





ARTICLE



<https://doi.org/10.1057/s41599-025-05392-9>

OPEN

Unpacking online hate speech in Portuguese social media: a social-psychological and linguistic-discursive approach

Rita Guerra^{1,10}, Paula Carvalho^{2,3,10}, Catarina Marques⁴, Margarida Carmona⁵, Rodrigo Sarroeira⁴, Fernando Batista^{5,9}, Ricardo Ribeiro^{5,9}, António Fonseca⁶, Sérgio Moro^{6,7} & Cláudia Silva⁸

Building on social psychology and language sciences, this research identified core social psychological, and linguistic-discursive features of online hate speech targeting racialized, migrant and LGBTI+ communities in two social media platforms in Portugal: YouTube, and Twitter/X. The research was based on the analysis of two annotated corpora comprising 24,739 YouTube comments and associated replies, and 29,758 contextualized tweets retrieved from 2775 conversations. Overall, the results, based on the detailed annotation framework developed in this study, revealed that i) online hate speech was mainly expressed in subtle ways (i.e., indirect hate speech); ii) the main underlying process of discrimination in both direct and indirect hate speech was outgroup derogation; iii) stereotypes, threats, and dehumanization were frequently used as discursive strategies to express online hate speech; iv) specific features, like emotions, often overlooked in hate speech annotated corpora, varied in their expression depending on the specific target community; v) the use of some discursive strategies, such as realistic and symbolic threats, seem to be dependent not only on the target community but also the social media platform; vi) discursive strategies and emotions mobilized in hate speech were correlated with specific rhetorical devices and fallacies. These findings provide valuable insights into the complex landscape of online hate speech and highlight the importance of interdisciplinary, context and culturally sensitive approaches in understanding this phenomenon.

¹Iscte - Instituto Universitário de Lisboa, CIS, Lisboa, Portugal. ²Universidade de Aveiro, Centro de Línguas Literaturas e Culturas (CLLC), Aveiro, Portugal. ³INESC-ID Lisboa, INESC, Lisboa, Portugal. ⁴Iscte - Instituto Universitário de Lisboa, Business Research Unit (BRU-ISCTE), Lisboa, Portugal. ⁵Iscte - Instituto Universitário de Lisboa, Lisboa, Portugal. ⁶Iscte - Instituto Universitário de Lisboa, ISTAR, Lisbon, Portugal. ⁷University of Jordan, Amman, Jordan. ⁸ITI-LARsYS and IST, Lisbon, Portugal. ⁹Present address: INESC-ID Lisboa, INESC, Lisboa, Portugal. ¹⁰These authors contributed equally: Rita Guerra, Paula Carvalho. Unaffiliated: Margarida Carmona. ✉email: ana_rita_guerra@iscte-iul.pt

Introduction

Automated detection methods have received considerable attention as a methodological solution for identifying online hate speech on digital media platforms. However, these methods come with their own set of limitations, such as ethical issues related to privacy, censorship, bias, or limited linguistic and contextual knowledge (Parker and Ruths 2023; Udupa et al. 2023). Moreover, hate speech is often embedded in nuanced, subtle, indirect forms (Baider and Constantinou 2020; Carvalho et al. 2023; Parvaresh 2023), which poses significant technical challenges to its accurate automatic detection. While advanced language models have the potential to capture complex hate speech manifestations, their efficacy relies on access to extensive and diverse data, which is often limited, particularly in resource-scarce languages, like Portuguese. Additionally, existing data collections often exhibit biases and lack comprehensive coverage of complex and nuanced forms of hate speech (Sap et al. 2019; Schäfer 2023). Understanding hate speech, especially its subtle manifestations, requires a deep examination of its content within the social, historical, and cultural contexts (Udupa 2023). However existing hate speech detection systems often overlook the influence of social practice (Fortuna and Nunes 2018; Schmidt and Wiegand 2017). In this context, social practice is closely tied to the social and historical context, referring to the ways in which language is used in various social settings to construct, challenge, and perpetuate social realities (Wodak and Meyer 2006). Gaining insight into how individuals express hatred on social media is essential for understanding the multidimensional nature of hate speech and developing more effective strategies to address it.

In this work, we adopt a comprehensive, culturally sensitive approach, grounded in social psychology and language sciences, to characterize online hate speech (OHS) targeting vulnerable groups in Portugal, namely racialized (i.e., Roma and Afro-descendant), LGBTI+ and migrant communities. Our goal is to identify the core social psychological, and linguistic-discursive features of online hate speech, which can then inform the development of accurate automated OHS detection systems. Our interdisciplinary approach extends existing knowledge on online hate speech in three ways: a) we distinguish different forms of expressing hatred, namely direct (i.e., overtly biased, derogatory language) and indirect hate speech (i.e., implicit, inferred, meaning); b) when unpacking direct and indirect hate speech, we examine core cognitive (e.g., stereotypes), motivational (e.g., processes of discrimination), and emotional features (e.g., hate, anger), going beyond generic classification models; finally, c) we examine different discursive strategies underlying hate speech, and explore how they manifest depending on the target communities and social media platforms. This approach extends previous research conducted in Portuguese (e.g., Aguiar and Barbosa 2024; Carvalho et al. 2023), offering a more granular annotation process, involving a range of novel dimensions and categories from the field of social psychology, and covering two social networks (YouTube and Twitter). From a computational perspective, the development of hate speech detection models based on such granular and theoretically informed approaches allows for a better understanding of the multifaceted content and expressions of hate speech, contributing to improving the accuracy of linguistic models and the explainability of automated results. Ultimately, this approach also facilitates the design of effective interventions, such as the creation of tailored counter-narratives that consider not only the targets but also the type of speech, and the main social psychological and linguistic phenomena underlying hate messages.

What is hate speech?

The absence of a clear and consensual definition of online hate speech poses a significant challenge for research on this topic, hindering the assessment and understanding of its multiple manifestations, triggers, and motivations (Siegel 2020). As highlighted by Siegel, scholar definitions of hate speech range from “extremely broad to fairly narrow” (Siegel 2020, p. 3). The most narrowed definitions emphasize the intention to harm and incitement to violence as defining features of hate speech, whereas the broader definitions highlight the use of biased, derogatory, language. The discussion of broader and narrower conceptions of hate speech and its expressions is also reflected in a recent Recommendation of the Council of Europe (CM/Rec/2022/16) that emphasizes the need to differentiate between criminalized hate speech, non-criminal but subject to civil or administrative law, and offensive expressions requiring alternative responses (e.g., counter-speech). While we refrain from presenting a solution for this conceptual challenge, we propose a working definition for online hate speech based on the definition recommended by the Council of Europe (i.e., “*hate speech is 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*”) and on solid social psychological literature.

Building on Allport (1954) seminal work on antilocutions and the proposal that derogatory speech is a reflection of social categorization processes, we approach hate speech as an intergroup phenomenon, targeting groups or individuals because of their perceived membership in certain social groups. Specifically, we defined online hate speech as bias-motivated, derogatory language that spreads, incites, promotes, or justifies hatred, exclusion, and/or violence/aggression, targeting groups or individuals based on their group membership (e.g., perceived characteristics as ethnicity, race, sexual orientation, etc.). Next, we unpack the social psychological theoretical foundations underlying our intergroup approach to hate speech.

A social psychological approach to hate speech content and expression

The notion of bias-motivated phenomena stems from the well-established social psychological concept of intergroup bias, which refers to a differential responsiveness to ingroup and outgroup members, in terms of cognitions (e.g., stereotyping), attitudes (e.g., prejudice), and behavior (discrimination) (Dovidio and Gaertner 2010; Tajfel and Turner 1979). Intergroup bias generally manifests in two main forms with distinct motivations: ingroup favoritism, which involves a more positive evaluation and response to one’s ingroup members; and outgroup derogation, which is characterized by negativity, hostility, and an intent to harm the outgroup (Brewer 1999). This social psychological conceptualization is also in line with the ideological square proposed by van Dijk (1993), which relies on the use of discursive strategies to portray a positive self-presentation (Us) and negative other presentation (Others). Thus, discriminatory behavior may take the form of either favoring the ingroup or derogating and harming the outgroup, resulting in significant negative consequences for the targeted individuals and groups (Brewer 1999). A recent typology by Brewer (2016) proposed a more complex approach to intergroup bias, differentiating an additional form of discrimination: Type 1 refers to ingroup favoritism, (i.e., biased

treatment favoring the ingroup); Type 2 refers to outgroup derogation (i.e., negativity, antagonism, and harm directed against the outgroup); and Type 3 encompasses a more complex form involving both favoring the ingroup (Type 1) and derogating the outgroup (Type 2). This last type is associated with perceiving zero-sum situations, where gains for one group mean losses for the other and perceiving the outgroup as a threat to ingroup integrity, interests, and values.

In the context of online hate speech, where the intent to cause harm is a key feature, we conceptualize it as bias-motivated language, centered on the derogation of others based on their group membership, without necessarily a motivation to favor the ingroup. This conceptualization aligns with Brewer's definitions of Type 2 and/or Type 3 discrimination, as both involve the intent to harm or derogate the targeted outgroup.

Another core element of our conceptualization of online hate speech involves justifying, spreading, or inciting hatred. To unpack the hate within hatred speech, we rely on a well-established functional approach to hate (Fischer et al. 2018). Emotions are powerful predictors of intergroup attitudes and behaviors, and different emotions have different motivational functions, predicting differential behaviors (e.g., approach/confront, avoidance) (Cottrell and Neuberg 2005; Mackie and Smith 2015). Building on this, we examine if hate and other related emotions (e.g., anger, disgust) are mobilized in online hate speech. Hate is a powerful negative emotion (short term), or sentiment (long term), driven by appraisals that others hold malevolent intentions to inflict harm to one's group, coupled with perceptions of danger, and a sense of powerlessness (Fischer et al. 2018). This emotion is a strong predictor of self-defense actions such as causing harm, eliminating, and potentially even annihilating the target of hate, either at psychological (e.g., humiliation), social (e.g., exclusion, neglect), or physical (e.g., killing, torturing, aggression) levels (Fischer et al. 2018). The cognitive appraisals that others have malevolent nature and malicious intent to harm are unique to the experience of hate and are in line with the idea that online hate speech takes the form of bias against an outgroup perceived, to some extent, as a threat to the ingroup.

Finally, our intergroup approach to hate speech considers the need for annotation guidelines that are socio-culturally tailored to ensure its accurate detection, considering the underlying social psychological features of hate speech as well as the socio-cultural context in which the language corpora were gathered. In the current research, we focused on social groups that are more frequently targets of online hate speech in Portugal, specifically racialized communities, including Afro-descendants and Roma, migrants and LGBTI+ communities (Bayer and Bárd 2020; EUAFR 2021; Reynders 2020). Considering the specific features of racist, xenophobic, and sexual prejudice is important to unpack the multifaceted content and forms of online hate speech and tailor effective strategies to prevent it.

Hate speech multifaceted content: racism, xenophobia and sexual prejudice. The process of racialization involves targeting specific groups and subjecting them to a constructed racialized and ethnicized identity. Racism, in general, involves the belief in the superiority of one's own racialized and ethnicized ingroup, justifying a differential treatment of the outgroup based on this presumption of superiority (i.e., racial prejudice and discrimination as a set of discriminatory or derogatory attitudes and behaviors). This justification not only makes privileged dominance seem logical but also normalizes it (Jones 1997). In our study, we focus on two specific target groups: Afro-descendants and Roma communities. Considering Portugal's colonial history, our analysis of hatred towards racialized communities, especially

Afro-descendants, considers the influence of Luso-tropicalism, a set of shared beliefs positing that Portuguese colonizers had "special skills" for fostering harmonious relations with colonized populations, demonstrating adaptability to intercultural environments, and inherently lacking prejudice (Valentim and Heleno 2018). Historically used to legitimize colonialism, this ideology became integral to Portuguese national identity until today and is widely accepted by Portuguese society, shaping contemporary intergroup attitudes (e.g., more negative attitudes, more resistance to hiring immigrants). Research shows an association between these beliefs and the denial of racism (Vala et al. 2008), an implicit form of racial derogation visibly present in Portugal. Indeed, scholars have argued that prejudice has shifted from overt and blatant expressions to more implicit forms, suggesting a change in social desirability norms (Brown 2011). This shift represents a change from biological to cultural and identity considerations in the hierarchical categorization of social groups (e.g., Jones 1999; Kinder and Sears 1981; McConahay 1986; Pettigrew and Meertens 1995), which can be particularly mobilized in indirect, more subtle, forms of hate speech. Roma community has also 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, laziness, and resistance to integration into mainstream society (Breazu and Machin 2019, 2022; Chovanec 2021). In what concerns the Portuguese social context, the Roma community is still considered the most vulnerable minority group, being socially and economically excluded (Casa-Nova 2021; Maeso 2021, Magano and Mendes 2021).

Xenophobia refers to prejudice against, hatred towards, or fear of people from other countries or cultures, or people perceived as foreigners or strangers in general (Wicker 2001). It commonly involves the perception of immigrants, asylum seekers, or individuals of immigrant descent as outsiders, posing a potential threat to nationals (Sanchez-Mazas and Licata 2015). Indeed, perceived threat either regarding realistic (e.g., loss of power, economic resources) or symbolic aspects (e.g., culture, values, identity) is a powerful predictor of prejudice and discriminatory behaviors towards migrants (Esses 2021; Stephan and Stephan 2000). Symbolic threat is also conceptualized as an expression of symbolic (McConahay 1986) and cultural racism (vs. biological). Consequently, perceived threat becomes a crucial underlying aspect for comprehending both racial and xenophobic prejudice and hatred.

Finally, sexual prejudice involves holding negative attitudes towards an individual or group based on their perceived sexual orientation, gender identity, expression, or sex characteristics. This is often linked to sexual stigma—an ideology or shared belief asserting that non-heterosexual identities, feelings, or behaviors are deemed wrong, and inferior compared to heterosexual counterparts (Herek 2004). These perceptions relate to different phobias, such as homophobia, biphobia, transphobia, and intersexphobia (EUAFR 2020). Sexual prejudice and stigma are robust predictors of discrimination and aggression against LGBTI+ individuals, encompassing various forms such as verbal harassment, physical and sexual assault, as well as avoidance and social distancing (Katz-Wise and Hyde 2012). Recognizing the significance of perceived threats is also essential in understanding this phenomenon. Herek's (2004) integrative framework, referred to as the affordance management approach, builds on a socio-functional threat-based model (Cottrell and Neuberg 2005), and proposes that specific appraisals of threats and opportunities induce emotions and behavioral responses aimed at alleviating threats or achieving opportunities. Applied to sexual prejudice, this framework suggests that concerns that certain sexual

orientation groups violate ingroup values, cohesion, or functioning, trigger moral disgust and anger, leading to social exclusion and aggression against those deviating from norms to suppress such behavior, enforce group norms, and prevent social influence (Pirlott and Cook 2018).

Overall, “knowing the locus of the differential has significant implications for documenting and changing discriminatory behavior” (Brewer 2016, p. 92). As such, we will rely on the above-mentioned definitions and theoretical approaches to help us understand and characterize the specific features of online hate speech on Portuguese social media, aiming to provide valuable insights for more accurate detection models.

Hate speech multifaceted expression. Prejudice may manifest through either overt, direct, and blatant expressions or more covert, indirect, subtle, or implicit forms (Brown 2011). In direct hate speech, the speaker explicitly spreads or justifies hatred, exclusion, discrimination, and/or violence against a target group or individual based on perceived group membership. The message typically contains biased, inflammatory language, insults, and derogatory terms (e.g., “*Racismo o c@ralho! Se não fossem esses parasitas da sociedade que não querem fazer nada, Portugal era um paraíso /Racism, the fuck! If it weren’t for these parasites in society who don’t want to do anything, Portugal would be a paradise*”). In contrast, indirect hate speech avoids explicit derogatory or insulting language, with the spreading, promotion, or justification of hatred, exclusion, discrimination, or violence being implicit. The meaning in such cases is typically not literal, from a semantic point of view, and must also be pragmatically inferred, drawing on social and historical context (Assimakopoulos et al. 2017; Baider 2022, e.g., “*Já alguma vez viste um cigano a trabalhar?/Did you ever see a Roma working?*”). A variety of discursive, rhetorical strategies and logical fallacies often serve as vehicles for indirect and direct hate speech.

Discursive strategies. We analyze a range of discursive strategies underlying hatred messages, including negative stereotypes, denial of hate, dehumanization, role reversal, and symbolic and realistic group threats. Stereotypes are cognitive schemas that involve shared beliefs about the perceived attributes and characteristics of a group, which can be positive or negative, and are key determinants of how people perceive, feel, and behave towards others (Dovidio et al. 2010). Negative stereotyping, specifically, refers to negative, and inaccurate beliefs and characteristics associated with the targeted social groups or their members and is often employed to disparage or humiliate the target through fallacious negative generalizations (Paz et al. 2020; Sanguinetti et al. 2018). Denial of hate is a strategy aimed at positively presenting the ingroup, whereby the speaker protects their own arguments from accusations of hate by denying it, to maintain credibility, reduce the perception of hate, and preserve the legitimacy of their argument (van Dijk 1992). In the Portuguese context, it is closely intertwined with the hegemonic ideology of Lusotropicalism, as a strategy to convey the narrative of the “unique special skills” for harmonious intercultural relations. Similarly, role reversal, a strategy through which members of a privileged or dominant social group assert victimhood, claiming to be subject to discrimination or prejudice, while portraying outgroups as oppressors posing a threat to the ingroup (van Dijk 1992), also relates to Lusotropicalism beliefs. Indeed, this tactic aims to obscure systemic inequalities by reversing power dynamics, and undermine oppression claims by marginalized groups, appropriating their experiences and downplaying their struggles.

Dehumanization refers to the process of denying positive human traits to others, viewing them as less human, more animal-like, thereby removing moral considerations commonly extended to fellow human beings (Borinca et al. 2023). Lastly, as previously discussed, hate speech frequently involves perceptions and assertions that the targeted outgroup poses a threat to the ingroup, either realistic (to the ingroup’s power, resources, and general welfare, physical health and security) or symbolic (to the ingroup’s religion, values, belief system, ideology, philosophy, morality, or worldview; Stephan and Stephan 2000).

Rhetorical devices and fallacies. The aforementioned strategies often employ indirect strategies, encompassing various rhetorical devices such as verbal irony, rhetorical questions, metaphors, comparison, and hyperbole, which abound in user-generated content (Carvalho et al. 2009) and are systematically analyzed in our research. Verbal irony 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). Research 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 (Attardo 2000). In the current study, we use both 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); (ii) their intended meaning is indirect, and is only arrived at inferentially (Attardo 2000); (iii) both strategies may be (but not necessarily) cloaked in the mask of humor (Dynel 2018a); and (iv) both strategies can be used to express covert hate speech against a specific target (Baider 2023). Rhetorical questions are often used to implicitly associate negative stereotypes with a target (ElSherief et al. 2021). Such questions have the illocutionary force of an assertion of the opposite polarity from what is apparently asked (Han 2002), and 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). Metaphor, unlike comparison where the analogy between subjects is explicit, subtly transfers attributes between subjects in hate speech. This strategy enables the evocation of negative emotions and the perpetuation of stereotypes without overtly drawing comparisons. According to Critical Metaphor Analysis (Charteris-Black 2004), such metaphors expose speaker bias by revealing underlying thought patterns and ideological views. Essentially, metaphors function as mirrors that reflect speaker perceptions and societal biases. Hyperbole also serves as a potent tool in hate speech, amplifying negative depictions of targeted individuals or groups. Through extreme overstatements, dramatic claims, or sensationalized descriptions, hyperbole dehumanizes, vilifies, and incites hatred toward its targets, fostering fear, intimidation, and threats (Ignat and Vogel 2022).

Furthermore, the expression of hate is frequently characterized by implicit strategies aimed at manipulating the audience’s opinions and actions, often concealed within various fallacies and inappropriate uses of emotive language (Macagno 2022). These actions signify violations of the standards for critical discussion that are intended to guide reasonable argumentative discourse (van Eemeren and Garssen 2023). In this work, we focus on the fallacies of appeal to action and fear, which have proved to be effective in covertly promoting, spreading, or inciting hate speech (Carvalho et al. 2023). Call to action entails an explicit or implicit plea for action to revert a perceived negative situation, often delivered with emotional intensity. On the other hand, appeal to fear does not explicitly threaten but warns of negative

consequences if the recipient fails to undertake the recommended action, whether implicit or explicit (Tindale 2007; Walton 1996).

Current research

The current study aims to characterize the social psychological and linguistic-discursive features of online hate speech, encompassing both direct and indirect forms, directed at social groups more frequently targets of online hate speech in Portugal (i.e., racialized and Roma communities, LGBTI+ and migrant communities). We developed annotation guidelines to examine the prevalence of core social psychological, discursive, and rhetorical features often embedded within broader discursive strategies used to target marginalized groups: processes of discrimination (ingroup favoritism, outgroup derogation, and zero-sum), negative emotions (hate, anger, disgust, and fear), discursive strategies (stereotypes, denial of hate, dehumanization, role reversal, and realistic and symbolic threats), and specific rhetorical devices (metaphor, comparison, verbal irony, hyperbole, and rhetorical questions), and fallacies (appeal to fear, and call to action). The research is based on the analysis of two different Portuguese corpora composed of comments and replies to comments retrieved from YouTube and Twitter/X. These platforms offer unique value for data diversity due to differences in content structure, moderation practices, and user demographics. Before its 2022 acquisition, Twitter enforced active content moderation to address misinformation and hate speech, a concern shared by YouTube. However, Twitter's/X character-limited, text-based format enables real-time and often spontaneous communication, making it a valuable source for capturing unfiltered public reactions not necessarily tied to audiovisual content. In contrast, YouTube encourages more deliberate, long-form engagement. Demographically, Twitter's user base has typically been more politically engaged and highly educated, whereas YouTube attracts a broader and more entertainment-focused audience (Pew Research Center 2019).

Methods

The data collection process involved different phases tailored to each platform's structural and content characteristics. For YouTube, we started with a small list of videos deemed relevant (i.e., possibly containing online hate speech) that was signaled by representatives (i.e., local and national associations) of the target communities under study¹. Those videos served as seeds to explore semantically related content using the YouTube recommendation system. One hundred videos were automatically selected. To ensure relevance, we considered only those with over 100 comments and 1000 views. The annotation team manually validated the selected videos to confirm relevance and identify the primary target community. A final sample of 88 videos was chosen, comprising 24,739 YouTube comments and replies.

For Twitter/X, we adopted a lexical-based approach to capture tweets mentioning at least one of the potential target communities, leveraging a lexicon describing terms associated with the communities of interest (as outlined in Carvalho et al. 2022). Tweets containing clear and unambiguous mentions of the target communities (e.g., *cigano*, Roma), were directly included in the data collection. For ambiguous mentions of the target communities (e.g., *preto*, which in Portuguese can mean either the color adjective black or be used as a racial slur), we included only tweets where these mentions co-occurred with offensive and derogatory terms (see Carvalho et al. 2022). This strategy aimed to enhance the extraction of relevant content by prioritizing tweets where offensive or derogatory expressions were present alongside mentions of the target communities, increasing the likelihood of capturing instances of hate speech or related discourse. This

approach facilitated a comprehensive collection of tweets mentioning the target communities, ensuring the inclusion of both direct and indirect forms of hate speech. This involved gathering approximately 37 thousand geolocated tweets in Portugal, during 2021 and 2022. Conversations integrating these tweets (2775 conversations) were retrieved to gather contextual information. For manual annotation, only conversations with the first tweet geolocated in Portugal were considered, resulting in a total of 29,758 contextualized tweets.

Both corpora, YouTube and Twitter/X, underwent meticulous annotation by a team of trained annotators comprising social psychologists and linguists. These annotators were also actively involved in the creation of the annotation framework. Following the annotation process, we conducted an inter-annotator agreement (IAA) study using Krippendorff's alpha (α), a reliability coefficient to measure the agreement among annotators. Next, we conducted a statistical descriptive analysis of the manually annotated comments.

Annotation framework. The manual annotation process involved classifying each comment according to several social psychological and linguistic-discursive dimensions aimed at capturing core features of the expression of online hate speech and online counter-speech. In the current research, we explored seven dimensions and 26 variables directly associated with the expression of hate speech (see Table 1): (1) type of speech, distinguishing between direct hate speech (DHS), indirect hate speech (IHS); (2) targets, including racialized communities in general, and the Roma community in particular, migrant communities and LGBTI+ communities²; (3) discrimination processes, including ingroup favoritism, outgroup derogation, and zero-sum; (4) discursive strategies, including stereotypes, dehumanization, symbolic and realistic threats, the denial of hate, and role reversal; (5) rhetorical mechanisms, such as metaphor, comparison, hyperbole, verbal irony, rhetorical questions, as well as specific fallacious arguments (6), including appeal to fear and the call to action; and, finally, (7) a set of emotions to characterize hate speech (hate, anger, disgust and fear).

Annotation process. The YouTube corpus annotation was performed by four annotators. Each annotator was responsible for annotating around 6000 comments. Except for one annotator, all were also engaged in annotating the Twitter/X corpus, with each annotating about 7000 tweets. Additionally, an identical subset from both collections was assigned to all annotators to assess the inter-annotator agreement. Each comment in both corpora was analyzed according to the dimensions of the annotation guidelines. All variables were binary, indicating the presence or absence of the specific feature within each comment. Comments could contain multiple features within each dimension; thus, the binary variables are not mutually exclusive.

After annotation, the data underwent thorough checks to detect errors such as missing values, duplicates, and inconsistencies between variables. To maintain data quality, rules outlined in the annotation guidelines were applied to both YouTube and Twitter/X datasets: comments with classification discrepancies (e.g., being labeled as relevant but lacking information on the type of hate speech) were flagged for review and annotations associated with flagged comments were corrected by annotators to ensure accuracy. This rigorous quality control aimed to enhance the reliability of the annotated data for subsequent analysis.

Inter-annotator agreement. The inter-annotator agreement (IAA) rate was calculated using Krippendorff's alpha (α), based

Table 1 Distribution and reliability of annotation results by social media platform.

Dimension	Variable	YouTube		Twitter/ X	
		Alpha	Freq. (%)	Alpha	Freq. (%)
Type of speech	Hate speech	0.549	66.1	0.355	19.1
	Direct hate speech	0.394	28.3	0.195	2.8
	Indirect hate speech	0.175	38.1	0.211	16.4
Target community	Racialized	0.713	21.3	0.571	9.7
	Roma	0.839	14.1	0.799	5.0
	LGBTI+	0.811	16.2	0.582	12.5
	Migrants	0.808	25.6	0.583	6.4
Discrimination	Ingroup favoritism	0.131	8.6	0.151	1.0
	Outgroup derogation	0.481	55.8	0.309	16.6
Rhetorical devices	Zero-sum	0.185	5.1	0.105	0.6
	Metaphor	0.262	12.9	0.099	2.0
	Comparison	0.377	15.2	0.323	7.2
	Hyperbole	0.107	14.5	-0.003	4.4
Discursive strategies	Rhetorical question	0.655	9.8	0.544	5.6
	Verbal irony	0.415	27.2	0.326	16.6
	Stereotypes	0.268	26.9	0.342	8.6
	Denial of hate	0.265	4.7	0.371	1.8
	Dehumanization	0.508	9.8	0.306	1.5
	Realistic threat	0.142	12.4	0.125	3.9
	Symbolic threat	0.186	12.3	0.114	1.8
Fallacies	Role reversal	0.234	12.3	0.084	1.5
	Appeal to fear	0.268	13.8	0.195	2.1
	Call to action	0.620	21.9	0.289	4.2
Emotions	Hate	0.194	19.6	0.076	1.1
	Anger	0.254	18.1	0.229	5.9
	Disgust	0.207	1.9	0.033	1.3
	Fear	0.290	3.9	0.120	0.5

on two data subsamples randomly extracted from YouTube and Twitter/X corpora. Krippendorff's alpha (α) ranges from 0 to 1, with 0 indicating no agreement and 1 indicating perfect agreement. An alpha value above 0.4 indicates moderate agreement, above 0.6 substantial, and above 0.8 almost perfect agreement. The first subsample included 825 YouTube comments that were annotated by all four annotators, whereas the second subsample comprised 805 tweets annotated by three annotators.

The YouTube annotations demonstrated higher agreement across most categories compared to the Twitter/X annotations (see Table 1). However, apart from the identification of the target communities, which reached substantial agreement among annotators, all other variables achieved moderate to low agreement. Overall, results underscored the difficulty in unpacking core features of hate speech, and the potential discrepancy depending on the data characteristics. However, the overall moderate/low inter-annotator agreement rate may not only stem from the difficulty in finely classifying subjective data; it could also be influenced by the limited variability present in the dataset. Binary variables in which one of the values is infrequent (such as the presence of a phenomenon, coded as 1 in our study) often exhibit low variability. Consequently, even if there is agreement in the annotation, the resulting alpha value may be low (Krippendorff 2011). This is commonly referred to as the "paradox of high agreement but low reliability" (Krippendorff 2013).

Results

This section provides a descriptive analysis of both annotated corpora, focusing exclusively on comments deemed relevant by at least one annotator. In the YouTube corpus, 79% of the total observations were identified as relevant for the analysis, resulting in a final dataset of 19,468 relevant comments. Of these, 16,128

Table 2 Distribution of direct and indirect online hate speech on the annotated corpora.

	YouTube (%)	Twitter/ X (%)
Direct hate speech	31.0	20.9
Indirect hate speech	68.1	79.1

were classified as containing hate speech. The proportion of non-relevant observations in the Twitter/X corpus was substantially higher, with about 75% of observations identified as non-relevant, leaving 5624 relevant observations for analysis, with 2621 tweets classified as containing hate speech.

Hate speech type. Indirect hate speech was more prevalent than direct hate speech on both annotated corpora, with occurrences ranging from 68% on YouTube to 79% Twitter/X %, as shown in Table 2. It is important to note that direct and indirect hate speech can coexist within the same comment, as illustrated in Example [1], extracted from the YouTube corpus:

[1] *Um dia talvez podemos atirar este lixo anti branco fora do pais (preferencialmente fora do um helicóptero como o grande pinochet fazia) Força Mario*

'One day we might be able to throw this anti-white trash out of the country (preferably out of a helicopter like the great Pinochet did) Go Mario

In this comment, the speaker overtly employs derogatory language by referring to the target group (racialized communities) as "anti-white trash", dehumanizing them and making extreme actions easier to justify. The mention of Pinochet and

Table 3 Distribution of social psychological and linguistic-discursive features by hate speech type and social media platform.

Dimension	Variable	YouTube (%)		Twitter/ X (%)	
		DHS	IHS	DHS	IHS
Discrimination types	Ingroup favoritism	5.8	12.0	1.1	3.9
	Outgroup derogation	87.3	82.0	88.9	84.8
	Zero-sum	10.4	5.5	3.1	3.1
Discursive strategies	Stereotype	32.2	27.0	39.2	30.9
	Denial of hate	4.6	6.9	7.3	15.3
	Dehumanization	20.1	6.6	14.4	2.5
	Realist threat	13.5	11.3	16.2	20.5
	Symbolic threat	24.8	16.4	12.4	9.5
Rhetorical strategies	Role reversal	13.9	17.1	5.5	9.0
	Metaphor	16.2	14.1	9.3	6.5
	Comparison	23.6	15.9	22.0	17.7
	Verbal irony	43.8	32.9	57.7	54.5
	Hyperbole	15.8	15.7	11.1	18.9
Fallacies	Rhetorical question	15.0	11.7	10.9	14.1
	Appeal to fear	18.1	14.6	11.8	11.5
	Call to action	34.6	24.0	25.5	13.4
Emotions	Hate	59.8	12.9	33.7	4.2
	Anger	37.4	21.6	19.7	20.1
	Disgust	5.1	2.2	7.3	3.4
	Fear	4.8	3.7	1.6	2.2

DHS Direct Hate Speech, IHS Indirect Hate Speech.

the indirect reference to the violent practice of “death flights” during his regime glorifies historical violence and implies that such actions are admirable or should be emulated. This comment carries an implied threat, suggesting that the target group should be removed by violent means. By referencing Pinochet positively, and invoking Mario (Machado), a Portuguese neo-Nazi convicted of various hate crimes, it normalizes extreme measures as acceptable solutions to eradicate the target group.

Hate speech materialization. All descriptive statistics are presented in Table 3. As expected, outgroup derogation was the most frequent process of discrimination for both direct and indirect hate speech, in both social media platforms. This strategy is illustrated in Example [2], where the speaker overtly discriminates against the Roma community with derogatory language, ridicules their online presence, and perpetuates negative stereotypes.

[2] *Ciganos nas redes sociais sao a merda mais cringe de sempre e pqp, existe posts do facebook de 2013/14 q matam qualquer um de constrangimento. Gipsy skills.*

‘Gypsies on social media are the cringiest shit ever, and damn, there are Facebook posts from 2013/14 that would embarrass anyone. Gipsy skills.’

Regarding discursive strategies, stereotypes (illustrated in Example [3]) emerged as the most prevalent both in indirect and direct hate speech. In this example, the speaker overtly portrays the Roma community as engaging in culturally, legally, and ethically unacceptable behaviors, such as child marriage, reinforcing societal biases and contributing to the stigmatization and dehumanization of this target group.

[3] *Coitad@. Vcs deitam se com primos, tios, irmãos etc., praticam incesto entre família, casam crianças com adultos e as forçam a engravidar.*

‘Poor thing@. You sleep with cousins, uncles, brothers, etc., practice incest among families, marrying children to adults and forcing them to get pregnant.’

The presentation of the outgroup as symbolic and realistic threats was also prevalent in both types of hate speech. However, symbolic threats were more frequent on YouTube (see Example [4] as a threat to the glorified national ingroup’s past and worldview), whereas realistic threats were more common on Twitter/X (see Example [5]).

[4] *Como ela [Joacine Katar Moreira] não sabe fazer nada ainda quer destruir o que os nossos antepassados fizeram*

‘Since she [Joacine Katar Moreira] doesn’t know how to do anything, she still wants to destroy what our ancestors did.’

[5] *Esta canalha vive à conta de quem trabalha e ainda se queixa. Montados em grandes máquinas, não pagam renda nem luz vivem à grande e à francesa. De facto não há um problema com a ciganada.*”

‘These scoundrels live off the work of others and still complain. Riding big machines, they don’t pay rent or electricity and live large. Indeed, there’s no problem with the gypsies.’

In Example [4], the target group (Joacine Katar Moreira, a former Afro-descendant Portuguese Member of the Parliament) is portrayed as a threat to the cultural and historical heritage of the ingroup (i.e., white Portuguese nationals). In Example [5], the speaker expresses the perception that the Roma community exploits resources and benefits without contributing to society, which is a common materialization of realistic threat.

Despite these similarities, direct and indirect hate speech differed regarding two core discursive strategies: the use of dehumanization was more prevalent in direct hate speech (see Example [6]), whereas role reversal and denial of hate were more prevalent in indirect hate speech (see Examples [7] and [8], respectively).

[6] *Ciganos são como javalis, são animais selvagens*

‘Gypsies are like wild boars, they are wild animals’

[7] *Estamos num Mundo e muito particularmente Portugal onde assumir a normalidade de carácter, e de natureza é logo histericamente criticado e adjectivado, se falamos e assumimos a nossa natural condição de sermos heterossexuais somos homofóbicos se mencionamos quaisquer problemática de certas minorias(já tenho dúvidas que o sejam) saltam da cartola esses mirabolantes adjectivos tais como xenófobos, racistas etc etc, longe vão os tempos da Trilogia : DEUS, PÁTRIA e FAMÍLIA.*

‘We are in a world, and particularly in Portugal, where assuming normality of character and nature is immediately hysterically criticized and adjectivized. If we talk about and assume our natural condition of being heterosexual, we are homophobic. If we mention any problems with certain minorities (I already doubt that they are), these crazy adjectives jump out of the hat, such as xenophobic, racist etc etc. Long gone are the days of the trilogy: GOD, COUNTRY and FAMILY’.

[8] *Omg eu juro, nao sou racista mas se tivesse de tirar uma raça da Europa seria esta*

‘Omg I swear, I’m not racist but if I had to take one race out of Europe it would be this one’

In Example [6], dehumanization is starkly demonstrated by explicitly comparing the target community to wild animals. Such similes, commonly used to equate humans with animals, are highly productive in hate speech and can serve as immediate indicators of dehumanizing language, reinforcing harmful stereotypes and portraying the target group in an inferior light. In Example [7] the speaker portrays heterosexual and white majority people as the new oppressed group, thus reversing historical roles and minimizing the existing discrimination faced by LGBTI+ and racialized communities. In Example [8] the speaker uses a disclaimer to downplay or deny racism but then follows with a blatantly racist remark about wanting to remove a specific race from Europe.

Regarding rhetorical strategies, verbal irony (illustrated in Example [9]) emerged as the main strategy in direct and indirect hate speech, regardless of the social media platform. The use of this strategy often minimizes overt aggression while reinforcing stereotypes and advocating for exclusion. In the illustrated example it subtly undermines Mamadu Ba’s activism and character, conveying hostility and prejudice through a seemingly humorous tone.

[9] *Afinal o homem [Mamadu Ba] tem que promover e denunciar o racismo, senão vai para as obras, ou com sorte, servir à mesa. Fica mais barato mandá-lo para casa, apoio.*

‘After all, the man [Mamadu Ba] has to promote and denounce racism, otherwise he’ll end up on a construction site or, with luck, serving tables. It’s cheaper to send him home, I support that.’

When it comes to fallacies, the data indicates that “call to action” is also widely employed, particularly in direct hate speech (cf. Example [10]). The comment clearly conveys an explicit call to action, urging others to take drastic and harmful measures against the target group (the Roma community). It explicitly references historical atrocities and serves to inflame hatred and incite violence. The remaining rhetorical devices were less prevalent and did not differ substantially between direct and indirect hate speech forms.

[10] *Mandem nos todos para os campos de concentração na China!!! Lá de certeza que vão saber o que é trabalho forçado...*

‘Send them all to the concentration camps in China!!! There they will surely learn what forced labor is...’

Finally, regarding emotions, hate was the most prevalent negative emotion in direct hate speech, as expected, and relatively low in indirect hate speech, both in YouTube and Twitter/X corpora (see Example [11]). On the contrary, anger was more prevalent than hate in indirect hate speech relative to direct hate speech (see Example [12]).

[11] *MATAR O CARALHO..VAI PRÁ TUA TERRA.MAS AFINAL DE QUE TRIBO ÉS..MAMADOU. E NÃO HÁ QUÊM LHE.TRATE DA SAÚDINHA.*

‘KILL MY ASS. GO BACK TO YOUR LAND.BUT AFTER ALL, WHAT TRIBE ARE YOU FROM..MAMADOU.AND THERE’S NO ONE THAT KILLS HIM.’ (“tratar da

Table 4 Distribution of target communities by social media platform.

	YouTube (%)	Twitter/ X (%)
Racialized	39.0	26.1
Roma	16.9	5.7
LGBTI+	17.4	52.7
Migrants	23.0	16.4

saudinha” idiomatic expression that implies physical harm and in some cases to kill, authors’ translation)

[12] *Força Mário, fizeste muito bem. Essa escumalha vai ter que aprender a respeitar os Portugueses e parar de gozar com os cidadãos de bem. Viva Portugal ‘Go ahead Mario, you’ve done very well. This scum will have to learn to respect the Portuguese and stop making fun of good citizens. Long live Portugal’.*

Hate speech by target communities. Racialized communities were the most frequently targeted group in YouTube, whereas in Twitter/X the LGBTI+ community was the primary target group (Table 4). Considering that the corpora is not balanced in terms of the representation of each community, any comparisons should be taken with caution. Table 5 shows that indirect hate speech was more prevalent across all target communities in both social media platforms.

As shown in Table 6, various aspects of hate speech manifestation were similarly prevalent across all targeted communities, regardless of the specific type of hate speech. For instance, outgroup derogation consistently emerged as the most frequently employed form of discrimination across all target communities. Additionally, discursive strategies such as negative stereotyping and dehumanization were also similarly utilized, with the latter more prominent on direct hate speech.

Despite these similarities, there were also noticeable trends in the differential use of analyzed strategies among target communities. For instance, concerning discursive strategies, symbolic threats were more prevalent than realistic across all target groups on YouTube, except for the Roma community. Conversely, on Twitter/X, realistic threats were more prevalent, particularly targeting racialized, Roma, and migrant communities.

Regarding rhetorical devices, while verbal irony appears to be highly used across various target communities, hyperbole was most frequently employed in comments targeting LGBTI+ communities, particularly on YouTube. In terms of fallacies, their usage did not vary significantly across different communities. However, when analyzing emotions, an interesting pattern emerged: hate was prevalent across all communities, except for LGBTI+ on YouTube (e.g., *São todos os paineleiros [...] só tinha de mata Los, juntamente com os vossos deuses, escumalha de merda, nem defendem família/You’re all faggots [...] all I had to do was kill them, along with your gods, you fucking scum, you don’t even defend your family*), where anger predominated (e.g., *A VOZ DESTE PANASCA IRRITA-ME... PORQUE É QUE OS DEIXARAM SAIR DO ARMARIO ??? FAZ-TE UM HOMEM ASERIO QUE NÃO FALA A BEBÉ.../ THIS FAGGOT’S VOICE IRRITATES ME... WHY DID YOU LET THEM OUT OF THE CLOSET ??? MAKE YOURSELF A REAL MAN WHO DOESN’T TALK LIKE A BABY...*). Conversely, on Twitter the pattern was reversed.

Finally, we explored the association between the most mobilized discursive strategies and emotions in hatred

Table 5 Distribution of OHS type by target communities and social media platform.

	YouTube (%)				Twitter/ X (%)			
	Racialized	Roma	LGBTI+	Migrants	Racialized	Roma	LGBTI+	Migrants
Direct hate speech	25.5	33.4	29.1	33.2	7.7	15.1	10.9	15.4
Indirect hate speech	66.2	59.6	64.2	51.4	47.8	53.4	46.2	46.5

Table 6 Distribution of social psychological and linguistic features by hate speech type, target communities and social media platforms.

	YouTube (%)								Twitter/ X (%)								
	Racialized		Migrants		Roma		LGBTI+		Racialized		Migrants		Roma		LGBTI+		
	DHS	IHS	DHS	IHS	DHS	IHS	DHS	IHS	DHS	IHS	DHS	IHS	DHS	IHS	DHS	IHS	
Discrimination types																	
Ing. Favoritism	8.0	16.0	8.9	16.1	3.5	4.5	1.9	9.6	2.3	6.8	0.9	4.4	0.0	1.5	0.4	2.3	
Out. Derogation	83.2	79.8	82.6	78.1	89.7	87.1	93.9	83.9	81.8	82.7	86.6	85.1	84.6	92.7	92.9	85.9	
Zero-sum	11.3	5.0	15.1	6.1	10.1	6.0	9.1	7.1	9.1	3.4	4.5	3.8	10.3	2.2	2.0	2.6	
Discursive strategies																	
Stereotype	18.4	23.6	41.2	33.6	43.3	32.1	33.6	27.8	52.3	23.0	33.9	33.3	61.5	46.7	30.4	30.4	
Denial of Hate	6.8	10.8	5.7	5.6	1.6	2.4	6.1	5.6	10.2	25.3	2.7	18.1	5.1	6.6	10.3	12.6	
Dehumanization	20.4	7.5	18.2	4.5	26.3	7.5	18.6	5.5	10.2	2.3	16.1	1.8	15.4	1.5	15.8	2.0	
Realist threat	9.3	11.5	13.1	10.6	21.4	13.5	15.9	13.5	19.3	31.6	25.0	32.2	30.8	33.6	15.0	13.8	
Symbolic threat	25.5	18.6	33.3	18.3	17.7	9.3	23.6	19.5	6.8	7.7	12.5	11.4	7.7	10.9	14.6	11.0	
Role reversal	13.4	17.8	12.6	17.1	10.6	17.4	23.2	19.2	11.4	11.8	5.4	8.5	5.1	12.4	5.1	8.2	
Rhetorical strategies																	
Metaphor	13.5	15.1	13.7	10.8	20.5	16.0	20.0	14.3	9.1	5.0	8.0	5.6	5.1	9.5	10.3	7.0	
Comparison	26.0	15.6	28.4	17.3	21.5	16.9	21.7	16.9	22.7	23.9	36.6	23.7	25.6	17.5	16.6	13.7	
Verbal irony	49.5	27.6	49.3	33.7	38.8	39.4	36.7	37.8	59.1	50.8	55.4	54.7	38.5	51.8	62.1	55.7	
Hyperbole	7.3	13.2	19.2	22.4	13.6	7.2	29.2	24.2	15.9	24.1	18.8	22.2	30.8	16.8	7.5	18.4	
Rhet. question	18.1	11.7	14.0	12.5	13.9	13.2	16.2	11.2	10.2	16.0	12.5	16.1	15.4	16.8	11.1	12.7	
Fallacies																	
Appeal to fear	14.3	13.3	17.8	15.7	23.1	14.7	22.1	19.2	9.1	17.1	10.1	19.3	28.2	16.8	9.5	7.5	
Call to action	40.5	29.7	34.8	21.1	41.0	23.9	24.2	18.5	21.6	14.1	39.3	16.4	23.1	9.5	24.9	13.6	
Emotions																	
Hate	77.6	14.2	64.5	12.9	65.8	24.7	21.9	2.0	21.6	4.5	17.0	2.0	20.5	2.9	47.8	3.7	
Anger	36.1	20.1	38.2	26.8	34.0	17.6	47.1	24.1	25.0	28.5	26.8	26.0	20.5	18.2	16.6	17.5	
Disgust	3.6	1.6	4.2	1.7	7.6	2.7	7.8	3.8	4.5	2.3	13.4	3.2	10.3	1.5	5.1	3.8	
Fear	3.0	3.4	7.7	4.0	5.0	2.9	4.5	6.2	2.3	3.7	5.4	4.4	5.1	5.1	0.8	1.2	

comments and the specific rhetorical devices and fallacies that were used to express them (see Table 7). We used the Phi coefficient (Φ) to measure the association between our variables of interest, considering these were binary data and focused on the YouTube corpora only considering the proportion of relevant observations was higher than in Twitter/X. Like Pearson’s correlation, this measure ranges from -1 to 1 . Positive values indicate that the two variables tend to appear together, while negative values indicate that the presence of one variable concurs with the absence of the other. Most correlations were significant but weak (<0.29), with some being moderate. For instance, the use of metaphors as a rhetorical device concurred with the mobilization of dehumanization of target communities, whereas hyperbole co-occurred with other discursive strategies such as stereotyping, role reversal and realistic threats. Additionally, the use of fallacies such as appeal to fear co-occurred with the mobilization of both symbolic and realistic threats, as well as fear. Finally, some discursive strategies and emotions were also associated, albeit generally weakly: dehumanization and hate, stereotypes and realistic threat, stereotypes and role reversal. Overall, this highlights the multifaceted nature of hate speech and how these different facets are connected in the mobilization of hatred.

Discussion

Building on social psychology and language sciences, this research identified core social psychological and linguistic-discursive features of online hate speech targeting racialized, migrant and LGBTI+ communities in two social media platforms in Portugal. Relying on an interdisciplinary lens to unpack the multifaceted expression of hate speech, our findings revealed that: i) online hate speech was predominantly expressed in subtle, nuanced ways (i.e., indirect hate speech); ii) regardless of being expressed directly (i.e., overtly) or indirectly (i.e., subtly or implicitly), the main underlying process of discrimination of target communities was outgroup derogation; iii) hate speech was often manifested through the use of stereotypes, threats, and dehumanization; iv) emotions, often overlooked in hate speech annotation, were mobilized in hatred discourse, specifically hate and anger; (v) the main discursive strategies underlying hateful messages can vary depending on the target communities, and the social media platform; (vi) discursive strategies and emotions manifested through the use of specific rhetorical devices and fallacies. Overall, these findings provide valuable insights into the complex landscape of online hate speech targeting vulnerable communities in Portugal, highlighting the need for nuanced approaches in both understanding and tackling this pervasive issue. By employing a comprehensive and interdisciplinary lens grounded in social

Table 7 Phi correlations between most prevalent discursive strategies, emotions, rhetorical strategies and fallacies (YouTube).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1.Out. derogation	-														
2. Stereotype	0.066*	-													
3. Denial of hate	0.00	0.051*	-												
4. Dehumanization	0.067*	0.01	-0.051*	-											
5. Realist threat	-0.034*	0.231*	0.025*	0.019	-										
6. Symbolic threat	-0.052*	0.144*	0.031*	0.00	0.040*	-									
7. Role reversal	-0.038*	0.147*	0.134*	-0.037*	0.209*	0.133*	-								
8. Metaphor	0.020	0.077*	0.01	0.293*	0.146*	0.091*	0.059*	-							
9. Comparison	-0.019	0.132*	0.122*	0.01	0.098*	0.130*	0.129*	0.01	-						
10. Verbal irony	0.088*	0.066*	0.066*	-0.037*	-0.031*	-0.029*	-0.034*	-0.032*	0.047*	-					
11. Hyperbole	-0.016	0.279*	0.041*	0.00	0.265*	0.147*	0.208*	0.074*	0.063*	0.029*	-				
12. Rhet. question	0.00	0.083*	0.044*	0.01	0.036*	0.045*	0.057*	0.01	0.054*	0.104*	0.018	-			
13. Appeal to fear	-0.016	0.132*	0.01	0.00	0.384*	0.202*	0.188*	0.081*	0.071*	-0.093*	0.190*	0.00	-		
14. Call to action	-0.019	-0.018	-0.040*	0.075*	0.046*	0.046*	-0.01	0.042*	-0.049*	-0.109*	-0.034*	-0.022*	0.00	-	
15. Hate	0.054*	-0.084*	-0.067*	0.251*	-0.088*	0.081*	-0.113*	0.022*	0.082*	0.168*	-0.143*	0.034*	-0.041*	0.147*	-
16. Anger	0.01	0.116*	0.052*	0.044*	0.148*	0.148*	0.163*	0.088*	0.100*	0.033*	0.162*	0.117*	0.146*	0.126*	-0.071*

Significance level: *p < 0.001.

psychology and language sciences, we contributed to existing research by uncovering nuanced trends in the manifestation of hate speech across different online platforms and target communities.

In line with previous research across different linguistic and cultural contexts, our research showed that indirect hate speech prevailed over direct hate speech, regardless of the specific target community or social media platform (Baider 2022; Carvalho et al. 2023; Rieger et al. 2021). This poses significant challenges for existing automatic detection systems that have strong limitations in accounting for language and contextual knowledge (Udapa et al. 2023). In addition, our results suggested that many of the analyzed rhetorical strategies typically associated with indirect hate speech expression, such as verbal irony, were also prevalent in comments classified as conveying direct hate speech. Consistent with previous research, verbal irony and sarcasm were predominantly employed to insult, humiliate, and ridicule the target communities (Baider and Constantinou 2020; Carvalho et al. 2023). These rhetorical devices served to promote negative emotions toward the target communities, often perpetuating negative stereotyping.

Nevertheless, the use of rhetorical strategies like verbal irony, traditionally associated with indirect hate speech, in the expression of direct hate speech, challenges existing assumptions and underscores the complexity of online hate speech dynamics. One potential explanation for these findings is that online users are aware of the potential consequences of explicit hate speech, such as removal by social media platforms for violating community guidelines. Consequently, they may employ discursive and rhetorical strategies to mask or attenuate explicit and direct hateful content. This strategic adaptation presents significant challenges for automatic detection and intervention efforts, as indirect hate speech may evade explicit detection while still perpetuating harmful ideologies and attitudes.

In both social media corpora, biased language primarily manifested through outgroup derogation (Type 2 discrimination, Brewer 2016), characterized by negativity, hostility, and an intent to harm the target community. This finding highlights the importance of understanding hate speech as an intergroup phenomenon, where biased language underlies processes of discrimination with a clear motivation to harm and derogate the other. The derogation of the target community, without mobilizing processes of ingroup favoritism, was similar in both direct and indirect forms of hate speech. This finding highlights that the covert expression of indirect hate speech is similarly motivated by the intent to harm and derogate others, and thus not less dangerous than overt expressions of hatred. It is also closely intertwined with the significant prevalence of negative stereotyping, that emerged as the most prevalent discursive strategy used to undermine, diminish, or ridicule all target communities. Indeed, racial, xenophobic and sexual prejudice and stereotypes were mobilized in hatred speech, together with dehumanization of target communities and portraying them as posing realistic and symbolic threats to society. These findings are in line with social psychological research showing that symbolic and realistic threats are powerful predictors of prejudice and discrimination (Esses 2021; Stephan and Stephan 2000) as well as with findings illustrating the association of dehumanization and intergroup harming (Haslam and Loughnan 2016). In addition, the prevalence of call to action as the most utilized rhetorical strategy in both types of hate speech, particularly in direct speech, underscores the reliance of online hate speech on flawed argumentation intentionally used to manipulate the audience's opinion. In essence, this strategy serves as a catalyst for turning prejudiced thoughts into harmful actions, exacerbating societal divisions and injustices against the target groups.

Our findings also illustrated that negative emotions like hate and anger are prevalent in online hate speech, extending existing research that generally did not consider specific emotions in online hate speech annotation, but rather relied on sentiment analysis to classify emotional valence and intensity (Schmidt and Wiegand 2017). Our findings revealed that different negative emotions were mobilized differently depending on the type of hate speech: hate and anger were the most frequent negative emotions in direct hate speech, whereas anger predominates in indirect hate speech. By including systems capable of recognizing emotions, we can enhance the accuracy of identifying hate speech and distinguishing between direct and indirect forms. Moreover, incorporating emotion recognition capabilities into automated detection systems can provide a deeper understanding of the underlying emotional dynamics of hate speech. This enhanced understanding can inform the development of more effective counter-speech strategies, thereby improving our ability to mitigate the harmful effects of hate speech and foster a safer online environment.

Finally, we also showed that the most mobilized discursive strategies and emotions were significantly associated with specific rhetorical devices and fallacies. Indeed, two rhetorical devices were primarily associated with hatred content: metaphor, associated with the dehumanization of target communities, and hyperbole, that occurred in comments mobilizing threats, stereotypes and role reversal. These findings not only validate foundational decisions in hate speech identification guidelines but also offer insights for refining these guidelines in future research. Analyzing prevalence helps prioritize or exclude specific categories in future studies. For instance, a deeper analysis of the distinction between hate and anger, rather than exploring fewer common emotions, could enhance our understanding of emotional dynamics in hate speech contexts. Moreover, examining associations across dimensions reveals that while widely used categories like verbal irony may serve as general indicators of hate speech, others such as metaphor and hyperbole provide nuanced insights into specific discursive strategies employed, such as dehumanization or the mobilization of realistic and symbolic threats, respectively. Indeed, exploring specific associations deepens our understanding of how these phenomena manifest. For instance, understanding the link between hyperbole and role reversal illuminates how hate speech distorts perceptions of victimhood and oppression.

While it is possible to extract overarching trends that are consistent across target communities and social media platforms analyzed, our findings also demonstrated some variability in the manifestation of online hate speech. For example, whereas racialized communities were predominantly targeted on YouTube, the LGBTI+ community faced more hatred on Twitter/X. Variations in rhetorical strategies were also evident, with appeal to fear being prominent in comments directed at Roma communities. Regarding emotions, hate prevailed in direct hate speech toward racialized, Roma, and migrant communities on YouTube, whereas anger was more prevalent in indirect hate speech on Twitter/ X. In general, these findings highlight the influence of the target community, social context, and source of data, emphasizing the need for comprehensive analysis before generalizations can be made. Notwithstanding, these nuances should be taken with caution as the corpora was not balanced across target communities. Future research should examine these differences relying on well-powered, balanced samples, that allow a thorough analysis and comparison of the social psychological and linguistic-discursive features of hate speech towards different target communities. Indeed, the expression of hate speech, as other forms of discrimination, is shaped by historical, political and social contexts. This highlights the need to further compare the content and expression of hate speech across different target communities

through comparable corpora representing multiple targets, as well as the complex interplay between intersectional belongings that consider the complex power dynamics and social hierarchies that contribute to the perpetuation of hate speech. This is in line with recent approaches suggesting the need to develop people-centric approaches for content moderation that consider the key role of cultural context and target communities (Udupa et al. 2023), as well as, with critical frameworks of racism and discrimination such as Critical Race Theory, that offer a complimentary lens to understand how hate speech is situated within the broader context of systemic racism. Indeed, this perspective highlights how hate speech is not just an isolated act of individual prejudice but is deeply rooted in systemic racism that permeates societal structures and institutions (Adams and Omar 2024). Future annotation guidelines could include specific categories to assess the systemic facets of discrimination besides the ones included in the current research (e.g., denial of hate, role reversal).

Conclusion

Our findings highlight how linguistic-discursive strategies, cognitions, motivations and emotions may play a role in improving existing models of automatic hate speech detection systems in social media platforms. By developing theoretically grounded language resources, we offer valuable insights for both theoretical and applied research in online hate speech and counter-speech, particularly in the Portuguese context. Ultimately, our study underscores the importance of interdisciplinary collaboration and nuanced, context and culturally sensitive approaches in understanding and fighting online hate speech.

Data availability

Data sets used for the current study are declared as Sensitive (not public) under the Grant Agreement with the funding agency (EU). We made available at OSF <https://osf.io/mpej3/> a subset of our data, corresponding to the test files (Golden Set) used in our experiments. For ethical compliance, the annotated comments are identified only by their Tweet or YouTube IDs. Original content must be retrieved through Twitter/X or YouTube APIs.

Received: 3 July 2024; Accepted: 20 June 2025;

Published online: 11 November 2025

Notes

- 1 Details about the local and national representatives are available at <https://knowhate.eu/partners/>
- 2 A category for intersectional targets was also included but later excluded from all analyses considering the extremely low representation in the corpora: 0.2% in X/ Twitter, and 0.75% in YouTube.

References

- Achim V (2004) *The Roma in Romanian History*. Central European University Press
- Adams G, Omar SM (2024) Confronting racism-evasive ignorance in standard pedagogy of hegemonic social psychology. *J Soc Issues* 80(2):607–628. <https://doi.org/10.1111/josi.12618>
- Aguiar J, Barbosa P (2024) Emotional deixis in online hate speech. In: Ermida I (ed) *Hate speech in social media: linguistic approaches*. Springer Nature, pp. 139–164
- Albelda Marco M (2022) Rhetorical questions as reproaching devices. *J Lang Aggression Confl* 11(2):176–199. <https://doi.org/10.1075/jlac.00077.alb>
- Allport GW (1954) *The nature of prejudice*. Addison-Wesley
- Assimakopoulos S, Baider FH, Millar S (2017) *Online hate speech in the European Union: a discourse-analytic perspective*. Springer Nature
- Attardo S (2000) Irony as relevant inappropriateness. *J Pragmat* 32(6):793–826. [https://doi.org/10.1016/S0378-2166\(99\)00070-3](https://doi.org/10.1016/S0378-2166(99)00070-3)
- Baider F (2022) Covert hate speech, conspiracy theory and anti-semitism: Linguistic analysis versus legal judgement. *Int J Semiotics Law Rev* 35:2347–2371. <https://doi.org/10.1007/s11196-022-09882-w>

- Baider F (2023) Accountability issues, online covert hate speech, and the efficacy of counter-speech. *Politics Gov* 11(2):249–260. <https://doi.org/10.17645/politics.gov.1112.6465>
- Baider F, Constantinou M (2020) Covert hate speech: a contrastive study of Greek and Greek Cypriot online discussions with an emphasis on irony. *J Lang Aggression Confl* 8(2):262–287. <https://doi.org/10.1075/jlac.00040.bai>
- Bayer J, Bárd P (2020) Hate speech and hate crime in the EU and the evaluation of online content regulation approaches. Policy Department for Citizens' Rights and Constitutional Affairs of the European Parliament. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/655135/IPOL_STU%282020%29655135_EN.pdf
- Borinca I, Van Assche J, Gronfeldt B, Sainz M, Anderson J, Taşbaş EHO (2023) Dehumanization of outgroup members and cross-group interactions. *Curr Opin Behav Sci* 50:101247. <https://doi.org/10.1016/j.cobeha.2023.101247>
- Brewer MB (1999) The psychology of prejudice: Ingroup love and outgroup hate? *J Soc Issues* 55(3):429–444. <https://doi.org/10.1111/0022-4537.00126>
- Brewer MB (2016) Intergroup discrimination: ingroup love or outgroup hate? In Sibley CG, Barlow FK (eds), *The Cambridge handbook of the psychology of prejudice*. Cambridge University Press, pp. 90–110. <https://doi.org/10.1017/9781316161579.005>
- Breazu P, Machin D (2019) Racism toward the Roma through the affordances of Facebook: bonding, laughter and spite. *Discourse Soc* 30(4):376–394
- Breazu P, Machin D (2022) Using humor to disguise racism in television news: the case of the Roma. *Humor* 35(1):73–92
- Brown R (2011) *Prejudice: Its social psychology*. John Wiley & Sons
- Carvalho P, Caled D, Silva C, Batista F, Ribeiro R (2023) The expression of hate speech against Afro-descendant, Roma, and LGBTQ+ communities in YouTube comments. *J Lang Aggress Conflict*. <https://doi.org/10.1075/jlac.00085.car>
- Carvalho P, Cunha B, Santos R, Batista F, Ribeiro R (2022) Hate speech dynamics against African descent, Roma and LGBTQI communities in Portugal. In Proceedings of the 13th Conference on Language Resources and Evaluation, Marseille, 2362–2370. <http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.253.pdf>
- Carvalho P, Sarmiento L, Silva MJ, Oliveira E (2009) Clues for detecting irony in user-generated contents: oh...!! it's so easy;-). In Proceedings of the 1st international CIKM Workshop on Topic-sentiment analysis for mass opinion, pp. 53–56
- Casa-Nova MJ (2021) Reflecting on public policies for Portuguese Roma since implementation of the NRIS: theoretical and practical issues. *J Contemp Eur Stud* 29(1):20–32
- Charteris-Black J (2004) Critical metaphor analysis. In *Corpus approaches to critical metaphor analysis*. Palgrave Macmillan, pp. 243–253. https://doi.org/10.1057/9780230000612_12
- Chovanec J (2021) 'Re-educating the Roma? You must be joking...': Racism and prejudice in online discussion forums. *Discourse Soc* 32(2):156–174
- Cottrell CA, Neuberg SL (2005) Different emotional reactions to different groups: a sociofunctional threat-based approach to "prejudice". *J Personal Soc Psychol* 88(5):770–789. <https://doi.org/10.1037/0022-3514.88.5.770>
- Dovidio JF, Gaertner SL (2010) Intergroup bias. In Fiske ST, Gilbert DT, Lindzey G (eds) *Handbook of social psychology*, 5th ed. John Wiley & Sons, Inc., pp. 1084–1121. <https://doi.org/10.1002/9780470561119.socpsy002029>
- Dovidio JF, Hewstone M, Glick P, Esses VM (2010). Prejudice, stereotyping, and discrimination: theoretical and empirical overview. In: *The SAGE handbook of prejudice, stereotyping, and discrimination*. Sage, London, England, pp. 3–29
- Dynel M (2018a) Irony, deception and humour: seeking the truth about overt and covert untruthfulness. *De Gruyter Mouton*. <https://doi.org/10.1515/9781501507922>
- Dynel M (2018b) Deconstructing the myth of positively evaluative irony. In: Jobert M, Sorlin S (eds), *The pragmatics of irony and banter*. John Benjamins Publishing Company, pp. 41–57. <https://doi.org/10.1075/lal.30.03.dyn>
- Dynel M (2019) Ironic intentions in action and interaction. *Lang Sci* 75:1–14. <https://doi.org/10.1016/j.langsci.2019.06.005>
- ElSherief M, Ziems C, Muchlinski D, Anupindi V, Seybolt J, De Choudhury M, Yang D (2021) Latent hatred: a benchmark for understanding implicit hate speech. *ArXiv*. <https://doi.org/10.48550/arXiv.2109.05322>
- Esses VM (2021) Prejudice and discrimination toward immigrants. *Annu Rev Psychol* 72:503–531. <https://doi.org/10.1146/annurev-psych-080520-102803>
- EUAFR - European Union Agency for Fundamental Rights (2020) A long way to go for LGBTQ equality. <https://fra.europa.eu/en/publication/2020/eu-lgbti-survey-results>
- EUAFR - European Union Agency for Fundamental Rights (2021) *Fundamental rights report – 2021*. <https://fra.europa.eu/en/publication/2021/fundamental-rights-report-2021>
- Fischer A, Halperin E, Canetti D, Jasini A (2018) Why we hate. *Emot Rev* 10(4):309–320. <https://doi.org/10.1177/1754073917751229>
- Fortuna P, Nunes S (2018) A survey on automatic detection of hate speech in text. *ACM Comput Surv* 51(4):1–30. <https://doi.org/10.1145/3232676>
- Han CH (2002) Interpreting interrogatives as rhetorical questions. *Lingua* 112(3):201–229. [https://doi.org/10.1016/S0024-3841\(01\)00044-4](https://doi.org/10.1016/S0024-3841(01)00044-4)
- Haslam N, Loughnan S (2016) How dehumanization promotes harm. In Miller AG (ed) *The social psychology of good and evil*, 2nd ed. Guilford, pp. 140–158
- Herek GM (2004) Beyond "homophobia": thinking about sexual prejudice and stigma in the twenty-first century. *Sexuality Res Soc Policy* 1(2):6–24. <https://ssrn.com/abstract=1142860>
- Ignat S, Vogel C (2022) Features and categories of hyperbole in cyberbullying discourse on social media. In: Proceedings of the second international workshop on resources and techniques for user information in abusive language analysis, Marseille, 25–31. <https://aclanthology.org/2022.restup-1.4>
- Jones JM (1997) *Prejudice and racism*, 2nd ed. McGraw-Hill, New York
- Jones JM (1999) Cultural racism: the intersection of race and culture in intergroup conflict. In: Prentice DA, Miller DT (eds) *Cultural divides: understanding and overcoming group conflict*. Russell Sage Foundation, pp. 465–490
- Katz-Wise SL, Hyde JS (2012) Victimization experiences of lesbian, gay, and bisexual individuals: a meta-analysis. *J Sex Res* 49(2-3):142–167. <https://doi.org/10.1080/00224499.2011.637247>
- Kinder DR, Sears DO (1981) Prejudice and politics: symbolic racism versus racial threats to the good life. *J Personal Soc Psychol* 40(3):414–431. <https://doi.org/10.1037/0022-3514.40.3.414>
- Krippendorff K (2011) Agreement and information in the reliability of coding. *Commun Methods Measures* 5(2):1–20. <https://doi.org/10.1080/19312458.2011.568376>
- Krippendorff K (2013) Commentary: a dissenting view on so-called paradoxes of reliability coefficients. *Ann Int Commun Assoc* 36(1):481–499. <https://doi.org/10.1080/23808985.2013.11679143>
- Macagno F (2022) Argumentation profiles and the manipulation of common ground: the arguments of populist leaders on Twitter/ X. *J Pragmat* 191:67–82. <https://doi.org/10.1016/j.pragma.2022.01.022>
- Magano O, Mendes MM (2021) Structural racism and racialization of Roma/Ciganos in Portugal: the case of secondary school students during the COVID-19 pandemic. *Soc Sci* 10(6):203
- Maeso RS (2021) O Estado do Racismo em Portugal: Racismo antinegro e anticiganismo no direito e nas políticas públicas. *Tinta-da-China*
- Mackie DM, Smith ER (2015) Intergroup emotions. In: Mikulincer M, Shaver PR, Dovidio JF, Simpson JA (eds) *APA handbook of personality and social psychology*, vol. 2. Group processes. American Psychological Association, pp. 263–293. <https://doi.org/10.1037/14342-010>
- McConahay JB (1986) Modern racism, ambivalence, and the Modern Racism Scale. In: Dovidio JF, Gaertner SL (eds) *Prejudice, discrimination, and racism*. Academic Press, pp. 91–125
- Parker S, Ruths D (2023) Is hate speech detection the solution the world wants? *Proc Natl Acad Sci* 120(10):e2209384120. <https://doi.org/10.1073/pnas.2209384120>
- Parvareh V (2023) Covertly communicated hate speech: a corpus-assisted pragmatic study. *J Pragmat* 205:63–77. <https://doi.org/10.1016/j.pragma.2022.12.009>
- Paz MA, Montero-Díaz J, Moreno-Delgado A (2020) Hate speech: a systematized review. *SAGE Open*, 10(4). <https://doi.org/10.1177/2158244020973022>
- Pettigrew TF, Meertens RW (1995) Subtle and blatant prejudice in western Europe. *Eur J Soc Psychol* 25(1):57–75. <https://doi.org/10.1002/ejsp.2420250106>
- Pew Research Center (2019) <https://www.pewresearch.org/short-reads/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/>
- Pirlott AG, Cook CL (2018) Prejudices and discrimination as goal activated and threat driven: the affordance management approach applied to sexual prejudice. *Psychological Rev* 125(6):1002–1027. <https://doi.org/10.1037/rev0000125>
- Reynders D (2020) Countering illegal hate speech online: 5th evaluation of the Code of Conduct (Report). Directorate-General for Justice and Consumers. https://commission.europa.eu/system/files/2020-06/codeofconduct_2020_factsheet_12.pdf
- Rieger D, Kümpel AS, Wich M, Kiening T, Groh G (2021) Assessing the extent and types of hate speech in fringe communities: a case study of alt-right communities on 8chan, 4chan, and Reddit. *Soc Media+ Soc*, 7(4). <https://doi.org/10.1177/20563051211052906>
- Sanchez-Mazas M, Licata L (2015) Xenophobia: social psychological aspects. *Int Encycl Soc Behav Sci* 25:802–807. <https://doi.org/10.1016/B978-0-08-097086-8.24031-2>
- Sanguinetti M, Poletto F, Bosco C, Patti V, Stranisci M (2018) An Italian Twitter corpus of hate speech against immigrants. In: Proceedings of the 11th International Conference on Language Resources and Evaluation, Miyazaki, 2798–2805. <https://aclanthology.org/L18-1443>
- Sap M, Card D, Gabriel S, Choi Y, Smith NA (2019) The risk of racial bias in hate speech detection. In: Proceedings of the 57th Annual Meeting Of the Association for Computational Linguistics, Florence, 1668–1678. <https://doi.org/10.18653/v1/P19-1163>

- Schäfer J (2023) Bias mitigation for capturing potentially illegal hate speech. *Datenbank-Spektrum* 23(1):41–51. <https://doi.org/10.1007/s13222-023-00439-0>
- Schmidt A, Wiegand M (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, Valencia*, 1–10. <https://aclanthology.org/W17-1101.pdf>
- Siegel AA (2020) Online hate speech. In: Persily N, Tucker JA (eds) *Social media and democracy: the state of the field, prospects for reform*. Cambridge University Press, pp. 56–88. <https://doi.org/10.1017/9781108890960>
- Stephan WG, Stephan CW (2000) An integrated threat theory of prejudice. In: Oskamp S (ed), *Reducing prejudice and discrimination*. Lawrence Erlbaum Associates Publishers, pp. 23–46
- Tajfel H, Turner JC (1979) An integrative theory of intergroup conflict. In: Worchel S, Austin WG (eds), *The social psychology of intergroup relations*. Brooks/Cole, pp. 33–47
- Tindale CW (2007) *Fallacies and argument appraisal*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511806544>
- Udupa S (2023) Extreme speech. In: Strippel C, Paasch-Colberg S, Emmer M, Trebbe J (eds), *Challenges and perspectives of hate speech research*. Berlin, pp. 233–248 <https://doi.org/10.48541/dcr.v12.14>
- Udupa S, Maronikolakis A, Wisioerek A (2023) Ethical scaling for content moderation: extreme speech and the (in)significance of artificial intelligence. *BigData Soc*, 10(1). <https://doi.org/10.1177/20539517231172424>
- Vala J, Lopes D, Lima M (2008) Black immigrants in Portugal: Luso-tropicalism and prejudice. *J Soc issues* 64(2):287–302. <https://doi.org/10.1111/j.1540-4560.2008.00562.x>
- Valentim JP, Heleno AM (2018) Luso-tropicalism as a social representation in Portuguese society: variations and anchoring. *Int J Intercultural Relat* 62:34–42. <https://doi.org/10.1016/j.ijintrel.2017.04.013>
- van Dijk TA (1992) Discourse and the denial of racism. *Discourse Soc* 3(1):87–118. <https://doi.org/10.1177/0957926592003001005>
- Van Dijk TA (1993) Principles of critical discourse analysis. *Discourse Soc* 4(2):249–283
- van Eemeren FH, Garssen B (2023) The pragma-dialectical approach to the fallacies revisited. *Argumentation* 37(2):167–180. <https://doi.org/10.1007/s10503-023-09605-w>
- Walton DN (1996) Practical reasoning and the structure of fear appeal arguments. *Philos Rhetor* 29(4):301–313. <https://www.jstor.org/stable/40237910>
- Wicker H-R (2001) Xenophobia. In: Smelser NJ, Baltes PB (eds), *International Encyclopedia of the Social & Behavioral Sciences*. Pergamon, pp. 16649–16652. <https://doi.org/10.1016/B0-08-043076-7/00980-3>
- Wodak R, Meyer M (2006) *Critical discourse analysis*. Qualitative Research Practice: Concise Paperback Edition. SAGE Publications

Acknowledgements

This research was funded by the European Union: CERV-2021-EQUAL (101049306). Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or kNOWHATE Project. Neither the European Union nor the kNOWHATE Project can be held responsible for them. This work was also financially supported by ITI-LARSyS, funded by FCT projects 10.54499/LA/P/0083/2020; 10.54499/UIDP/50009/2020 & 10.54499/UIDB/50009/2020; by Centro de Línguas, Literaturas e Culturas (CLLC), Universidade de Aveiro, funded by national funds, FCT - Fundação para a Ciência e a Tecnologia, I.P., project UIDB/04188/2020; by national funds through FCT - Fundação para a Ciência e a Tecnologia, under project UIDB/50021/2020 (DOI:10.54499/UIDB/50021/2020); by FCT - Fundação para a

Ciência e a Tecnologia, grant UIDB/00315/2020 (DOI: 10.54499/UIDB/00315/2020). and by FCT - Fundação para a Ciência e Tecnologia within projects: UIDB/04466/2020 and UIDP/04466/2020.

Author contributions

Rita Guerra: Funding acquisition, Supervision, Conceptualization, Project Administration, Writing – Original Draft Preparation, Writing – Review & Editing; Paula Carvalho: Supervision, Conceptualization, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; Catarina Marques: Data curation, Formal analysis (lead), Writing – Original Draft Preparation; Margarida Carmona: Writing – Original Draft Preparation, Conceptualization; Rodrigo Sarroeira: Data curation, Formal analysis; Fernando Batista: Software, Data curation (lead), Supervision, Writing – Review & Editing; Ricardo Ribeiro: Software, Data curation (lead), Supervision, Writing – Review & Editing; António Fonseca: Software, Data curation, Sérgio Moro: Software, Data curation, Cláudia Silva: Writing – Review & Editing.

Competing interests

The authors declare no competing interests.

Ethical approval

This article does not contain research conducted with human participants, thus ethical approval was not mandatory and no informed consent was used.

Informed consent

This article does not contain research conducted with human participants, thus no informed consent was used.

Additional information

Correspondence and requests for materials should be addressed to Rita Guerra.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025