speech is not a simple outburst of verbal violence; it must be understood as the result of a cognitive and affective process. Considerable research in sociology and psychology has investigated the stages that lead people to violence against a community. Whether overt or covert, hate speech can be analyzed as part of a social process of dehumanization, which occurs in stages and is animated by intermediate feelings (with disgust and contempt at the heart of this process), both expressed by different speech acts (Baider 2019, 2020). Therefore, on a broader societal level, while it is important to answer hateful messages, if the social reasons underlying the violent comment are ignored, this intervention can only be a short-term solution. Fairclough, adopting a Foucauldian perspective on discourse, advised that the text studied should always be anchored within the macro context. We found that reframing hate speech within the macro context is quite rare in the field of hate studies. And while in our first section we dealt exclusively with the huge influx in migration wave in 2015 in Europe, we would suggest that this is only the *meso* context of the data. The *macro* context would involve analysis of the social reasons underlying the use of irony or sarcasm (Jobert and Sorlin 2018: 3). For that matter, it seems that investigating, understanding and addressing these social reasons are key to countering hate speech. Indeed, as Bouvier (2020) pointed out, racist outbursts on the part of ordinary citizens must be neither individualised nor decontextualized. She observed that, in the name of social and political justice, many Facebook posts demonize the hate-filled person by "posturing and showing their own relative morality." Often, they also miss the point by asking: how can ordinary-- and perhaps otherwise apparently decent people-- hold such radical views? Many earlier studies devoted to online communication have underlined that these outbursts may be due to the spontaneity and speed of the medium, the deindividuation effect of having pixels and not a human being in front of you, the self-promotion typical of such posts with the aim of attracting likes, shares and clicks (cf. our footnote 1). However, as Bouvier (2020) aptly remarks, it may also well be the response "to the affective flows of emotion around simplified narratives and symbols." This 'affective flow' is exacerbated if the comment is addressed to the personality of the commentator (e.g., labelling the person as racist or hateful) rather than identifying the remark as racist or hateful, which is what our data concluded is the function /impact of sarcastic remarks. We can only conclude that further research focused on the impact of counter-speech (and alternative narratives) is required .

References

Allport, Gordon Walter. 1954. *The Nature of Prejudice.* Cambridge, MA: Addison-Wesley.

Archakis, Argiris & Villy Tsakona. 2005. Analyzing conversational data in GTVH terms: A new approach to the issue of identity construction via humor. *Humor: International Journal of Humor Research* 18 (1). 41–68.

Archakis, Argiris & Villy Tsakona. 2019. Racism in recent Greek migrant jokes. *Humor* 32 (2). 267-287.

Attardo, Salvatore. 2000. Irony as relevant inappropriateness. *Journal of pragmatics* 32(6). 793-826.

Attardo, Salvatore, Jodi Eisterhold, Jennifer Hay & Isabella Poggi. 2003. Multimodal markers of irony and sarcasm. *Humor* 16 (2). 243-260.

Baider, Fabienne. 2019. Double Speech Act: Negotiating Inter-cultural Beliefs and Intra-cultural Hate Speech among the Youth, *Journal of Pragmatics* 151. 155-166.

Baider, Fabienne. 2020. Pragmatics lost? Overview, synthesis and proposition in defining online hate speech. *Pragmatics and Society* 11 (2). 196-218

Baider, Fabienne. 2013. Hate: Saliency features in cross-cultural semantics. In Istvan Kecskes & Jesus Romero-Trillo (eds), *Linguistic Aspects of Intercultural Pragmatics*, 7-27. Berlin, Boston: de Gruyter Mouton.

Baider, Fabienne & Maria Constantinou. 2020. Covert hate speech: A contrastive study of Greek and Greek Cypriot online discussions with an emphasis on irony. *Journal of Language Aggression and Conflict* 8 (2). 262-287.

Baider, Fabienne & Anna Bobori. 2020. Mitigating the frame SEXUAL THREAT in anti-migration discourse online. In Darja Fisher and Philippa Smith (eds), *The Dark Side of Digital Platforms: Linguistic Investigations of Socially Unacceptable Online Discourse Practices*, 86-113. Ljubljana: Ljubljana University Press.

Bartlett, Jamie, Jeremy Reffin, Noelle Rumball & Sarah Williamson. 2014. Antisocial Media. *DEMOS*, 1-51

Benesch, Susan. 2014. Countering dangerous speech: New ideas for genocide prevention. *Washington, DC: United States Holocaust Memorial Museum*. Retrieved from http://www.ushmm.org/m/pdfs/20140212-benesch -countering-dangerous-speech.pdf

Billig, Michael. 2005. *Laughter and Ridicule: Towards a Social Critique of Humour*. London: Sage Publications.

Binny, Mathew, Navish Kumar, Ravina, Pawan Goyal & Animesh Mukherjee. 2018. Analyzing the hate and counter speech accounts on Twitter. *arXiv preprint arXiv:1812.02712*.

Bourdieu, Pierre. 1977. The economics of linguistic exchanges. *Social Science Information* 16 (6). 645-668.

Bouvier, Gwen. 2020. Racist call-outs and cancel culture on Twitter: The limitations of the platform's ability to define issues of social justice. *Discourse, Context & Media* 38. 100431.

Braddock, Kurt & John Horgan. 2016. Towards a guide for constructing and disseminating counternarratives to reduce support for terrorism. *Studies in Conflict & Terrorism* 39 (5). 381-404.

Chakraborti, Neil. 2015. Re-thinking hate crime: Fresh challenges for policy and practice. *Journal of Interpersonal Violence* 30 (10). 1738-1754.

Charteris-Black, Jonathan. 2006. Britain as a container: Immigration metaphors in the 2005 election campaign. *Discourse & Society* 17 (5). 563-581.

Chovanec, Jan. 2018. Irony as counter positioning. In Manuel Jobert & Sandrine Sorlin (eds), *The Pragmatics of Irony and Banter*, 165-194. Amsterdam: John Benjamins.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, & Marco Guerini. 2019. CONAN--COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech. *Proceedings of ACL 2019*, 2819-2829. arXiv preprint arXiv:1910.03270.

Citron, Danielle. 2014. *Hate Crimes in Cyberspace*. Cambridge, MA: Harvard University Press.

Culpeper, Jonathan. 2011. Politeness and impoliteness. In Karin Aijmer & Gisle Andersen (eds), *Pragmatics of Society. Handbooks of Pragmatics*, 391-436. Berlin: de Gruyter Mouton.

Culpeper, Jonathan. 2021. Impoliteness and hate speech: Compare and contrast. *Journal of Pragmatics* 179. 1-11 https://doi.org/10.1016/j.pragma.2021.04.019

Dixon Lucas, John Li, Nithum Thain, Jeffrey Sorensen & Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 67-73.
https://www.aies-conference.com/2018/contents/papers/main/AIES_2018_paper_9.pdf.

Dovidio John F., Miles Hewstone, Peter Glick & Victoria M. Esses. 2010. Prejudice, stereotyping, and discrimination: Theoretical and empirical overview. In *The SAGE Handbook of Prejudice, Stereotyping and Discrimination*, 1-26. Sage Publications.

Dynel, Marta. 2017. The irony of irony: Irony based on truthfulness. *Corpus Pragmatics* 1. 3–36.

Erjavec, Karmen & Melita Poler Kovačič. 2012. "You Don't Understand, This Is A New War!" Analysis of Hate Speech in News Web Sites' Comments. *Mass Communication and Society* 15 (6). 899-920. doi:10.1080/15205436.2011.619679.

Fairclough, Norman. 1995. *Critical discourse analysis. The critical study of language*. London: Longman.

Gemmerli, Tobias. 2015. *Avoid the pitfalls of counter narratives*. Danish Institute for International Studies, accessed 14 March 2020, https://www.diis.dk/node/6900;

Grice Paul H. 1975. Logic and conversation. In Peter Cole, & Jerry L. Morgan (eds.), *Syntax and Semantics, Vol. 3, Speech Acts,* 41-58. New York: Academic Press.

Hamilton, David L. & Tina K. Trolier. 1986. Stereotypes and stereotyping: An overview of the cognitive approach. In John F. Dovidio & Samuel L. Gaertner (eds), *Prejudice, Discrimination, and Racism,* 127–163. Academic Press.

Hardacker, Claire. 2010. Trolling in asynchronous computer mediated communication: From user discussions to academic definitions. *Journal of Politeness Research* 6 (2). 215-242.

Hakoköngäs Eemeli, Otto Halmesvaara, & Inari Sakki. 2020. Persuasion Through Bitter Humor: Multimodal Discourse Analysis of Rhetoric in Internet Memes of Two Far-Right Groups in Finland. *Social Media + Society,* 1–11.

Herring, Susan, Kirk Job-Sluder, Rebecca Scheckler & Sasha Barab. 2002. Searching for Safety Online: Managing 'Trolling' in a Feminist Forum. *Information Society* 18. 371-384.

Hill, Jane. 2008. *The Everyday Language of White Racism*. Malden, MA: Wiley-Blackwell.

Kuisma, Mikko. 2013. "Good" and "bad" immigrants: The economic nationalism of the true Finns' immigration discourse. In Umut Korkut, Gregg Bucken-Knapp, Aidan McGarry, Jonas Hinnfors, & Helen Drake (eds) *The Discourses and Politics of Migration in Europe,* 93-108. New York: Palgrave Macmillan.

Kumar, Ritesh, Atul Kr Ojha, S. Malmasi & Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018),* 1-11.

Lippmann, Walter. 1922. *Public Opinion.* New-York: Harcourt, Brace and co.

Ljubešić Nikola, Darja Fišer & Tomaž Erjavec. 2019. The FRENK Datasets of Socially Unacceptable Discourse in Slovene and English. *Text, Speech, and Dialogue 2019,*103–119. DOI:10.1007/978-3-030-27947-9_9

Musolff, Andreas. 2015. Dehumanizing metaphors in UK immigrant debates in press and online media. *Journal of Language Aggression and Conflict* 3 (1). 41-56.

O'Brien, Gerald V. 2003. Indigestible food, conquering hordes, and waste materials: Metaphors of immigrants and the early immigration restriction debate in the United States. *Metaphor and Symbol* 18 (1). 33-47.

Sap, Maarten, Dallas Card, Saadia Gabriel, Yejin Choi, & Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1668-1678.

Shelley 2004

Silva, Leandro, Mainack Mondal, Denzil Correa, Fabricio Benevenuto & Ingmar Weber. 2016. Analyzing the targets of hate in online social media. In *Proceedings of the International AAAI Conference on Web and Social Media* 10, 1.

Speckhard, Anne & Ardian Shajkovci. 2018. Challenges in Creating, Deploying Counter-Narratives to Deter Would-be Terrorists. *ICSVE*.

Sternberg, Robert J. & Karin Sternberg. 2008. *The Nature of Hate*. Cambridge: Cambridge University Press.

Strossen, Nadine. 2018. *Hate: Why we should resist it with free speech, not censorship*. New York: Oxford University Press.

Tuck, Henry & Tania Silverman. 2016. The counter-narrative handbook. *Institute for Strategic Dialogue*, https://www.isdglobal.org/isd-publications/the-counter-narrative-handbook/

Van Dijk, Teun. 1993. *Elite Discourse and Racism*. Sage Publications.

Warrington, Ann. 2017. Countering violent extremism via de-securitisation on Twitter. *Journal of Deradicalization* 11. 258-280.

Weaver, Simon. 2011. Liquid racism and the ambiguity of Ali G. *European Journal of Cultural Studies* 14 (3). 249–264.

Weaver, Simon. 2016. *The Rhetoric of Racist Humour: US, UK and Global Race Joking*. London: Routledge.

Wodak, Ruth. 2015. *The Politics of Fear: What right-wing populist discourses mean*. Sage Publications.

Wyszynski, Mia Caroline, Rita Guerra, & Kinga Bierwiaczonek. 2020. Good refugees, bad migrants? Intergroup helping orientations toward refugees, migrants, and economic migrants in Germany. *Journal of Applied Social Psychology* 50 (10). 607-618.

Xia Mengzhou, Anjalie Field & Yulia Tsvetkov. 2020. Demoting Racial Bias in Hate Speech Detection. *Association for Computational Linguistics*. 7–14 DOI 10.18653/v1/2020.socialnlp-1.2 .