The expression of hate speech against Afro-descendant, Roma, and LGBTQ+ communities in YouTube comments

Paula Carvalho,[1,2] Danielle Caled,[1,2] Cláudia Silva,[2,3]
Fernando Batista,[1,4] Ricardo Ribeiro[1,4]

[1]INESC-ID Lisboa, Portugal | [2]Instituto Superior Técnico, Universidade de Lisboa, Portugal | [3]ITI-LARSyS, Lisboa, Portugal | [4]Instituto Universitário de Lisboa (ISCTE-IUL), Portugal

This paper addresses the specificities of online hate speech against the Afro-descendant, Roma, and LGBTQ+ communities in Portugal. The research is based on the analysis of CO-HATE, a corpus composed of 20,590 YouTube comments, which were manually annotated following detailed guidelines that were created for that purpose. We applied methods from corpus linguistics to assess the prevalence of overt and covert hate speech, counter-speech, and offensive speech, considering different grounds of discrimination, and to investigate the main linguistic and rhetorical strategies underlying hatred messages. The research results highlight the importance of tackling covert hate speech, a recurring phenomenon often anchored in irony and fallacious argumentation, including the emotional appeal to fear and the implicit call to action. We believe this study will aid in advancing the analysis of online hate speech, while promoting the development of efficient automated detection models, specifically regarding the Portuguese language.

Keywords: overt hate speech, covert hate speech, counter-speech, Afrophobia, Romaphobia, LGBTQphobia

1. Introduction

In recent years, Online Hate Speech (OHS) has become an emerging research topic, due to its widespread presence across various social media platforms and the detrimental effects it has on individuals and society at large (Paz et al. 2020; Siegel 2020; Ullmann and Tomalin 2020). To prevent and counter the spread of OHS, several automated detection approaches, combining methods from Natural Language Processing (NLP) and Artificial Intelligence (AI), have been proposed in the literature (see Fortuna and Nunes 2018; Schmidt and Wiegand 2019). In addition, various language resources, particularly annotated corpora, have been developed to support automated approaches (Fortuna and Nunes 2018; Poletto et al. 2021). Nevertheless, the absence of a universal, unambiguous, and easily operationalized definition of hate speech (Benesch et al. 2016; Sellars 2016; Siegal 2020) hinders the usefulness of those resources and makes their comparison challenging, as they may cover completely distinct phenomena. Moreover, since hate speech is intrinsically dependent on both the target communities and social practice (i.e., the social and historical context), existing resources cannot be directly transferable or easily adapted to other linguistic and pragmatic contexts (Baider 2020, 2022; Pohjonen and Udupa 2017).

Current corpora are often created through generic lexical-based approaches, retrieving words or expressions with negative polarity, typically epithets and slurs, which may incite hatred or violence against individuals or groups (Poletto et al. 2021). Although keyword strategies can be effective at identifying potentially offensive terms, they are inaccurate at identifying hate speech (Davidson et al. 2017). On the one hand, hatred messages may not contain explicit hatred words or expressions; on the other, potentially neutral words may be used disparagingly, to attack a specific target (Benesch et al. 2016; Hill 2008). Besides, this selection method leaves out an immense set of potentially relevant hatred content, including covert, subtle or indirect forms of hate speech (Assimakopoulos, Baider and Millar 2017; Baider and Constantinou 2020; Bhat and Klein 2020; Rieger et al. 2021; van Dijk 1992, 1993; Wodak and Reisigl 2015). The terms *overt* and *covert* hate speech are used in this study to distinguish between explicit and implicit forms of discriminatory or hateful content that primarily target individuals based on their perceived social group membership. In overt hate speech, the speaker explicitly spreads, incites, promotes, or justifies hatred, exclusion, discrimination, and/or violence/aggression against a target group or individual because of group membership, and the message often conveys overtly biased, inflammatory speech, insults, abusive and derogatory language. Conversely, in covert hate speech, the message does not usually contain derogatory or insulting terms, and the spreading, promotion or justification of hatred, exclusion, discrimination, or violence against the target is not explicit; instead, its meaning needs to be inferred (Assimakopoulos, Baider and Millar 2017; Baider 2022). Irony, sarcasm, humor (Baider and Constantinou 2020; Billig 2001; Dynel 2018a), euphemisms (Magu and Luo 2018) and rhetorical questions (Albeda Marco 2022; Krobová and Zàpotocký 2021) are examples of rhetorical strategies that proved to be effective in spreading covert hate speech. Covert forms of hate speech can also be concealed within various types of fallacies that are employed to manipulate the audience's opinions or perceptions regarding a situation (Macagno 2022; Serafis et al. 2023).

From a legal perspective, unlike overt hate speech (often mixed with the concepts of *hard* or *illegal* speech), covert hate speech (also referred in the literature as *soft* hate speech) does not comprise prosecutable forms prohibited by law, which may justify the prevalence of such type of speech both in mainstream and social media (Assimakopoulos, Baider and Millar, 2017; Kumar et al. 2018). Nevertheless, as argued by Assimakopoulos, Baider and Millar (2017, 88), it manifests an illocutionary dimension akin to overt hate speech (the intention to harm) and may have the same effect on the audience, at the perlocutionary level.

Although there is increasing interest in studying covert hate speech, most annotated corpora – particularly for Portuguese – do not explicitly include information on the overt/covert character of hate speech, thus preventing an in-depth understanding of the nature and extent of this phenomenon.

To overcome such limitations, we created CO-HATE (**C**ounter, **O**ffensive and **Hate** speech), a finely grained annotated corpus for Portuguese.[1] CO-HATE comprises 20,590 comments posted by 8,485 different online users on a collection of 39 YouTube videos potentially targeting the Afro-descendant, Roma, and LGBTQ+ communities. Despite the lack of official reports on hate speech against racialized groups in Portugal, several studies have revealed that the Portuguese Afro-descendant and Roma communities are among the most targeted ethnic minorities regarding hate speech (Cádima et al. 2021; Magano and Mendes 2021; Torres da Silva 2021). Discrimination and violence against people based on their sexual orientation, gender identity, gender expression, and sexual characteristics are also poorly documented and underrepresented in official indicators of crimes and discriminatory incidents in Portugal. However, hate speech targeting LGBTQ+ individuals is still prevalent on social media in the European context, underscoring the importance of including this group in our study.[2]

By using a corpus linguistics approach, which involves computer-based empirical analyses (both quantitative and qualitative) of naturally occurring language use from usually large and electronically available collections of texts (McEnery and Hardie 2011; McEnery and Brezina 2022), this study seeks to investigate the potential specificities of OHS, expressed both overtly and covertly, aimed at the targeted communities. In particular, the CO-HATE corpus will help us answer the following research questions:

RQ1: How does OHS against the Afro-descendant, Roma, and LGBTQ+ communities materialize in the Portuguese social context?

RQ2: Which are the main linguistic and rhetorical aspects underlying the expression of covert hate speech?

The main contributions of this research can be summarized as follows: (i) the creation of the first finely-grained annotated corpus for European Portuguese, critical both for studying and supporting the detection of OHS, particularly against the Afro-descendant, Roma and LGBTQ+ communities on social media; (ii) the analysis of the potential specificities underlying hatred comments targeting the aforementioned groups, combining quantitative and qualitative research, based on methods from corpus linguistics; and (iii) the inter-annotator agreement study focused on OHS and semantically related phenomena.

2. Hate speech corpora

The complexity and multidimensional nature of hate speech poses diverse challenges on its modeling and automated detection. This problem is compounded by the fact that hate speech is often mixed with other instances of offensive language (Davidson et al. 2017; Wiegand, Siegel and Ruppenhofer 2018), or language aggression (Basile et al. 2019; Kumar et al. 2018), which is reflected on the heterogeneity of language resources specifically created to assist automatic OHS detection (Poletto et al. 2021).

[1] The corpus is available on: https://hate-covid.inesc-id.pt.

[2] Anti-gypsyism (or Romaphobia), xenophobia (including anti-migrant hatred) and sexual orientation have been the most commonly reported grounds of OHS in the scope of the monitoring rounds of the Code of Conduct performed by the European Commission: http://tiny.cc/4rd6vz.

The heterogeneity among the existing hate speech corpora is also explained by the diversity of categories and attributes being considered by researchers in their annotation experiments. While some studies are mainly concerned with distinguishing generic grounds of discrimination, such as *racism* or *sexism* (Waseem 2016), others have adopted complex hierarchical labeling schemata, including dozens of categories and subcategories (e.g., Fortuna et al. 2019), rendering their comparison difficult or even impossible.

Moreover, although there are several resources and benchmark corpora for many different languages, we have found only four annotated hate speech corpora for Portuguese, and none of them focuses exclusively on European Portuguese. Pelle and Moreira (2017) developed a corpus with 1,250 comments randomly extracted from *O Globo*, which is one of the most popular Brazilian news media outlets. These comments focus on political and sports news, whose topics could potentially generate more controversy and hate speech. Each comment was labeled as being offensive or non-offensive, and the former was also categorized into one of the following classes: *xenophobia*, *homophobia*, *sexism*, *racism*, *cursing* and *religious intolerance*. Fortuna et al. (2019) have compiled a corpus of 5,668 Portuguese tweets, posted by 115 different users, which were manually labeled as conveying hate speech or not; hatred messages were then labeled according to its target, following a hierarchical multiple label scheme, including 81 categories. The tweets were retrieved by applying a list of offensive keywords and by selecting the users who usually post hateful comments. Leite et al. (2020) created a corpus composed of 21,000 Brazilian Portuguese tweets. These posts were retrieved by applying a filtering list of offensive words, and keywords related to influential Brazilian users that could be victims of hate speech or abuse. The tweets were assigned with one of the following categories: *LGBTQphobia*, *obscene*, *insult*, *racism*, *misogyny*, and *xenophobia*. Lastly, Vargas et al. (2022) present a corpus of 7,000 comments extracted from Instagram posts of six Brazilian political personalities. The messages in the corpus were classified following different layers of analysis: first, the messages were classified as being offensive or non-offensive; offensive messages were then classified according to the intensity of the offense and the following semantic classes: *xenophobia*, *racism*, *homophobia*, *sexism*, *religious intolerance*, *partyism*, *apology for the dictatorship*, *antisemitism*, and *fatphobia*.

The usefulness of these resources is quite limited for our study, as they do not cover the target groups nor the social and historical context (Portugal) we are particularly interested in monitoring. Moreover, they do not address covert hate speech, critical for understanding the real expression of this phenomenon on social media.

3. Methods

Our research combines quantitative and qualitative analyses, commonly used in corpus linguistics (McEnery and Hardie 2011; Tognini-Bonelli 2001), which is recognized as a suitable approach to study hate speech and related phenomena (Baker and McEnery 2005; Baker et al. 2008; Brindle, McEnery and Hoey 2016; Geyer, Bick and Kleene 2022). We will investigate: (i) word frequency, to search for the most frequent content words or expressions, (ii) concordances, also referred to as key word in context (KWIC), to examine the context of specific words or lexico-syntactic patterns, and (iii) collocations, to find co-occurrence patterns of words, by exploring specifically adjacent combinations of two words (bigrams) and three words (trigrams) in the corpus. We will also examine the diversity of annotations in the corpus to identify the most prevalent codes and analyze the linguistic constructions where those codes (co)occur. This section provides a detailed account of the data collection procedures and outlines the protocol that was developed for annotating the corpus.

3.1 Data collection

The focus of this study is a collection of 39 YouTube videos that generated a total of 20,590 comments (comprising 795,111 tokens) from 8,485 unique online users. These comments, which constitute the CO-HATE corpus, were imported into MAXQDA,[3] where they were automatically coded as either independent comments or replies to comments. The corpus was then divided into five subsets, each consisting of approximately 4,000 comments from an average of seven videos, and each subset was randomly assigned to a different annotator (Section 3.2). Additionally, the annotators were assigned to a common subset of 534 comments from two additional videos that were randomly selected from the initial list of videos. The common subset was used to measure inter-annotator agreement (IAA), as described in Section 4.

YouTube videos were selected based on the following criteria: the video title and description make a direct or indirect reference to the Afro-descendant, Roma, and LGBTQ+ communities, and the topic approached can potentially trigger polarized content and hatred against the previously mentioned groups, as illustrated in Examples (1) - (3):

(1) *A história de Portugal é racista*
‘The history of Portugal is racist’
(2) *A comunidade cigana vive numa bolha de impunidade*
‘The Roma community lives in a bubble of impunity’
(3) *Preferias ter um filho homossexual ou ladrão? Experiência social*
‘Would you rather have a homosexual child or a thief child? A Social Experiment’

It must be noted that the comments associated with the selected videos were not subject to any data selection or filtering. This strategy allowed us to both assess, for each video, the real distribution of hate speech, and investigate other related phenomena, in particular counter-speech and offensive speech. Since we are particularly interested in analyzing these phenomena within the Portuguese context, we restricted our selection to videos posted by Portuguese authors, including either public or anonymous figures, news organizations, or independent channels.

3.2 Annotators profile

The corpus annotation was performed by five recruited annotators, who were enrolled in a bachelor’s or a master’s degree in communication or social sciences. To account for the potential impact of individual and social differences on hate speech detection, the annotation team was composed by individuals belonging to the communities monitored

[3] VERBI Software. (2019). MAXQDA 2020 [computer software]. Berlin, Germany: VERBI Software. Available at https://www.maxqda.com.

in this study (one annotator from each targeted community), and by others who do not belong to any potentially marginalized group. More specifically, it includes Portuguese citizens as follows: a female cisgender of African descent, a White cisgender male who identifies himself as part of the LGBTQ+ community, a female cisgender of Roma descent, a White cisgender hetero male, and a White cisgender hetero female.

3.3 Annotation guidelines

The main tasks underlying the annotation process, which lasted 7 months, are illustrated in the timeline presented in Figure 1.

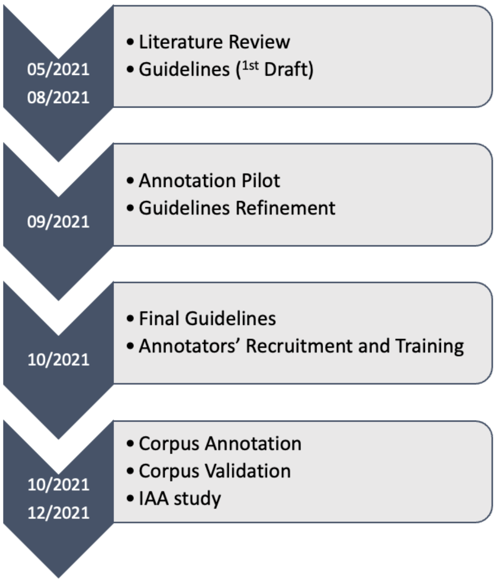


Figure 1. Annotation process timeline

Prior to coding, the annotators were provided with detailed guidelines, created by the senior members of the project team. They were required to watch the selected video first, followed by a careful reading of every comment associated with the multimodal content. Moreover, they were instructed to follow the order of the comments and replies to comments (the minimum unit of analysis), because some messages (especially, replies to comments) could only be interpreted when considering the entire conversation.

The messages were labeled considering four dimensions of analysis: speech acts (Section 3.3.1); the main grounds of discrimination from which hate speech emerges in our corpus (Section 3.3.2); rhetorical strategies (Section 3.3.3); and sentiment polarity and intensity (Section 3.3.4). The categories and subcategories assigned to each dimension are presented in Table 1.

Table 1. List of categories and subcategories described in the guidelines.

Speech Acts	Hate speech	Overt Covert
	Counter-speech	
	Offensive	
Grounds of Discrimination	Specific	Afrophobia Romaphobia LGBTIphobia
	Generic	Racism Xenophobia
	Other	
Rhetorical/Discursive Strategies	Irony/Sarcasm	
	Rhetorical questions	
	Negative stereotypes	
	Fallacies	Appeal to action Appeal to fear Personal attack
	Other	
Sentiment	Negative	[1-2]
	Neutral	[3]
	Positive	[4-5]
Non-Relevant		

Whenever a message did not convey hate, counter-speech or offensive speech, the annotators were instructed to label it as "non-relevant". Apart from sentiment, the remaining categories are not mutually exclusive. This means that annotators could assign as many labels as they saw as relevant. The selection of the categories and subcategories considered was inspired by previous annotation experiments and was tailored to fit the specific goals of this research. For example, we distinguish, like Kumar et al. (2018), overt from covert forms of hate; as Sanguinetti et al. (2018), we discern hate speech from offensiveness; and like Mathew et al. (2018) and we include counter-speech as a category intrinsically related to hate speech. Regarding the linguistic and discursive annotation schemes, we include aspects such as irony and sarcasm, often relying on negative stereotyping (ElSherif 2021; Sanguinetti et al. 2018). Like Sanguinetti et al. (2018), we also consider sentiment intensity; however, our annotation scheme is not specifically reserved to hate speech comments.

3.3.1 Speech acts

Following the guidelines provided by the Council of Europe in its latest Recommendation (CM/Rec/2022/16), hate speech is generically understood as "all types of expression that incite, promote, spread or justify violence, hatred or discrimination against a person or group of persons, or that denigrates them, by reason of their real or attributed personal characteristics or status such as 'race', colour, language, religion, nationality, national or ethnic origin, age, disability, sex, gender identity and sexual orientation". Within this study, hate speech is specifically operationalized through the following coexisting conditions:

i. Hate speech is an intergroup phenomenon, targeting groups or individuals because of their perceived social group membership. Therefore, this work does not encompass hateful expressions that occur at an individual/interindividual level.

ii. In general, the speaker is a member of a privileged group, while the target of hate is a member of a minority, vulnerable or stigmatized group in a specific social context.
iii. Hate speech can be expressed either overtly, where the meaning is explicitly expressed, or covertly, where the meaning is implicit, and must be inferred.

According to these premises, Examples (4) and (5) should be labeled as conveying overt and covert hate speech, respectively.

(4) *Racismo o c@ralho! se não fossem esses parasitas da sociedade que não querem fazer nada, Portugal era um paraíso.*
'F@ck racism! If it were not those social parasites that don't want to do anything, Portugal was a paradise'

(5) *Coitadinhos dos "feirantes", vão ficar sem os benefícios.*
'Poor "market vendors" [reference to Roma], they will lose their [social] benefits.'

While in Example (4), the Roma community (metaphorically mentioned as *social parasites*) is directly attacked, in Example (5) the speaker uses irony to indirectly harm this target, metaphorically mentioned as *market vendors*. In both examples, the speaker's intention is to harm/marginalize the target group, reinforcing common stereotypes and prejudice associated with this community, often portrayed as lazy and profiteer (Breazu and Machin 2019, 2022; Buturoiu and Corbu 2020; Chovanec 2021; Erjavec 2001). On the contrary, the comment in Example (6) is simply classified as "offensive" because it does not target the individual based on their perceived characteristics and group membership, but rather because of the perspective they expressed in a previous comment.

(6) *Lendo o seu texto outra vez percebe-se que é bastante inculta. Deve ter nascido num pardieiro.*
'Reading your text again it's clear that you are quite uneducated. You were certainly born in a dump.'

By considering the conversations in the corpus, one can find spontaneous reactions to hateful speech, including counter-speech, i.e., a direct reply to hateful or harmful content, aiming at undermining it (Benesch et al. 2016; Mathew et al. 2018). Within this study, annotators were asked to both label the cases like the one illustrated in Example (7) as counter-speech.

(7) *Conheço tantos Portugueses a fazer o mesmo lá fora...!*
'I know so many Portuguese people doing the same abroad...!'

In this case, the speaker calls attention to the hypocrisy conveyed in previous comments where afro-descendants are accused of being social benefit dependents. This discursive strategy is often deployed in racist exchanges, as described in the ideological square proposed by van Dijk (1993), which rely on the positive self-presentation (*Us*) and negative other presentation (*Others*). To deconstruct this fallacy, the author claims that there are also Portuguese emigrants (intended as the Portuguese White people) taking advantage of social benefits.

3.3.2 Grounds of discrimination

Our annotation schema takes into consideration three specific grounds of discrimination:

Afrophobia: specific form of racism towards people of African descent. It can also be generally understood as the manifestation of racism towards Black people, intended as all those individuals, groups and communities that define themselves as 'Black' (European Commission 2018a).

Romaphobia: specific form of racism against social groups identified under the stigma 'gypsy' or other related terms (European Commission 2018b).

LGBTQphobia: any type of hatred or discrimination targeting people based on their sexual orientation, gender identity and sex characteristics. It includes concepts such as *homophobia*, *biphobia*, *transphobia*, and *intersexphobia* (European Union Agency for Fundamental Rights 2020).

In addition, we included the broader categories of *racism* and *xenophobia* to capture instances of discrimination that are common in the corpus but do not fall under the specific categories we are investigating in this study.

Racism: term generically used to express the belief that a perceived characteristic such as race, color, language, religion, nationality or national or ethnic origin justifies contempt for a person or a group of persons, or the notion of superiority of a person or a group of persons (ECRI 2018).

Xenophobia: term generically used to express the belief that immigrants are outsiders or foreigners to the community, society, or national identity (Migration and Home Affairs n.d.).

If a message was found to contain hate speech or counter-speech but did not pertain to any of the categories mentioned earlier, it was categorized as "other".

3.3.3 Rhetorical devices

Annotators were asked to identify the messages conveying the following rhetorical and discursive strategies:

Irony and sarcasm: This rhetorical strategy relies on the use of words to utter something different (often the opposite) of their literal meaning. It is typically used to express an intentionally negative evaluation towards a specific target (Attardo 2000; Dynel 2018b), being often employed to disseminate hate speech, albeit covertly (Baider and Constantinou 2020), as illustrated in Examples (8) and (9).

(8) *Até a cerveja mudou de nome, antes era preta agora é stout.*
'Even the beer has changed its name, before it was black, it is now stout.'

(9) *Olha, coitado, este é vitima de racismo, va nos trabalhamos para ele!*
'Oh, poor guy, he is a victim of racism, let's work for him!'

As noted by Attardo (2000), the literature has drawn attention to the explicitly aggressive nature of sarcasm, in comparison to other forms of irony, and its deliberate aim to offend or hurt a specific target. In this work, we use these terms interchangeably, identifying the common aspects underlying both strategies: (i) they are intentionally produced by the speaker to be understood by the hearer (Dynel 2019, 3); (ii) their intended meaning is indirect, and is only arrived at inferentially (Attardo 2000, 823); (iii) both strategies may be (but not necessarily) cloaked in the mask of humor (Dynel 2017, 70); and (iv) both strategies can be used to express covert hate speech against a specific target (Baider and Romain forthcoming).

Rhetorical questions: Such questions have the illocutionary force of an assertion of the opposite polarity from what is explicitly asked (Han 2002). They can be used as reproaches, where the speaker appeals to their interlocutor's moral conscience, creating the expectation of a duty that should have been carried out by the interlocutor (Albelda Marco 2022). In hate speech, rhetorical questions are often used to implicitly associate negative stereotypes with a target (ElSherief et al. 2021), as illustrated in Example (10).

(10) *Já alguma vez viste um cigano a trabalhar?*
'Have you ever seen a gypsy working?'

Negative stereotypes: Stereotypes are commonly defined as beliefs about the attributes of social groups (Stangor 2016). Negative stereotyping is often used to disparage or humiliate the members of a vulnerable community, based on fallacious negative generalizations (Paz et al. 2020; Sanguinetti et al. 2018), as illustrated in Example (10).

Fallacies: Fallacies can be defined as violations of the standards for critical discussion that must guide reasonable argumentative discourse (van Eemeren and Garssen 2023). Despite fallacies being a common feature of the argumentation inventory of online discussion (Habernal et al. 2018; Krobová and Zàpotocky 2021; Macgano 2022), they have not usually been considered in previous annotation experiments. In this study, we have considered the following fallacies, which can covertly promote, spread or incite hate speech: (i) personal attack (or *ad hominem argument*) – attacking the opponent instead of providing counterarguments against a specific position or argumentation (Tindale 2007); (ii) appeal to fear (or *ad baculum argument*) – relying on an implicit warning that some bad outcome will occur if the receiver does not carry out the recommended action (Tindale 2007; Walton 1996); and (iii) appeal to action – involving an explicit appeal for action to revert the negative state of affairs, carrying on an emotionally charged tone. These strategies are illustrated in Examples (11), (12) and (13), respectively.

(11) *SOS Racismo é uma organização parasita subsidiada pelo Estado (todos nós) para atacar os portugueses étnicos no seu próprio país.*
'SOS Racism is a parasitic organization subsidized by the State (all of us) to attack the ethnic Portuguese in their own country.'

(12) *Minorias? Minoria somos nós se olharmos em termos globais...Passámos de 30% da população mundial em 1930 para menos de 11% e a continuar a trajetória em 30 anos seremos menos de 7%...*
'Minorities? We are a minority if we look at global terms... We went from 30% of the world population in 1930 to less than 11% and if we continue the trajectory in 30 years we will be less than 7%...'

(13) *Agradeça a União Europeia pela destruição da Europa e de Portugal! Parem de votar em políticos de esquerda!*

‘Thank the European Union for the destruction of Europe and Portugal! Stop voting for left-wing politicians!’

3.3.4 Sentiment

Hate speech detection and sentiment analysis are closely related tasks since hate speech messages usually convey a negative sentiment (Schmidt and Wiegand 2019). In addition to sentiment polarity, which is often defined as positive, negative or neutral, some annotation experiments have also considered the sentiment intensity (Sanguinetti et al. 2018). This aspect is particularly relevant in the context of hate speech detection, since it can provide important clues on the patterns associated with the most polarized sentiment and opinions, and on the evolution of patterns of hate intensity among the discussion threads (Dahiya et al. 2021). The annotators were asked to classify each comment according to the following scale: 1 (very negative), 2 (negative), 3 (neutral), 4 (positive), and 5 (very positive). According to the scale adopted in the guidelines, Example (14), which involves indirect incitement to violence, should be classified as very negative (1), and Example (15) as negative (2), because despite being offensive it is comparatively less harmful.

(14) *BANDIDO BOM É BANDIDO MORTO!*
‘A GOOD BANDIT IS A DEAD BANDIT!’

(15) *Esse mamadu é muito estúpido, como quer ele colocar cotas raciais, num país de maioria caucasiana?*
‘This mamadu is very stupid, how does he want to put racial quotas in a country with a Caucasian majority?[4]

4. Inter-annotator agreement

To assess the task complexity and subjectivity, we performed an IAA study, based on a data sample composed of 534 comments that were labeled by all the annotators. We measured the IAA using *Krippendorff’s alpha* (α), a reliability coefficient that enables us to measure the agreement among annotators. We calculated the IAA for (i) all the recruited annotators (ALL), (ii) annotators belonging to the target groups (A-B-C), and (iii) annotators that do not belong to those groups (D-E), as represented in Table 2. With that, we tried to inspect how individual and social differences may potentially influence subjective tasks, such as hate speech detection, figurative language recognition and sentiment analysis.

To perform the IAA analysis, we aggregated the sentiment scores in terms of (i) polarity, by both considering all the classes (negative vs. positive vs. neutral) and restricting the analysis to the positive and negative classes, and (ii) intensity (very negative vs. very positive). We also grouped overt and covert hate speech to calculate the agreement on the identification of hate speech as a broader category.

Table 2 - IAA among all the annotators (ALL), the annotators belonging to the target groups (A-B-C), and the remainder annotators (D-E).

[4] Mamadu Ba is a Portuguese-Senegalese anti-racist political activist in Portuguese society.

Category	**ALL**	**A-B-C**	**D-E**
Hate speech	0.478	0.360	0.735
» Overt HS	0.416	0.383	0.548
» Covert HS	0.237	0.145	0.421
Counter-speech	0.419	0.358	0.762
Offensive	0.143	0.005	0.472
Non-Relevant	0.305	0.202	0.594
Afrophobia	0.322	0.227	0.421
Romaphobia	0.299	0.131	0.818
LGBTQphobia	0.082	-0.001	0.666
Racism	0.139	-0.017	0.351
Xenophobia	0.231	0.164	0.305
Other	0.008	0.033	-0.005
Appeal to fear	0.172	0.108	0.346
Appeal to action	0.418	0.297	0.566
Personal attack	0.206	0.120	0.269
Stereotype	0.252	0.142	0.293
Irony	0.239	0.234	0.356
Rhetorical question	0.449	0.382	0.522
Other	0.208	0.173	0.388
Negative vs. Neutral vs. Positive	0.516	0.515	0.791
Negative vs. Positive.	0.568	0.574	0.818
Very Negative. vs. Very Positive	0.936	0.947	n.a.

Except for sentiment, where the agreement among all the annotators ranges from moderate (in terms of sentiment polarity) to almost perfect (in terms of sentiment intensity), the agreement for the remaining dimensions and classes is relatively low, being in line with the results reported in the literature for related tasks (Poletto et al. 2017; Fortuna et al. 2019). This reinforces the task subjectivity and potential overlapping between categories.

Surprisingly, the annotators belonging to the communities targeted in this study tend to disagree more with each other than the annotators not belonging to the target groups. In fact, in contrast to annotators A, B, and C, the annotators D and E achieved a higher agreement rate for almost all the categories considered. In some cases, the agreement between these two annotators is almost perfect (i.e., ≥ 0.8). A tentative explanation for this discrepancy is that the comments analyzed by the annotators belonging to the target communities are close to their own experience, and it depends much more on their personal sensitivity. Nevertheless, it would be imprudent to make any generalization based on such a small group of annotators.

Identifying overt hate speech appears to be easier than identifying covert hate speech, as anticipated. On the other hand, the identification of counter-speech appears to present the same challenges as hate speech. In both cases, annotators D and E had substantial agreement (i.e., > 0.7). By contrast, offensive speech seems harder to recognize; in particular, for annotators A, B, and C the agreement is almost nonexistent. Although the adopted guidelines tried to distinguish hate speech from offensive language, the results obtained suggest that annotators tend to conflate hate speech and offensive language, a problem already reported in previous studies (Davidson et al. 2017; Poletto et al. 2017).

Regarding hate speech targets, a surprisingly low (and, in some cases, inverse) agreement is observed, particularly among annotators A, B and C. This may be explained by the fact that the same comment can be potentially labeled as targeting different marginalized groups, which points at the complexity of identities (often hidden under generic perceived characteristics). Indeed, some individuals can be affected by multiple sources of discrimination (e.g., race and sexuality), rendering the task of identifying the hate speech target more difficult. As pointed out by Hancock Alfaro (2022), social identity markers do not exist independently of each other, often creating a complex hate speech target group, rather than being directed at isolated and distinct groups. Finally, the lack of agreement, which is critical among the annotators belonging to the target groups, may support the idea that hate speech detection highly depends on personal perception and can be affected by several variables, including the individual's social identity.

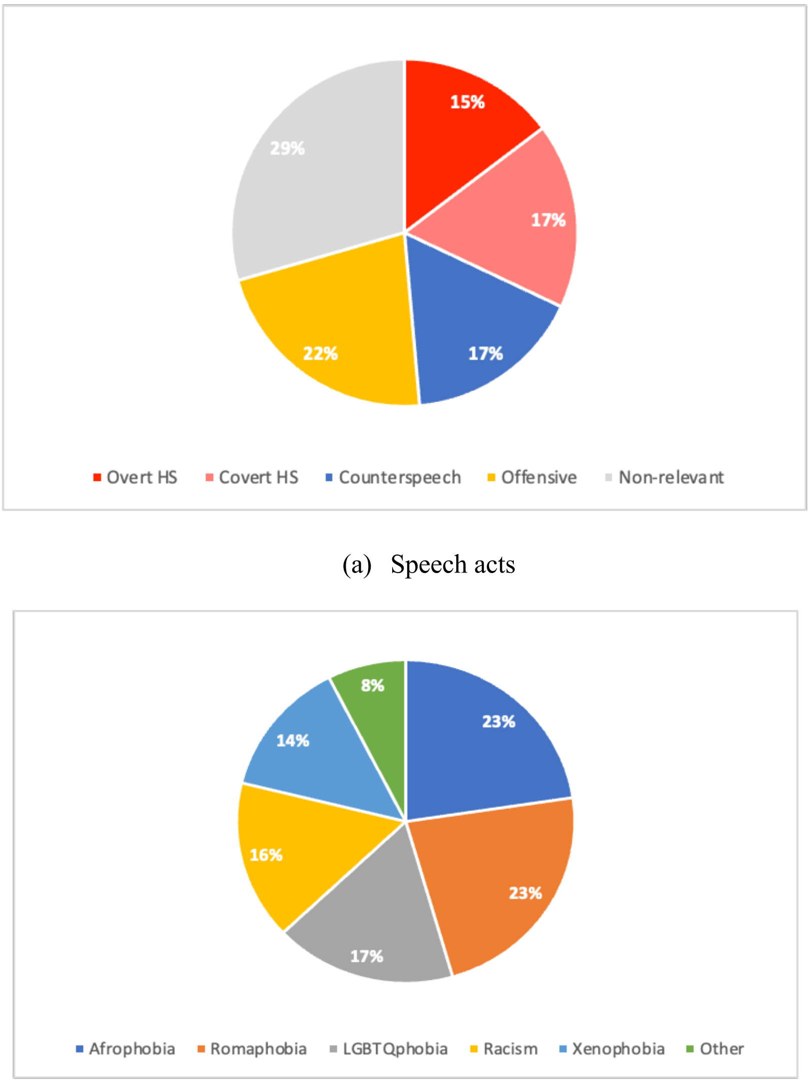
Concerning the rhetorical strategies considered in this study, the highest agreement achieved concerns the identification of rhetorical questions and the appeal to action fallacy. As expected, the agreement rate on most of the rhetorical devices is very low.

5. Results and discussion

In this section, descriptive statistics on the annotated corpus are presented, along with an exploration of techniques from corpus linguistics to identify patterns that may be associated with the materialization of OHS.

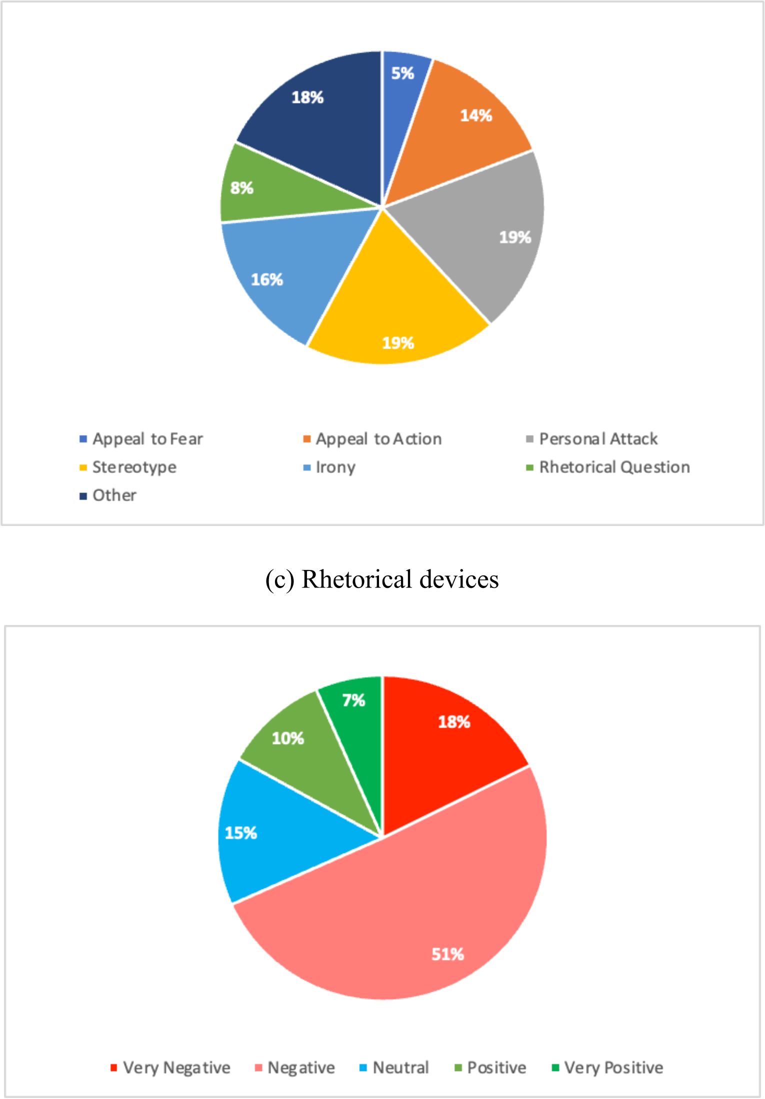
5.1 Categories distribution

The statistics presented in this section are derived from the annotations assigned to comments in the entire corpus, excluding the comments labeled by all the annotators, which resulted in exactly 20,056 comments. As illustrated in Figure 2a, half of the comments in CO-HATE were classified as conveying hate speech or counter-speech, which reinforces the importance of considering both categories in hate speech annotation studies. Moreover, 22% of the messages were classified as offensive, and 29% were considered non-relevant by annotators. Regarding the expression of hate speech, we noted that covert hate speech is highly frequent in the corpus, even surpassing overt hate speech.



(a) Speech acts

(b) Grounds of discrimination



(c) Rhetorical devices

(d) Sentiment

Figure 2 - Distribution of speech acts, grounds of discrimination, rhetorical strategies, and sentiment in the CO-HATE corpus.

Globally, the most representative hate speech targets in the CO-HATE corpus are the Roma and Afro-descendant communities, followed by the LGBTQ+ group, respectively (Figure 2b). It must be noticed that these results are not completely aligned with the results provided by the 6th and 7th EU monitoring exercises, which do not include Afrophobia among the most reported grounds of hate speech. About 38% of the comments focus on other targets, either directly related to racism and xenophobia or other grounds of discrimination that were not considered in the annotation scheme. Nevertheless, any generalization of results on the targets' distribution cannot be made since the data selection does not guarantee an identical or proportional distribution of the target communities. Regarding rhetorical strategies (Figure 2c), the presence of negative stereotyping (19%) and personal attack (19%) is notable. Verbal irony and sarcasm, as well as other rhetorical strategies often used in fallacious arguments, such as appeal to action, are also frequent in the corpus. As expected, negative polarity is the prevailing sentiment class, with about 70% of the comments being classified as negative or very negative (Figure 2d).

By comparing the comments identified as targeting specifically the Afro-descendant, Roma, and LGBTQ+ communities, some differences emerge. As illustrated in Figure 3, the Roma community was the target of the highest percentage of hate speech, and the LGBTQ+ community the lowest one. Except for LGBTQ+, covert hate speech surpasses overt hate speech in most cases. For this group, the most representative categories concern counter speech and offensive speech, respectively. Regarding counter-speech, it is comparatively much less common in comments targeting the Roma community.

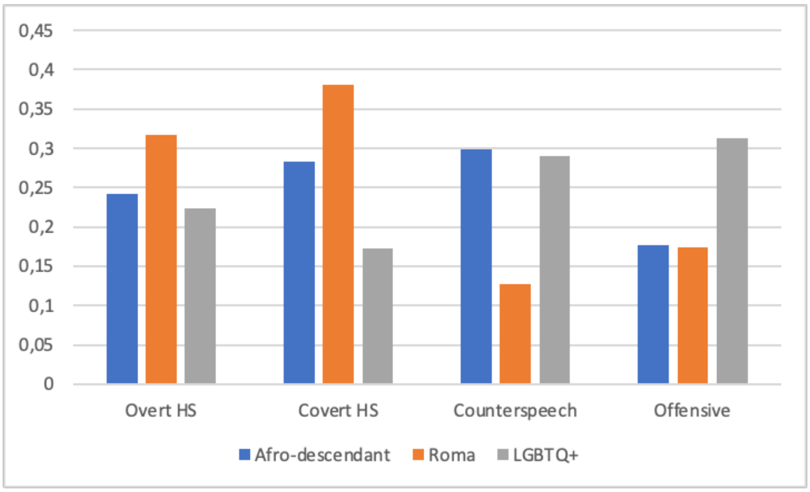


Figure 3 - Distribution of overt and covert speech, counter-speech and offensive speech in comments targeting Afro-descendant, Roma and LGBTQ+ communities.

Although interesting, these results are not completely surprising. In fact, the Roma community has experienced centuries of discrimination and social exclusion across Europe (Achim 2004; Maeso 2021), perpetuated both on social and conventional media, through normalized discourses alluding to Roma criminality, illiteracy, immorality, promiscuity, laziness and resistance to integration into mainstream society (Breazu and Machin 2019, 2022; Chovanec 2021; Erjavec 2001). In what concerns the Portuguese social context, the Roma community is still considered the most vulnerable group, being victims of poverty and social exclusion (Casa-Nova 2021; Maeso 2021, Magano and Mendes 2021). Indeed, the low rate of counter-speech targeting Roma, in comparison with the remaining groups, reinforces its vulnerable position in social media platforms, which can be measured either by their lack of active voice to respond to hate speech or the lack of support by potential allies.

Regarding rhetorical strategies, the prevalence of negative stereotyping in messages targeting either the Afro-descendant or Roma communities should be pointed out, as represented in Figure 4.

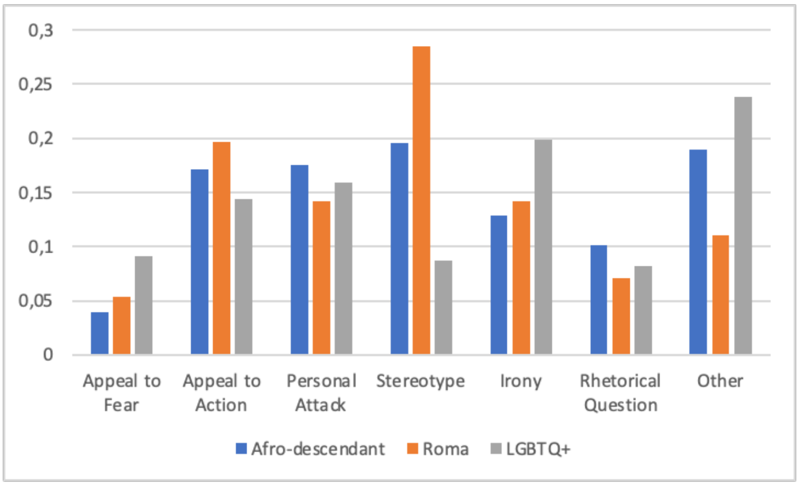


Figure 4 - Distribution of rhetorical strategies in comments targeting Afro-descendant, Roma and LGBTQ+ communities.

On the other hand, irony is the most common rhetorical strategy in messages targeting the LGBTQ+ community. Regarding fallacious argumentation, Afrophobia is often expressed by means of personal attack and appeal to action; the latter is also highly frequent to express Romaphobia. The differences emerging from the data suggest that hate speech can indeed materialize differently, depending on the target groups, highlighting the importance of considering the specificities of the target groups, rather than handling hate speech as a generic phenomenon.

Unsurprisingly, negative polarity is prevalent in comments targeting either Afro-descendant, Roma, or LGBTQ+ groups, as illustrated in Figure 5. However, the highest proportion of negative messages targets the Roma community, who have also received the highest proportion of hate speech. Specifically, about 85% of the comments targeting this group were classified as negative. Inversely, the LGBTQ+ community, who have received the lowest proportion of hate speech, also presents the highest percentage of positive polarity and intensity.

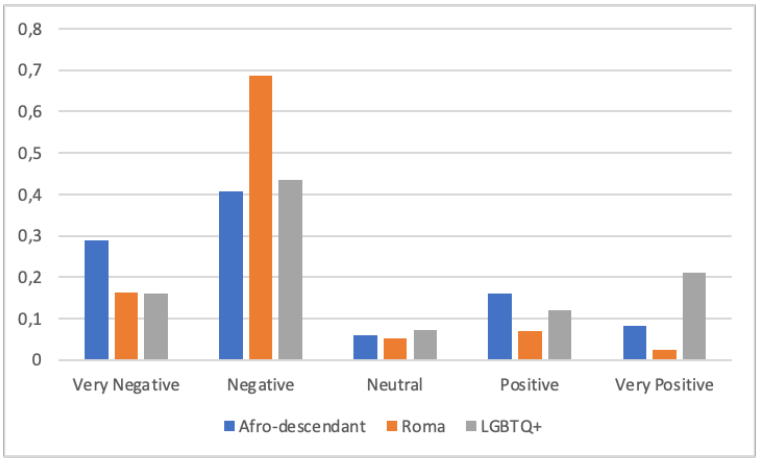


Figure 5 - Distribution of sentiment in comments targeting Afro-descendant, Roma and LGBTQ+ communities.

Although the comments targeting the Afro-descendant group seem to convey the most extreme negative sentiment, it must be noted that the messages conveying a positive sentiment correspond to 24% of the comments targeting this group. These results support the idea that, although LGBTQ+ individuals are still victims of hate speech in Portugal, the identified structural racism targeting the racialized groups in the Portuguese context seems comparatively more critical.

5.2 Linguistic realization of hate speech

To study the linguistic realization of OHS, we started by exploring the most frequent nouns, verbs, and adjectives (unigrams) and collocations (bigrams and trigrams) in comments targeting each studied group.[5] Overall, the presence of words such as *Portugal* and *portugueses* ('Portuguese') must be noted. Other proper nouns and adjectives that are frequently used to reference specific identity factors, such as ethnicity (e.g., *ciganos* 'Roma' or 'Gypsy'), nationality (e.g., *Brasil* and *brasileiros* 'Brazil' and 'Brazilians'), color (e.g., *branco*, *negro* and *preto* 'White', 'Black' and 'N-word'[6]), descent (e.g., *África* and *africano* 'Africa' and 'African') and sexual orientation (e.g., *gay* and *homossexual* 'gay' and 'homosexual'), are also very common. The analysis of these occurrences in context shows that, in most cases, the terms associated with the dominant *in-group* (in this case, White cisgender heterosexual people) co-occur with the terms associated with the *out-groups* (including the Afro-descendant, the Roma, and the LGBTQ+ communities), as illustrated in Example (16). In this case, the speaker intentionally reverses the positions of dominance and power between the dominant (the *Portuguese* and the *Europeans*) and the dominated (*Africans*, *LGBT*, *gypsies*, *Third World people*, *minorities*) social groups, a commonly used discursive strategy associated with the denial of racism (van Dijk 1992).

(16) *Os portugueses têm fama de povo sereno, e hoje quem vale é os africanos, os LGBT, e os ciganos também, em breve os europeus estarão de joelhos, a pedir desculpas por apanhar dos povos do 3º Mundo, e minorias em geral...*
'The Portuguese are famous for being a serene people, and now the Africans, the LGBT and the gypsies are the greatest ones; soon the Europeans will be on their knees, apologizing for being beaten up by Third World people, and minorities in general...'

It is also important to stress the explicit mention, in the comments targeting the Afro-descendant and the Roma groups, of the noun *racismo* ('racism') and the adjective *racista* ('racist'). Again, context examination shows that these terms usually occur in constructions where the author denies individual and collective racism. As noted by van Dijk (1992), one of the strongest forms of denial is reversal, illustrated in example (17). In this case, the members of the out-group (i.e., the Black people) tend to be represented as the ones who are intolerant, and the ones belonging to the in-group (i.e., the White people) as the victims.

(17) *99.9% dos negros são racistas fanaticos sempre estao esperando uma oportunidade pra fazer mal a um branco.*

[5] The most frequent words and collocations are available at: https://hate-covid.inesc-id.pt/?page_id=265.

[6] When referring to people of African descent in Portugal, the term "preto" is often viewed as offensive, while "negro" is generally accepted within the Afro-descendant community. To convey the offensive connotation in our translation, we have used the term 'N-word.'

> '99.9% of black people are fanatical racists; they are always waiting for an opportunity to hurt a White person.'

Racial and sexual stereotyping and prejudice may also be signaled by the frequent use of several nouns in all the subsets. For example, the nouns *problema* ('trouble') and *polícia* ('policeman') may evoke the criminality stereotyping and over-policing towards the Afro-descendant group (Council of Europe 2021). Regarding the Roma community, in addition to the noun *problema* ('trouble'), the presence of terms like *lei* ('law'), *direito* ('rights') and *medo* ('fear') is remarkable, associated with Anti-Roma rhetoric across Europe, which perpetuates the stereotypes that the Roma are dangerous, and live perpetually outside the law (Breazu and Machin 2019, 2022; Buturoiu and Corbu 2020; Chovanec 2021; Erjavec 2001).

With respect to the LGBTQ+ community, it is worth mentioning the presence of pejorative nouns like *vergonha* ('shame'), and negative adjectives such as *mentiroso* ('liar'), and *vagabundo* ('vagabond'). The use of these terms primarily serves to strengthen the metaphor of amorality, which suggests a deficiency of ethical or moral conduct (Baider, Constantinou, and Petrou 2017; Munt 2019). This reinforces the notion that the dominant social structure in Europe continues to discriminate against individuals who do not conform to the heteronormative standard (Ben Chikha 2021).

The most frequent proper nouns in comments targeting the Afro-descendant group are *Mário* (*Machado*) and *Mamadu* (*Ba*). Mário Machado was considered one of the most influential far-right advocates in Portugal, who has been convicted several times of various hate crimes. On the other hand, *Mamadu Ba* is a prominent Portuguese-Senegalese anti-racist, political activist. While the former is mostly positively mentioned in hatred messages, the latter is often associated with negative sentiment (see Examples 18 and 19, respectively).

(18) *grande Mário. votava em ti para presidente deste país . devias pensar nisso. tamos entregues a bicharada. abraco*
'great Mário. I would vote for you for president of this country. you should think about it. we are in the hands of bugs. hug'

(19) *O porco do mamadu ? Devíamos fazer o mesmo expulsar tudo daqui para fora*
'The slob mamadu ? We should do the same kicking everything out of here'

Regarding the comments targeting the Roma community, we observe the high occurrence of named entities such as *Chega*, a national conservative, far-right populist political party in Portugal, and *André Ventura*, the leader of said party, who has been publicly discriminating the minorities in Portugal, particularly the Roma community (Magano and Mendes 2021).

With respect to verbs, in addition to auxiliary (e.g., *ser* and *estar* 'to be'), support (e.g., *ter* 'to have', *fazer* 'to make', and *haver* 'to exist') and modal (e.g., *poder* 'can' and *dever* 'should') verbs, the occurrence of volitive (*querer* 'want') and locative verbs, including static (e.g., *viver* 'to live', *existir* 'to exist') and dynamic verbal forms (e.g., *ir* 'to go', *vir* 'to come', *deixar* 'to leave', *partir*, 'to go') is noteworthy. Locative verbs are frequently found in the scope of frozen and semi-frozen idiomatic expressions, such as *vai para a tua terra*, ('go back home'), *volta para África* ('go back to Africa') or *volta para selva* ('go back to the jungle'), especially in comments labeled as conveying overt hate speech. In this case, locative expressions are used to invoke a certain Portuguese identity, which is deeply shaped by the legacy of colonialism, reinforcing the exclusion of out-groups.

When looking at the most frequent bigrams, there is a remarkable prevalence of negative constructions in comments attacking all the target groups, such as *não ter* ('not to have'), *não ser* ('not to be'), and *não poder* ('not to be able'). If we restrict our selection to the main verbs, the presence of *não saber* ('not to know') in all the collections is also noteworthy.

The inspection of trigrams shows common patterns in comments targeting all groups. In this case, those expressions may be related to the textual realization of racial stereotypes; for example, Afro-descendants are often stereotyped as being poor (*não ter dinheiro*, 'do not have money'), and lazy (*não fazer nada* 'do nothing', *não querer trabalhar* 'do not want to work', *ninguém fazer nada* 'nobody does anything'). In addition to these stereotypes, Roma people are also commonly portrayed as beggars, criminals and profiteers, which may be inferred from expressions such as *não pagar impostos* ('do not pay taxes').

Moreover, it is interesting to observe the generalized use of expressions supporting the leader of the far-right wing political party Chega, André Ventura (*força andré ventura* 'go for it André Ventura', *votar andré ventura* 'vote andré ventura'), and the explicit negation of racism (*não haver racismo* 'racism does not exist', *não ser racista*, 'not to be racist').

Indeed, the denial of racism is quite visible in Portuguese public and political spheres (Magano and Mendes 2021), and it has dramatically increased with the representation of the right-wing populist party Chega in the national Parliament, with several social movements being organized around the motto of "Portugal não é racista" ('Portugal is not racist').

5.3 Rhetorical and discursive strategies underlying covert hate speech

To investigate the realization of covert hate speech in our corpus, the cases labeled by most annotators as conveying covert hate speech were examined. Two major patterns emerged from the data:

i. Covert hate speech is expressed through fallacious argumentation, where the poster tries to manipulate the audience by means of an emotional appeal to fear, and the (implicit) command to take an action.
ii. Covert hate speech is often reproduced through figurative language, such as verbal irony and sarcasm, reinforcing negative stereotyping associated with those groups.

Both strategies are used to promote beliefs that could be used to legitimate hatred against the targeted groups.

4.3.1 Appeal to action

Covert hate speech is often anchored in poor and wrong argumentation, deliberately used for manipulating the audience's opinion. According to pragma-dialectical theory, argumentation is part of an explicit or implicit dialogue, in which one participant aims at convincing the other of the acceptability of their standpoint (van Eemeren and Grootendorst 2004). When the dialogue is implicit, the speaker must elaborate on by foreseeing any doubts or criticisms that the hearers may have. The analysis proposed by pragma-dialectics is based on a dialog model, in which an evaluation of critical discussion is required, for instance, to detect fallacious arguments (van Eemeren and Grootendorst 1987; van Eemeren and Garssen 2023). In Examples (20) and (21), the poster tries to manipulate the audience by emotionally appealing to fear (Wodak 2020), which is regarded as one of the major causes of prejudice and discrimination (Stephan and Stephan 2000). In fact, some discursive and argumentation strategies characterizing the populist right-wing rhetoric (Serafis et al. 2023; Wodak 2015, 2020) seem to be effective in spreading covert hate speech in our corpus. As exemplified in Example (20), the suggestion of threat is often anchored in the concept of reversal racism, which captures the perception of White victimhood. Frequently, the speaker uses hyperboles to emphasize their arguments and highlight the existence of an impending threat that requires action, as exemplified in Example (21).

(20) *Informo-te então que eles têm mais direitos e menos deveres do que um português trabalhador.*
‘So I then inform you that they have more rights and fewer duties than a Portuguese worker.’

(21) Eles sao mais a votar do que “Nós”. já passamos o ponto sem retorno
‘They are more voting than “Us”. we are past the point of no return’

As illustrated in Example (22), in our corpus, the appeal to act is often materialized through the invocation of Portuguese political parties and actors (e.g., *Chega* and *PNR*) that espouse, albeit subtly, White supremacy, and share the core ideological characteristics of the populist radical right family, including *nativism*, *authoritarianism* and *populism* (Krzyżanowski and Ekström 2022).

(22) *CHEGA ou PNR! Ou os dois! Mas rápido!*
‘CHEGA or PNR! Or both! But fast!’

By posting that message, we can infer that the speaker endorses the apparent promotion of hostility and discrimination against the Afro-descendant community that is portrayed either in the video or its caption. The proposed solution to address the reported problems is to vote for the far-right political parties Chega and PNR, who try to eradicate these communities.

This appeal is often reproduced by means of elementary exclamatory and imperative sentences, frequently intensified by (marked) punctuation (especially exclamation marks) and capitalization. Other forms of intensification include the use of quantifiers, where the *out-groups* are usually overestimated, and the *in-groups* underestimated (see Examples 20 and 21).

4.3.2 Irony, sarcasm and negative stereotyping

In covert hate speech, the targeted groups are often reduced to their perceived characteristics, which are frequently explored “creatively” to intentionally deny racism, as illustrated in Example (23).

(23) *As vacas são racistas porque só dão leite branco*
´Cows are racist because they only give white milk’

This is an example of indirect hostility and discrimination towards people perceived as Black. To deny the racism that this community often faces, the speaker parodies the situation, claiming that cows should be also accused of being racist, since they only give white (instead of black) milk. This statement suggests that the use of animal metaphors in certain jokes may be interpreted as a form of subtle dehumanization. Such jokes create a negative image of the targeted group, portraying them as inferior or stupid. This reinforces racial hierarchies and may lead to social bonding among those who share similar views (Billig 2001; Hodson and MacInnis 2016; Weaver 2011).

In fact, users often resort to irony and sarcasm to diminish or ridicule the target groups, through the exploitation of negative stereotyping commonly associated with those groups. For example, in the comments illustrated in Examples (24) and (25), the targeted groups are characterized as lazy, dishonest and social benefit dependent. In both cases, the Portuguese Government, and respective representatives (sharing left-wing ideologies), are also addressed in those comments, and explicitly or implicitly accused of being their allies and condoning social injustice. In Example (24), the speaker suggests that the target group (in this case, the African descendant community) only procreates to receive social benefits. Argumentation is based on a subtle and ironic provocation, relying on the idea that anyone who does not want to work should have children and apply for social integration income. For its part, in Example (25), the message is characterized by a more hurtful and aggressive language, being chiefly offensive to some of the targets indirectly mentioned in the message. In particular, the speaker uses the racist term *monhé*[7] to refer to the Portuguese prime minister, reducing him to his perceived group membership. Moreover, the speaker uses a combination of words to refer to the former Afro-descendant Portuguese deputy, Joacine Katar Moreira. Specifically, the adjective “gaga” (‘stutterer’ in Portuguese), found in the discontinuous mention of Lady Gaga, was chosen by the author to disparage the deputy, who experiences speech disfluency. In addition, the reference to her phenotype is concretized using a recurrent dehumanizing metaphor, comparing (or downgrading) afro-descendants to monkeys.

(24) *fazer o quê? é legal. tens filho e não queres fazer pisso, recebes rsi*
‘what to do? it’s legal. you have a son and you don’t want to do anything, you receive a social integration income’

(25) *Agora ja esta la mais uma para os defender... a lady macaca gaga, é ativista veemente, ainda vai propor ao monhé que lhes seja atribuido subsidio de ferias e subsidio de natal*
‘Now is another person there to defend them [people receiving social integration grants]... lady monkey gaga [reference to an Afro-descendant Portuguese deputy], she is a vehement activist, she will propose to the Paki [reference to the Portuguese prime-minister, who is of Indian-descent] to give them holiday and Christmas bonuses’

In terms of classification, the comment illustrated in Example (25) was assigned both to the labels of overt and covert hate speech; the first category applies to the individuals

[7] In Portugal, this insulting and contemptuous term is used to refer to people of Indian heritage.

who are explicitly denigrated because of their attributed group membership; the latter applies to African descent community, who is negatively stereotyped as lazy, and get privileged access to limited socio-economic resources, and benefits that are not available to the *in-group.*

In summary, our corpus suggests that irony and sarcasm, much like the findings of Baider and Constantinou (2020), can have distinct socio-pragmatic purposes. Specifically, they can be used to insult, humiliate, and ridicule targeted communities, and promote negative sentiments and emotions toward them through negative stereotyping. This can, in turn, legitimate hate speech and contribute to its normalization.

6. Concluding remarks

This study enabled us to address the research questions outlined in the Introduction. On the one hand, we aimed at understanding the characteristics of OHS against the Afro-descendant, Roma, and LGBTQ+ communities in Portugal. On the other, we sought to identify the primary rhetorical mechanisms utilized in covert hate speech. Rather than presenting the distribution of hate speech across all groups, our corpus analysis revealed significant differences between them. Specifically, the data showed a higher prevalence of hate comments targeting the Roma community compared to the other groups. Moreover, the resistance and opposition (i.e., the counter-narratives) to hate speech are much less expressive in comments targeting the Roma community than in the remaining target groups. Indeed, this study suggests that the Roma community, who has been targeted as the most rejected minority group in all European countries, is also a victim of severe racism in the Portuguese online context. The analysis of hate comments directed at the Roma community reinforces the use of generalized negative stereotypes associated with this ethnic group, such as criminality and the receipt of undeserved benefits, which are quite common in Europe. Despite being recognized as one of the primary targets of OHS in the European context, the LGBTQ+ community received the smallest proportion of hate comments in our corpus. In fact, more than half of the comments referring to this group were either offensive (often arising from exchanges of insults among social media users) or aimed at countering hate speech directed at this community.

The discrimination faced by people of African descent is rooted in the negative racial stereotypes perpetuated against them, which contribute to the normalization of hate speech directed at this community in both mainstream and social media. In comparison to the other targeted groups, people of African descent have received the highest percentage of extremely negative comments. This is consistent with concerns raised by the United Nations Working Group of Experts on People of African Descent regarding the prevalence of racial discrimination and human rights violations in Portugal, which were reported during their last official visit. However, our analysis of counter-speech aimed at this group indicates that they are more likely to respond to OHS.

Although there is a growing effort to create large and fine-grained datasets that explicitly cover implicit or subtle forms of hate (e.g., ElSherief et al. 2021), detecting covert hate speech remains a challenging task for NLP systems. In fact, the recognition

of hate speech will fail if systems are not capable of identifying not only what is uttered, but also what is intended by that, and what effect it has on the targeted groups.
Our study shows that, with the exception of the comments targeting the LGBTQ+ community, covert hate speech is the most representative type in CO-HATE, which reinforces the importance of investigating the main discursive and rhetorical strategies underpinning covert hate speech. Despite its inherent complexity, the efforts to model covert hate speech are crucial, because it has the same intention to discriminate against a target and has potentially the same effect on the recipient as overt verbal harassment (Baider 2022).

Importantly, the analysis of covert hate speech messages targeting each community reveals that covert hate speech is often reproduced by means of rhetorical strategies, including irony and sarcasm. These rhetorical devices rely on negative stereotyping and generalizations about hate speech targets, cloaked in the mask of humor. In addition, covert hate speech is often expressed using (direct and indirect) appeals to the audience's emotion, namely by invoking political entities associated with extreme right-wing populist ideology, rising across Europe and beyond. As in overt hate speech, the polarization of in-groups and out-groups is realized through the explicit and implicit reference to the dominant and the dominated (or vulnerable) groups, whose positions of dominance and power are intentionally reversed.

Based on the idea that perceptions of racism, and other related beliefs based on prejudice and discrimination, can be influenced by group membership (Carter and Murphy 2015), we deliberately included in our annotation team members belonging to the targeted groups. The lack of agreement among annotators highlights the fragility of existing hate speech automated detection systems, which usually do not take into consideration the multiplicity of perspectives, particularly the ones of the members belonging to targeted groups. Moreover, this study can also provide important clues to investigate potential annotation bias (linked to the annotators' profile).

Following the *Perspectivist Data Manifesto*,[8] in CO-HATE disagreement will not be treated as noise; instead, it reflects potential nuances in interpretation of subjective data, which may be directly related to the background of the annotators and group membership. In line with the results from the experiments reported by Akhtar et al. (2020), we think that the performance of hate speech detection models can be improved by considering inclusive approaches that contemplate the multiplicity of perspectives on such a complex and subjective phenomenon.

Furthermore, we believe the information here reported provides important insights on the basis of which to approach OHS detection and will allow for a deeper understanding of the dynamics of online hate speech in Portugal, particularly regarding the most representative marginalized groups. Moreover, the corpus created will be an important resource for those interested in developing methods for detecting both overt and covert hate speech, and other related phenomena, like counter-speech and offensive speech, in Portuguese.

[8] https://pdai.info/

Future research could involve extending this study to other social media platforms (e.g., Twitter) and targeted communities (e.g., migrants and refugees). Moreover, we intend to include more annotators in future annotation experiments, taking into consideration their social identity, aiming at assessing inter-community agreement.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This research was funded by Fundação para a Ciência e a Tecnologia (FCT) under the Project Hate-Covid (Ref. 759274510), and is supported by INESC-ID (UIDB/50021/2020), and ITI-LARSyS (UIDP/50009/2020-IST-ID). D. Caled is also funded by FCT, under a PhD scholarship (SFRH/BD/145561/2019).

References

Achim, Viorel. 2004. *The Roma in Romanian History*. Budapest, Hungary: Central European University Press. doi:10.1515/9786155053931.

Akhtar, Sohail, Valerio Basile, and Viviana Patti. 2020. “Modeling Annotator Perspective and Polarized Opinions to Improve Hate Speech Detection.” In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 151–54. doi:10.1609/hcomp.v8i1.7473.

Albelda Marco, Marta. 2022. "Rhetorical Questions as Reproaching Devices." *Journal of Language Aggression and Conflict*. doi:10.1075/jlac.00077.alb

Assimakopoulos, Stavros, Fabienne Baider, and Sharon Millar. 2017. *Online Hate Speech in the European Union: A Discourse-analytic Perspective*. New York: Springer Nature. doi:10.1007/978-3-319-72604-5.

Attardo, Salvatore. 2020. "Irony as Relevant Inappropriateness." *Journal of Pragmatics 32*(6): 793-826. doi:10.1016/S0378-2166(99)00070-3.

Baider, Fabienne. 2020. "Pragmatics lost?: Overview, Synthesis and Proposition in Defining Online Hate Speech." *Pragmatics and Society* 11(2): 196-218. doi:10.1075/ps.20004.bai.

Baider, Fabienne. 2022. "Covert Hate Speech, Conspiracy Theory and Anti-semitism: Linguistic Analysis versus Legal Judgment." *International Journal for the Semiotics of Law* 35(6): 2347-2371. doi:10.1007/s11196-022-09882-w.

Baider, Fabienne, and Christina Romain. *Forthcoming*. “Irony, Sarcasm and other Playful Devices Used in Online Covert Hate Speech.”, *Language and Social Life*, de Gruyter.

Baider, Fabienne, Anna Constantinou, and Anastasia Petrou. 2017. “Metaphors Related to Othering the Non-natives”. In *Online Hate Speech in the European Union: A Discourse-Analytic Perspective,* edited by Stavros Assimakopoulos, Fabienne H. Baider, and Sharon Millar, 38–42. Berlin: Springer.

Baider, Fabienne, and 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. doi:10.1075/jlac.00040.bai.

Baker, Paul, and Tony McEnery. 2005. "A Corpus-based Approach to Discourses of Refugees and Asylum Seekers in UN and Newspaper Texts." *Journal of Language and Politics* 4(2): 197-226. doi:10.1075/jlp.4.2.04bak.

Baker, Paul, Costas Gabrielatos, Majid KhosraviNik, Michał Krzyżanowski, Tony McEnery, and Ruth Wodak. 2008. "A Useful Methodological Synergy? Combining Critical Discourse Analysis and Corpus Linguistics to Examine Discourses of Refugees and Asylum seekers in the UK Press." *Discourse & Society* 19(3): 273-306. doi:10.1177/095792650808896.

Basile, Valerio, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. “SemEval-2019 Task 5: Multilingual Detection of Hate Speech against Immigrants and Women in Twitter.” In *Proceedings of the 13th International Workshop on Semantic Evaluation*, 54–63. Minneapolis, Minnesota, USA: Association for Computational Linguistics. doi:10.18653/v1/S19-2007.

Ben Chikha, Fourat. 2021. *Combating Rising Hate against LGBTI People in Europe*. http://tiny.cc/kod6vz

Benesch, Susan, Derek Ruths, Kelly Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016. “Counterspeech on Twitter: A Field Study.” A Report for Public Safety Canada under the Kanishka Project. doi:10.15868/socialsector.34066.

Bhat, Prashanth, and Ofra Klein. 2020. "Covert Hate Speech: White Nationalists and Dog Whistle Communication on Twitter." In *Twitter, the Public Sphere, and the Chaos of Online Deliberation*, edited by Gwen Bouvier, and Judith E. Rosenbaum, 151-172. Cham: Palgrave Macmillan. doi:10.1007/978-3-030-41421-4_7.

Billig, Michael. 2001. "Humour and Hatred: The Racist Jokes of the Ku Klux Klan." *Discourse & Society* 12(3): 267-289.

Breazu, Petre, and David Machin. 2019. "Racism toward the Roma through the Affordances of Facebook: Bonding, Laughter and Spite." *Discourse & Society* 30(4): 376-394. doi:10.1177/09579265198373.

Breazu, Petre, and David Machin. 2022. "Using Humor to Disguise Racism in Television News: The Case of the Roma." *HUMOR* 35(1): doi:10.1515/humor-2021-0104.

Brindle, Andrew, Tony McEnery, and Michael Hoey. 2016. *The Language of Hate: A Corpus Linguistic Analysis of White Supremacist Language*. New York and London: Routledge.

Buturoiu, Dana Raluca, and Nicoleta Corbu. 2020. "Exposure to Hate Speech in the Digital Age. Effects on Stereotypes about Roma People." *Journal of Media Research* 13(2). doi:10.24193/jmr.37.1.

Cádima, Francisco Rui, Carla Baptista, Marisa Silva, and Patrícia Abreu. 2021. *Monitoring Media Pluralism in the Digital Era: Application of the Media Pluralism Monitor in the European Union, Albania, Montenegro, The Republic of North Macedonia, Serbia & Turkey in the Year 2020. Country Report: Portugal*. doi:10.2870/818313.

Carter, Evelyn R., and Mary C. Murphy. 2015. "Group-based Differences in Perceptions of Racism: What Counts, to Whom, and Why?." *Social and Personality Psychology Compass* 9(6): 269-280. doi:10.1111/spc3.12181.

Casa-Nova, Maria José. 2021. "Reflecting on Public Policies for Portuguese Roma since Implementation of the NRIS: Theoretical and Practical Issues." *Journal of Contemporary European Studies* 29(1): 20-32. doi:10.1080/14782804.2021.1877119.

Chovanec, Jan. 2021. "'Re-educating the Roma? You Must Be Joking...': Racism and Prejudice in Online Discussion Forums." *Discourse & Society* 32(2): 156-174. doi:10.1177/095792652097038.

Council of Europe. 2021. *Combating racism and racial discrimination against people of African descent in Europe. Round-table with Human Rights Defenders Organised by the Office of the Council of Europe Commissioner for Human Rights*. http://tiny.cc/ood6vz.

Dahiya, Snehil, Shalini Sharma, Dhruv Sahnan, Vasu Goel, Emilie Chouzenoux, Víctor Elvira, Angshul Majumdar, Anil Bandhakavi, and Tanmoy Chakraborty. 2021. "Would Your Tweet Invoke Hate on the Fly? Forecasting Hate Intensity of Reply Threads on Twitter." In *Proceedings of the ACM SIGKDD 2021*, 2732–42. Virtual Event Singapore: ACM. doi:10.1145/3447548.3467150.

Davidson, Thomas, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. "Automated Hate Speech Detection and the Problem of Offensive Language." In *Proceedings of the International AAAI Conference on Web and Social Media*, 512–15. Montreal, Canada: AAAI. doi:10.1609/icwsm.v11i1.14955.

Dynel, Marta 2017. "Academics vs. American Scriptwriters vs. Academics: A Battle over the Etic and Emic 'Sarcasm' and 'Irony' Labels." *Language & Communication* 55: 69–87. doi:10.1016/j.langcom.2016.07.008.

Dynel, Marta. 2018a. *Irony, Deception and Humour: Seeking the Truth about Overt and Covert Untruthfulness.* Berlin: De Gruyter Mouton. doi:10.1515/9781501507922.

Dynel, Marta. 2018b. "Deconstructing the Myth of Positively Evaluative Irony." In *The Pragmatics of Irony and Banter*, edited by Manuel Jobert, and Sandrine Sorlin, 1-17. Berlin: John Benjamins.

Dynel, Marta. 2019. "Ironic Intentions in Action and Interaction." *Language Sciences* 75: 1-14. doi:10.1016/j.langsci.2019.06.005.

ElSherief, Mai, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. "Latent Hatred: A Benchmark for Understanding Implicit Hate Speech." In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 345–63. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. doi:10.18653/v1/2021.emnlp-main.29.

Erjavec, Karmen. 2001. "Media Representation of the Discrimination against the Roma in Eastern Europe: The Case of Slovenia." *Discourse & Society* 12(6): 699-727. doi:10.1177/0957926501012006001.

European Commission. 2018a. *Afrophobia: Acknowledging and Understanding the Challenges to Ensure Effective Responses*. http://tiny.cc/lpd6vz.

European Commission. 2018b. *Antigypsyism: Increasing its Recognition to Better Understand and Address its Manifestations*. http://tiny.cc/rpd6vz.

ECRI. 2018. *ECRI General Policy Recommendation N°7 (revised) on National Legislation to Combat Racism and Racial Discrimination.* http://tiny.cc/uqd6vz.

European Union Agency for Fundamental Rights. 2020. *A Long Way to Go for LGBTQ Equality*. http://tiny.cc/3qd6vz.

Fortuna, Paula, João Silva, Juan Soler Company, Leo Wanner, and Sérgio Nunes. 2019. "A Hierarchically-Labeled Portuguese Hate Speech Dataset." In *Proceedings of the*

Third Workshop on Abusive Language Online, 4–104. Florence, Italy: Association for Computational Linguistics.

Fortuna, Paula, and Sérgio Nunes. 2018. "A Survey on Automatic Detection of Hate Speech in Text." *ACM Computing Surveys* 51(4): 1-30. doi:10.1145/3232676.

Geyer, Klaus, Eckhard Bick, and Andrea Kleene. 2022. "'I Am No Racist, but…': A Corpus-Based Analysis of Xenophobic Hate Speech Constructions in Danish and German Social Media Discourse." In *The Grammar of Hate: Morphosyntactic Features of Hateful, Aggressive, and Dehumanizing Discourse*, edited by Natalia Knoblock, 241-261. Cambridge: Cambridge University Press.

Habernal, Ivan, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. "Before Name-calling: Dynamics and Triggers of Ad Hominem Fallacies in Web Argumentation." In *Proceedings of the NAACL HLT 2018*, 386–396, New Orleans, Louisiana: Association for Computational Linguistics.

Han, Chung-hye. 2002. "Interpreting Interrogatives as Rhetorical Questions." *Lingua* 112(3): 201-229. doi:10.1016/S0024-3841(01)00044-4.

Hancock Alfaro, Ange-Marie. 2022. "When Words don't Disappear: An Intersectional Analysis of Hate Speech" In *Citizenship on the Edge: Sex/Gender/Race*, edited by Nancy J. Hirschmann, and Deborah A. Thomas, 19-40. Philadelphia: University of Pennsylvania Press.

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

Hodson, Gordon, and Cara C. MacInnis. 2016. "Derogating Humor as a Delegitimization Strategy in Intergroup Contexts." *Translational Issues in Psychological Science* 2(1): 63-74. doi:10.1037/tps0000052.

Krobová, Tereza, and Jan Zàpotocký. 2021. "'I Am Not Racist, But...': Rhetorical Fallacies in Arguments about the Refugee Crisis on Czech Facebook." *Journal of Intercultural Communication* 21(2): 58-69. doi:10.36923/jicc.v21i2.14.

Krzyżanowski, Michał, and Mats Ekström. 2022. "The Normalization of Far-right Populism and Nativist Authoritarianism: Discursive Practices in Media, Journalism and the Wider Public Sphere/s." *Discourse & Society* 33(6): 719-729. doi:10.1177/09579265221095406.

Kumar, Ritesh, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. "Benchmarking Aggression Identification in Social Media." In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, 1-11, Santa Fe, New Mexico: USA. Association for Computational Linguistics.

João Leite, Diego Silva, Kalina Bontcheva, and Carolina Scarton. 2020. "Toxic Language Detection in Social Media for Brazilian Portuguese: New Dataset and Multilingual Analysis". *arXiv preprint arXiv:2010.04543*.

Maeso, R. Silvia. 2021. *O Estado do Racismo em Portugal: Racismo Antinegro e Anticiganismo no Direito e nas Políticas Públicas*. Lisbon: Tinta-da-China.

Macagno, Fabrizio. 2022. "Argumentation Profiles and the Manipulation of Common Ground. The Arguments of Populist Leaders on Twitter." *Journal of Pragmatics* 191: 67-82. doi:10.1016/j.pragma.2022.01.022.

Magano, Olga, and Maria Manuela Mendes. 2021. "Structural Racism and Racialization of Roma/Ciganos in Portugal: The Case of Secondary School Students during the COVID-19 Pandemic." *Social Sciences* 10(6): 1-14. doi:10.3390/socsci10060203.

Magu, Rijul, and Jiebo Luo. 2018. "Determining Code Words in Euphemistic Hate Speech Using Word Embedding Networks." In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, 93–100. Brussels, Belgium: Association for Computational Linguistics. doi:10.18653/v1/W18-5112.

Mathew, Binny, Navish Kumar, Pawan Goyal, Pawan, and Animesh Mukherjee. 2018. “Analyzing the Hate and Counter Speech Accounts on Twitter.” arXiv preprint arXiv:1812.02712.

McEnery, Tony, and Andrew Hardie. 2011. *Corpus Linguistics: Method, Theory and Practice*. Cambridge Textbooks in Linguistics. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511981395.

McEnery, Tony, and Vaclav Brezina. 2022. *Fundamental Principles of Corpus Linguistics*. Cambridge: Cambridge University Press. doi:10.1017/9781107110625.

Mendes, Mariana S. 2021."‘Enough’of What? An Analysis of Chega’s Populist Radical Right Agenda." *South European Society and Politics* 26(3): 329-353. doi:0.1080/13608746.2022.2043073.

Migration and Home Affairs (n.d.) *Unesco Glossary of Migration-related Terms*. http://tiny.cc/4rd6vz.

Munt, Sally R. 2019. "Gay Shame in a Geopolitical Context." *Cultural Studies* 33(2): 223-248. doi:10.1080/09502386.2018.1430840.

Paz, María Antonia, Julio Montero-Díaz, and Alicia Moreno-Delgado. 2020. "Hate Speech: A Systematized Review." *Sage Open* 10(4). doi:10.1177/2158244020973022.

De Pelle, Rogers Prates, and Viviane Moreira. 2017. “Offensive Comments in the Brazilian Web: A Dataset and Baseline Results.” In *Proceedings of BraSNAM*. Porto Alegre. SBC. doi:10.5753/brasnam.2017.3260.

Pohjonen, Matti, and Sahana Udupa. 2017. "Extreme Speech Online: An Anthropological Critique of Hate Speech Debates." *International Journal of Communication* 11: 1173-1191.

Poletto, Fabio, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. "Resources and Benchmark Corpora for Hate Speech Detection: A Systematic Review." *Language Resources and Evaluation* 55(2): 477-523. doi:0.1007/s10579-020-09502-8.

Poletto, Fabio, Marco Stranisci, Manuela Sanguinetti, Viviana Patti, and Cristina Bosco. 2017. “Hate Speech Annotation: Analysis of an Italian Twitter Corpus.” In *Proceedings of CLiC-It 2017*, edited by Roberto Basili, Malvina Nissim, and Giorgio Satta, 263–68. Academia University Press. doi:10.4000/books.aaccademia.2448.

Rieger, Diana, Anna Sophie, Maximilian Wich, Toni Kiening, and Georg Groh. 2021. "Assessing the Extent and Types of Hate Speech in Fringe Communities: A Case Study of Alt-right Communities on 8chan, 4chan, and Reddit." *Social Media + Society* 7(4). doi:10.1177/20563051211052906.

Sanguinetti, Manuela, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. "An Italian Twitter Corpus of Hate Speech against Immigrants." In *Proceedings of the LREC 2018*, 2798-2805. Miyazaki, Japan: ELRA.

Schmidt, Anna, and Michael Wiegand. 2017. “A Survey on Hate Speech Detection Using Natural Language Processing.” In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 1–10. Valencia, Spain: Association for Computational Linguistics. doi:10.18653/v1/W17-1101.

Sellars, Andrew. 2016. "Defining Hate Speech." Berkman Klein Center Research Publication 2016-20: Boston Univ. School of Law, Public Law Research. doi:10.2139/ssrn.2882244.

Serafis, Dimitris, Franco Zappettini, and Stavros Assimakopoulos. 2023. "The Institutionalization of Hatred Politics in the Mediterranean: Studying Corpora of

Online News Portals during the European 'Refugee Crisis'." *Topoi* (2023): 1-20. doi:10.1007/s11245-023-09890-w.

Siegel, Alexandra A. 2020. "Online Hate Speech." In *Social Media and Democracy: The State of the Field, Prospects for Reform*, edited by Nathaniel Persily, and Joshua A. Tucker, 56–88. Cambridge: Cambridge University Press.

Stangor, Charles. 2016. The study of Stereotyping, Prejudice, and Discrimination within Social Psychology: A Quick History of Theory and Research. In *Handbook of Prejudice, Stereotyping, and Discrimination*, edited by Todd Nelson, 3–27. New York: Psychology Press.

Stephan, Walter S., and Cookie White Stephan. 2013. "An Integrated Threat Theory of Prejudice." In *Reducing Prejudice and Discrimination*, edited by Stuart Oskamp, 33-56. New York: Psychology Press.

Tindale, Christopher W. 2007. *Fallacies and Argument Appraisal*. 1st ed. Cambridge: Cambridge University Press.doi:10.1017/CBO9780511806544.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam/Philadelphia: John Benjamins.

Torres da Silva, Marisa. 2021. *Discurso de Ódio, Jornalismo e Participação das Audiências. Enquadramento, Regulação e Boas Práticas*. Lisboa: Almedina ERC.

Ullmann, Stefanie, and Marcus Tomalin. 2020. "Quarantining Online Hate Speech: Technical and Ethical Perspectives." *Ethics and Information Technology* 22(1): 69-80. doi:10.1007/s10676-019-09516-z.

van Dijk, Teun A. 1992. "Discourse and the Denial of Racism." *Discourse & Society* 3(1): 87-118.

van Dijk, Teun A. 1993. "Principles of Critical Discourse Analysis." *Discourse & Society* 4(2): 249-283.

van Eemeren, Frans, and Rob Grootendorst. 1987. "Fallacies in Pragma-dialectical Perspective." *Argumentation* 1: 283-301. doi:10.1007/BF00136779.

van Eemeren, Frans H., and Rob Grootendorst. 2004. *A Systematic Theory of Argumentation: The Pragma-dialectical Approach*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511616389.

van Eemeren, Frans., and Bart Garssen. 2023. "The Pragma-Dialectical Approach to the Fallacies Revisited." *Argumentation* 1-14. doi:10.1007/s10503-023-09605-w.

Vargas, Francielle, Isabelle Carvalho, Fabiana Góes, Thiago Pardo, and Fabrício Benevenuto. 2022. "HateBR: A Large Expert Annotated Corpus of Brazilian Instagram Comments for Offensive Language and Hate Speech Detection." In *Proceedings of LREC 2022*, 7174–83. Marseille, France: ELRA. https://aclanthology.org/2022.lrec-1.777.

Walton, Douglas N. 1996. "Practical Reasoning and the Structure of Fear Appeal Arguments". *Philosophy & Rhetoric* 29(4): 301-313.

Waseem, Zeerak. 2016. "Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter." In *Proceedings of the First Workshop on NLP and Computational Social Science*, 138–42. Austin, Texas: Association for Computational Linguistics. doi:10.18653/v1/W16-5618.

Weaver, Simon. 2011. "Jokes, Rhetoric and Embodied Racism: A Rhetorical Discourse Analysis of the Logics of Racist Jokes on the Internet." *Ethnicities* 11(4): 413-435. doi:10.1177/1468796811407755.

Wiegand, Michael, Melanie Siegel, and Josef Ruppenhofer. 2019. "Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language." In *Proceedings of GermEval 2018 Workshop*, 14th Conference on Natural Language

Processing (KONVENS 2018), 1-10. Vienna, Austria: Austrian Academy of Sciences.

Wodak, Ruth. 2015. *The Politics of Fear: What Right-wing Populist Discourses Mean*. London: Sage. doi.org/10.4135/9781446270073.

Wodak, Ruth, and Martin Reisigl. 2015. "Discourse and Racism." In *The Handbook of Discourse Analysis*, 2nd Edition, edited by Deborah Tannen, Heidi Hamilton, and Deborah Schiffrin, 576-596. West Sussex: John Wiley and Sons.

Wodak, Ruth. 2020. "The Politics of Fear: The Shameless Normalization of Far-right Discourse." *The Politics of Fear*. London: Sage.

Paula Carvalho
INESC-ID Lisboa
R. Alves Redol 9, 1000-029 Lisboa, Portugal
paula.c.carvalho@inesc-id.pt
ORCID: 0000-0003-2884-1250

Danielle Caled
INESC-ID Lisboa
R. Alves Redol 9, 1000-029 Lisboa, Portugal
dcaled@inesc-id.pt
ORCID: 0000-0003-1397-531X

Cláudia Silva
ITI-LARSyS - IST
Av. Rovisco Pais 1, 1049-001 Lisboa, Portugal
claudiasilva01@tecnico.ulisboa.pt
ORCID: 0000-0002-5334-3424

Fernando Batista
ISCTE - Instituto Universitário de Lisboa
Av. das Forças Armadas, 1649-026 Lisboa, Portugal
fernando.batista@iscte-iul.pt
ORCID: 0000-0002-1075-0177

Ricardo Ribeiro
ISCTE - Instituto Universitário de Lisboa
Av. das Forças Armadas, 1649-026 Lisboa, Portugal
ricardo.ribeiro@iscte-iul.pt
ORCID: 0000-0002-2058-693X

Date received: 23 July 2022
Date accepted: 20 April 2023