COOPERHATE

COUNTERING HATE SPEECH

Hate Speech Glossary & Handbook

Co-funded by the European Union

COOPERHATE.EU

Authors

Rita Guerra (CIS-Iscte), Raquel António (CIS-Iscte), Paula Carvalho (cllc, Universidade de Aveiro)

Contributing authors

Mónica Gonçalves & Vânia Sampaio (IPS_Innovative Prison Systems), Tânia Azevedo (ILGA Portugal), Anizabela Amaral & Beatriz Realinho (SOS Racismo)

COOPERHATE: Multidisciplinary cooperation approach to prevent and counter hate crime and hate speech

Funding statement: Cooperhate, under the Grant Agreement Nº. 101213938, is funded by the European Union. Views and opinions expressed in this document are, however, those of the authors only and do not necessarily reflect those of the European Union. The European Union cannot be held responsible for them.

How to cite this report:

Guerra, R., António, R., & Carvalho, P. (2026). Hate Speech: Digital Glossary & Handbook. ISBN: 978-989-584-299-5.

Table of Contents

Co-funded by the European Union

BACK TO TABLE OF CONTENTS

Co-funded by the European Union

1. Introduction

The **Hate Speech Glossary & Handbook** provides a structured and accessible reference tool to support a shared, evidence-based understanding of hate speech in the Portuguese context. It clarifies key concepts, explains how hateful discourse manifests within broader ecosystems of hate, and compiles relevant terms, expressions, and symbolic references commonly associated with hate speech.

This handbook builds on and expands the knowledge developed under the kNOwHATE project (CERV-2021-EQUAL 101049306), incorporating updated theoretical and empirical findings and refining a contextual interpretation of hate speech, in light of the most recent research. It aligns terminology and analytical insights with the broader evidence-based approach developed within COOPERHATE, while remaining a public-facing, educational resource.

By combining solid conceptual clarification and an overview of targeted groups and recurring patterns, the Hate Speech Glossary & Handbook aims to i) promote shared understanding and evidence-based knowledge among civil society organisations, public authorities, and the wider public; ii) strengthen consistency in training, awareness-raising and prevention efforts; iii) foster cooperation among relevant stakeholders involved in tackling the growing threats posed by hateful discourses; iv) and improve recognition of harmful patterns, supporting more effective responses to hate speech and its multiple impacts.

Importantly, the Hate Speech Glossary & Handbook is intended to promote informed, contextual and responsible interpretation of language, rather than to function as a blocklist or a legal classification tool.

2. Disclaimer and Ethical Use

2.1. Why Context and Patterns Matter

Language does not exist in isolation. As emphasised in Critical Discourse Studies, meaning is socially constructed and inherently context-dependent, shaped by historical, political, and ideological factors (van Dijk, 2023; Wodak, 2015). Beyond the content of the message itself, meaning and potential harm are shaped by the broader socio-historical and discursive context in which it is produced and disseminated, as well as by communicative intent and speaker positioning, including factors such as tone and repetition (e.g., Calderón et al., 2021; Theofilopoulos, 2024). The potential for harm is further influenced by the medium of transmission and the susceptibility of the audience (Benesch, 2023).

Hate speech often operates through patterns rather than isolated elements, including combinations of linguistic and symbolic resources, the persistence of narratives over time, and the targeting of specific social groups (e.g., Mannocci et al., 2024). Accordingly, the inclusion of a word, phrase, or symbol in this handbook does not automatically imply that its use constitutes hate speech; rather, it should be understood as a potential indicator requiring an integrated contextual interpretation.

The handbook is designed to foster awareness and informed reflection, while recognising that each situation requires careful, context and culturally sensitive approaches and, where relevant, appropriate legal consideration (see section 8).

2.2. Evolving Language and Digital Environments

Language evolves rapidly, and this evolution is particularly accelerated in digital environments. Expressions may shift in meaning, new coded references may emerge, and symbols may be reused or reinterpreted across platforms and communities (e.g., Magu & Luo, 2018). In these environments, users may also

employ irony, humour, euphemisms and so-called "dog whistles"[1] to obscure, soften or conceal harmful meanings (Baider & Constantinou, 2020). These indirect or covert forms of hate speech rely on shared knowledge within specific communities, making them harder to identify as discourse evolves (Baider, 2022; Baider & Constantinou, 2020). The dynamic, context-dependent nature of language on social media makes the detection and classification of hate speech increasingly complex (Geetanjali & Kumar, 2025). In addition, hate speech is no longer confined to written text. As digital platforms evolve, it appears across multiple formats and modes, including images, memes, audio and video (Prabhu & Seethalakshmi, 2025). This multimodal character further complicates efforts to identify and respond to harmful discourse.

For these reasons, this handbook should be understood as a non-static resource. It reflects current knowledge and research findings, but it does not aim to provide a fixed and comprehensive catalogue of hate speech expressions.

2.3. Ethical and Responsible Use of This Handbook

This handbook is designed for educational, preventive, awareness-raising and capacity-building purposes. It should not be used as a standalone tool for the detection or classification of hate speech, nor for any related actions, including content removal or legal classification. Furthermore, the examples included in this handbook must not be reproduced or disseminated in ways that could contribute to amplifying or normalising harmful content.

When discussing or referencing terms and examples described in this handbook, particular care should be taken to avoid re-victimisation or unnecessary exposure of targeted communities. The authors have carefully balanced the decision of reproducing hateful content and the need to increase awareness and knowledge about this dangerous form of speech. Selected examples include both direct and indirect hateful, derogatory, and offensive

[1] Messages created to be understood by a specific audience while remaining ambiguous or deniable to the broader public (Åkerlund, 2021)

language, retrieved from publicly available social media data. This decision was made to minimise the risk of harm and follow the APA Code of Conduct for research, which recommends that "*the risk of harm must be no greater than in ordinary life, i.e. individuals should not be exposed to risks greater than or additional to those encountered in their normal lifestyles*". Exposing people to hate speech may trigger unwarranted negative consequences, and these are discussed in section 5, alongside existing strategies to counter it (section 7).

Finally, a responsible use requires sensitivity, proportionality and respect for fundamental rights. This includes freedom of expression, a fundamental right guaranteed by the Constitution of the Portuguese Republic in Article 37(1). However, freedom of expression is not an absolute right. This right is enshrined in international and regional human rights treaties, including in the case law of the Inter-American Court and the Inter-American Commission on Human Rights, particularly in Article 13(5). There are also limits to the right to freedom of expression established by Portuguese law. This occurs when freedom of expression infringes upon the rights of others by promoting discrimination, violence and incitement to hatred against a person or group on grounds related to their ethnicity, origin, religion, nationality or sexual orientation, amongst others, as stated in Article 240 of the Portuguese Penal Code (see Section 8 for a detailed discussion). Distinguishing between clear instances of hate speech and potentially legitimate expression, including satire or sarcasm, requires careful, context-sensitive judgment.

By fostering contextual awareness and ethical engagement, this handbook seeks to contribute to more informed responses to hate speech while safeguarding democratic values and human dignity.

PART I

Understanding Hate Speech

3. What is Hate Speech?

The handbook builds on the conceptual approach and empirical findings from the **kNOwHATE project**, specifically the linguistic-discursive and socio-psychological features of online hate speech, as well as its main target groups.

3.1. Working Definition

There is no single, universally accepted definition of **hate speech**; however, for the purposes of COOPERHATE, we rely on the working definition of hate speech developed within kNOwHATE, which draws on the Council of Europe's recommendations and on social psychological literature (Guerra et al., 2025). Hate speech is, therefore, approached 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)*" (Guerra et al., 2025, p. 2). Importantly, hate speech manifests in multiple ways, involving both overt and direct expressions and more covert and subtle forms.

3.2. Different Types of Hate Speech

Following the approach developed by Guerra and colleagues (2025), it is important to distinguish between two types of hate speech based on their expression: direct and indirect.

In **direct hate speech**, there is an explicit spread or justification of hatred, exclusion, discrimination, and/or violence against a target group or individual based on perceived group membership. It typically contains biased, inflammatory language, insults, and derogatory terms. The following example illustrates this pattern, in which the target group is constructed as an outgroup

and dehumanised through the metaphor "parasites", alongside explicit reinforcement of negative stereotypes ("don't want to do anything"):

> *"Racism my ass! If it weren't for these parasites of society who don't want to do anything, Portugal would be a paradise."*

Indirect hate speech avoids the use of explicit derogatory or insulting language; instead, the expression of spreading, promoting, or justifying hatred, exclusion, discrimination, or violence is typically implicit and subtle. From a semantic point of view, its meaning is often not literal and must be pragmatically inferred, drawing on social and historical context (Assimakopoulos et al., 2017; Baider, 2022). The following example illustrates this pattern, involving strategies such as ingroup victimisation and the denial of racism.

> *"If a white person gets angry with another white person, that's fine! If a white person gets angry with a black person, that's racism!"*

3.3. Prevalent Forms of Online Hate Speech in Portugal

There is no official monitoring of online hate speech in Portugal, as stated in the 2025 report of the European Commission Against Racism and Intolerance (ECRI): "***While there is a lack of official and disaggregated data on incidents of hate speech in Portugal****, several credible reports from civil society organisations and other independent institutions point to a sharp rise of hate speech in the country*" (ECRI, 2025). Notwithstanding, findings from scientific projects conducted in Portugal (kNOwHATE; HateCovid; Racism and Xenophobia in Portugal: The Normalization of Hate Speech in the Public Sphere of the Internet) converge with the core conclusions of ECRI and other civil society organisations (e.g., #MigraMyths Hate Speech and immigration in Portugal, 2025; 8th Monitoring Exercise of Online Hate Speech, 2026; Sex-based cyberviolence in Portugal: perspectives of children, youth, teachers, and specialised technical staff, 2026): online hate speech became increasingly

prevalent and normalised. Importantly, findings also demonstrate that, in the Portuguese online context, **indirect (subtle or covert) forms of hate speech** are more common than direct (or explicit) hate speech (e.g., Carvalho et al., 2023; Guerra et al., 2025). This pattern was observed across different target groups and social media platforms. Similar trends have been identified in previous research conducted in other cultural contexts (e.g., Baider, 2023).

Rather than relying primarily on overt slurs or direct calls for exclusion or violence, online hate speech is mostly conveyed in subtle and nuanced ways, involving denial of racism or role reversal (Guerra et al., 2025). These forms often employ rhetorical devices, such as irony, sarcasm, rhetorical questions, metaphors, and hyperbole (Carvalho et al., 2023; Guerra et al., 2025).

The prevalence of indirect and nuanced expressions makes it particularly important to understand how hate speech operates beyond explicit insults or threats. The following section examines the recurring discursive strategies through which both direct and indirect hate speech are constructed and sustained.

Recommended readings:

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. *Journal of Language Aggression and Conflict*. Advance online publication. https://doi.org/10.1075/jlac.00085.car

Guerra, R., Carvalho, P., Marques, C., Carmona, M., Sarroeira, R., Batista, F., Ribeiro, R., Fonseca, A., Moro, S., & Silva, C. (2025). Unpacking online hate speech in Portuguese social media: a social-psychological and linguistic-discursive approach. *Humanities and Social Sciences Communications, 12*(1). https://doi.org/10.1057/s41599-025-05392-9

Silva, C., & Carvalho, P. (2023). When can compliments and humour be considered hate speech? A perspective from target groups in Portugal. *Comunicação e sociedade*, *43*, e023006. https://doi.org/10.17231/comsoc.43(2023).4135

4. How Hate Speech Manifests and Operates

Direct hate speech is often easier to identify, as it involves explicit derogatory language, such as insults or openly discriminatory statements. Indirect hate speech, however, tends to be more subtle. It may not rely on overt hostility, but instead operates through implication, framing and strategic ambiguity (Baider, 2023).

In these cases, understanding the message requires attention to **underlying linguistic-discursive strategies**: what is suggested rather than directly stated. It is important to note that direct and indirect hate speech can coexist within the same comment (Guerra et al., 2025).

Both direct and indirect hate speech mobilise a range of discursive strategies, including dehumanisation, negative stereotyping, and threats, as well as the expression of negative emotions, such as hate and anger. These are often reinforced through rhetorical devices (e.g., metaphor, comparison, and verbal irony) and fallacious reasoning, particularly appeals to fear and calls to action. Indirect hate speech, however, further relies on additional mechanisms, such as the denial of hate and role reversal (Carvalho et al., 2023; Guerra et al., 2025).

The sections below outline some of the most recurrent patterns observed in online hate speech, drawing on and expanding the social psychological and linguistic-discursive dimensions developed by the kNOwHATE consortium (Carvalho & Guerra, 2023; Guerra et al., 2025). Concrete examples of linguistic, symbolic, and contextual markers associated with these patterns are provided in Appendix A.

4.1. Dehumanisation and Negative Stereotyping

One of the most powerful mechanisms through which hate speech operates is **dehumanisation**, whereby individuals or groups are denied positive human

traits and are portrayed as less human, more animal-like, thus removing moral considerations commonly extended to fellow human beings (Borinca et al., 2023; Haslam, 2006). Dehumanisation often materialises in comparisons and metaphors that reduce people to animals, objects or automata (Bahador, 2023; Kteily & Bruneau, 2017), as illustrated in the example below:

> *"Gypsies are like wild boars, they are wild animals."*

In this case, the target group is compared to "wild boars" and "wild animals", symbolically excluding them from the category of equal human beings. Such discursive strategies reinforce harmful stereotypes and portray targeted groups as inferior to the ingroup. Research shows that deliberate dehumanisation increases perceptions of threat, reduces empathy, and strengthens support for harmful actions against target groups (Borinca et al., 2023; Ghenai et al., 2025). In the context of hate speech, dehumanisation often appears alongside negative emotions, such as fear or anger (Ghenai et al., 2025; Guerra, et al., 2025).

Stereotypes are shared beliefs about a group's perceived attributes and characteristics, and they can be positive or negative. They shape how people perceive, interpret and respond to others (Dovidio et al., 2010). **Negative stereotyping** involves attributing negative, inaccurate, and unfair beliefs and characteristics to targeted social groups and is used to disparage or humiliate them through fallacious generalisations (Paz et al., 2020; Silva & Carvalho, 2023). In the context of hate speech, these stereotypes often appear alongside negative emotions such as fear or anger toward the targeted group, reinforcing perceptions of inferiority and social subordination (Guerra et al., 2025; Papcunová et al., 2023).

In the following example, the speaker portrays the Roma community as engaging in culturally, legally, and ethically unacceptable behaviours, such as incest and child marriage, thereby reinforcing societal biases and contributing to the stigmatisation and dehumanisation of this group.

> *"Poor thing. You guys sleep with cousins, uncles, brothers, etc., practice incest within the family, marry children to adults, and force them to get pregnant."*

Understanding dehumanisation and stereotyping is essential because they often precede more explicit forms of hostility (e.g., Haslam & Loughnan, 2014). By recognising these patterns early, it becomes possible to identify harmful dynamics before they escalate into open incitement or violence.

To know more about dehumanisation and stereotyping:

Dovidio, J. F., Hewstone, M., Glick, P., & Esses, V. M. (2010). Prejudice, stereotyping and discrimination: Theoretical and empirical overview. In *The SAGE handbook of prejudice, stereotyping, and discrimination* (pp. 3–29). London, England: Sage.

Haslam, N., & Loughnan, S. (2014). Dehumanization and infrahumanization. *Annual Review of Psychology, 65*(1), 399-423. https://doi.org/10.1146/annurev-psych-010213-115045

Sousa, Y., & Cabecinhas, R. (2025). Estereótipos Sociais. *Psicologia Social, Comunicação e Cultura, 69*. https://doi.org/10.21814/uminho.ed.157.6

4.2. Narratives of Threat

Hate speech often operates through narratives that portray certain groups as a **threat** to the ingroup (Guerra et al., 2025). These threats may be framed as **realistic**, suggesting that a group is a threat to the ingroup's power, resources, and general welfare, health and security, or as **symbolic**, claiming that a group threatens cultural values, religion, traditions, belief system, ideology, philosophy, morality, or worldview (Stephan & Stephan, 2000).

Below, the first example reflects a form of realistic threat, constructing the target group as associated with crime and insecurity through negative stereotyping. In

contrast, the second illustrates a symbolic threat, framing the target group (i.e., LGBTI+) as undermining cultural values and collective identity through the perceived misuse or appropriation of national symbols.

> *"When gypsies rob, kill, or attack security forces, they are not referred to as gypsies, but as young people. When it comes to praise, however, they are referred to as gypsies."*

> *"I feel offended when my country's flag is "twisted", altered, and used for a "leftist" cause. Go to bed with whomever you want, but respect and order in our symbols cannot be tampered with or alienated. You want respect... show respect!!!!"*

Research on intergroup relations shows that perceived threats are strongly associated with negative intentions and behaviours toward outgroups, including discrimination and aggression (Stephan & Stephan, 2016). Intergroup threats also distort perception: they shape how ingroup members interpret the outgroup, reinforce prejudice and negative stereotyping, and undermine the ability to engage in balanced and constructive intergroup interactions (Stephan & Stephan, 2016).

In the Portuguese context, racial, xenophobic and sexual prejudice have been often mobilised through narratives that construct targeted communities as posing both realistic and symbolic threats to society, typically grounded in fallacious reasoning such as *appeals to fear* and *calls for action* (Almeida & Pereira, 2026; Carvalho et al., 2023; Guerra et al., 2025; see section 4.6).

To know more about narratives of threat:

Pereira, C. R., & Souza, L. E. C. D. (2016). Fatores legitimadores da discriminação: Uma revisão teórica. Psicologia: teoria e pesquisa, 32(2), e322222. https://doi.org/10.1590/0102-3772e322222

Stephan, W. G., & Stephan, C. W. (2000). An integrated threat theory of prejudice. In S. Oskamp (Ed.), *Reducing prejudice and discrimination*, (pp. 23-46). Lawrence Erlbaum Associates Publishers.

Stephan, W. G., & Stephan, C. W. (2016). Intergroup Threats. *The Cambridge Handbook of the Psychology of Prejudice*, 131–148. https://doi.org/10.1017/9781316161579.00

4.3. Denial of Hate and Role Reversal

Another common strategy in hate speech is the **denial of hate**. Speakers may present their statements as neutral, factual or merely expressing concern, while explicitly rejecting accusations of prejudice. This strategy helps protect the speaker's credibility, reduce the perception of hate and maintain the legitimacy of their claims (van Dijk, 1992). In the example below, 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.

> *"OMG, I swear, I'm not racist, but if I had to remove one race from Europe, it would be this one."*

Closely related is the use of **role reversal**, one of the strongest forms of denial (van Dijk, 1992). In this case, the members of socially dominant or majority groups portray themselves as the true victims of discrimination or prejudice, while depicting minority groups as unfairly privileged or threatening (Guerra et al., 2025). In the example below, the members of the outgroup (i.e., Black people) tend to be represented as the ones who are intolerant, and the ones belonging to the ingroup (i.e., White people) as the victims.

> *"99.9% of black people are fanatical racists; they are always waiting for an opportunity to hurt"*

In the Portuguese context, such strategies are often linked to the enduring influence of **Lusotropicalism**: an ideology that portrays Portuguese society as historically tolerant and uniquely capable of harmonious intercultural relations (Bastos, 2019; Valentim, 2011). References to supposed "special skills" in managing diversity may be used to deny the existence of racism or systemic inequality, thereby minimising the experiences of marginalised communities and reframing discussions of discrimination as exaggerated or unjustified.

Understanding these mechanisms is essential because they often appear more subtle and socially acceptable than explicit insults. Narratives of threat, denial and role reversal can normalise exclusion while maintaining an appearance of reasonableness, making them particularly powerful within contemporary public debate.

To know more about denial of hate, role reversal and colonial narratives:

Valentim, J. P., & Heleno, A. M. (2018). Luso-tropicalism as a social representation in Portuguese society: Variations and anchoring. International *Journal of Intercultural Relations, 62*, 34-42. https://doi.org/10.1016/j.ijintrel.2017.04.013

van Dijk, T. A. (1992). Discourse and the Denial of Racism. *Discourse & Society, 3*(1), 87–118. https://doi.org/10.1177/0957926592003001005

Van Nieuwenhuyse, K., Bentrovato, D., & Valentim, J. P. (2026). The colonial past and/in history textbooks: a literature review. *Current Opinion in Psychology, 68*,102264. https://doi.org/10.1016/j.copsyc.2025.102264 .

4.4. Euphemisms and Code Words

Hate speech is often communicated through **euphemisms and coded language,** linguistic strategies that allow speakers to express hostility while softening its appearance or avoiding moderation and social sanction (e.g., Magu & Luo, 2018). Although these strategies overlap, they differ in how they communicate hidden meaning.

Euphemisms involve replacing openly discriminatory expressions with terms that sound more neutral, technical or even humorous (e.g., Magu & Luo, 2018). This can make exclusionary ideas less explicit and therefore more socially acceptable, while still allowing the intended meaning to be understood by those who recognise it. For example, the expression "send him home" in the case below may be understood as a euphemistic reference to exclusion or removal from the country:

> *After all, the man [Black anti-racist activist] 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."*

However, euphemisms rarely occur in isolation. In this example, the expression does not mainly serve to soften the message. Instead, the apparent softening is strategic: it coexists with and supports an exclusionary message that is reinforced by the broader discourse. In this sense, euphemistic language is used less to soften meaning and more to indirectly express discriminatory ideas while maintaining plausible deniability.

Code words are expressions that carry a hidden or secondary meaning within particular communities, including extremist or far-right networks (Åkerlund, 2021; Calderón et al., 2021; Magu & Luo, 2018). On the surface, these terms may seem innocuous. However, in specific contexts, they function as signals to conspiracy narratives, racial hierarchies, or hostility toward particular groups. Their ambiguity allows speakers to evade detection systems and deny hateful intent if challenged. In the example below, the speaker uses the term "Google"

to refer to Afro-descendant communities, with the meaning made explicit through the hashtag #backtoafrica:

> *"@user I'm sick of these worthless googles »#backtoafrica"*
>
> *Code Word: Google*
>
> *Actual Word: Black people*

In digital environments, coded language evolves rapidly. As certain terms become widely recognised and moderated, new variations or substitutions emerge. This highlights the importance of contextual interpretation: identifying hate speech requires attention not only to the literal wording, but also to patterns of use, audience interpretation and the broader discursive context (e.g., Calderón et al., 2021).

Therefore, recognising coded references is essential for understanding how contemporary hate speech adapts, persists and circulates, often in ways that appear subtle, ironic or indirect on the surface.

To know more about euphemisms and code words:

Magu, R., & Luo, J. (2018). Determining Code Words in Euphemistic Hate Speech Using Word Embedding Networks. *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. https://doi.org/10.18653/v1/w18-5112

4.5. Socio-Historical References

Hate speech can draw on selective interpretations of history to legitimise exclusion in the present. **Socio-historical references** may invoke colonial nostalgia, authoritarian or dictatorship-era rhetoric, the denial or distortion of well-documented historical events, and be mobilised in conspiracy theories (e.g., Almeida et al., 2023; Baider, 2022; NCTV, 2024; Papcunová et al., 2023).

These references often present the past as a time of order, unity or national greatness, implicitly or explicitly contrasting it with a present described as weakened by diversity, migration or social change. By idealising certain historical periods, such narratives can normalise hierarchical worldviews and reinforce the idea that some groups "naturally" belong while others do not (e.g., Couperus et al., 2023; Homolar & Löfflmann, 2021; Kentmen-Cin, 2025). Research shows that these narratives work by evoking a past as glorious and victorious, then framing its loss as a shared humiliation directed against those considered not to belong to the "true" people (i.e., the ingroup) (Homolar & Löfflmann, 2021).

In other cases, these references rely on historical denial or minimisation, such as downplaying past discrimination, racism, colonial violence or systemic oppression, to delegitimise contemporary claims for equality and recognition (Castelo, 2021; van Dijk, 1992). In the case of Portugal, these references are often connected to the endorsement of Lusotropicalism (see section 4.3). The use of historical references in hateful discourse usually mobilises discursive strategies such as the denial of racism (see section 4.3), which effectively disables resistance (van Dijk, 1992): when past and present injustice is denied, calls for anti-racist policies or legal redress are rendered unnecessary. Right-wing populist discourse, for instance, has been shown to rehabilitate national glory by downplaying collective crimes (e.g., colonialism) through selective reappraisal or fabrication of historical accounts (Couperus et al., 2023).

The reinterpretation of past historical events, justifying suspicion, exclusion or hostility toward specific groups (e.g., the Holocaust; Williams, 2022), is often present in conspiracy theories. Drawing on collective memory and historical references, conspiracy theories frequently repurpose long-standing intergroup histories of distrust, conflict, marginalisation, and institutional betrayal to lend coherence to new claims about present-day threats (Wagoner et al., 2026). These narratives create a sense of continuity between past and present threats, reinforcing fear and distrust.

The example below illustrates the intersection between conspiracy thinking and hate speech through the reinterpretation of colonial history. Rather than acknowledging historical asymmetries, the statement reframes colonialism as a

benevolent process, attributing "infrastructure", "technology" and "culture" to Portuguese colonial presence. This constitutes a form of historical revisionism, which is typical of conspiracy narratives that repurpose collective memory to construct simplified accounts of past and present relations (Inwood & Zappavigna, 2023; Wagoner et al., 2026).

> *"If there were to be any kind of reparation, it should be compensation to Portugal for its former colonies, for all the infrastructure and technology left behind by the Portuguese, and for the spirit and culture they gave to the world.*
>
> *I will not allow them to shame our country any longer!"*

The example combines conspiracy-based historical reinterpretation with hate speech strategies, such as collective blame reversal, ingroup glorification, and implicit exclusion of the outgroup, thereby normalising an exclusionary narrative without explicit derogatory language.

Understanding the use of socio-historical references is important because they give hate speech a deeper layer of meaning. By anchoring prejudice in seemingly shared memories or national identity, such narratives can make exclusion appear justified, inevitable or even patriotic. Recognising how history is mobilised helps to identify the broader discursive frameworks within which hate speech operates. Examples of such symbolic and historically embedded markers can be found in Appendix A.

To know more about socio-historical references:

Almeida, P., Pereira, J., & Candido, D. (2023). Online hate speech on social media in Portugal: extremism or structural racism? *Social Identities, 29*(5), 419–435. https://doi.org/10.1080/13504630.2024.2324277

Castelo, C. (2021). *"Portuguese non racism": On the historicity of an invented tradition*. https://cesa.rc.iseg.ulisboa.pt/afroport/portuguese-non-racism-on-the-historicity-of-an-invented-tradition/

Wagoner, B., Jørgensen, M. S., & Pahuus, K. (2026). Conspiracy theories through the lens of collective memory. *Current Opinion in Psychology, 68*, 102227. https://doi.org/10.1016/j.copsyc.2025.102227

4.6. Fallacies

Hate speech is often expressed indirectly, using subtle strategies that shape how audiences interpret situations and respond to them. Rather than relying solely on explicit hostility, messages draw on **fallacies** and emotionally charged language, departing from the standards of balanced and critical discussion. Two fallacious reasoning are particularly relevant: *appeal to fear* and *call to action*, both of which can contribute to the spread and reinforcement of hate speech (Carvalho et al., 2023; Guerra et al., 2025).

Appeal to fear is a persuasive strategy that does not involve an explicit threat, but rather a warning that a negative outcome will occur if the receiver does not adopt the (explicit or implicit) recommended action (Tindale, 2007), as illustrated in the following example:

> *"Minorities? We are the minority if we look at things globally... We went from 30% of the world population in 1930 to less than 11%, and if the trend continues, in 30 years we will be less than 7%..."*

In this comment, demographic statistics are selectively mobilised to construct a narrative of demographic decline and loss of majority status. Population data are reframed as an implicit warning of future marginalisation of the ingroup, thereby eliciting insecurity and anxiety through the projection of numerical inversion.

Call for action is a closely related strategy that involves an explicit or implicit request for behavioural response aimed at reversing a perceived negative situation, typically conveyed through an emotionally charged tone, as illustrated in the following example:

> *"Thank the European Union for destroying Europe and Portugal! Stop voting for left-wing politicians!"*

Here, direct imperatives ("Stop voting") and evaluative language construct a narrative of societal decline attributed to political actors and supranational institutions associated with outgroups. The emotionally charged framing and attribution of blame function to mobilise the audience towards concrete behavioural change, particularly voting behaviour, grounded in a discourse of political and cultural deterioration.

To know more about fallacies:

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. *Journal of Language Aggression and Conflict*. Advance online publication. https://doi.org/10.1075/jlac.00085.car

Tindale, C. W. (2007). *Fallacies and Argument Appraisal*. Cambridge University Press. https://doi.org/10.1017/cbo9780511806544

4.7. Negative Emotions

Emotions play a key role in shaping intergroup attitudes and behaviours, with different emotions linked to distinct motivational tendencies, such as confrontation or avoidance (Cottrell & Neuberg, 2005; Mackie & Smith, 2015). Certain **negative emotions**, such as hate and anger, play a central role in the formation, expression, and justification of online hate speech (Ghenai et al., 2025; Guerra et al., 2025).

Hate is a powerful negative emotion or, over time, a more stable sentiment that emerges when individuals or groups are perceived as having malicious intentions to harm the ingroup (Fischer et al., 2018). In hate speech, this emotion is reflected in the portrayal of targeted groups as inherently immoral, dangerous and unchangeable (Guerra et al., 2025). Hate is frequently associated with processes of dehumanisation, silencing and realistic and symbolic threats. It is a strong predictor of harmful intentions and behaviours. In fact, hate motivates actions aimed not only at hurting, but at excluding, humiliating, or even eliminating the target at psychological, social or physical levels (Fischer et al., 2018). In this sense, hate goes beyond momentary hostility, often involving a desire for punishment, revenge or the removal of the perceived source of harm.

As illustrated in the examples below, this emotion is reinforced through several discursive strategies, including humiliation, dehumanisation, explicit calls for action (namely, exclusion or even annihilation of the target group), and expressions of intense hostility:

> *“KICK THEM OUT OF THE COUNTRY LIKE THE STRAY DOGS THEY ARE!!! I'M SICK OF THESE PEOPLE!!! ENOUGHHHH”*
>
> *“KILL MY ASS. GO BACK TO YOUR LAND.BUT AFTER ALL, WHAT TRIBE ARE YOU FROM.[Black anti-racist activist].AND THERE'S NO ONE THAT KILLS HIM."*

Anger is an emotion that arises when the actions of an outgroup are perceived as unfair, unjustified or deviating from established social norms (Fischer et al., 2018). Unlike hate, which is often rooted in perceptions of inherent malicious intent, anger is typically directed at specific behaviours and situations.

In hate speech, anger is reflected in messages that attribute responsibility to the outgroup and frame them as obstacles to the ingroup’s goals, values or well-being (Guerra et al., 2025). These may include perceived threats to economic

resources, social order, or rights. Anger often motivates confrontation, criticism or attempts to correct or challenge others.

Anger can involve a willingness to harm, but unlike from hate, this is usually linked to a desire to respond to perceived injustice or to remove obstacles, rather than to eliminate the target. As illustrated in the example below, anger is thus associated with narratives of realistic and symbolic threat, as well as with discursive strategies such as role reversal (Carvalho & Guerra, 2023; Guerra et al., 2025).

> *"Go ahead [Far-Right extremist], you've done very well. These scum will have to learn to respect the Portuguese and stop making fun of good citizens. Long live Portugal".*

To know more about hate and anger:

Fischer, A., Halperin, E., Canetti, D., & Jasini, A. (2018). Why we hate. *Emotion Review, 10*(4), 309-320. https://doi.org/10.1177/1754073917751229

Halperin, E. (2011). The emotional roots of intergroup aggression: The distinct roles of anger and hatred. In P. R. Shaver & M. Mikulincer (Eds.), *Human aggression and violence: Causes, manifestations, and consequences* (pp. 315–331). American Psychological Association. https://doi.org/10.1037/12346-017

4.8. Context, Interaction and Escalation

Online communication is organised through conversational exchanges, including message threads, replies, reposts, likes, hashtags and shared images. Thus, online hate speech's meaning and potential harm often become clearer when examined within the broader conversational context in which it is embedded (Calderón et al., 2021).

For instance, an ambiguous or potentially inoffensive comment can be disambiguated when read alongside previous messages, repeated references or

coordinated responses. These interaction patterns may reveal coded hostility, shifting meanings or indirect forms of aggression that remain opaque when considered in isolation (e.g., Calderón et al., 2021; Fonseca et al., 2024).

The following indicators may help identify how hate speech emerges, intensifies and spreads:

- **Patterns of repetition:** the repeated use of slurs, memes (e.g., far-right memes; NCTV, 2024), or coded language/code words (e.g., Calderón et al., 2021; Magu & Luo, 2018), including letter repetitions and symbolic elements (e.g., Mansur et al., 2024), can reinforce hateful narratives and may indicate coordinated or persistent targeting of groups (e.g., Arce-García et al., 2025; Ghenai et al., 2025; Magu & Luo, 2018; Mansur et al., 2024).

> *For instance, angry, laughing or vomit emojis, often appearing in sequences of three or more, or as combinations presented in a patterned manner, were found to be common in hate speech targeting gender minorities (Theofilopoulos, 2024).*

- **Coordinated posting**: Multiple accounts or group members sharing similar or synchronised hateful content (e.g., through hashtags; Mannocci et al., 2024) may indicate organised amplification or bot[2] activity, increasing the visibility and reach of hateful messages (e.g., Arce-García et al., 2025; Pontes et al., 2024).

> *One recent example in the Portuguese context is the spread of specific disinformation narratives (e.g., that "Portugal is being invaded", "Islamisation of Portugal") during elections in 2024 and 2025 that involved coordination between anonymous accounts to amplify content aligned with the Chega party's discourse. The accounts had a strong overlap of followers and*

[2] Bots are algorithms developed to automatically generate content and interact with human users (Uyheng & Carley, 2020).

> *synchronised sharing and reached hundreds of thousands of views at key moments (Cardoso et al., 2025).*

- **Network signals**: Patterns of interaction between users can help identify how hate speech spreads and gains visibility online. Rather than focusing only on individual messages, it is important to consider how content circulates across networks, for example, through repeated sharing, coordinated posting or the activity of highly visible accounts (this can be identified via Social Network Analysis[3]). Research shows that hateful content often forms tightly connected clusters, which facilitates its persistence and dissemination within online spaces (Ghenai et al., 2025). These interconnected networks allow harmful narratives to circulate more easily, reinforcing shared beliefs and increasing their reach. Looking at broader interaction patterns, such as who shares content, how often it is repeated, and how groups of users are connected, can help identify the complex dynamics of hate speech propagation online (Fonseca et al., 2024; Pontes et al., 2024).

- **Escalation in speech**: Hate speech can escalate over time through repeated interactions, conversations, and group dynamics, evolving progressively shifting from subtle or indirect expressions (e.g. negative stereotypes) to explicit hostility, calls for action, or incitement to violence. This escalation is illustrated in the Pyramid of Hate (Williams, 2022, adapted from Gordon Allport, The Nature of Prejudice, 1954), where, under specific conditions, verbal rejection progresses to social exclusion, violence and genocide.

- **Escalation dynamics and online-offline correlations**: While online hate does not directly cause offline violence, there is currently solid evidence of how online and offline violence correlate, and how the spread of hate speech

[3] Social Network Analysis (SNA) is a methodological approach and research methodology that delves deep into the patterns of relationships between interconnected actors (i.e., individuals, groups, organisations, or other units). SNA maps out these relationships in the form of networks where 'nodes' stand for actors and 'edges' symbolise connections between them.

on social media precedes incidents of hate crimes in the physical world (e.g., Calderón et al., 2024; Madriaza et al., 2025; Williams et al., 2020). For example, research shows that anti-refugee posts on the Facebook page of the far-right party Alternative für Deutschland (AfD) were associated with violent offline crimes against immigrants in Germany (Müller & Schwart, 2021). In the UK, online hate speech targeting Black and Muslim communities correlated with racial and religious hate crimes in London, and preceded the attacks (Williams et al., 2020). Similar findings were found in the USA and Spain. In the United States, online hate speech targeting racialised and LGBTI+ communities was found to correlate with offline hate crimes against these groups (Bozhidarova et al, 2023). Likewise, an analysis of Twitter and Facebook posts published in Spain between 2016 and 2018 revealed correlations between online hate speech and hate crimes targeting migrant communities and LGBTI+ individuals (Calderón et al., 2024).

- **Contextual Drivers of Hate Speech**
  - **Social, Political and Crisis-Related Triggers**: Major societal events can act as catalysts for hate speech, with increases observed following terrorist attacks, as well as during highly polarising political moments such as Brexit in the United Kingdom or the Peace Accords in Colombia, often amplified through disinformation (Madriaza et al., 2025). In moments of crisis, societies search for explanations or scapegoats, and minority groups can become convenient targets for blame, particularly when misinformation or conspiracy narratives circulate widely (Šrol et al., 2022).

> *For instance, this was especially visible during the COVID-19 pandemic, described by the United Nations as triggering a "tsunami of hate" online, marked by the proliferation of xenophobic, misogynistic, transphobic, and conspiratorial content (Vergani et al., 2024).*

  - **Information Environments and Crisis Amplification:** These dynamics can be more pronounced in conflict-affected or high-risk contexts,

where fragile institutions, limited access to reliable information, and restrictions on freedom of the press create particularly vulnerable information environments. In such settings, heightened uncertainty and low trust facilitate the rapid spread of rumours, misinformation, disinformation (see section 4.9), and hate speech, often intensifying existing divisions and amplifying their real-world consequences (Wardle, 2024).

- **Political Instrumentalisation of Hate Speech:** Hate is also often overused or strategically misused for political purposes (Bilewicz & Soral, 2020; Williams, 2022). Political debate can become a powerful catalyst for hate speech, particularly during election periods, legislative reforms or public controversies involving migration, gender equality or minority rights (Kentmen-Cin, 2025; Said-Hung et al., 2023). Research shows that online hate is used strategically to advance political aims, such as mobilising support, reinforcing ingroup and outgroup boundaries, and influencing public discourse and policy (Kentmen-Cin, 2025).

For instance, in the United States, contemporary political developments have been associated with the increased dehumanisation of immigrants, the working class, and women across both social media platforms and news outlets (Karantzas & Simpson, 2026).

In Portugal, an analysis of over 10 000 posts by key figures of the far-right political party CHEGA on X, between 2019 and 2024, showed the prevalence of topics related to identity and immigration converged with electoral cycles. This suggests the strategic use of emotionally charged xenophobic content for political mobilisation (Cardoso et al., 2025).

Overall, hateful statements by politicians can gradually desensitise people to derogatory language, making stereotyping and discrimination more likely and contributing to a broader climate of intergroup hostility. At the same time, such rhetoric can reshape norms of acceptable

behaviour, weakening existing social norms that protect minority groups (Bilewicz & Soral, 2020). Digital platforms further amplify these narratives, enabling far-right actors to extend beyond their immediate networks and create political echo chambers (see section 4.9) that help spread and normalise more radical viewpoints (Kentmen-Cin, 2025; Williams et al., 2020).

- **Sport-Related Triggers of Online Hate Speech**: Sports events can also become triggers that fuel online hate speech. Rivalries between teams, strong group identification and emotional investment may contribute to aggressive or discriminatory language (Kearns et al., 2022; Miranda et al., 2023). While much sports-related hostility may be framed as rivalry, it can intersect with racism, xenophobia, homophobia, and misogyny (Kearns et al., 2022; Miranda et al., 2023; Montesinos-Cánovas et al., 2023). In fact, hate speech can manifest during sports events, such as the case of FC Porto player Moussa Marega, who left the pitch in 2020 after being subjected to racist chants from opposing fans, or more recently, the racist and homophobic slurs targeting Vinicius Junior during a Benfica vs. Real Madrid game, leading to a suspension for discriminatory conduct. Social media has extended these dynamics beyond stadiums, and sporting events have repeatedly been identified as triggers for spikes in online hate (e.g., Carvalho et al., 2022).

For example, the UEFA EURO 2020 final between Italy and England was followed by a surge of racist abuse targeting three Black England players, a pattern also observed during subsequent tournaments such as the Women's EURO 2022 and the 2022 World Cup.

Such incidents illustrate how moments of high visibility and emotional intensity can quickly translate into targeted hostility online. Similar dynamics have also emerged in debates around the participation of transgender athletes in sport, where online discussions are frequently

accompanied by transphobic narratives, biological essentialism and exclusionary rhetoric (Murib, 2022).

> *For instance, during the Paris 2024 Olympic Games, where the case of Imane Khelif generated a wave of transphobic and racialised discourse across platforms (Taha & Sailofsky, 2025).*

Together, these indicators highlight that hate speech is not only about what is said, but also how it circulates, evolves, and gains meaning over time. The interaction between individual expression and broader social contexts shapes hate speech. Political debate, crisis, even sports events, and digital platform structures all converge to influence how hostility emerges, circulates and escalates. Recognising these ecosystems allows for more informed and context-sensitive responses. Thus, addressing hate speech effectively requires not only identifying harmful expressions, but also understanding the contexts in which they emerge, circulate and become normalised.

In line with recent recommendations emphasising that online hate analysis must move *beyond detection,* we recommend that hate speech be evaluated considering not only the presence of specific linguistic features, but also by their **combination with extra linguistic factors,** including **repetition, escalation over time, and broader contexts** (e.g., platform dynamics, trigger events).

To know more about context, interaction and escalation:

Bilewicz, M., & Soral, W. (2020). Hate Speech Epidemic. The Dynamic Effects of Derogatory Language on Intergroup Relations and Political Radicalization. *Political Psychology, 41*(S1), 3–33. https://doi.org/10.1111/pops.12670

Magu, R., & Luo, J. (2018). Determining Code Words in Euphemistic Hate Speech Using Word Embedding Networks. *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. https://doi.org/10.18653/v1/w18-5112

Williams, M. (2022). *The Science of Hate: How Prejudice Becomes Hate and What We Can Do to Stop It*. Faber & Faber, Ltd

4.9. Polarisation, Disinformation and the Normalisation of Hate

As hate speech circulates, repeats and escalates across interactions, it can become embedded in broader narratives that shape how groups and social issues are understood. In online environments, it often interacts with processes of **polarisation** and **disinformation**, including **conspiracy theories**, contributing to increasingly divided and hostile public discourse.

Polarisation can take different forms in online and social contexts. It may involve individuals adopting increasingly extreme positions (individual polarisation), groups reinforcing more radical viewpoints through internal discussion (group polarisation), opposing groups becoming more hostile and distant from one another (intergroup or political polarisation), or growing ideological and issue-based divisions on social or political issues (issue-based polarisation). Thus, polarisation should not be understood as a fixed condition, but rather as a dynamic, multilevel process shaped by disagreement and social interaction (Bliuc et al., 2024).

Intergroup polarisation, specifically, tends to reinforce “us versus them” dynamics (i.e., othering), where individuals or groups are depicted as fundamentally different, inferior, or even threatening to the ingroups (Smith et al., 2024). **Othering** is a discursive strategy often deployed in hate speech, as described in the ideological square proposed by van Dijk (1992), where individuals who promote or spread hate speech rely on positive self-presentation (*Us*) and negative other presentation (*Others*). In such contexts, related strategies such as negative stereotyping, threat narratives and dehumanising representations become more visible, more accepted and more difficult to challenge (e.g., Morales et al., 2025).

Polarisation is also related to the formation of **echo chambers** (or ‘filter bubbles’), networks of like-minded users where information circulates largely

unchallenged, with limited exposure to opposing views, partly due to ranking algorithms that filter out any contradictory posts (e.g., Gallacher et al., 2021; Harel et al., 2020; Williams, 2022). Algorithms often reinforce polarised information exposure and, in turn, influence how debates unfold and how people act online (Williams, 2022).

Within these networks of like-minded users, polarisation, **disinformation**[4], **misinformation**[5] **and malinformation**[6] may spread more easily because they are less likely to be questioned or contested (Harel et al., 2020; Williams et al., 2020; Zhang, 2018).

These mechanisms can further foster hate speech by providing seemingly factual justification for hostility, giving prejudiced and threat narratives an appearance of legitimacy (Bradshaw, 2024; Carvalho et al., 2025; Madriaza et al., 2025). While digital platforms play a significant role in shaping and accelerating these processes, they often build on longstanding patterns of prejudice and social division (Udupa, 2025).

In particular, **conspiracy theories** play an important role in these dynamics by reinforcing prejudice and hostility between groups (Bilewicz, 2025). By presenting unfounded claims in a narrative structure that resembles factual explanation, they blur the boundary between opinion and "evidence", thereby increasing the perceived credibility of hostile narratives (Burnham et al., 2022; Carvalho et al., 2025; Ghenai et al., 2025). In some cases, hate speech directly draws on conspiracy narratives, contributing to the spread of exclusionary and harmful worldviews (Bilewicz, 2025). A prominent example is the "Great Replacement" conspiracy theory, which falsely claims that white populations are being deliberately replaced by non-white immigrants as part of a plot orchestrated by elites. Such narratives reinforce rhetoric of invasion, symbolic and realistic threats, and "otherness", and are strategically used by extremist actors to intensify online polarisation, with the expectation that this may

[4] Intentional spread of inaccurate information, intended to deceive and shared to cause serious harm (Wardle, 2024).
[5] Unintentional spread of inaccurate information shared in good faith by people unaware that they are spreading falsehoods (Wardle, 2024).
[6] Information based on reality, but used out of context to inflict harm (Greene, 2025).

translate into offline outcomes, such as financial support, participation in rallies, or even hate crimes (Williams, 2022). Research shows that such theories are particularly likely to emerge in polarised social contexts characterised by hostile or ideologically dissimilar outgroups, and that the mere existence of a distrusted outgroup is often sufficient to generate unfounded conspiratorial accusations against it (van Prooijen & Douglas, 2017; van Prooijen, 2021).

Repeated exposure to hate speech, conspiracy theories and other forms of disinformation can contribute to the **normalisation of hate** and to processes of **desensitisation** (e.g., Madriaza et al., 2025; Soral et al., 2018, see sections 5.2; 5.3).

Overall, hate speech can both reflect and reinforce polarisation, fostering hostility between groups and contributing to climates of social fear and division (Ghenai et al., 2025; Madriaza et al., 2025). These dynamics highlight how hate speech operates not only at the level of individual expression, but also as part of broader social processes that shape perceptions, attitudes and behaviours over time. These processes are often amplified in broader social and institutional contexts, where they contribute to the normalisation of hate (see Section 4.8).

Recommended readings:

Bilewicz, M., & Soral, W. (2020). Hate Speech Epidemic. The Dynamic Effects of Derogatory Language on Intergroup Relations and Political Radicalization. *Political Psychology, 41*(S1), 3–33. https://doi.org/10.1111/pops.12670

Carvalho, P., Caled, D., Silva, M. J. (2025). The Thin Line Between Conspiracy Theories and Opinion: Why Humans and AI Struggle to Differentiate Them. *International Journal of Communication, 19*, 565–591. https://ijoc.org/index.php/ijoc/article/view/22182

Smith, L. G. E., Thomas, E. F., Bliuc, A.-M., & McGarty, C. (2024). Polarization is the psychological foundation of collective engagement. *Communications Psychology, 2*(1). https://doi.org/10.1038/s44271-024-00089-2

Wardle, C. (2024). *A conceptual analysis of the overlaps and differences between hate speech, misinformation and disinformation*. United Nations, Department of Peace Operations and Office of the Special Adviser on the Prevention of Genocide. https://peacekeeping.un.org/sites/default/files/report_-_a_conceptual_analysis_of_the_overlaps_and_differences_between_hate_speech_misinformation_and_disinformation_june_2024.pdf

5. Impacts of Hate Speech

Online hate speech is not necessarily less harmful than physical acts. Literature shows that it has direct impacts on targeted **individuals**, **communities**, and **society as a whole** (Hassan et al., 2022; Madriaza et al., 2025; Silva & Carvalho, 2023). Because of the anonymity that the internet provides, perpetrators are likely to engage in more hate speech, which is more severe in nature due to a lack of inhibition (Williams, 2022).

5.1. Impacts on Individuals

For those directly targeted, hate speech can lead to feelings of fear, distrust, anxiety, depressive symptoms, stress, humiliation, and isolation (e.g., Dreißigacker et al., 2024; Hassan et al., 2022; Kentmen-Cin, 2025; Madriaza et al., 2025; Wachs et al., 2022). Experiences of victimisation have also been associated with self-harming and suicidal ideation and behaviour (Madriaza et al., 2025). Repeated experiences of hate reduce life satisfaction and are associated with increased social fear (Madriaza et al., 2025).

Online environments can intensify these effects (e.g., de Roos & Caon, 2026; Madriaza et al., 2025). The speed, visibility and permanence of digital communication mean that harmful messages can reach large audiences quickly and remain accessible over time (e.g., Ghenai et al., 2025). Victims may experience harassment across multiple platforms, creating a sense of constant exposure and insecurity that transfers from online to offline spaces (Dreißigacker et al., 2024). Those who produce hateful content are also affected by it: users who post hate speech show significantly higher levels of anger, anxiety, and negative emotions, suggesting that online hate environments shape the emotional states of those who engage in them, not only those who are targeted (Ghenai et al., 2025).

5.2. Impacts on Target Communities and Bystanders

Hate speech not only affects those directly targeted but it also impacts entire communities, as exposure to hate speech in social media reinforces prejudice, stereotypes, polarised views, and deepens social divisions (e.g., Kentmen-Cin, 2025). Namely, hate speech reinforces people's negative attitudes and beliefs about social groups and is directly linked to an increase in participation in online hate speech, as well as violent behaviour in real life (de Roos & Caon, 2026). Members of targeted groups, even if they are not directly victimised, also feel its negative consequences (i.e., vicarious victimisation; Madriaza et al., 2025; Pickles, 2020).

Over time, frequent exposure to hate content can desensitise bystanders (i.e., those who witness hate speech without being directly targeted), reduce empathy, normalise discrimination, and reduce intervention in these situations (Kentmen-Cin, 2025; Madriaza et al., 2025; Soral et al., 2018). Hence, as harmful language becomes more widespread, it may be perceived as more acceptable or less extreme, shaping social norms, whereby individuals come to believe that hate speech is common and socially tolerated simply because they encounter it regularly (Bilewicz & Soral, 2020). For instance, research shows that frequent exposure to derogatory language attenuates sensitivity to others' pain, suggesting that hate speech erodes empathy (Pluta et al., 2023). Over time, desensitisation and normalisation can reduce people's ability to recognise hate speech and its consequences, while simultaneously weakening the social norms that would otherwise discourage it (Bilewicz & Soral, 2020).

5.3. Impacts on Society and Democracy

At a societal level, hate speech can erode democratic values and undermine social cohesion (Hassan et al., 2022). Public debate depends on respect, pluralism and equal participation. When certain groups are systematically portrayed as inferior, dangerous or illegitimate, their voices may be marginalised or silenced (Kentmen-Cin, 2025). Drawing on the spiral of silence theory, research suggests that as hate speech intensifies into more extreme forms of

abuse, it can silence and isolate targeted individuals and communities, discouraging them from expressing their views and participating in public discourse, thereby deepening political marginalisation and polarisation (Kentmen-Cin, 2025).

Hate speech exposure is also related to changes in political beliefs and can influence individuals' willingness to support radical or extremist political parties (e.g., right-wing) and foster agreement with extremist ideologies (Madriaza et al., 2025). Sustained online hostility can contribute to broader patterns of social tension and, in some contexts, may precede increases in real-world aggression or hate crimes (e.g., Calderón et al., 2024; Hassan et al., 2022; Madriaza et al., 2025). While online speech may not automatically cause violence, it can help shape attitudes and social norms that make violence more acceptable.

As the digital landscape continues to evolve, understanding and responding to the psychological and societal consequences of hate speech remains an urgent priority (Dreißigacker et al., 2024).

Recommended readings:

Hassan, G., Rabah, J., Madriaza, P., Brouillette-Alarie, S., Borokhovski, E., Pickup, D., Varela, W., Girard, M., Durocher-Corfa, L., & Danis, E. (2022). PROTOCOL: Hate online and in traditional media: A systematic review of the evidence for associations or impacts on individuals, audiences, and communities. *Campbell Systematic Reviews, 18*(2). Portico. https://doi.org/10.1002/cl2.1245

Madriaza, P., Hassan, G., Brouillette-Alarie, S., Mounchingam, A. N., Durocher-Corfa, L., Borokhovski, E., Pickup, D., & Paillé, S. (2025). Exposure to hate in online and traditional media: A systematic review and meta-analysis of the impact of this exposure on individuals and communities. *Campbell Systematic Reviews, 21*, e70018. https://doi.org/10.1002/cl2.70018

PART II

Targeted Groups and Patterns of Hate

6. Targeted Groups and Common Forms of Hate in Portugal

According to recent EU reports and research conducted in Portugal, the social groups that are most frequently targeted by online hate speech in the country, are **racialised communities** (including Afro-descendants and Roma), **migrants**, **LGBTI+ communities, and women** (Carvalho & Guerra, 2023; Carvalho et al., 2024; EUAFR, 2025; European Commission Against Racism and Intolerance, 2025; European Institute for Gender Equality, 2025; Guerra et al., 2025). Understanding the nuances in content and motivations underlying hate speech towards different social groups is crucial to developing effective responses to prevent and counter it. Although there are some similarities, it is important to consider that there are specific processes of discrimination underlying these patterns of hate speech.

6.1. Processes of Racialisation

The **process of racialisation** involves targeting specific groups by constructing racialised identities. Racism reflects the belief in the superiority of one's own racialised ingroup, which justifies discriminatory and derogatory behaviour toward the outgroup, normalising the privileged (Jones, 1997). Within this framework, two main target groups are often targeted in Portugal: Afro-descendants and Roma communities (Almeida & Pereira, 2026; Almeida et al., 2023; Carvalho et al., 2023; Guerra et al., 2025). The colonial legacy and the shared ideology of Lusotropicalism (see section 4.3) shape intergroup attitudes and behaviours, including subtle forms of racial derogation. For instance, colonial nostalgia and exceptionalism are prevalent narratives in hate speech targeting racialised groups, and the denial of racism is a highly relevant strategy mobilised in hateful discourse (Almeida et al., 2023; Guerra et al., 2025). Similarly, the Roma community faces centuries of discrimination and social

exclusion, echoed in social and mainstream media discourses that perpetuate stereotypes of criminality and social marginalisation. Accordingly, research shows that online hate speech against the Roma in Portugal is highly normalised and dehumanising, portraying these communities as a threat to society (Almeida & Pereira, 2026; Guerra et al., 2025).

Research further shows that users tend to cluster different forms of prejudice within the same discourse, as illustrated in the example below:

> *"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, apologising for being beaten up by Third World people, and minorities in general..."*

In this case, terms associated with the dominant ingroup (in the example, Portuguese white, cisgender, heterosexual individuals) co-occur with terms referring to multiple outgroups, including Afro-descendant, Roma, and LGBTI+ communities. In this comment, the speaker strategically reverses the positions of dominance and power between dominant groups (Portuguese/Europeans) and marginalised groups (Africans, LGBTI+ individuals, Roma, "Third World" populations, and minorities more broadly). By portraying dominant groups as potential victims, such discourse obscures structural inequalities and reframes existing power relations as unjustly inverted.

6.2. Xenophobia

Xenophobia involves prejudice, fear, or hostility toward people perceived as foreigners or outsiders, often **targeting migrants and individuals of immigrant descent** (Sanchez-Mazas & Licata, 2015). These dynamics are often rooted in ethnocentric beliefs that position one's own culture, values, or way of life as superior to others, reinforcing ingroup favouritism[7] and outgroup derogation[8].

[7] Unequal treatment through displaying favouritism towards the ingroup.
[8] Discrimination against the outgroup.

For instance, migrants are frequently viewed as threats to national identity or economic stability, with both realistic (e.g., economic resources) and symbolic (e.g., cultural or value-based) perceived threats fuelling prejudice and discrimination, as illustrated in the example below (see section 4.2). These dynamics play a crucial role in contemporary patterns of hate speech directed at migrant communities (Guerra et al., 2025).

> *"A few years ago, before the wave of immigration, television rarely reported cases of violence against civilians or law enforcement – maybe one story per month. Now, if you turn on CMTV, there are dozens of such reports every day. It cannot be a mere coincidence."*

In this example, xenophobic discourse is constructed through an implicit causal link between immigration and rising violence. By contrasting a "before" and "after" scenario, the comment frames migrants as a threat to public safety, relying on selective evidence to support a broader generalisation. This reflects the mobilization of realistic threats (i.e., violence), while maintaining plausible deniability.

Notably, hate speech may also target individuals based on their perceived **religious identity**, particularly where religion intersects with **migration and ethnicity**. For instance, in Portugal, anti-Muslim and Islamophobic rhetoric often intersects with xenophobic and racialised narratives (European Commission Against Racism and Intolerance, 2025). Accordingly, recent research shows that islamophobia was highly prevalent online, with refugees and immigrants being perceived as Muslim regardless of their diverse ethnicities and religions (Almeida et al., 2023).

6.3. Sexual Prejudice

Sexual prejudice involves negative attitudes toward people based on their perceived sexual orientation, gender identity, gender expression, or sex

characteristics. This prejudice is closely linked to sexual stigma - the belief that non-heterosexual identities and behaviours are inherently wrong or inferior, and perceived threats (Herek, 2004). Manifestations of sexual prejudice include homophobia, biphobia, transphobia, and intersexphobia, which predict discrimination and aggression such as verbal harassment, physical and sexual violence, and social exclusion (Katz-Wise & Hyde, 2012). The example below illustrates how perceptions of threat can be expressed in exaggerated and generalised ways, reflecting underlying fears about social change.

> *"These LGBTI+ people are taking over everything"*

Specifically**, trans people** in Portugal face distinct challenges related to both social prejudice and institutional barriers (e.g., Neves et al., 2023). While legal gender recognition is available through self-determination from age 16[9], trans individuals remain particularly vulnerable to hate speech, harassment, and social exclusion (e.g., Neves et al., 2023). When people believe that certain groups challenge their values or way of life, this can lead to feelings of disgust or anger and result in exclusion or aggression to defend those norms (Guerra et al., 2025). Addressing the specific realities of trans communities is essential for comprehensive detection and understanding of anti-LGBTI+ hate speech in Portugal.

> *"Faggots and transvestites... Come on, let's prepare the ovens [reference to the crematorium in Nazi concentration camps]"*

Gender-based violence and discrimination also remain persistent challenges in Portugal, particularly affecting women and girls through forms of online hate speech, harassment, and sexualised abuse. Despite progress in legal rights and political representation, women continue to experience high rates of gender-

[9] Between the ages of 16 and 18, it is possible, but only with the authorisation of their legal representatives. However, a legislative process to repeal Law No. 38/2018 is currently ongoing in the Portuguese Parliament. Notwithstanding, the law has not yet been officially repealed, and its legal procedures remain active.

based violence, including domestic violence and femicide (e.g., European Institute for Gender Equality, 2025). A 2026 report (Silva & Brázia, Gender-Based Cyberviolence in Portugal: Perspectives of Children and Youth, Schools, and Law Enforcement) shows that online violence against women is deeply connected to broader structures of sexism and gender inequality and that online hate speech targeting women is sustained by persistent sexist stereotypes, hypersexualisation, and victim-blaming attitudes that normalise aggression against women in digital spaces. International reports (e.g., FRA, 2023) converge with these findings, concluding that online spaces increasingly reflect and amplify **misogynistic** narratives (i.e., hostility or prejudice specifically targeting women), perpetuating stereotypes and hostile attitudes toward women. The example below illustrates incitement to violence and hatred targeting a woman.

> *"Victimisation. Something very common among women. Besides being overweight, she's weak."*
>
> *"I wasn't even totally anti-abortion, but after all this shit, I finally realised that you're just a bunch of whores who want to get fucked without any consequences. Abortion isn't a contraceptive."*

6.4. Intersectionality

Finally, individuals can be targeted based on their perceived belonging to multiple marginalised social groups. **Intersectionality** highlights how individuals are positioned within historically rooted systems of oppression and domination linked to racism, culture, gender, class, and other social categories. These identities intersect to produce unique, combined experiences that cannot be understood by examining each identity in isolation. Social identity markers rarely operate independently, often giving rise to complex, overlapping hate speech targets rather than isolated, separate groups (Carvalho & Guerra, 2023). The example below illustrates an intersectional post targeting women and religion.

> *"You are psychotically retarded, don't f****ng believe you dare to show your face, lol, you do not have a nice future ahead, there will be significant consequences for traitors to the country and Islamic-hugging wh***s. Just wait."*

Recommended readings:

Almeida, P. (2022). *Relatório do projeto de Investigação "Racismo e Xenofobia em Portugal: a normalização dos discursos de ódio no espaço público da internet"*. https://racismoexenofobia.cria.org.pt/wp-content/uploads/2022/07/Relato%CC%81rio_Racismo_Xenofobia.pdf

European Commission Against Racism and Intolerance. (2025). *Sixth report on Portugal* [Report]. Council of Europe.

Guerra, R., Carvalho, P., Marques, C., Carmona, M., Sarroeira, R., Batista, F., Ribeiro, R., Fonseca, A., Moro, S., & Silva, C. (2025). Unpacking online hate speech in Portuguese social media: a social-psychological and linguistic-discursive approach. *Humanities and Social Sciences Communications, 12*(1). https://doi.org/10.1057/s41599-025-05392-9

PART III

Prevention and Legal Framework

7. Preventive and Awareness Practices

Addressing online hate speech requires more than identifying and responding to harmful content after it circulates, it must be a continuous process rather than an isolated intervention. Given the complexity of hate speech manifestations and the ecosystems where it operates, efforts must cross multiple levels and spheres and be strongly articulated with public policies. These should encompass legal and regulatory measures, educational policies, both targeting digital literacy as well as human rights education, and raising societal awareness and empowering target communities.

7.1. Awareness Raising: Public Campaigns

Public campaigns play an important role in signalling that online hate speech has detrimental consequences and in encouraging reporting and reinforcing the message that digital behaviour carries social and legal implications. In Portugal, for example, the recent campaign "Hate online, kills offline", promoted by the Judiciary Police, highlighted the danger and consequences of extremism and hate content online. Other examples of awareness and prevention campaigns in Portugal include #CortarOMalPelaRaiz [Nip It In The Bud] campaign by the kNOwHATE project, #respectbattles Movement: Fighting Hate with Respect by APAV (Portuguese Association for Victim Support), and "Hate Speech is Not an Argument" by the European Anti-Poverty Network Portugal.

Whereas awareness campaigns alone cannot eliminate hate speech, they contribute to shaping social norms and increasing collective sensitivity to harmful discourse.

7.2. Educational and Training Initiatives

Another important set of strategies for addressing and limiting the spread of online hate focuses on **education and training** (Isasi & Juanatey, 2016; Madriaza

et al., 2025; Mossou & Lane, 2018). These approaches offer a way to balance responsibility with freedom of expression and active participation in digital spaces, and are fundamental for equipping people, especially youth, with the skills needed to recognise and challenge online hate speech (Isasi & Juanatey, 2016). Although all sectors of society engage with such content, young people are particularly exposed due to their intensive use of social media, making them a key focus of many awareness initiatives. While not exhaustive, these initiatives can be broadly grouped into two categories: those aimed at the **general public**, which seek to foster forms of "digital citizenship"[10] by developing relevant knowledge and skills; and those designed for individuals already engaged with the issue, including **practitioners and activists**, who require more targeted tools to respond effectively (Isasi & Juanatey, 2016).

There are also examples in the **Portuguese context**, such as the References Manual. This handbook was designed to support the Movement Against Hate Speech and the Council of Europe's Youth Campaign Against Online Hate Speech, and is **useful for educators** working on this issue, both within and outside the formal education system. The handbook is intended for use with young people aged between 13 and 18, although the activities can be adapted for other age groups and learner profiles (Council of Europe, 2016).

Initiatives that promote digital literacy, "digital citizenship", critical thinking and media literacy can help individuals:

- Recognise disinformation and manipulative narratives;
- Identify coded or indirect forms of hate speech;
- Recognise instances of racism, sexism, xenophobia, among others, mobilised in hate speech in online spaces;
- Reflect on the psychological and social impacts of hate speech;
- Reflect before sharing one's own content or that of others.

[10] Adapting the concept of education for citizenship to encompass the knowledge and skills necessary for responsible interaction in digital contexts (Isasi & Juanatey, 2016).

Training programmes for educators, journalists, public officials and community leaders are also particularly important (e.g., UNESCO & United Nations Office on Genocide Prevention and the Responsibility to Protect, 2023), as these actors often play a mediating role in public discourse, either reinforcing or challenging harmful narratives.

Ultimately, preventive education and training initiatives do not aim to suppress debate or freedom of expression, but to foster responsible communication, democratic engagement, and awareness of fundamental rights (Kentmen-Cin, 2025; Zendeli, 2017).

7.3. Counter-Speech and Positive Narratives

Recent research has increasingly focused on how new social norms are emerging in online communication, including the growing use of counter-speech to challenge hate speech (Bilewicz & Soral, 2020). **Counter-speech** refers to responses that challenge and undermine hate speech by introducing alternative messages that promote inclusion, factual clarification or solidarity (e.g., Bilewicz & Soral, 2020; Gagliardone et al., 2015; Williams, 2022). Rather than silencing harmful voices, counter-speech introduces different perspectives into the conversation (e.g., content removal; Baider, 2023).

Research shows that counter-speech can help reduce the spread of online hate, for example, by using inclusive hashtags (Williams, 2022). Online counter-speech may also emerge spontaneously or be more organised, such as coordinated hashtag campaigns (e.g., #ichbinhier, #brasileirasnaosecalam), and can take different forms, including text and images (Benesch et al., 2016). Evidence suggests that certain approaches tend to be more effective, such as highlighting the potential consequences of harmful speech and drawing on positive emotions, such as empathy or humour (Benesch et al., 2016; Chung et al., 2024). These strategies can either encourage a shift in tone from the original speaker or influence the wider audience (i.e., bystanders), making them less likely to engage in or support hate speech.

Whereas research highlights the potential of counter-speech (e.g., Baider, 2022; Gennaro et al., 2025), its effectiveness is still not consistent, and research shows that not all forms are equally productive (e.g., Baider, 2023; Wang et al., 2026; Williams, 2022). For instance, it may unintentionally increase the visibility, engagement, and emotional intensity of hateful content, especially considering that social media algorithms are engineered to maximise engagement. The use of insults or a hostile tone when responding to hate speech also often fails, escalating the interaction and leading to further harmful content (Baider, 2023; Williams, 2022). Accordingly, research suggests that counter-speech is more effective when it follows a few key principles: avoiding insulting or hateful language, presenting clear and consistent arguments, and asking for evidence when claims are false or questionable. It can also be helpful to encourage others to engage in counter-speech and, in cases where an account appears to be fake or automated (bot), to report it to the platform for review (Chung et al., 2024; Williams, 2022).

Previous findings of the kNOwHATE project, based on two annotated corpora composed of comments and associated replies retrieved from YouTube and Twitter/ X in Portugal between 2021-2022, revealed the most prevalent counter-speech strategies rely on counter-stereotypes, empathy, inclusive identities, and appeals to legal and social norms.

↘ Counter-stereotypes

Counter-stereotypes refer to beliefs or representations that challenge widely held cultural beliefs about a group (Čehajić-Clancy & Bilewicz, 2021). By highlighting information that goes against common stereotypes, they help question biases and encourage more accurate and nuanced perceptions of social groups (Čehajić-Clancy & Bilewicz, 2021). In the example below, the speaker explicitly contests the stereotype that the Roma community disproportionately benefits from state subsidies:

> *"Have you ever tried to find out the true facts about the subsidies declared for the Roma community? I don't think so, but you have several websites that*

> *can prove your gigantic stupidity disguised as ethnic prejudice. If you need me to, I can send you some"*

↘ Empathy

Empathy also plays an important role in countering hate speech (Bilewicz & Soral, 2020) and is one of the most tested counter-speech strategies, including the use of counter-speech bots (e.g., Gennaro et al., 2025; Wang et al., 2026). Empathy refers to both emotional responses, such as feeling concern for others, and cognitive elements, such as seeing situations from another person's perspective, and is associated with greater tolerance and a stronger willingness to help (Batson, 2009). The example below illustrates an empathetic counter-speech strategy:

> *"Try dressing as a woman and go look for a job and find housing to see if it's easy! Sorry, but you shouldn't be so backward when it comes to supporting people [...]"*

In this comment, the speaker invites perspective-taking by encouraging the addressee to imagine the lived experiences and structural difficulties faced by women. This appeal to empathy is used to challenge prejudiced or dismissive views by foregrounding inequality and social barriers.

↘ Inclusive identities

This strategy is highly effective for reducing prejudice and intergroup conflict (e.g., Gaertner & Dovidio, 2000; Gaertner et al., 2016). It involves highlighting shared identities, encouraging people to move beyond "us versus them" distinctions and instead see themselves as part of a broader, more inclusive "we". This change can be achieved by highlighting cooperation, interdependence or commonalities, thus redirecting ingroup favouritism towards a more positive perspective (e.g., greater empathy) towards those who were previously regarded as "them" (Gaertner & Dovidio, 2000; Gaertner et al.,

2016). For example, the statement presented below illustrates an appeal to shared identity. This type of discourse reframes differences (e.g., skin colour, nationality) as secondary to a common identity, thereby promoting unity across diverse social groups.

> *"Quite right. The Portuguese must unite. Whether we're white, black, gypsy or whatever, we have one thing in common: we're Portuguese".*

↘ Social norms

Social norms are shared expectations about appropriate attitudes and behaviours within a group or society (McDonald & Crandall, 2015). They are "*a kind of grammar of social interactions. Like a grammar, a system of norms specifies what is acceptable and what is not in a society or group*" (Bicchieri et al., 2023). Hence, social norms are powerful standards that guide individuals' perceptions, judgments, and behaviours.

When social norms frame hateful language as socially unacceptable, they can discourage individuals from engaging in hate speech and limit its spread. In this way, social norms also support collective responses, especially among those who recognise hate speech as a form of injustice and are motivated to challenge it (Bilewicz & Soral, 2020). Referring to legal norms can also help deter online hate speech by signalling that such behaviour may have consequences. At the same time, these regulations reinforce social norms that frame hateful language as unacceptable, discouraging its use and supporting collective efforts to challenge it (Bilewicz & Soral, 2020).

The examples provided illustrate how social norms can be mobilised to promote inclusion and regulate behaviour.

> *"Whoever is born in Portugal is Portuguese. It doesn't matter if you're ethnically European, African, American, Asian, etc..."*

> *"No one should be harassed on the street in this disgusting and threatening way; both of them should be fined or arrested for what they did. This is not the way to "make themselves heard", this is extremism. Do they want to speak out? Let them do so through democratic means."*

The first comment reinforces an inclusive norm of national belonging. By emphasising that Portuguese identity is not dependent on ethnicity, this discourse contributes to establishing a shared expectation that exclusion based on origin is inappropriate or even illegitimate. Similarly, the second comment illustrates the enforcement of both social and legal norms against harassment and aggression. It explicitly frames such behaviour as unacceptable and calls for consequences, thereby reinforcing the idea that hate-related actions violate both societal standards and legal boundaries.

Together, these examples demonstrate how social norms operate not only by promoting inclusive understandings of group membership but also by discouraging harmful behaviours and legitimising collective responses to hate speech.

Finally, counter-speech is a strategy that goes beyond victims of hate speech, as it highlights the important role of bystanders (Wang et al., 2026; Gennaro et al., 2025). For instance, research indicates that fewer than half of bystanders intervene when they witness prejudice or hate, whether to support the target or challenge the perpetrator (Williams, 2022). Specifically in online settings, there are several ways to respond, including reporting harmful content or engaging in counter-speech that challenges the message while avoiding escalation. Ultimately, collective responses are particularly important as counter-speech tends to be more effective when multiple users reinforce norms of acceptable behaviour (Williams, 2022).

7.4. Platform-Level Measures

Digital platforms play a central role in the prevention and management of online hate speech (Weber et al., 2023). Reporting mechanisms, content moderation systems and transparency policies can influence how quickly harmful content is addressed (Isasi & Juanatey, 2016). Comment sections may also serve as spaces for counter-speech, and recommendation algorithms can be designed to mitigate filter bubbles and introduce more diverse viewpoints (Weber et al., 2023).

However, in practice, reporting, moderation, and removal mechanisms do not always function effectively. Recent research highlights persistent limitations in both automated moderation tools and human review systems, particularly in detecting indirect forms of hate speech, including misogynistic and racist content (FRA, 2023). An analysis of 1,500 social media posts that had already passed through platform moderation systems found that more than half were still considered hateful by human coders, illustrating the difficulty of consistently identifying harmful content (FRA, 2023). At the same time, moderation systems may also incorrectly remove legal content, raising additional concerns regarding transparency and accountability. These inconsistencies may reflect the challenges platforms face in moderating large volumes of content, including reliance on automated systems and the pressure on human review teams. The scale, complexity, and speed of online communication make content moderation extremely complex (FRA, 2023). Platforms like Twitter/X process hundreds of millions of posts each day, with activity sometimes rising even higher. Even with automated systems and human moderation, managing this volume of content is a huge challenge. It also involves difficult decisions, as platforms must balance moderation efforts with concerns around freedom of expression and censorship (Isasi & Juanatey, 2016).

Research further suggests that reporting mechanisms themselves may discourage user engagement. Reporting interfaces are often perceived as opaque, impersonal and difficult to navigate, with limited feedback, insufficient victim support and little transparency regarding moderation decisions (Cover et

al., 2025; European Observatory of Online Hate, 2024). Evidence further suggests that reporting rates remain relatively low, partly due to distrust in platform responses and uncertainty about whether action will be taken.

In response to these challenges, recent European initiatives such as the Digital Services Act (DSA) seek to strengthen transparency, accountability and cooperation between platforms, civil society organisations and law enforcement agencies. Alongside platform-based reporting systems, third-party organisations and police reporting may provide additional support for victims, advocacy services, and alternative channels for reporting harmful content (European Observatory of Online Hate, 2024).

Despite these limitations, reporting hateful content on social media remains an important step. While procedures vary across platforms, reporting tools are usually accessible through policies, complaint forms, or "Report a problem" sections. Even if the outcomes are not always as expected, using these formal channels remains a meaningful way to challenge and limit the spread of hate speech (Silva et al., 2024).

In sum, preventing hate speech is a collective responsibility. Institutions, educators, IT companies, platform providers, civil society organisations, law enforcement agencies, and citizens all play a role in shaping a safe and inclusive digital environment. Importantly, effective prevention does not rely solely on restriction or punishment. It combines awareness, education, ethical responsibility and institutional safeguards to reduce harm and strengthen democratic resilience.

Recommended readings:

European Union Agency for Fundamental Rights - FRA (2023). *Online content moderation - Current challenges in detecting hate speech (Catalogue No. TK-04-23-883-EN-C).* Publications Office of the European Union. https://fra.europa.eu/en/publication/2023/online-content-moderation

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

UNESCO & United Nations Office on Genocide Prevention and the Responsibility to Protect. (2023). *Addressing hate speech through education: A guide for policy-makers*. https://unesdoc.unesco.org/ark:/48223/pf0000384872.

Williams, M. (2022). *The Science of Hate: How Prejudice Becomes Hate and What We Can Do to Stop It*. Faber & Faber, Ltd

8. Legal and Institutional Framework in Portugal

Beyond understanding what hate speech is and how it manifests and operates, it is crucial to distinguish between what constitutes a potentially criminal offence under Portuguese law and what falls outside that scope.

The legal analysis of hate speech in Portugal must necessarily begin with an examination of what falls under freedom of expression and information, a right enshrined in Article 37 of the Constitution of the Portuguese Republic and in other international legal instruments, such as the Universal Declaration of Human Rights (Article 19), the European Convention on Human Rights (Article 10), the International Covenant on Civil and Political Rights (Article 19), and the American Convention on Human Rights (Article 13 (5)). According to Article 37 of the Constitution of the Portuguese Republic, "*everyone has the right to freely express and disseminate their thoughts through speech, images, or any other means, as well as the right to inform, to be informed, and to receive information, without impediments or discrimination*".

However, freedom of expression is not an absolute right, a view upheld by Portuguese case law, as in the ruling of the Porto Court of Appeal[11], which emphasises that "*the right to freedom of expression, like many others, is not an absolute right that can be exercised without restriction or limit, since there are limits to the exercise of the right to freely express and disseminate thoughts and opinions.*" Article 18 of the Constitution of the Portuguese Republic also reinforces this position, further stating that restrictions must be "*limited to what is necessary to safeguard other constitutionally protected rights or interests*," such as the right to personal identity, good name and reputation, image, and legal protection against all forms of discrimination (Article 26 (1)). These rights are often called into question when discussing incidents of hate speech, as

11 Acórdão do Tribunal da Relação do Porto de 7 de junho de 2023 (Proc. n.º 5551/19.0T9LSB-A.P1)

there is a fine line between what is considered freedom of expression and what is classified as hate speech.

Understanding these constitutional limits helps us understand how the Portuguese legal system addresses this phenomenon.

Legally, the **Portuguese Penal Code establishes hate speech as an autonomous legal offence**, namely the crime of discrimination and incitement to hatred and violence (Article 240 of the Penal Code).

Article 240 – Discrimination and Incitement to Hate and Violence

1 – Any person who:

a) Founds or establishes an organisation or engages in propaganda activities that incite or encourage discrimination, hatred, or violence against a person or group of people based on their ethnic or racial origin, national or religious origin, colour, nationality, ancestry, territory of origin, religion, language, sex, sexual orientation, gender identity or expression, or sexual characteristics, or physical or mental disability; or

b) Participates in the organisations referred to in the preceding subparagraph, in the activities they undertake, or assists them, including their financing;

is punished with imprisonment for a term of 1 to 8 years.

2 - **Any person who, publicly, through any means intended for dissemination**, including through the advocacy, denial, or gross trivialisation of crimes of genocide, war crimes, or crimes against peace and humanity:

a) **Incites acts of violence against a person or group of people because of their** ethnic or racial origin, national or religious origin, colour, nationality, ancestry, territory of origin, religion, language, sex, sexual orientation, gender identity or expression, or physical or mental disability;

b) **Defames or insults a person or group of people because of their** ethnic or racial origin, national or religious origin, colour, nationality, ancestry, territory of origin, religion, language, sex, sexual orientation, gender identity or expression, or sexual characteristics, or physical or mental disability;

c) **Threaten a person or group of people because of their** ethnic or racial origin, national or religious origin, colour, nationality, ancestry, territory of origin, religion, language, sex, sexual orientation, gender identity or expression, or sexual characteristics, or physical or mental disability; or

d) **Incite discrimination, hatred, or violence against a person or group of people because of their** ethnic or racial origin, national or religious origin, colour, nationality, ancestry, territory of origin, religion, language, sex, sexual orientation, gender identity or expression, or sexual characteristics, or physical or mental disability;

is punished with imprisonment for a term of 6 months to 5 years.

3 - When the crimes provided for in the preceding paragraphs are committed through a computer system, the court may order the deletion of computer data or content.

Upon analysis, it becomes clear that certain requirements must be met for a behaviour to be considered hate speech; therefore, only specific incidents are criminalised, given the potential harm they cause. Thus, **the following criteria[12] must be considered**:

1. Criminalisation **does not apply to conduct that takes place in private**. The definition of this criminal offence **requires that the punishable conduct take place in a public space and involve any means intended for dissemination**, which includes the use of verbal speech, leaflets, graffiti, the posting of posters, the use of the press and websites, as well as the posting of messages on the internet outside the scope of closed groups.
2. It is a prerequisite for the commission of the crime that the perpetrator's public use of such media be **intended to "glorify, deny, or grossly trivialise crimes of genocide, war crimes, or crimes against peace and humanity".**
3. It is required that the use of media intended to glorify or deny crimes against peace and humanity **have a concrete discriminatory effect or result**, manifested in the incitement of acts of violence, the commission of crimes of insult or defamation, or threats and incitement to violence or hatred against "a person or group of people because of their race, colour, ethnic or national origin, ancestry, religion, sex, sexual orientation, gender identity, or physical or mental disability."

12 Diário da República (n.d.). *Crime de incitamento ao ódio e à violência* [Crime of incitement to hatred and violence]. Lexionário. https://diariodarepublica.pt/dr/lexionario/termo/crime-incitamento-ao-odio-a-violencia

Furthermore, Article 240 provides a **closed, albeit extensive, list of bias-based motivations that can constitute the crime** of discrimination and incitement to hatred and violence, with a violation of at least one of these being required for such a crime to have been committed. This conclusion is important because it shows that **other bias-based motivations not featured in the article cannot be criminally punished**, such as, hatred motivated by political reasons or associated with subcultures.

This is because, according to the principle of criminal legality outlined in Article 1 of the Penal Code, individuals may only be criminally punished for a conduct "*described and declared punishable by a law in effect prior to the time of its commission*," and "*it is not allowed to resort to analogy to classify an act as a crime*". These behaviours may, however, be criminally punishable under other offences, such as Defamation (Article 180) and insult (Article 181), provisions of the same code, thus disregarding their potential biased motivation.

Below are presented two real-life cases for illustrative purposes. The aim is to illustrate what may or may not be considered hate speech under Portuguese law.

Example 1[13]

- **Context**: Two defendants, AA and BB, were convicted in the court of first instance for posting comments on Twitter/X, such as: "*Forced prostitution of the Bloco [left-wing political party] women*", "*Everything, like a mass sexual assault*", "*I agree. Include the women from the PCE, MRRP, MAS, and PS [left-wing political parties]*".

- **Trial Court Ruling**: The court found that the remarks were not humorous but rather directed at left-wing women with the intent to offend, humiliate, and associate them with prostitution. As such, it sentenced AA to 2 years and 10 months of actual imprisonment and BB to 1 year and 8

13 Acórdão do Tribunal da Relação de Lisboa, de 05/12/2024, processo n.º 1633/22.0T9LSB.L1-5. https://www.dgsi.pt/jtrl.nsf/33182fc732316039802565fa00497eec/f9ef3628cf2e149080258bf80055a9c9?OpenDocument&Highlight=0,discrimina%C3%A7%C3%A3o,e,incitamento,ao,%C3%B3dio

months of imprisonment, suspended for 2 years with probation and payment of €750 to APAV [Portuguese Association for Victim Support].

- **Appeal**: Not satisfied with the outcome of the trial court's ruling, AA filed an appeal, arguing that the statements were not serious but rather humorous; that they did not intend to offend all women, only political activists; that there was an error in the assessment of the evidence; and that they should be eligible for a suspended sentence. BB, on the other hand, argued that it had not been proven that the account holder was BB; that no computer forensic analysis had been conducted; that there had been a violation of the principle of in *dubio pro reo*; and mentioned the need to adjust the sentence.

- **Decision of the Court of Appeal:** The court dismissed both appeals and upheld the lower court's judgments.

- **Reasons for the Court of Appeal's Decision**: The court confirmed the authorship of the posts, concluding that the AA and BB accounts did in fact belong to the defendants, based on evidence such as messages, emails, and photographs, and upheld the use of circumstantial evidence through natural presumptions. It rejected the argument that the expressions were humorous in nature, stating that phrases such as "*forced prostitution of the Bloco girls*" are objectively degrading and directed at a specific group of left-wing women, thus constituting the crime provided for in Article 240(2)(b) of the Penal Code. It further clarified that it is not necessary to offend all women; it is sufficient to target a group defined by sex and ideology and concluded that the defendants acted with direct intent. Finally, the court upheld the prison sentence imposed on AA, emphasising their extensive criminal record, the short time span between the end of their probation and the events in question, and the lack of remorse or empathy demonstrated during the trial

Example 2[14]

- **Context:** The defendant published an opinion piece in the newspaper Público in which she used generalising and derogatory language about African and Roma people, associating them with deviant behaviour, incivility, violence, and cultural inferiority. The SOS Racismo Movement filed a complaint, arguing that the text constituted a crime of discrimination and incitement to hatred and violence (Article 240 of the Penal Code).

- **Decision by the Investigating Judge:** The Public Prosecutor's Office dismissed the investigation, ruling that the text was protected by freedom of speech. The assistant (SOS Racismo Movement), in response, requested that a formal investigation be opened.

- **Pre-Trial Decision:** The investigating judge dismissed the motion to open a preliminary investigation, finding that the motion did not describe sufficient facts to constitute a crime, that the defendant's text constituted mere opinion, protected by freedom of expression, and that the elements of Article 240 of the Penal Code had not been met.

- **Appeal:** Not satisfied with the decision, the assistant appealed to the Court of Appeal, arguing that the complaint described sufficient and specific instances and that the defendant's statements were objectively discriminatory. He further added that the investigating judge had confused freedom of expression with hate speech and that the order violated Article 287 of the Code of Criminal Procedure and Article 240 of the Penal Code. Considering this, he requested that the Court of Appeal revoke the order and open a preliminary investigation.

- **Decision of the Court of Appeal:** The Court granted the appeal, overturned the order, and ordered the opening of the preliminary investigation, finding that the assistant's petition was formally valid and

14 Acórdão do Tribunal da Relação de Lisboa, de 06/07/202, processo n.º 5551/19.0T9LSB.L1-5. https://www.dgsi.pt/jtrl.nsf/33182fc732316039802565fa00497eec/53d43a27fb12dafc802587480047ae12?OpenDocument&Highlight=0,discrimina%C3%A7%C3%A3o,e,incitamento,ao,%C3%B3dio

contained sufficient facts. The court emphasised that freedom of expression does not justify discriminatory speech, noting that the investigating judge had interpreted this fundamental right too broadly.

In sum, the Portuguese legal framework seeks to balance the protection of freedom of expression with the need to prevent discrimination, hatred and violence. As the examples presented illustrate, this distinction is not always straightforward, particularly in digital environments where harmful discourse may be framed as humour, opinion or political commentary. Not all offensive or discriminatory speech constitutes a criminal offence under Portuguese law, as specific legal criteria must be met for conduct to fall within Article 240 of the Penal Code. At the same time, the legal system recognises that freedom of expression cannot be used to legitimise speech that undermines the dignity, safety and rights of individuals or groups. Understanding these legal boundaries is therefore essential for identifying hate speech, recognising its potential consequences, and promoting more informed and responsible participation in public and online spaces.

Conclusion

Online hate speech is not a marginal phenomenon. It reflects and shapes broader social dynamics, including inequality, polarisation, disinformation and shifting social norms. It impacts victims, bystanders and societal functioning by normalising exclusion, desensitising audiences and eroding shared democratic norms. By exploring how hate speech operates, through dehumanisation, threat narratives, coded language or escalation, and by situating it within specific social, political and digital contexts, this handbook highlights the importance of moving beyond isolated expressions toward a contextual and cumulative understanding of harm.

Hate speech does not emerge in isolation; it evolves within ecosystems shaped by polarisation, disinformation and social and political tensions. Recognising these dynamics is essential for proportionate and responsible responses.

At the same time, addressing hate speech requires balance. Democratic societies depend on freedom of expression, open debate and pluralism. Protecting these principles must go hand in hand with safeguarding human dignity, equality and non-discrimination. Legal tools, institutional mechanisms and public awareness initiatives all play a role, but prevention also depends on education, critical thinking and collective responsibility.

Ultimately, preventing and combating hate speech is not only about restricting harmful content. It is about strengthening democratic coexistence and respect for human rights. By fostering informed recognition, ethical engagement and inclusive public discourse, societies can reduce harm while reinforcing the values that sustain them.

Co-funded by the European Union

References

Allport, G. W. (1954). *The Nature of Prejudice*. Addison-Wesley.

Almeida, P., & Pereira, J. (2026). Anticiganismo no Facebook: discursos de ódio e racismo quotidiano em Portugal. *Etnográfica. Revista do Centro em Rede de Investigação em Antropologia*.

Almeida, P., Pereira, J., & Candido, D. (2023). Online hate speech on social media in Portugal: extremism or structural racism? *Social Identities, 29*(5), 419–435. https://doi.org/10.1080/13504630.2024.2324277

Arce-García, S., Said-Hung, E., & Montero-Díaz, J. (2025). Unmasking coordinated hate: Analysing hate speech on Spanish digital news media. *New Media & Society, 27*(10), 5848–5868. https://doi.org/10.1177/14614448241259715

Åkerlund, M. (2021). Dog whistling far-right code words: the case of 'culture enricher' on the Swedish web. *Information, Communication & Society, 25*(12), 1808–1825. https://doi.org/10.1080/1369118x.2021.1889639

Assimakopoulos, S., Baider, F.H., & Millar, S. (2017) Online hate speech in the European Union: a discourse-analytic perspective. Springer Nature

Bahador, B. (2023). Monitoring hate speech and the limits of current definition. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 291-298). Berlin https://doi.org/10.48541/dcr.v12.17

Baider, F. (2022). Covert hate speech, conspiracy theory and anti-semitism: Linguistic analysis versus legal judgement. International Journal for the Semiotics of Law-Revue, 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 and Governance, 11*(2), 249-260. https://doi.org/10.17645/pag.v11i2.6465

Baider, F., & Constantinou, M. (2020). Covert hate speech. *Journal of Language Aggression and Conflict, 8*(2), 262–287. https://doi.org/10.1075/jlac.00040.bai

Bastos, C. (2019). Luso-tropicalism debunked, again. Race, racism, and racialism in three Portuguese-speaking societies. In W. Anderson, R. Roque, & R. Ventura Santos (Eds.), Luso-tropicalism and its discontents: the making and unmaking of racial exceptionalism (pp. 243–264). Berghahn. https://doi.org/978-1-78920-113-0

Batson, C. D. (2009). These things called empathy: Eight related but distinct phenomena. In J. Decety & W. Ickes (Eds.), *The social neuroscience of empathy* (pp. 3–15). Boston Review. https://doi.org/10.7551/mitpress/9780262012973.003.0002

Benesch, S. (2023). Dangerous speech. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), Challenges and perspectives of hate speech research (pp. 185-197). Berlin https://doi.org/10.48541/dcr.v12.11

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

Bicchieri, C., Muldoon, R. & Sontuoso, A. (2023) "Social Norms", *The Stanford Encyclopedia of Philosophy* (Winter 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), https://plato.stanford.edu/archives/win2023/entries/social-norms/

Bilewicz, M. (2025). How Appraisal Model Allows to Distinguish Intergroup Conspiracy Theories from Other Forms of Hate Speech. *Psychological Inquiry, 35*(3-4), 216–222. https://doi.org/10.1080/1047840x.2024.2442919

Bilewicz, M., & Soral, W. (2020). Hate Speech Epidemic. The Dynamic Effects of Derogatory Language on Intergroup Relations and Political Radicalization. *Political Psychology, 41*(S1), 3–33. https://doi.org/10.1111/pops.12670

Bliuc, A.-M., Betts, J. M., Vergani, M., Bouguettaya, A., & Cristea, M. (2024). A theoretical framework for polarization as the gradual fragmentation of a divided society. *Communications Psychology, 2*(1). https://doi.org/10.1038/s44271-024-00125-1

Borinca, I., Van Assche, J., Gronfeldt, B., Sainz, M., Anderson, J., & Taşbaş, E. H. O. (2023). Dehumanization of outgroup members and cross-group

interactions. *Current Opinion in Behavioral Sciences, 50*, 101247. https://doi.org/10.1016/j.cobeha.2023.101247

Bozhidarova, M., Chang, J., Ale-Rasool, A., Liu, Y., Ma, C., Bertozzi, A. L., Brantingham, P. J., Lin, J., & Krishnagopal, S. (2023). Hate speech and hate crimes: a data-driven study of evolving discourse around marginalized groups. 2023 IEEE International Conference on Big Data (BigData), 3107–3116. https://doi.org/10.1109/bigdata59044.2023.10386312

Bradshaw, S. (2024). *Disinformation and identity-based violence: Analysis & new insights*. Stanley Center for Peace and Security.

Burnham, S. L. F., Arbeit, M. R., & Hilliard, L. J. (2022). The Subtle Spread of Hateful Memes: Examining Engagement Intentions Among Parents of Adolescents. *Social Media + Society, 8*(2). https://doi.org/10.1177/20563051221095100

Calderón, F. H., Balani, N., Taylor, J., Peignon, M., Huang, Y.-H., & Chen, Y.-S. (2021). Linguistic Patterns for Code Word Resilient Hate Speech Identification. *Sensors, 21*(23), 7859. https://doi.org/10.3390/s21237859

Calderón, C. A., Holgado, P. S., Gómez, J., Barbosa, M., Qi, H., Matilla, A., Amado, P., Guzmán, A., López-Matías, D., & Fernández-Villazala, T. (2024). From online hate speech to offline hate crime: the role of inflammatory language in forecasting violence against migrant and LGBT communities. *Humanities and Social Sciences Communications, 11*(1). https://doi.org/10.1057/s41599-024-03899-1

Cardoso, C., Moreno, J., Narciso, I., Couraceiro, P., & dos Santos, J.G.B. (2025). Portuguese General Elections 2025 – Information and Disinformation on Social Media. OberCom – Observatório da Comunicação. https://doi.org/10.5281/zenodo.15781756

Cardoso, M., Ribeiro, E., & Batista, F. (2025). Portuguese Far-Right Discourse on Social Media: Insights from Topic Modeling. In 14th Symposium on Languages, Applications and Technologies (SLATE 2025). Open Access Series in Informatics (OASIcs), 135, 12:1-12:16, Schloss Dagstuhl – Leibniz-Zentrum für Informatik. https://doi.org/10.4230/OASIcs.SLATE.2025.12

Carvalho, P., Caled, D., Silva, M. J. (2025). The Thin Line Between Conspiracy Theories and Opinion: Why Humans and AI Struggle to Differentiate Them. *International Journal of Communication, 19*, 565–591. https://ijoc.org/index.php/ijoc/article/view/22182

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. *Journal of Language Aggression and Conflict, 12*(2), 171-206. . https://doi.org/10.1075/jlac.00085.car

Carvalho, P., Cunha, B., Santos, R., Batista, F., & Ribeiro, R. (2022, June). Hate speech dynamics against African descent, Roma and LGBTQI communities in Portugal. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 2362-2370).

Carvalho, P. & Guerra, R. (2023). "D3.2/D3.3 annotation guidelines OHS & OCS," Knowhate Project—CERV-2021-EQUAL (101049306), Private/Sensitive—Limited Under the Conditions of the Grant Agreement, ISCTE-Inst. Univ. de Lisboa, Lisbon, Portugal, Tech. Rep., May 2023.

Castelo, C. (2021) https://cesa.rc.iseg.ulisboa.pt/afroport/portuguese-non-racism-on-the-historicity-of-an-invented-tradition/.

Čehajić-Clancy, S., & Bilewicz, M. (2021). Moral-exemplar intervention: A new paradigm for conflict resolution and intergroup reconciliation. *Current Directions in Psychological Science*, *30*(4), 335-342. https://doi.org/10.1177/09637214211013001

Chung, Y.-L., Abercrombie, G., Enock, F., Bright, J., & Rieser, V. (2024). Understanding Counterspeech for Online Harm Mitigation. *Northern European Journal of Language Technology, 10*(1). https://doi.org/10.3384/nejlt.2000-1533.2024.5203

Conselho da Europa (2016). *REFERÊNCIAS: Manual para o combate do discurso de ódio online através da educação para os direitos humanos* (Ed. revista). Direção-Geral da Educação. https://www.dge.mec.pt/sites/default/files/ECidadania/educacao_Direitos_Humanos/documentos/referencias_manual_para_o_combate_do_discurso_de_odio_online.pdf.

Cottrell, C. A., & Neuberg, S. L. (2005). Different emotional reactions to different groups: A sociofunctional threat-based approach to "prejudice". *Journal of Personality and Social Psychology*, *88*(5), 770–789. https://doi.org/10.1037/0022-3514.88.5.770

Couperus, S., Rensmann, L., & Tortola, P. D. (2023). Historical legacies and the political mobilization of national nostalgia: Understanding populism's relationship to the past. *Journal of Contemporary European Studies, 31*(2), 253–267. https://doi.org/10.1080/14782804.2023.2207480

Cover, R., Beckett, J., Brevini, B., Lumby, C., Simcock, R., & Thompson, J. D. (2025). Reporting online abuse to platforms: Factors, interfaces and the potential for care. *Convergence: The International Journal of Research into New Media Technologies, 32*(1), 142–158. https://doi.org/10.1177/13548565251324508

Dare to be Grey. (2024). *Reporting models of online hate*. European Observatory of Online Hate. https://eooh.eu/articles/reporting-models-online-hate-speech

de Coster, S., Veilleux-Lepage, Y., Amarasingam, A., & Abbas, T. (2024). Uncovering the Bias and Prejudice in Reporting on Islamist and Non-Islamist Terrorist Attacks in British and US Newspapers. *Perspectives on Terrorism, 18*(3). https://doi.org/10.19165/2024.3034

de Roos, M. S., & Caon, G. (2026). The radicalisation loop: A layered linguistic model of disengagement and re-engagement in an incel forum. *Current Psychology, 45, 358*. https://doi.org/10.1007/s12144-025-08944-z

Dovidio, J. F., Hewstone, M., Glick, P., & Esses, V. M. (2010). Prejudice, stereotyping and discrimination: Theoretical and empirical overview. In *The SAGE handbook of prejudice, stereotyping, and discrimination* (pp. 3–29). London, England: Sage.

Dreißigacker, A., Müller, P., Isenhardt, A., & Schemmel, J. (2024). Online hate speech victimization: consequences for victims' feelings of insecurity. *Crime Science, 13*(4). https://doi.org/10.1186/s40163-024-00204-y

European Union Agency for Fundamental Rights - FRA (2023). *Online content moderation - Current challenges in detecting hate speech* (Catalogue No.

TK-04-23-883-EN-C). Publications Office of the European Union. https://fra.europa.eu/en/publication/2023/online-content-moderation

EUAFR - European Union Agency for Fundamental Rights (2025). Fundamental rights report – 2025. https://fra.europa.eu/en/publication/2025/fundamental-rights-report-2025

European Commission Against Racism and Intolerance. (2025). *Sixth report on Portugal* [Report]. Council of Europe.

European Institute for Gender Equality. (2025). Country profile for Portugal. https://eige.europa.eu/gender-based-violence/countries/portugal?language_content_entity=en

Fischer, A., Halperin, E., Canetti, D., & Jasini, A. (2018). Why we hate. *Emotion Review*, *10*(4), 309-320. https://doi.org/10.1177/1754073917751229

Fonseca, A., Pontes, C., Moro, S., Batista, F., Ribeiro, R., Guerra, R., Carvalho, P., Marques, C., & Silva, C. (2024). Analyzing hate speech dynamics on Twitter/X: Insights from conversational data and the impact of user interaction patterns. *Heliyon, 10*(11), e32246. https://doi.org/10.1016/j.heliyon.2024.e32246

Fortuna, P., & Nunes, S. (2018). A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys, 51*(4), 1–30. https://doi.org/10.1145/3232676

Gaertner, S. L., & Dovidio, J. F. (2000). *Reducing intergroup bias: The common ingroup identity model*. Psychology Press.

Gaertner, S. L., Dovidio, J. F., Guerra, R., Hehman, E., & Saguy, T. (2016). A common ingroup identity: Categorization, identity, and intergroup relations. In T. D. Nelson (Ed.), *Handbook of prejudice, stereotyping, and discrimination* (2nd ed., pp. 433–454). Psychology Press.

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

Gallacher, J. D., Heerdink, M. W., & Hewstone, M. (2021). Online Engagement Between Opposing Political Protest Groups via Social Media is Linked to

Physical Violence of Offline Encounters. *Social Media + Society, 7*(1), 1-16. https://doi.org/10.1177/2056305120984445

Geetanjali, & Kumar, M. (2025). Exploring hate speech detection: challenges, resources, current research and future directions. *Multimedia Tools and Applications, 84*, 38423–38459. https://doi.org/10.1007/s11042-025-20716-2

Gennaro, G., Derksen, L., Abdelrahman, A. *et al.* (2025). Counterspeech encouraging users to adopt the perspective of minority groups reduces hate speech and its amplification on social media. *Scientific Reports,* 15, 22018. https://doi.org/10.1038/s41598-025-05041-w

Ghenai, A., Noorian, Z., Moradisani, H., Abadeh, P., Erentzen, C., & Zarrinkalam, F. (2025). Exploring hate speech dynamics: The emotional, linguistic, and thematic impact on social media users. *Information Processing & Management, 62*(3), 104079. https://doi.org/10.1016/j.ipm.2025.104079

Greene, J. (2025). *Malinformation*. EBSCO Research Starters. https://www.ebsco.com/research-starters/information-technology/malinformation#full-article

Guerra, R., Carvalho, P., Marques, C., Carmona, M., Sarroeira, R., Batista, F., Ribeiro, R., Fonseca, A., Moro, S., & Silva, C. (2025). Unpacking online hate speech in Portuguese social media: a social-psychological and linguistic-discursive approach. *Humanities and Social Sciences Communications, 12*, 1709. https://doi.org/10.1057/s41599-025-05392-9

Harel, T. O., Jameson, J. K., & Maoz, I. (2020). The Normalization of Hatred: Identity, Affective Polarization, and Dehumanization on Facebook in the Context of Intractable Political Conflict. *Social Media + Society, 6*(2), 1-10. https://doi.org/10.1177/2056305120913983

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and social psychology review*, *10*(3), 252-264. https://doi.org/10.1207/s15327957pspr1003

Haslam, N., & Loughnan, S. (2014). Dehumanization and infrahumanization. *Annual Review of Psychology*, *65* (1), 399-423. https://doi.org/10.1146/annurev-psych-010213-115045

Hassan, G., Rabah, J., Madriaza, P., Brouillette-Alarie, S., Borokhovski, E., Pickup, D., Varela, W., Girard, M., Durocher-Corfa, L., & Danis, E. (2022). PROTOCOL: Hate online and in traditional media: A systematic review of the evidence for associations or impacts on individuals, audiences, and communities. *Campbell Systematic Reviews, 18*(2), e1245. Portico. https://doi.org/10.1002/cl2.1245

Herek, G. M. (2004). Beyond "homophobia": Thinking about sexual prejudice and stigma in the twenty-first century. *Sexuality Research and Social Policy*, *1*(2), 6–24. https://ssrn.com/abstract=1142860

Homolar, A., & Löfflmann, G. (2021). Populism and the Affective Politics of Humiliation Narratives. *Global Studies Quarterly, 1*(1), 1–11. https://doi.org/10.1093/isagsq/ksab002

Inwood, O., & Zappavigna, M. (2023). Conspiracy Theories and White Supremacy on YouTube: Exploring Affiliation and Legitimation Strategies in YouTube Comments. *Social Media + Society, 9*(1), 1-16. https://doi.org/10.1177/20563051221150410

Isasi, A. C., & Juanatey, A. G. (2016). *Hate speech in social media: A state-of-the-art review*. Ajuntament de Barcelona. https://ajuntament.barcelona.cat/bcnvsodi/wp-content/uploads/2017/02/Informe_discurso-del-odio_ES-en-GB.pdf

Karantzas, G. C., & Simpson, J. A. (2026). The Perpetration of Dehumanization: A Systematic Review. *Current Opinion in Psychology, 70,* 102309. https://doi.org/10.1016/j.copsyc.2026.102309

Katz-Wise, S. L., & Hyde, J. S. (2012). Victimization experiences of lesbian, gay, and bisexual individuals: A meta-analysis. *Journal of Sex Research, 49*(2-3), 142–167. https://doi.org/10.1080/00224499.2011.637247

Kearns, C., Sinclair, G., Black, J., Doidge, M., Fletcher, T., Kilvington, D., Liston, K., Lynn, T., & Rosati, P. (2022). A Scoping Review of Research on Online Hate and Sport. *Communication & Sport, 11*(2), 402–430. https://doi.org/10.1177/21674795221132728

Kentmen-Cin, C. (2025). Hate Speech on Social Media: A Systemic Narrative Review of Political Science Contributions. *Social Sciences, 14*(10), 610. https://doi.org/10.3390/socsci14100610

Kteily, N., & Bruneau, E. (2017). Backlash: The politics and real-world consequences of minority group dehumanization. *Personality and Social Psychology Bulletin, 43*(1), 87–104. https://doi.org/10.1177/0146167216675334

Mackie, D. M., & Smith, E. R. (2015). Intergroup emotions. In M. Mikulincer, P. R. Shaver, J. F. Dovidio, & J. A. Simpson (Eds.), *APA handbook of personality and social psychology, Vol. 2. Group processes* (pp. 263–293). American Psychological Association. https://doi.org/10.1037/14342-010

Madriaza, P., Hassan, G., Brouillette-Alarie, S., Mounchingam, A. N., Durocher-Corfa, L., Borokhovski, E., Pickup, D., & Paillé, S. (2025). Exposure to hate in online and traditional media: A systematic review and meta-analysis of the impact of this exposure on individuals and communities. *Campbell Systematic Reviews, 21*, e70018. https://doi.org/10.1002/cl2.70018

Magu, R., & Luo, J. (2018). Determining Code Words in Euphemistic Hate Speech Using Word Embedding Networks. *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. https://doi.org/10.18653/v1/w18-5112

Mannocci, L., Mazza, M., Monreale, A., Tesconi, M., & Cresci, S. (2024). Detection and Characterization of Coordinated Online Behavior: A Survey. arXiv preprint arXiv:2408.01257. https://doi.org/10.48550/arXiv.2408.01257

Mansur, Z., Omar, N., Tiun, S., & Alshari, E. M. (2024). A normalization model for repeated letters in social media hate speech text based on rules and spelling correction. *Plos One, 19*(3), e0299652. https://doi.org/10.1371/journal.pone.0299652

McDonald, R. I., & Crandall, C. S. (2015). Social norms and social influence. *Current Opinion in Behavioral Sciences*, *3*, 147-151. http://dx.doi.org/10.1016/j.cobeha.2015.04.006

Miranda, S., Gouveia, C., Di Fátima, B., & Antunes, A. C. (2023). Hate speech on social media: behaviour of Portuguese football fans on Facebook. *Soccer & Society, 25*(1), 76–91. https://doi.org/10.1080/14660970.2023.2230452

Montesinos-Cánovas, E., García-Sánchez, F., García-Díaz, J. A., Alcaraz-Mármol, G., & Valencia-García, R. (2023). Spanish hate-speech detection in football. *Procesamiento del Lenguaje Natural*, 71, 15-27.

Morales, E., Hodson, J., O'Meara, V., Gruzd, A., & Mai, P. (2025). Online toxic speech as positioning acts: Hate as discursive mechanisms for othering and belonging. *New Media & Society,* 1-19. https://doi.org/10.1177/14614448251338493

Mossou, S., & Lane, A. (2018). *Anti-migrant hate speech*. Quaker Council for European Affairs. https://www.qcea.org/wp-content/uploads/2018/06/Hate-Speech-Report_final.pdf

Müller, K., & Schwarz, C. (2021). Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*, *19*(4), 2131-2167. https://doi.org/10.1093/jeea/jvaa045

Murib, Z. (2022). Don't Read the Comments: Examining Social Media Discourse on Trans Athletes. *Laws, 11*(4), 53. https://doi.org/10.3390/laws11040053

National Coordinator for Counterterrorism and Security (NCTV). (2024). *Memes as an online weapon. An analysis into the use of memes by the far right*. https://english.nctv.nl/publications/reports/2024

Neves, S., Borges, J., Ferreira, M., Correia, M., Sousa, E., Rocha, H., Silva, L., Allen, P., & Vieira, C. P. (2023). A literature review on violence and discrimination against trans people in Portugal: Are we still living in a dictatorship? *Sexualities*, 28(1–2), 349–364. https://doi.org/10.1177/13634607231197059

Papcunová, J., Martončik, M., Fedáková, D., Kentoš, M., Bozogáňová, M., Srba, I., Moro, R., Pikuliak, M., Šimko, M., & Adamkovič, M. (2021). Hate speech operationalization: a preliminary examination of hate speech indicators and their structure. *Complex & Intelligent Systems, 9*(3), 2827–2842. https://doi.org/10.1007/s40747-021-00561-0

Paz, M. A., Montero-Díaz, J., & Moreno-Delgado, A. (2020). Hate speech: A systematized review. *SAGE Open, 10*(4), 1-15. https://doi.org/10.1177/2158244020973022

Pickles, J. (2020). Sociality of hate: The transmission of victimization of LGBT+ people through social media. *International Review of Victimology, 27*(3), 311–327. https://doi.org/10.1177/0269758020971060

Pontes, C., Fonseca, A., Moro, S., Batista, F., Ribeiro, R., Marques, C., Carvalho, P., Silva, C., & Guerra, R. (2024). Unveiling Patterns of Hate Speech in the

Portuguese Sphere: A Social Network Analysis Approach. *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, 70–81. https://doi.org/10.1007/978-3-031-73997-2_7

Prabhu, R., & Seethalakshmi, V. (2025). A comprehensive framework for multi-modal hate speech detection in social media using deep learning. *Scientific Reports, 15*(1). https://doi.org/10.1038/s41598-025-94069-z

Said-Hung, E., Moreno-López, R., & Mottareale-Calvanese, D. (2023). Promotion of hate speech by Spanish political actors on Twitter. *Policy & Internet, 15*(4), 665–686. Portico. https://doi.org/10.1002/poi3.353

Sanchez-Mazas, M., & Licata, L. (2015). Xenophobia: Social psychological aspects. *International Encyclopedia of the Social & Behavioral Sciences*, *25*, 802-807. https://doi.org/10.1016/B978-0-08-097086-8.24031-2

Silva, C., & Carvalho, P. (2023). When can compliments and humour be considered hate speech? A perspective from target groups in Portugal. *Comunicação e sociedad*e, 43, e023006. https://doi.org/10.17231/comsoc.43(2023).4135

Smith, L. G. E., Thomas, E. F., Bliuc, A.-M., & McGarty, C. (2024). Polarization is the psychological foundation of collective engagement. *Communications Psychology, 2*(1). https://doi.org/10.1038/s44271-024-00089-2

Soral, W., Bilewicz, M., & Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior, 44*(2), 136–146. Portico. https://doi.org/10.1002/ab.21737

Sousa, Y., & Cabecinhas, R. (2025). Estereótipos Sociais. *Psicologia Social, Comunicação e Cultura*, *69*. https://doi.org/10.21814/uminho.ed.157.6

Šrol, J., Čavojová, V., & Ballová Mikušková, E. (2022). Finding Someone to Blame: The Link Between COVID-19 Conspiracy Beliefs, Prejudice, Support for Violence, and Other Negative Social Outcomes. *Frontiers in Psychology, 12*. https://doi.org/10.3389/fpsyg.2021.726076

Stephan, W. G., & Stephan, C. W. (2000). An integrated threat theory of prejudice. In S. Oskamp (Ed.), Reducing prejudice and discrimination, (pp. 23-46). Lawrence Erlbaum Associates Publishers.

Stephan, W. G., & Stephan, C. W. (2016). Intergroup Threats. The Cambridge Handbook of the Psychology of Prejudice, 131–148. https://doi.org/10.1017/9781316161579.00

Taha, R., & Sailofsky, D. (2025). 'But she's not even trans!': A rhetorical analysis of 'liberal feminist' defences of Imane Khelif amid Olympic transvestigations. *International Review for the Sociology of Sport*. https://doi.org/10.1177/10126902251371324

Theofilopoulos, T. (Ed.). (2024). *Online hate speech patterns in media platforms' comments sections: Cyprus, France, Greece, Italy* [Report]. Symplexis. https://cesie.org/media/chase-online-hate-speech-patterns.pdf

Tindale, C. W. (2007). *Fallacies and Argument Appraisal*. Cambridge University Press. https://doi.org/10.1017/cbo9780511806544

Udupa, S. (2025). *Extreme speech and hate hide in everyday chats*. Ludwig-Maximilians-Universität München. https://www.lmu.de/en/newsroom/news-overview/news/extreme-speech-and-hate-hide-in-everyday-chats-7805fcd7.html

Uyheng, J., & Carley, K. M. (2020). Bots and online hate during the COVID-19 pandemic: case studies in the United States and the Philippines. *Journal of Computational Social Science, 3*(2), 445–468. https://doi.org/10.1007/s42001-020-00087-4

UNESCO & United Nations Office on Genocide Prevention and the Responsibility to Protect. (2023). *Addressing hate speech through education: A guide for policy-makers*. https://unesdoc.unesco.org/ark:/48223/pf0000384872.

Valentim, J. P. (2011). Social psychology and colonialism: Luso-tropicalism as a social representation in the context of contemporary Portuguese society. Em J. P. Valentim (Ed.). Societal approaches in social psychology (pp.179-194). Berne: Peter Lang.

van Dijk, T. A. (1992). Discourse and the Denial of Racism. *Discourse & Society, 3*(1), 87–118. https://doi.org/10.1177/0957926592003001005

van Dijk, T. A. (2023). (Anti)Racist discourse. The Routledge Handbook of Discourse Analysis, 244–260. https://doi.org/10.4324/9781003035244-20

van Prooijen, J.-W. (2021). Injustice Without Evidence: The Unique Role of Conspiracy Theories in Social Justice Research. *Social Justice Research, 35*(1), 88–106. https://doi.org/10.1007/s11211-021-00376-x

van Prooijen, J.-W., & Douglas, K. M. (2017). Conspiracy theories as part of history: The role of societal crisis situations. *Memory Studies, 10*(3), 323-333. https://doi.org/10.1177/1750698017701615

Vergani, M., Perry, B., Freilich, J., Chermak, S., Scrivens, R., Link, R., Kleinsman, D., Betts, J., & Iqbal, M. (2024). Mapping the scientific knowledge and approaches to defining and measuring hate crime, hate speech, and hate incidents: A systematic review. *Campbell Systematic Reviews, 20*(2). Portico. https://doi.org/10.1002/cl2.1397

Wagoner, B., Jørgensen, M. S., & Pahuus, K. (2026). Conspiracy theories through the lens of collective memory. *Current Opinion in Psychology, 68*, 102227. https://doi.org/10.1016/j.copsyc.2025.102227

Wang, M., Ma, S., Li, N., Zhang, P., Li, C., Gu, N., & Lu, T. (2026). Echoes of Norms: Investigating Counterspeech Bots' Influence on Bystanders in Online Communities. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems* (pp. 1-23).

Wardle, C. (2024). *A conceptual analysis of the overlaps and differences between hate speech, misinformation and disinformation*. United Nations, Department of Peace Operations and Office of the Special Adviser on the Prevention of Genocide. https://peacekeeping.un.org/sites/default/files/report_-_a_conceptual_analysis_of_the_overlaps_and_differences_between_hate_speech_misinformation_and_disinformation_june_2024.pdf

Weber, I., Vandebosch, H., Poels, K., & Pabian, S. (2023). Features for Hate? Using the Delphi Method to Explore Digital Determinants for Online Hate Perpetration and Possibilities for Intervention. *Cyberpsychology, Behavior, and Social Networking, 26*(7), 479–488. https://doi.org/10.1089/cyber.2022.0195

Williams, M. (2022). *The Science of Hate: How Prejudice Becomes Hate and What We Can Do to Stop It*. Faber & Faber, Ltd

Williams, M. L., Burnap, P., Javed, A., Liu, H., & Ozalp, S. (2020). Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime. *The British Journal of Criminology, 60,* 93-117. https://doi.org/10.1093/bjc/azz049

Wodak, R. (2015). The Politics of Fear: What Right-Wing Populist Discourses Mean. London: Sage. https://doi.org/10.4135/9781446270073

Zendeli, E. (2017). The right to education as a fundamental human right. *Contemporary Educational Researches Journal, 7*(4), 158–166. https://doi.org/10.18844/cerj.v7i4.2718

Zhang, C. (2018). *WeChatting American Politics: Misinformation, Polarization, and Immigrant Chinese Media* (TOW Reports). Tow Center for Digital Journalism. https://www.cjr.org/tow_center_reports/wechatting-american-politics-misinformation-polarization-and-immigrant-chinese-media.php/

Appendix

A- Recognising Hate Speech: Linguistic, Symbolic, and Contextual Markers

This appendix complements the analytical dimensions presented in section 4 by focusing on the practical recognition of hate speech in real-world contexts. It brings together a range of **linguistic, symbolic, and contextual markers** that can signal the presence of hateful or discriminatory expression, including both widely used and context-specific forms. These markers should not be understood as fixed or exhaustive indicators, but as evolving elements whose meaning depends on usage, intent and context. The examples provided aim to support awareness and interpretation, helping readers identify how hate speech may appear in different communicative environments.

Table A1. General Cross-Platform Markers of Hate Speech Expression

PLATFORM CATEGORY	MARKER TYPE	DESCRIPTION	ILLUSTRATIVE EXAMPLE
VIDEO-SHARING PLATFORMS (YOUTUBE, TWITCH, DISCORD, ROBLOX)	Transcribed or Captioned Hate Language	Hateful expressions appearing in video titles, descriptions, subtitles, auto-generated captions, or manually transcribed speech used for analysis.	Derogatory remarks targeting a group in video subtitles or auto-captions.
	Engagement-Driven / Comment-Triggered Hate	Hate speech emerging in comment sections as a reaction to video content, titles, or influencer framing; often amplified through replies, likes, or	Hateful or dehumanising comments targeting groups referenced in the video intensified through comment threads after publication.

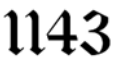
		repeated commenting.	
	Visual Symbolism	Logos, flags, colour schemes, or visual memes signalling ideological identity or exclusion.	Use of extremist insignia or historic flag imagery.
SOCIAL MEDIA NETWORKS (X, INSTAGRAM, TIKTOK, FACEBOOK)	Hashtags, Emojis & Numeric codes	Encoded communication using hashtags, emojis, numbers, or combinations that signal ideological affiliation or exclusionary narratives.	#1488, 🦃, or emojis combinations repeatedly used in hate clusters.
	Visual-Text Pairing	Image-caption combinations conveying sarcasm or superiority.	Meme contrasting “us” vs. “them.”
	Usernames, Group Naming & Description/Bios	Account names or profile descriptions containing extremist references, coded hate terms, or exclusionary slogans relevant to the posted content.	Usernames including “Reconquista” [Reconquest], “Portugueses Primeiro” [Portuguese First], “1143”, “Blood & Honour”, “IncelsExit” or similar coded titles.
INSTANT MESSAGING PLATFORMS (TELEGRAM, WHATSAPP, DISCORD GROUPS)	Forwarded Media	Circulation of edited videos, stickers, or memes reinforcing stereotypes.	GIFs implying dehumanisation.

FORUMS & BLOGS	Narrative Framing	Long-form posts using moral panic, pseudoscience, or conspiracy narratives.	"They control the media and the banks."
	User Signatures & Avatars	Visual identity markers or quotations revealing ideological perspective.	Historical fascist imagery.
NEWS & MEDIA COMMENT SECTIONS	Framing & Reaction Patterns	Comments reproducing hate in reaction to specific news topics.	"Next time, it's better if the police don't go...let them kill each other." "If we deport the Africans, crime goes down 40% in Portugal!"

Table A2. Context-Specific Markers: Examples from Portugal

This table documents terms, codes, and references that have been associated with extremist or hate-related discourse in the Portuguese context. Inclusion is strictly for research, analytical, and prevention purposes and does not imply endorsement. The presence of any marker alone is not sufficient to classify content as hate speech; contextual and cumulative analysis is required.

SYMBOL / MARKER	CONTEXT OF USE IN PORTUGAL	ILLUSTRATIVE NOTE
CELTIC CROSS (TEXTUAL REFERENCE)	Referenced in usernames, group names, slogans, or written descriptions	Appears in posts or bios combined with exclusionary slogans such as "Portugal aos

 **(E.G., CLUBE ÉTNICO PORTUGUEZ (PORTUGUESE ETHNIC CLUB) LOGO)**	linked to neo-Nazi or white nationalist groups.	Portugueses" [Portugal to the Portuguese], often signalling ideological alignment.
NUMERIC CODE "1143" 	Symbolises Portuguese independence and is used in usernames, hashtags, group titles, or slogans by nationalist or identitarian groups (e.g., references to "Grupo 1143").	Functions as a coded nationalist identifier; requires contextual co-occurrence with exclusionary narratives to be flagged.
"14 – 88" CODES	Numeric shorthand appearing in hashtags, handles, or comments associated with white supremacist ideology.	Strong ideological signal (e.g., "#1488" hashtag).
IDENTITARIAN SYMBOL / "ESCUDO IDENTITÁRIO" 	Symbol of the Portuguese identitarian youth movement. Used in group names, campaign slogans, or protest-related text.	Flyers against what they call "gender ideology" and its "danger".
"PATRIOTS88" NAMING PATTERN OR PATRIOTS NETWORK/ PATRIOTS FOR EUROPE	Channel names, hashtags, or self-descriptions referencing nationalist or transnational extremist networks.	@PatriotsNet_Off X page and Patriots Foundation conference in Portugal.

GROUP "BLOOD AND HONOUR PORTUGAL" 	A neo-nazi group that organises international events, such as concerts, which essentially serve as sites of radicalisation, recruitment, and the financing of its activities, including the production of propaganda.	An attack on the cast of the play *"Amor é um fogo que arde sem se ver"*, outside the A Barraca theatre in Lisbon, was reportedly carried out by the violent neo-Nazi group Blood & Honour.
"INCEL" / INCEL-RELATED TERMINOLOGY	Used mainly in online spaces and youth digital culture, particularly on social media, forums and gaming platforms. Associated with misogynistic and anti-feminist narratives.	References to "involuntary celibates" ("incels"), where some users express hostility towards women and feminism, often through coded language, memes, emojis or ironic humour. Examples of emojis: 💯 – Tied to the "80/20 rule," the belief that 80% of women are only attracted to 20% of men. 💥 – An "exploding red pill," meaning someone is a radicalised incel. 🫘 – Used by incel communities to identify themselves and to label others who share their views.
"RED PILL", "BLUE PILL", AND "BLACK	Used in adolescent and young-adult "incel" online communities (e.g., TikTok, Discord,	TikTok or YouTube videos/ and YouTubers claiming men must

PILL" SYMBOLISM/ IDEOLOGY	manosphere forums) to promote anti-feminist, gender-hierarchical worldviews.	"wake up" from feminism. 💊 Blue Pill: means remaining ignorant to the 'real world' experienced by Incels. 💊 Red Pill: means 'waking up' to the real world, where women have an advantage and female oppression is a myth. 💊 Black pill: means that they adhere to ideas behind the 'Red Pill', but don't believe that society will change or that Incels lives can be improved.
MANOSPHERE / ANTI-FEMINIST IDEOLOGY 💊 **(USED TO INDICATE AFFILIATION WITH THE "MANOSPHERE")**	Used in YouTube comments, TikTok, and Telegram groups to delegitimise feminist activism, portray women as manipulative or morally inferior, and for the promotion of actual or symbolic violence against women, but always of a sexual nature.	Quotes such as "Women who are dating don't go out at night" are used on PT TikTok pages. Following or being a fan of influencers such as Numeiro, whose philosophy of life is intertwined with misogynistic ideas spread online.
MISOGYNISTIC EMOJIS, SLURS AND CODED TERMS / "ADOLESCENCE"	Gendered slurs, slang, or seemingly "humorous" or innocent phrases or symbols	🍆 🌭 🍌 🌽 : an eggplant, a hot dog, a banana, or a corn cob are the emojis used to

(NETFLIX SERIES) REFERENCES	circulating in youth-oriented platforms.	refer to the male sexual organ. 🌷🌮🍣🐱: the female organ is "represented" by a flower, a taco, a piece of sushi, or a cat.

***Sources**: Global Project Against Hate and Extremism (2025); Sic Notícias; Público; Expresso; Diário de Notícias; Notícias ao Minuto; UN Women; nit; HM Government*.

COOPERHATE

COUNTERING HATE SPEECH

Cooperhate, under de Grant Agreement Nº. 101213938, is funded by the European Union.

Co-funded by the European Union