Politics and Governance (ISSN: 2183–2463)
2023, Volume 11, Issue 2, Pages 249–260
https://doi.org/10.17645/pag.v11i2.6465

Article

Accountability Issues, Online Covert Hate Speech, and the Efficacy of Counter-Speech

Fabienne Baider

Department of French and European Studies, University of Cyprus, Cyprus; fabienne@ucy.ac.cy

Submitted: 8 October 2022 | Accepted: 4 January 2023 | Published: 17 May 2023

Abstract

Concerning individual or institutional accountability for online hate speech, research has revealed that most such speech is covert (veiled or camouflaged expressions of hate) and cannot be addressed with existing measures (e.g., deletion of messages, prosecution of the perpetrator). Therefore, in this article, we examine another way to respond to and possibly deflect hate speech: counter-speech. Counter-narratives aim to influence those who write hate speech, to encourage them to rethink their message, and to offer to all who read hate speech a critical deconstruction of it. We created a unique set of parameters to analise the strategies used in counter-speech and their impact. Upon analysis of our database (manual annotations of 15,000 Twitter and YouTube comments), we identified the rhetoric most used in counter-speech, the general impact of the various counter-narrative strategies, and their specific impact concerning several topics. The impact was defined by noting the number of answers triggered by the comment and the tone of the answers (negative, positive, or neutral). Our data reveal an overwhelming use of argumentative strategies in counter-speech, most involving reasoning, history, statistics, and examples. However, most of these argumentative strategies are written in a hostile tone and most dialogues triggered are negative. We also found that affective strategies (based on displaying positive emotions, for instance) led to a positive outcome, although in most cases these narratives do not receive responses. We recommend that education or training—even machine learning such as empathetic bots—should focus on strategies that are positive in tone, acknowledging grievances especially.

Keywords

accountability; argumentative strategies; counter-speech; covert hate speech; emotional appeal

Issue

This article is part of the issue "Hate Speech, Demonization, Polarization, and Political Social Responsibility" edited by Luis M. Romero-Rodríguez (Rey Juan Carlos University), Pedro Cuesta-Valiño (University of Alcalá), and Bárbara Castillo-Abdul (Rey Juan Carlos University).

1. Introduction

Concerning individual or institutional accountability for online hate speech, we argue that, because most online hate speech is covert, current measures regulating hate speech are insufficient. For example, the 2016 Code of Conduct from the EU Commission falls short in several regards (Konikoff, 2021), including concerns about the qualifications of those deleting hate messages and the fact that artificial intelligence models used to detect hate speech are 1.5 times more likely to flag tweets written by specific communities (Silva et al., 2016). Covert hate speech entails even more problems, as it uses implicit meaning and indirect discursive strategies to express hatred, including derogative metaphors (Musolff, 2015), inferences (Baider, 2022), and humor (Weaver, 2016). Such covert expressions fall outside the legal definitions of hate speech, and thus purveyors of such hateful speech remain unaccountable before the law. Examination of how hate speech is regulated by social media platforms such as Facebook and YouTube (Fortuna & Nunes, 2018; Hietanen & Eddebo, 2022) underscores the diverse interpretations of what constitutes hate speech. For this reason, we argue for greater emphasis on counter-speech rather than censorship (Strossen, 2018) as the best way to deflect or halt hate speech.

We suggest the use of counter-narratives, which we define as any form of expression that aims to influence those who sympathise with or take part in abusive speech. These narratives can encourage those who write hate speech to rethink their message, while at the same time, they offer a critical counter-argument to all who read the hate speech. They also offer another point of view and can potentially trigger positive feelings for victims of discriminatory narratives. The present study discusses the most frequent types of counter-narratives and their impact, based on analysis of our database of manual annotations of 15,000 Twitter and YouTube comments (collected within the IMsyPP EU program).

2. Addressing Online Hate Speech: Censorship vs. Dialogue

Since the 1990s, research targeting hate speech has noted the prevalence of hostile and aggressive content in online platforms, which might suggest that the medium itself is partially to blame—insofar as it offers anonymity, instantiation of communication, depersonalisation, deindividuation, etc. (cf. Baider, 2020). In fact, Wodak (2015, p. 207, emphasis added) concluded that "the more anonymous the genre, the more *explicit exclusionary rhetoric* tends to be."

2.1. Overt and Covert Hate Speech

Indeed, many of the advantages of digitisation, e.g., connectivity, access to new knowledge, and the creation of new relationships, have led to the rapid rise in cyber hate across the internet. As recently as 20 years ago, social media platforms, online fora, and group discussions were found to be prime locations for the collection and analysis of (violent) discriminatory discourse (Herring et al., 2002, p. 371). Such discourses manifest in various ways, including Twitter mobbing, trolling, cyberbullying, and sexting—all of which may fall under the umbrella term "hate speech" depending on the definition applied.

And herein lies part of the problem: the many, often contradictory, official definitions of hate speech. As early as 1965, the United Nations General Assembly Resolution 2106, in their International Convention on the Elimination of All Forms of Racial Discrimination, defined hate speech as "the promotion of racial hatred and discrimination" based on "ideas or theories of superiority of one race or group of persons of one colour or ethnic origin," and as speech that would incite "racial discrimination, or acts of violence…against any race or group of persons of another colour or ethnic origin" (United Nations, 1965, p. 3). The International Covenant on Civil and Political Rights, adopted by the United Nations General Assembly Resolution 2200A (XXI) and commonly used in court cases, defines hate speech as an "any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence" (United Nations, 1966, article 20). Most research studies are based on the broad definition suggested by the Office for Democratic Institutions and Human Rights (2009, pp. 37–46), which is the principal institution of the Organisation for Security and Cooperation in Europe dealing with the "human dimension" of security: "The expression of hatred towards an individual or group of individuals on the basis of protected characteristics, where the term 'protected characteristics' denotes a member of some specific social group that could, on its own, trigger discrimination.''

In EU countries, all judgments and social network regulations are based on the 2008 European Union Framework Decision, which delineates hate speech as statements "publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin" and "publicly condoning, denying or grossly trivialising crimes of genocide, crimes against humanity and war crimes" (Council of the European Union, 2008, Article 1).

Further, there are some scholars (Gelber, 2019) who define hate speech as speech that causes harm to a degree sufficient to warrant government regulation, reasoning also followed by some legal judgments; for example, the case of a British National Party poster bearing the message "Islam out—Protect the British people," accompanied by a photo of the World Trade Center towers in flames and a sign with the Islamic crescent barred, which was legally judged as hate speech and banned (*Norwood v. the United Kingdom*, 2004). Another such example is *Leroy v. France* (2008), where Leroy was convicted of publicly condoning terrorism with his 2001 cartoon published in a Basque newspaper. The cartoon represented the attack on the World Trade Center with the caption: "We all dreamt of it….Hamas did it." Because the humorous use of a well-known catchphrase to express criticism against the United States was also glorifying extreme violence and death, the locutionary act was judged condemnable. However, the likelihood of causing harm, i.e., the perlocutionary effect, was also decisive in Leroy's condemnation since the Basque country is politically sensitive to terrorism. Thus, we can see that to determine whether the speech qualifies as hate speech, the courts must consider the social, historical, and cultural context, and "under what circumstances targets are vulnerable to harm" (*Leroy v. France*, 2008)—a definition that is very similar to that of the 2012 Rabat Plan of Action (United Nations, 2012).

Considering the various legal definitions, therefore, it is clear that covert hate speech is difficult to moderate and regulate, even though this disguised means of expressing hatred or calls for violence has sometimes been successfully addressed under present legislation (cf. *Norwood v. the United Kingdom* and *Leroy v. France*, among many). Indeed, it is not difficult to replace explicit stereotypes, which could be prosecuted, with implicit ones, and thus communicate hateful comments through other means. These covert means use disguised ways to express racism, sexism, homophobia, or any bias against a specific community that could incite violence (Baider, 2020; Ben-David & Fernández, 2016; Matsuda, 1989). They are based on the same stereotypes and harmful prejudices as found in overt hate speech but use indirect strategies to express hateful sentiments, and/or a very negative stance towards specific communities. These covert strategies include metaphors (Musolff, 2015), sarcastic remarks or humour (Hill, 2008; Weaver, 2016), conspiracy theories (Baider, 2022), dog-whistling strategies (a strategy that refers to the use of words, phrases, and terminology that mean one thing to the public at large, but that carry an additional, implicit meaning only recognised by a specific subset of the audience; Bhat & Klein, 2020, p. 168), and memes (Askanius, 2021); even absence or silence can be used to invite hateful inferences (Hill, 2008, p. 41). They function as "Othering" mechanisms that breed anger, disgust, contempt, and fear towards a specific community—all emotions that are core to hatred, which in turn is core to extremism. It is extremely difficult to legally address these indirect expressions of hate speech; often speakers can escape accountability by pleading an excuse such as "I did not mean it" in cases of sarcasm for instance, or by using hedges such as "no offence, but" when uttering an insulting remark, etc. This type of hate speech is the most common form of racism, sexism, homophobia, etc., found on social media, supplanting overt hate speech by a huge margin (Bhat & Klein, 2020). In our data, less than 10% of comments are overt hate speech.

If the last ten years have seen an ever-growing dependency on automatic detection mechanisms to identify hate speech (Fortuna & Nunes, 2018), it is clear that online participants who want to express extremely negative attitudes have found solutions to circumvent censorship.

2.2. Hate Speech Management

Hate speech is typically managed by one or more of four responses (Benesch et al., 2016; Citron & Norton, 2011): deletion or suspension, inaction, education, and counter-speech. The response of inaction implies not responding to the abusive message and can lead to one of two consequences. On the one hand, ignoring a hate-filled message can lessen its impact, as it neither encourages nor feeds the debate and its (possible) ensuing thread. On the other hand, it can imply that such speech is acceptable. The education response involves training, media literacy, and national or international campaigns to inform the public, especially the youth, about hate speech, its consequences, and the best ways to address the messages. While this solution is important, it is more of a long-term investment.

A more immediate measure would be to increase accountability within computer meditated communication. Brown (2020, p. 32) has argued for different levels of governance regarding hate speech: the moderation level, which would primarily concern social media companies; and the regulatory level, which would concern agencies. The regulatory level is typically assumed by governments or their agencies, which guide internet governance and enact legislation. In fact, there are already a number of legal measures that enforce some degree of communication etiquette in computer meditated communication—e.g., international instruments such as the 2019 UN Action Plan on hate speech and the 2016 Code of Conduct enacted by the EU Commission, the latter being the most drastic measure, which is also closely monitored (Pingen, 2021). This EU Code of Conduct mandates that social media companies remove or disable access to—and within 24 hours—what has been deemed illegal hate speech on the basis of the 2007 Framework definition.

According to Brown (2020, p. 32), responsibility for the second type of governance, moderation, is to be assumed by internet platforms and ordinary citizens. Social platforms have therefore delineated their own limitations on freedom of speech. As an example, Facebook has a complex set of rules determining what constitutes hate speech, and even considers covert hate speech in some cases. They will take down straightforward animal metaphors such as "migrants are filthy cockroaches" and well-known wordplay such as "refugees and rape-fugees." However, they will differentiate between "migrants are so filthy" (non-violating—ignore) and "all English people are dirty" (violating—delete), or "fucking migrants" (non-violating—ignore) and "fucking Muslims" (violating—delete). To understand their evaluation of the statements, we have to bear in mind that:

1. A statement such as "migrants are filth" is deleted since the metaphor "migrants are DIRT" is an established metaphor in racist discourse. This example reveals that covert hate speech (here, a metaphor), even if not always identified as such by Facebook, is nevertheless covered in its anti-hate speech rules.
2. The statement "all English people are dirty" is deleted because condemning people based on their nationality violates hate speech laws, a rule consistent with the Council of Europe definition specifying nationality as a criterion that warrants the label of hate speech.

Therefore, while it is the task of artificial intelligence mechanisms to detect hate speech, these mechanisms follow the social media regulators' understanding and definition of hate speech. This raises both questions and concerns over the legitimacy of anyone other than trained lawyers deciding what is hate speech. Indeed, to evaluate what qualifies as such is hard enough for human beings, never mind artificial intelligence systems. Moreover, it appears that such rules are devised by socially homogeneous teams (Baider, 2020) since artificial intelligence models have been found to flag tweets written by African Americans as offensive 1.5 times more often—in other words, a false positive—than tweets written by other communities (Sap et al., 2019). This is explained by the over-sensitivity of hate speech classifiers, e.g., “nigger” or “bitch,” which do not signal hate speech when used in specific settings; rather, they can signify relational proximity in some communities of practice (Baider, 2020; Culpeper, 2021).

However, although deleting a message limits its spread on a specific network, it encourages the author to post the same message on another, less-censored network. It does not challenge the arguments or the resentment expressed in such messages.

Moreover, critics argue that blocking free speech is a dangerous precedent and that these measures may restrict freedom of expression, which is recognised by the European Court of Human Rights as a fundamental human right and a basic condition of democratic societies, as well as necessary for individual development. Indeed, the danger with such laws is that they could be used to curb dissent and pursue “persecution of minorities under the guise of anti-incitement laws” (United Nations, 2012); for example, blasphemy laws can threaten inter-religious dialogue and ban legitimate debate. Clearly, transnational regulation of hate speech is not an easy task (Burnap & Matthew, 2015).

Most relevant to our study is the first solution, deletion or suspension, which we argue cannot address covert hate speech. It does not respond to the need for an immediate answer to the millions of messages exchanged every day.

We suggest that counter-speech would be the best solution, since debate “is nearly always preferable to censorship and removal of content, including when dealing with extreme or radical content, whatever its origin” (Bartlett & Krasodomski-Jones, 2016, p. 5). In fact, over 20 years ago Richards and Calvert (2000) argued that the best way to combat hate speech would be to add more speech, i.e., to use counter-speech to tackle hate speech. Indeed, the advantages of counter-speech are far greater than those of deletion—one example is using specific argumentation to respond in particular social contexts, hopefully destabilising the presupposition on which hate speech is based (McGowan, 2009). While counter-speech should respect freedom of speech, we found in our data that it is often violent, ultimately defeating the purpose of halting spiraling violence. A message that responds to the stereotyping that is core to hate speech offers readers another point of view and a chance to “take back” cyberspace. In fact, it may be our responsibility as forum participants to address such issues rather than let them pass unattended, as the use of automatic bots to respond to hate speech is a very recent solution to counter the massive number of message exchanges (Ashida & Komachi, 2022).

Counter-speech is not an entirely new subject, and to date, there are a number of studies examining its effect on hate speech. The most important studies will be discussed in the next section.

2.3. Counter-Speech: Definition and Impact

Definitions of counter-speech vary; for example, Bartlett and Krasodomski-Jones (2016, p. 5) propose “a crowd-sourced response to hateful messages,” which means a direct response to harmful speech using any form of expression, whether a text, a meme, a hyperlink, etc. The researchers focus on argumentation based on logic or affect, whereas counter-speech aims to deconstruct hate speech and weaken its impact. Other scholars posit that counter-speech can also take the form of an alternative narrative (Braddock & Horgan, 2016; Briggs & Feve, 2013). Alternative narratives make a deliberate choice to change the narrative, focusing on positive stories to promote tolerance and debunk the presupposition on which the hate speech is based. In any case, in this article, we use counter-speech as a hypernym that includes alternative narratives and counter-narratives.

Counter-narratives should attempt to affect the behavior and the thinking of those who sympathise with or take part in spreading prejudices. Most important, they should foster critical thinking, tackle the source of prejudice (McGowan, 2009), and provoke reactions (i.e., spark a dialogue even if it is fierce; Gemmerli, 2015; Silverman et al., 2016). At the same time, they should also point out the complexity of the issue, and facilitate exposure to alternative viewpoints (Bartlett & Krasodomski-Jones, 2016); they should encourage readers to condemn hateful comments, trigger positive feelings (such as empathy) for victims of discriminatory narratives, and/or trigger some doubt that could lead to a change in attitudes (Gemmerli, 2015; Silverman et al., 2016). While arguments exchanged between strangers may lead to a favourable change in discourse, this is very rare (Bartlett & Krasodomski-Jones, 2016; Benesch et al., 2016; Ernst et al., 2017; Konikoff, 2021; Schieb & Preuss, 2016; Wright et al., 2017). Most research is based on small experiments such as Munger’s (2017), which attest to the power of in-group norms and the need to tackle this phenomenon if we want to reduce racism. The studies above argue that the most effective messages do not lecture the audience, rather, they must offer something to think about and reflect on (Braddock & Horgan, 2016; Gagliardone et al., 2015).

For that matter, Benesch and his colleagues, who define hate speech as “dangerous speech” (see Benesch, 2014; Benesch et al., 2016) were the first to suggest a series of strategies for writing counter-speech and reducing the impact of hateful comments: (a) present the facts in order to correct misstatements or misperceptions; (b) point out hypocrisy or contradictions to discredit the accuser; (c) warn of offline or online consequences of such action; (d) claim some affiliation to give weight to the counter-speech; (e) denounce the speech as hateful; (f) use humour and sarcasm to deescalate conflict and encourage social cohesion; (g) adopt a positive tone to appeal to the other participants; (h) adopt hostile language to potentially persuade a participant to delete their message.

The few large-scale research projects that have focused on counter-narratives have used these parameters. For example, they were the basis for the Conan project in which Chung et al. (2019) created the first large-scale, multilingual dataset of hate speech and counter-narrative pairs, i.e., type of hate speech vs. type of counter-speech.

However, hate speech is foremost a type of argumentation, i.e., an attempt to persuade others that a specific community or individual is a danger to them; Stephan et al. (1999) examined hate-filled comments from a psychological perspective and concluded that the concept of "threat" was core to hate speech, especially racism.

Therefore, when writing a counter-narrative, it is important to first identify the strategies of argumentation and determine if they are dependent on the topic, and then identify their impact and focus on deconstructing the presuppositions of their comments. To determine the most effective type of counter-speech we will address two research questions that have not yet been answered with big data:

> RQ1: What counter-speech argumentative strategies are used in situ on a large-scale basis?
>
> RQ2: What is the impact of each of these different strategies in relation to the different identified topics?

3. Data, Methodology, and Results

To answer our research questions, we begin with a quantitative approach, drawing statistics from data that has been annotated. We focused on strategies of argumentation used in counter-speech to understand which strategies are widely used (RQ1) and the impact of these strategies (RQ2). Before presenting our results, we describe the data of our corpus and the annotation scheme we developed for the project.

3.1. Data and Teamwork

We worked with data available within the IMsyPP EU project (2020–2022): 15,000 annotated Facebook posts and YouTube comments focused on several topics known to trigger hate speech, i.e., migration, politics, and LGBTQ issues. The category "politics" is an umbrella term covering a variety of political topics (e.g., India helping Pakistan during the Covid-19 pandemic). The comments and posts referring to migration were collected from Facebook in 2015, when an unprecedented number of migrants flooded into Europe, while the comments related to LGBTQ and political issues were collected in 2020 from YouTube. Each comment was first annotated for triggering hate speech and offensive speech, whether it was covert or overt, resulting in 9,700 comments annotated by eight annotators working in pairs. All comments were then tagged twice for counter-speech and assessed for impact; ultimately, the idea was to offer recommendations. The datasets are all in English and are comparable, insofar as they have a similar number of comments. The datasets were all annotated by the same team for a period of one year (see Table 1 for a summary of the data).

3.2. Methodology

We had to first decide on a set of parameters to annotate the counter-narratives, so we turned to earlier research studies, notably Benesch (2014) and Benesch et al. (2016), whose parameters categorising counter-speech have been widely used (see, e.g., Braddock & Horgan, 2016; Chung et al., 2019; Tuck & Silverman, 2016). As noted earlier, these parameters are: presenting facts, pointing out hypocrisy, warning of consequences, claiming some affiliation, denouncing the speech as hateful, using humor and sarcasm, adopting a positive tone, and adopting hostile language. We also added using multimedia, as Benesch et al. (2016) advised.

As we noted above, counter-narratives must be tested and evaluated in terms of their strategies as well as their impact—for example, a measurable change in behavior. Therefore, we created a category titled impact, wherein we took note of the number of answers triggered by the comment and the tone of the answers (whether negative, positive, or neutral).

We next ran a two-week pilot study to test these criteria, which ultimately resulted in the creation of our own set of annotations. Our pilot study revealed several shortcomings in the criteria, as follows:

1. Some of Benesch's criteria were absent from our annotations, e.g., warning of consequences,

Table 1. Datasets used for annotations.

Data				
Number of comments	Number of annotated comments	Source of dataset	Topic	Language
5,873	3,700	Facebook	Migration	English
3,009	3,000	YouTube	Politics	English
5,979	3,000	YouTube	LGBTQ issues	English
Total: 14,861	Total: 9,700			

claiming some affiliation, and pointing out hypocrisy, or too difficult to be distinguished from other choices;

2. We found other elements being used, such as "acknowledging grievances," that were not present in Benesch's criteria;
3. The criteria "presenting facts" was found to be too broad, since merely offering data may not be sufficiently convincing. It does, however, show the audience that the accusations are not substantiated, so we subdivided the category by adding "using statistics," "using history," and "using examples or testimonies";
4. We found that emotional appeal should be annotated in its own right; it is a subdivision of classic rhetoric as is argumentation;
5. We included conspiracy theories, which came up as a variable in migration data and then in political debates.

Thus, we decided it was necessary to review criteria used in several other domains known for their work on counter-speech, i.e., the fields of data mining (for example, Fortuna & Nunes, 2018), psychology (Stephan, 1999; Stephan et al., 1999), discourse analysis and rhetoric (Baider, 2019, 2020; Wodak, 2015).

Eventually, we determined a new set of criteria, which we used to work on another set of annotations. The criteria now included:

1. The specific topic, since we had three data sets;
2. A rhetoric category, subdivided into argumentation (and further divided into logic or reasoning, statistics, examples, history, and other facts) and emotional appeal or affective rhetoric, to allow us to consider the affective dimension of rhetoric (subdivided into insult, personal attack, empathy with acknowledging grievances, displaying positive emotions, displaying negative emotions, and sarcasm);
3. A multimedia category, to identify the role of sharing links; we also included images and emoticons in this category;
4. An impact category (as mentioned above, this comprises the number of comments and the general tone of comments; however, we also annotated the tone of the counter-speech to have some correlation with the tone of answers to that specific counter-speech);
5. A notation of to which comment the counter-speech is addressed, so that at a later stage we can further correlate the variables of the comment and the counter-speech.

Weekly monitoring of the annotator results ensured consistency and coherence in the process. From a database of 14,861 comments, 1,500 (10%) were annotated as counter-narratives (Table 2).

4. Detailed Discussion of Results

4.1. Rhetoric Used in Counter-Speech Across Topics

In the next section are a number of graphs summarising our results. We keep the original (mis)spelling in the quotations. The first column of all graphs gives the statistics referring to the migration database (MIG); the second gives the results for the political issues database (POL); the third is for the LGBTQ database. The number before these abbreviations refers to the number of the comments in our database. In this section, we look at our results in terms of the questions we posed initially: What counter-strategy strategies are most effective? Are some strategies more effective for certain topics?

4.1.1. Use of Argumentation

The strategic use of reasoning is high across all topics discussed, with an average of 78% for all categories, and with the political issues database displaying the highest percentage (86%). This result is surprising, considering that most research into online speech has found high spontaneity and a lack of control (Herring et al., 2002; Yus, 2011). In response to this seeming inconsistency, we might suggest that those who engage in counter-speech will be less prone to outbursts in expressing their views, as perhaps they will have been educated or trained in the use of counter-narratives. The following examples show some of the reasoning strategies we found in the MIG and POL databases:

> You could be made a refugee at some point in your life, have some compassion, madam, or best stay silent. (a; 37 MIG)
>
> Sir, leave aside the jokes; state the truth, and use face masks and protect yourself and those near and dear to you from this harmful virus that is spreading. (b; 49 POL)

In (a) the argumentation uses the reversal of role tactic: "You could be in their shoes." In (b) several words of

Table 2. Numbers of annotated counter-narratives.

No. of comments analyzed	No. of comments annotated for triggering overt or covert hate speech	No. of annotated counter-narratives
14,861	9,700	1,500

advice are offered in an effort to halt a participant's sarcasm. The comment in (b) refers to a video posted by the government and its recommendations for Covid-19.

Statistics are predominant in the migration dataset (11%), which can be expected (Figure 1): The fact that the number of migrants or foreigners is hugely exaggerated is well-known (Wohlfeld, 2014). In the following quotation (c), the participant gives numbers to explain the plight of the asylum seekers from Syria and dismisses the stereotype of young men migrating to Europe from the Middle East with statistical facts:

> Most of them settled in Lebanon? You make it sound so nice. [In fact] 70% of the 1 m[illion] Syrian refugees in Lebanon live below the poverty line...36% of those entering Europe are children and women. But in your eyes even Muslim children are violent. (c; 77 MIG)

We find the use of examples, which are generally personal experiences, more prominent in the migration and LBGTQ datasets (Figure 2). The "history" argument is also quite common and is used to counter the idea that Muslims are prone to violence or that they condone ISIS violence, as in (d), or to counter LGBTQ stereotyping, as in (e):

> As a British Muslim, [I say to] people commenting on this, I thought I'd tell you all this has nothing to do with Islam! Islam goes against the killing of all innocent people! So do not think this is Islam! (d; 1466 MIG)

> I'm transgender. I served 14 years active duty. And I didn't do it for any purpose except to Serve My Country. (e; 510 LGBTQ)

In (f), by pointing out the history of the Vikings, the counter-argument dismisses the preceding allegations that Westerners bring civilisation to the countries they invade, and that today migrants only take advantage of the host nations. In (g), in response to hateful messages about gay marriage because of religious principles, the counter-speech argues that religion is a historical artifact and is based on tradition rather than truth or fact:

> You mean the Vikings who spent most of their time abroad raping and stealing all the goodies? And also took over a lot of other countries. :D :D (f; 4947 MIG)

> Religion is just a mix of history and the world in terms that humans can understand, especially for filling in the gaps. If an undiscovered tribe saw a helicopter fly over, they'd call it a flying beast because birds are the only thing they have to compare to. It wasn't any different 2000 years ago. It's just that now it's to do with tradition more than anything else. (g; 166 LGBTQ)

4.1.2. Use of Affect

Regarding the use of affect, Ernst et al. (2017) noted the degree of hostility often found in counter-speech. Our results point to a level of hostility and negativity in counter-speech equal to that of hate speech. If we

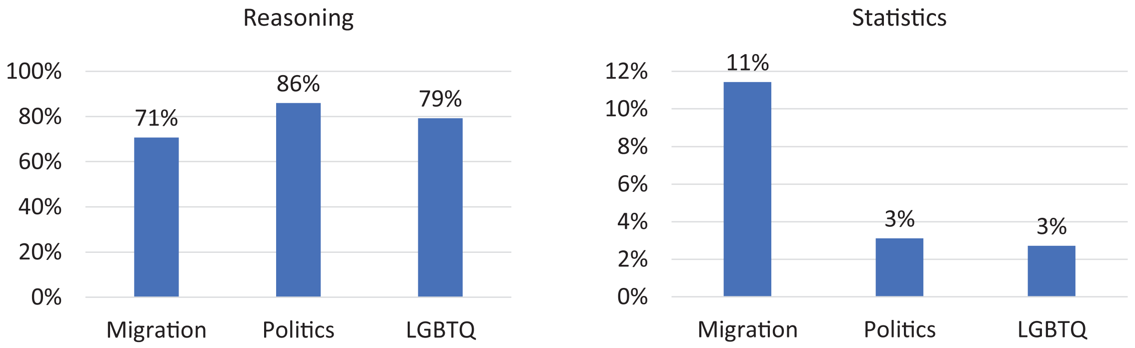

Figure 1. Reasoning vs. statistics used as arguments for the three topics.

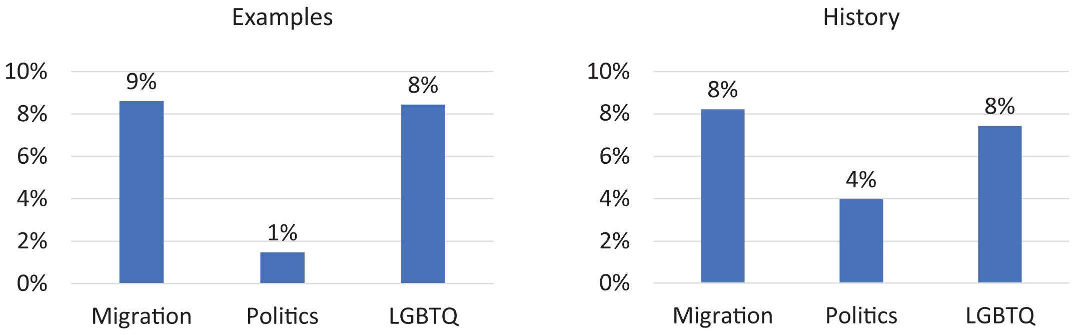

Figure 2. Examples and history used as arguments for the three topics.

group together the negative categories (insults, personal attacks, sarcasm, and display of negative emotions), we obtain an average of 73% for all topics of hostile counter-speech; respectively, 77% for migration, 66% for politics, and 76% for LGBTQ issues.

How hostility is evidenced, however, is specific to the topic debated (Figure 3): Insults are rare in political debates, while sarcasm prevails when countering LBGTQ "bashing" (33%), as seen in the following quotations:

> Have you been drinking? (20 LGBTQ)

> God can come have a chat with me about what people should have anger about [responding to a previous comment referring to the wrath of God]. (63 LGBTQ)

> Too absurd to bother commenting on. Number of gay people who have died as a result of the wrath of god 0. Find a sensible argument or give up. (73 LGBTQ)

Other signs of hostility, such as personal attacks and insults, were found in our migration data (Figure 4):

> You only see what you want to see. Pretty much like an ostrich. (h; 13 MIG)

> That's the most self-centered statement I've heard all day. No. We will continue to discuss them until they're safe from harm, have food and can work. (i; 71 MIG)

Although previous research targeting online speech generally, and online counter-speech specifically, has underlined the negativity of the messages or posts (Ernst et al., 2017), we nevertheless noted that almost a quarter of interventions were managed in a positive way. The rhetoric surrounding political issues is more often positive (34%), in comparison to migration (22%) or LGBTQ issues (24%); the two latter topics are more emotional and involve fundamental values such as religious values.

A more successful counter-speech strategy involves presenting positive emotions, as in the following examples. The speaker in (j) tries to appease verbal violence against a woman in a video wearing a veil in support of Muslim women, and in (k) the writer tries to derail racist rants against Pakistan by suggesting that India can help, if only on humanitarian grounds.

> Listen guys....The woman just wanted to show some love and solidarity with Muslim women. Don't make a big deal out of it. (j; 3309 MIG)

> On humanitarian grounds alone, we can help Pakistan also [the topic is about India giving medicine against Covid-19 to Pakistan]. (k; 739 POL)

To summarise the main counter-speech strategies, our data show a predominant use of argumentation, even though we know that the specific topics are better served by other types of counter-speech. We have also observed a notably limited use of statistics or historical

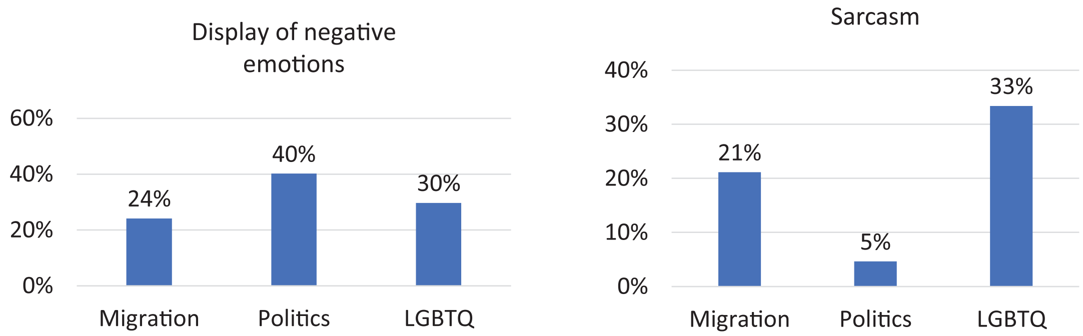

Figure 3. Hostile speech in counter-speech for the three topics.

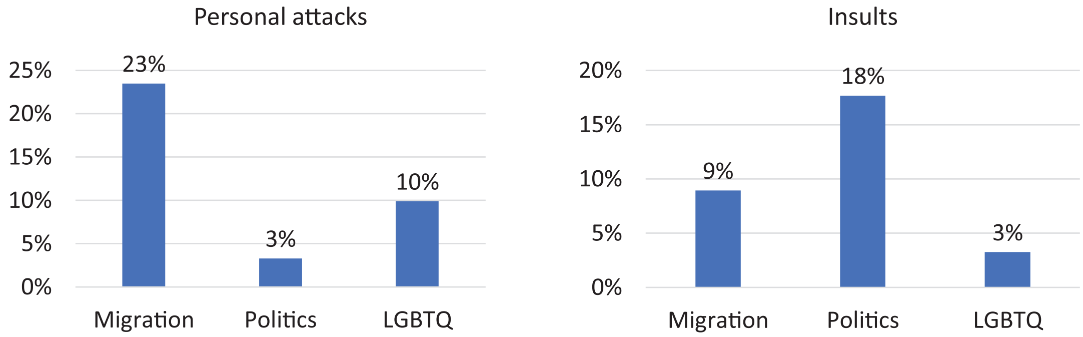

Figure 4. Personal attacks and insults used as arguments for the three topics.

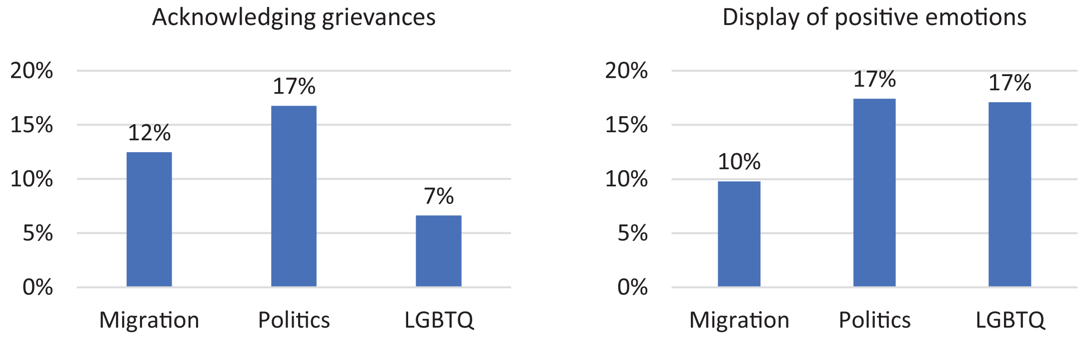

Figure 5. Positive comments used as arguments for the three topics.

examples when discussing political issues, but a high use of logical arguments. In contrast, the migration topic seems to favour the use of statistics, which should inform educational training programs in counter-speech: Knowing facts and statistics are important when responding to racist comments. We noted heavy use of sarcasm and personal attacks in responses to homophobic (and racist) comments.

The most-recorded strategy, whether in argumentation or use of affect, was a hostile stance. This is among the tactics recommended by Benesch et al. (2016), as it can make the commentator feel embarrassed because of their statement. The power of hostile comments to generate dialogue was confirmed in our study.

4.2. Impact of Counter-Speech Strategies

As explained in our methodology section, we evaluated the success of counter-speech strategies in relation to two variables and in line with the literature (de Latour et al., 2017; Silverman et al., 2016): (a) whether the comment initiated a dialogue, i.e., triggered a response (whether positive or negative is not important); and (b) what the tone of the responses triggered by the counter-speech was.

4.2.1. Number of Answers on Average

In broad terms, almost half the counter-speech strategies generated comments for all categories considered; moreover, we found no statistically significant difference among the three datasets concerning the impact of counter-speech. Therefore, we give the average numbers across all topics (see Table3).

Table 3. Average number of answers to counter-speech.

Number of answers to counter-speech	Percentage
0	46%
1–5	51%
5–10	0%
More than 10	0%

4.2.2. Correlations Between Strategies and Number of Answers

We found no evidence of a statistically significant relationship between the different argumentative strategies and generating a dialogue (i.e., triggering several answers between 1–5), although some emotional appeal (affective) strategies were found to be correlated with generating a dialogue.

We found a positive correlation between generating a dialogue and using personal attacks, while in contrast, counter-speech that is less aggressive, e.g., displaying positive emotions and acknowledging grievances, is less likely to generate a dialogue.

Additionally, the correlation coefficients suggest that there is no real relationship between generating dialogue and the strategies of using insults, sarcasm, or displaying negative emotions.

4.2.3. Impact of the Tone Used in Counter-Speech

We found that the tone of the counter-speech may influence the tone of the response: There is a positive correlation between a negative tone in the counter-narrative and a negative tone in the response; however, there is no correlation between a negative tone in the counter-narrative and a positive tone in its response. These correlations suggest that since most counter-narratives are classified as having a negative tone, they will generate a dialogue with negative answers.

In contrast, a counter-narrative with a positive tone correlates with a positive tone in the response. This finding indicates that, although counter-speech that is positive in tone is less likely to generate a dialogue, when it does, the resulting exchange is likely to be positive.

Counter-narratives that are classified as positive in tone are more likely to use statistical facts as part of their argumentative rhetoric, and to acknowledge grievances as part of their affective rhetoric (Table 4). We can observe in Table 4 a positive correlation between acknowledging grievances and generating answers as well as between acknowledging grievances and using a positive tone.

Table 4. Correlations for the strategy of acknowledging grievances.

Correlations for acknowledging grievances	
Number of answers	27%
Positive tone	71%

We conclude, therefore, that if a counter-speech aims to generate a positive dialogue, it would be most effective if it used facts, statistics, or acknowledged grievances while avoiding a hostile tone.

4.3. Summarising Results

Our first research question examined the counter-speech strategies used in situ in a large sample. Our analysis revealed that the majority of counter-narratives expose or ridicule the authors of offensive comments. We found this to be true for the majority of narratives, which used argumentative strategies (70% of the chosen strategies, on average for the three topics under investigation), as well as for the 30% of strategies that used affective rhetoric, where most narratives were highly negative, featuring insults, personal attacks, and negative emotions. We found that although humor is seldom used, sarcasm is prevalent in both covert hate speech and counter-speech. Thus, whether argumentative or affective, these strategies exacerbate verbal violence and fuel the negativity, further polarising the debate. Importantly, analysis of our findings led us to conclude that the particular strategy selected in counter-arguments is highly influenced by the social context as well as by the topic under discussion.

Our second research question measured, from a quantitative perspective, the impact of the various strategies on the different topics. The general results reveal that dialogue is rarely sparked: most often the counter-speech is ignored. Moreover, in the case that dialogue is generated by counter-speech, it is usually because of its hostile tone, especially if it contains a personal attack. It would appear that positive dialogues, the ultimate aim of counter-speech, are only generated by acknowledging grievances or displaying positive emotions, two strategies that are not often encountered in online heated debates.

5. Concluding Remarks

In summary, our results indicate that the tone of the counter-narrative is highly important and should be the first consideration when responding to hate speech. In contrast, we found, in our data, that most counter-speech took a hostile tone, and although this is a strategy recommended by Benesch et al. (2016), our results show that this is ineffective: It only puts the "opponent" on the defensive and often leads to continued verbal violence. Our results, therefore, confirm a number of earlier studies that found hateful posts were most often responded to with disagreement, conflict, and derision (Bartlett & Krasodomski-Jones, 2016; Maity et al., 2018). Nevertheless, we did identify some argumentative strategies that led to a positive outcome: the use of historical facts and/or personal examples correlated with generating dialogue, even when the tone was negative. Positive-toned responses—which we consider a marker of the effectiveness of a counter-narrative—resulted when the comment acknowledged the writer's grievances or used a positive emotional tone. Yet we found that very few counter-narratives (on average 10%) used these strategies. We, therefore, recommend that educational training—even machine learning and empathetic bots—should focus on such strategies.

Acknowledgments

This article is based on data collected within the Innovative Monitoring Systems and Prevention Policies of Online Hate Speech (IMsyPP) project, co-funded 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 its contents, which reflect the views of the authors alone. The Commission cannot be held responsible for any use which may be made of the information contained in this article.

Conflict of Interests

The author declares no conflict of interest.

References

Ashida, M., & Komachi, M. (2022). Towards automatic generation of messages countering online hate speech and microaggressions. In A. Narang, A. Davani, L. Mathias, B. Vidgen, & Z. Talat (Eds.), *Proceedings of the sixth workshop on online abuse and harms* (pp. 11–23). Association for Computational Linguistics.

Askanius, T. (2021). On frogs, monkeys, and execution memes: Exploring the humor-hate nexus at the intersection of neo-nazi and alt-right movements in Sweden. *Television & New Media*, *22*(2), 147–165.

Baider, F. (2019). Le discours de haine dissimulée; le mépris pour humilier [Covert hate speech; using contempt to humiliate]. *Déviance et Société*, *43*(1), 71–100.

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

Baider, F. (2022). Covert hate speech, conspiracy theory and anti-semitism: Linguistic analysis versus legal judgement. *International Journal for the Semiotics of*

Law, *35*, 1–25. https://doi.org/10.1007/s11196-022-09882-w

Bartlett, J., & Krasodomski-Jones, A. (2016). *Counter-speech on Facebook*. Demos. https://demosuk.wpengine.com/wp-content/uploads/2016/09/Counter-speech-on-facebook-report.pdf

Ben-David, A., & Fernández, A. M. (2016). Hate speech and covert discrimination on social media: Monitoring the Facebook pages of extreme-right political parties in Spain. *International Journal of Communication*, *10*, 1167–1193.

Benesch, S. (2014). *Countering dangerous speech: New ideas for genocide prevention*. United States Holocaust Memorial Museum.

Benesch, S., Ruths, D., Dillon, K. P., Saleem, H. M., & Wright, L. (2016). *Counterspeech on twitter: A field study*. The Dangerous Speech Project. https://dangerousspeech.org/counterspeech-on-twitter-a-field-study

Bhat, P., & Klein, O. (2020). Covert hate speech: White nationalists and dog whistle communication on Twitter. In G. Bouvier & J. Rosenbaum (Eds.), *Twitter, the public sphere, and the chaos of online deliberation*. Palgrave Macmillan. https://doi.org/10.1007/978-3-030-41421-4_7

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

Briggs, R., & Feve, S. (2013). *Review of programs to counter narratives of violent extremism. What works and what are the implications for government?* Institute for Strategic Dialogue. https://www.publicsafety.gc.ca/lbrr/archives/cn28580-eng.pdf

Brown, A. (2020). *Models of governance of hate speech*. Council of Europe.

Burnap, P., & Matthew, L. W. (2015). Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy Internet*, *7*(2), 223–242.

Citron, D. K., & Norton, H. (2011). Intermediaries and hate speech: Fostering digital citizenship for our information age. *Boston University Law Review*, *91*, 1436–1460.

Chung, Y. L., Kuzmenko, E., Tekiroglu,S., & Guerini, M. (2019). CONAN—COunter NArratives through Nichesourcing: A multilingual dataset of responses to fight online hate speech. In P. Nakov & A. Palmer (Eds.), *Proceedings of the 57th annual meeting of the Association for Computational Linguistics* (pp. 2819–2829). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1271

Council of the European Union. (2008). Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law. *Official Journal of the European Union*, *51*(L 328/55). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32008F0913

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

Ernst, J., Schmitt, J. B., Rieger, D., Beier, A. K., Vorderer, P., Bente, G., & Roth, H. J. (2017). Hate beneath the counter speech? A qualitative content analysis of user comments on YouTube related to counter speech videos. *Journal for Deradicalization*, *10*, 1–49.

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, *51*(4), 1–30.

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering online hate speech*. UNESCO Publishing. https://unesdoc.unesco.org/ark:/48223/pf0000233231

Gelber, K. (2019). Differentiating hate speech: A systemic discrimination approach. *Critical Review of International Social and Political Philosophy*, *24*(4), 393–414. https://doi.org/10.1080/13698230.2019.1576006

Gemmerli, T. (2015). *The challenges of propaganda war: A guide to counter-narratives in the prevention of radicalisation* (Policy Brief). Danish Institute for International Studies. https://www.diis.dk/node/6900

Herring, S., Kirk, J. S., Scheckler, R., & Barab, S. (2002). Searching for safety online: Managing trolling in a feminist forum. *Information Society*, *18*, 371–384.

Hietanen, M., & Eddebo, J. (2022). Towards a definition of hate speech with a focus on online contexts. *Journal of Communication Inquiry*. Advance online publication. https://doi.org/10.1177/01968599221124309

Hill, J. (2008). *The everyday language of white racism*. Wiley Blackwell.

Konikoff, D. (2021). Gatekeepers of toxicity: Reconceptualizing Twitter's abuse and hate speech policies. *Policy Internet*, *13*, 502–521.

de Latour, A., Perger, N., Salaj, R., Tocchi, C., & Viejo Otero, P. (2017). *WE CAN! Taking action against hate speech through counter and alternative narratives*. Council of Europe.

Leroy v. France, 36109/03 (2008).

Maity, S. K., Chakraborty, A., Goyal, P., & Mukherjee, A. (2018). Opinion conflicts: An effective route to detect incivility in twitter. In K. Karahalios, A. Monroy-Hernández, A. Lampinen, & G. Fitzpatrick (Eds.), *Proceedings of ACM on Human-Computer Interaction* (Vol. 2, pp. 1–27). Association for Computing Machinery.

Matsuda, M. (1989). Public response to racist speech: Considering the victim's story. *Michigan Law Review*, *87*(8), 2320–2381.

McGowan, M. K. (2009). Oppressive speech. *Australasian Journal of Philosophy*, *87*(3), 389–407. https://doi.org/10.1080/00048400802370334

Munger, K. (2017). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, *39*(3), 629–649.

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

Norwood v. the United Kingdom, 23131/03 (2004).

Office for Democratic Institutions and Human Rights (2009). *Hate crime laws.* https://www.osce.org/odihr/36426

Pingen, A. (2021). *EU Commission 6th evaluation of code of conduct on countering illegal hate speech online*. Eucrim. https://eucrim.eu/news/commission-6th-evaluation-of-code-of-conduct-on-countering-illegal-hate-speech-online

Richards, R. D., & Calvert, C. (2000). Counterspeech 2000: A new look at the old remedy for "bad" speech. *BYU Law Review*, *2*(2), 553–586.

Sap, M., Dallas, C., Saadia, G., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. In P. Nakov & A. Palmer (Eds.), *Proceedings of the 57th annual meeting of the Association for Computational Linguistics* (pp. 1668–1678). Association for Computational Linguistics.

Schieb, C., & Preuss, M. (2016, June 9-13). *Governing hate speech by means of counterspeech on Facebook* [Conference session]. 66th ICA annual conference, Fukuoka, Japan.

Silva, L., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016). Analyzing the targets of hate in online social media. *Proceedings of the International AAAI Conference on Web and Social Media*, *10*(1), 687–690.

Silverman, T., Stewart, C. J., Amanullah, Z., & Birdwell, J. (2016). *The impact of counter-narratives. Insights from a year-long cross-platform pilot study of counter-narrative curation, targeting, evaluation, and impact*. Institute for Strategic Dialogue. https://www.isdglobal.org/wp-content/uploads/2016/08/Impact-of-Counter-Narratives_ONLINE_1.pdf

Stephan, W., Stephan, C., & Gudykunst, W. (1999). Anxiety in intergroup relations: A comparison of anxiety/uncertainty management theory and integrated threat theory. *International Journal of Intercultural Relations*, *23*(4), 613–628.

Strossen, N. (2018). *Hate: Why we should resist it with free speech, not censorship*. Oxford University Press.

Tuck, H., & Silverman, T. (2016). *The counter-narrative handbook*. Institute for Strategic Dialogue. https://www.isdglobal.org/wpcontent/up

United Nations. (1965). *International convention on the elimination of all forms of racial discrimination.* https://www.ohchr.org/sites/default/files/cerd.pdf

United Nations. (1966). *International covenant on civil and political rights*. https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights

United Nations. (2012). *Rabat plan of action* (HRC/22/17/Add.4). https://www.ohchr.org/sites/default/files/Rabat_draft_outcome.pdf

Weaver, S. (2016). Humor and race. In J. Stone, R. M. Dennis, P. S. Rizova, A. D. Smith, & X. Hou (Eds.), *The Wiley Blackwell encyclopedia of race, ethnicity and nationalism* (pp. 1–3). John Wiley & Sons.

Wodak, R. (2015). *The politics of fear: What right-wing populist discourses mean*. SAGE.

Wohlfeld, M. (2014). Is migration a security issue? In O. Grech & M. Wohlfeld (Eds.), *Migration in the Mediterranean: Human rights, security and development perspectives* (pp. 61–77). Mediterranean Academy of Diplomatic Studies.

Wright, L., Ruths, D., Dillon, K. P., Saleem, H. M., & Benesch, S. (2017). Vectors for counterspeech on Twitter. In Z. Waseem, W. Hui Kyong Chung, D. Hovy, & J. Tetreault (Eds.), *Proceedings of the first workshop on abusive language online* (pp. 57–62). Association for Computational Linguistics.

Yus, F. (2011). *Cyberpragmatics. Internet-mediated communication in context*. John Benjamins.

About the Author

Fabienne Baider is a professor of linguistics and gender studies at the Department of French and European Studies, University of Cyprus. Her current research focusses on online hate speech from a linguistic perspective. Her recent work on this topic has been published in the *Handbook of Intercultural Pragmatics*, *Pragmatics and Society* as well as the *International Journal for the Semiotics of Law*.