ONLINE HATE SPEECH TRILOGY – VOL III

METHODS, TECHNIQUES AND AI SOLUTIONS IN THE AGE OF HOSTILITIES

BRANCO DI FÁTIMA [ED]

LABCOM COMUNICAÇÃO & ARTES

Editorial Universidad Icesi

ONLINE HATE SPEECH TRILOGY — VOL III

METHODS, TECHNIQUES AND AI SOLUTIONS IN THE AGE OF HOSTILITIES

BRANCO DI FÁTIMA [ED]

LABCOM COMUNICAÇÃO & ARTES

Editorial Universidad Icesi

Technical Specification

Title
Methods, Techniques and AI Solutions in the Age of Hostilities — Online Hate Speech Trilogy - vol III

Editor
Branco Di Fátima

LabCom Books & Editorial Universidad Icesi
www.labcom.ubi.pt
www.icesi.edu.co/editorial

Collection
Communication Books

Direction
Gisela Gonçalves (LabCom Books)
Adolfo A. Abadía (Editorial Universidad Icesi)

Graphic Design
Cristina Lopes

ISBN
978-989-9229-10-5 (print)
978-989-9229-05-1 (pdf)

Legal Deposit
538471/24

DOI
https://doi.org/10.18046/EUI/ohst.v1

Print
Print-on-demand

University of Beira Interior
Rua Marquês D'Ávila e Bolama
6201-001 Covilhã
Portugal
www.ubi.pt

Universidad Icesi
Calle 18 No. 122-135 (Pance)
760031, Cali - Colombia
www.icesi.edu.co/es

Covilhã, Portugal 2024
Cali, Colombia 2024

Contents

Preface

TOXIC LANGUAGE, DETECTION METHODS AND AI

Branco Di Fátima
/ LabCom – University of Beira Interior

This is the third book in the **Online Hate Speech Trilogy**. It focuses on presenting methods for detecting, analysing, and combating toxic language on the Internet. Alongside the legal dilemmas born from a desire to punish hate speech disseminators, identifying online hate speech is one of the biggest challenges in the field of studies on violent narratives and virtual attacks.

One of the primary epistemological problems is the lack of a universally accepted definition for hate speech (Tontodimamma et al., 2020). How can one measure the impact of a phenomenon that is not adequately defined? (Müller & Schwarz, 2021). Hate speech is generally understood to be a verbal or non-verbal attack on an individual or group, usually a social minority. However, this definition can be broader or even profoundly different depending on the values and cultural codes of each society (Matamoros-Fernández & Farkas, 2021).

Empirical research highlights the difficulties of detecting violent narratives online (Miranda et al., 2022). Haters mobilize numerous subterfuges to obscure their intentions, such as irony, humour, or sarcasm (Filibeli & Ertuna, 2021). With the popularization of image and video editing software, it has become easier to create very sophisticated hate messages. Haters also dehumanize their opponents by comparing them to repulsive animals such as wasps, snakes, or spiders, and it is common to find memes with these characteristics (Makhortykh & González-Aguilar, 2023; Ndahinda & Mugabe, 2022).

This book brings together chapters written by 18 authors, from 7 universities, who examine alternatives for identifying, analysing, and combating hate speech online. They achieve this by testing traditional and digital methods, cross-referencing quantitative and qualitative data, and exploring the intricacies of various digital platforms such as websites, instant messaging apps, and social media.

The authors analyse the challenges of identifying violent narratives through automation, the advantages of manually coding social media posts, and the opportunities offered by AI in this field of research. They also highlight the use of machine learning, Social Network Analysis, and large language models to map toxic language. Additionally, the authors present new classification and taxonomy models that can be replicated by other researchers, along with alternatives for combating virtual attacks through media literacy and video games.

Hate speech has become increasingly complex on the Internet, and methods for detecting it need to be rapidly improved (Di Fátima, 2023). The use of automated detection tools and computer languages is a viable alternative, albeit with limitations. When humans code discourse manually, it is crucial to also consider how violent narratives can impact the researchers themselves as they encounter them.

This book explores scientific methods for identifying, analysing, and combating toxic language on the Internet. Volumes 1 and 2 of the **Online Hate Speech Trilogy** delve into the close links between disinformation, polarization, and virtual attacks. The legal challenges of prosecuting hate crimes while safeguarding freedom of expression are also analysed. The aim is to provide a multicultural overview of one of the most pressing issues in contemporary society, which is responsible for undermining democratic values.

References

Di Fátima, B. (2023). *Hate speech on social media: A global approach*. LabCom Books & EdiPUCE.

Filibeli, T. E. & Ertuna, C. (2021). Sarcasm beyond hate speech: Facebook comments on Syrian refugees in Turkey. *International Journal of Communication*, 15, 2236–2259.

Makhortykh, M. & González-Aguilar, J. M. (2023). Is it fine? Internet memes and hate speech on Telegram in relation to Russia's war in Ukraine. In Di Fátima, B. (Ed.), *Hate speech on social media: A global approach* (pp. 73-94). LabCom Books & EdiPUCE.

Matamoros-Fernández, A. & Farkas, J. (2021). Racism, hate speech, and social media: A systematic review and critique. *Television & New Media*, 22(2), 205-224. https://doi.org/10.1177/1527476420982230

Miranda, S., Malini, F., Di Fátima, B., & Cruz, J. (2022). I *love to hate! The racist hate speech in social media*. Proceedings of the 9th European Conference on Social Media (pp. 137 145). Krakow: Academic Conferences International (ACI).

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

Ndahinda, F. M. & Mugabe, A. S. (2022). Streaming hate: Exploring the harm of anti-Banyamulenge and anti-Tutsi hate speech on Congolese social media. *Journal of Genocide Research*, 26(1), 48-72. https://doi.org/10.1080/14623528.2022.2078578

Tontodimamma, A., Nissi, E., Sarra, A., & Fontanella, L. (2020). Thirty years of research into hate speech: Topics of interest and their evolution. *Scientometrics*, 126(2021), 157-179. https://doi.org/10.1007/s11192-020-03737-6

DEFINING HATE SPEECH: CONSTITUTIVE RHETORIC AND THE MEANING OF HATE ON SOCIAL MEDIA

Reed Van Schenck
IE University, Spain
rvanschenck@faculty.ie.edu

This chapter proposes an addendum emphasizing constitutive rhetoric to academic definitions of hate speech. Constitutive rhetoric is discourse which identifies speaker, audience, and implied public listeners through tropes held or opposed in common. I argue that current legal, platform, and academic definitions acknowledge the immediate consequences of hate speech but overlook its ability to constitute specific others as inferior in relation to hate speakers. After offering an overview of key concepts in critical rhetoric, I interpret two extremist social media memes, triple-parentheses and "dindu nuffin," as hate speech by examining how they constitute Jews and black people, respectively, as inferior to humanity. I explicate how constitutive rhetoric may strengthen researchers' ability to identify hate speech on social media through recommendations for scholars of all fields to incorporate constitutive rhetoric into research design.

Keywords: hate speech, constitutive rhetoric, critical race theory, identification, social media

MONITORING HATE SPEECH AGAINST IMMIGRANTS IN SOCIAL MEDIA: A TAXONOMY AND A GUIDE TO DETECT IT

Berta Chulvi Ferriols
Universitat de València, Spain
berta.chulvi@uv.es

Paolo Rosso
Universitat Politècnica de València, Spain
prosso@prhlt.upv.es

Karoline Fernandez De La Hoz Zeitler
Ministry of Inclusion, Social Security and Migrations of the Spanish Government, Spain
oberaxe@inclusion.gob.es

The chapter presents hate speech as a new display of prejudice towards minorities and analyses which characteristics of digital societies facilitate its dissemination. The state of the art in automatic identification of hate speech is reviewed. In order to facilitate an empirical approach to hate speech monitoring a conceptual framework, a taxonomy and a guide for annotation of this discourse are proposed. Moreover, an empirical analysis of some aspects of the hate speech monitoring done by OBERAXE in 2021 is also provided.

Keywords: hate speech, immigrants, monitoring, taxonomy, data annotation

TRIALS AND CHALLENGES MEASURING ONLINE HATE

Andre Oboler
La Trobe University, Australia
a.oboler@latrobe.edu.au

This chapter explores how efforts to map hate speech can be assessed and measured. Drawing on past efforts in academia, civil society, the technology industry, and government, the chapter explores four approaches to mapping online hate: demonstrating hate, counting hate, manually coding hate, and modelling hate. The importance of both expert knowledge and reliability in replication for manual classification is discussed. The benefits and limitations of modelling hate with sampling, pattern matching, and supervised machine learning are considered. Metrics discussed in this chapter include the inter-coder agreement rate for manual coding, and confusion matrices, precision, recall, and F-score for evaluating models against a known source assumed to be true. Throughout the chapter the measurement aspect of past work is explored.
Keywords: online hate, hate speech, empirical research, antisemitism

HARNESSING ARTIFICIAL INTELLIGENCE TO COMBAT ONLINE HATE: EXPLORING THE CHALLENGES AND OPPORTUNITIES OF LARGE LANGUAGE MODELS IN HATE SPEECH DETECTION

Tharindu Kumarage
Arizona State University, USA
kskumara@asu.edu

Amrita Bhattacharjee
Arizona State University, USA
abhatt43@asu.edu

Joshua Garland
Arizona State University, USA
garland.joshua@gmail.com

Large language models (LLMs) excel in many diverse applications beyond language generation, e.g., translation, summarization, and sentiment analysis. One intriguing application is in text classification. This becomes pertinent in the realm of identifying hateful or toxic speech – a domain fraught with challenges and ethical dilemmas. In our study, we have two objectives: firstly, to offer a literature review revolving around LLMs as classifiers, emphasizing their role in detecting and classifying hateful or toxic content. Subsequently, we explore the efficacy of several LLMs in classifying hate speech: identifying which LLMs excel in this task as well as their underlying attributes and training. Providing insight into the factors that contribute to an LLM's proficiency (or lack thereof) in discerning hateful content. By combining a comprehensive literature review with an empirical analysis, our paper strives to shed light on the capabilities and constraints of LLMs in the crucial domain of hate speech detection.

Keywords: AI, large language models, text classification, hate speech detection

MAPPING THE HATE SPEECH ON TWITTER: POLITICAL ATTACKS ON JOURNALIST PATRÍCIA CAMPOS MELLO

Fábio Malini
Federal University of Espírito Santo, Brazil
fabiomalini@gmail.com

Jéssica do Nascimento Oliveira
Federal University of Espírito Santo, Brazil
jessicanoliveira3@gmail.com

Gabriel Herkenhoff Coelho Moura
Federal University of Espírito Santo, Brazil
gabriel.herkenhoff@gmail.com

This research focuses on the attacks suffered by journalist Patrícia Campos Mello, from Folha de São Paulo, motivated by Hans River's testimony to the Joint Parliamentary Commission of Inquiry (CPMI) on Fake News. Brazilian former president Jair Messias Bolsonaro, the target of the investigation led by Campos Mello on the use of *fake news* in the Brazilian presidential election of 2018, stimulated sexist comments and sexual insinuations against the journalist on social media. He used the deponent's speech and made the statement: "She wanted a scoop. She wanted to scoop the scoop at any price against me". Based on this episode, we aimed, through the analysis of comments on Twitter, to identify words and expressions that characterize hate speech against women in the digital environment. After filtering the content of the most shared tweets in the database, we labeled the linguistic material identified using the network discourse perspectives method. The team of coders grouped these words/expressions into categories. Then, this supervised database was applied to all retweets, taking topic modeling and machine learning techniques for classification algorithms as a starting point. It was possible to highlight some evidence that contributes to delineating the concept of hate speech and to discuss how this speech is operated against women. The methodological contribution of this work is to combine Social Network Analysis and Digital Discourse Analysis to reflect on topics

in the field of Linguistics. Furthermore, this work contributes to the disclosure of power relations created through language in situations of violence against women in the digital environment.
Keywords: hate speech, social media, digital platforms, perspectives, SNA, digital discourse analysis

OBSTACLES TO DETECTING AND SUPPRESSING ONLINE HATE SPEECH

Mariana Magalhães
University of Porto, Portugal
mariana@fpce.up.pt

Sara Alves
University of Porto, Portugal
up201304933@edu.fpce.up.pt

Márcia Bernardo
University of Porto, Portugal
oliviabernardo95@gmail.com

In the online world, hate speech is becoming increasingly prevalent and the real prevalence may be even higher than estimated. Detecting and suppressing hate speech is, thus, a priority of national and supranational authorities, including the European Union. However, some obstacles limit the efforts of detection and suppression of online hate speech. This chapter will explore these obstacles, which can be divided into two main types. The first relates to the limitations of existing social control mechanisms, namely the Portuguese and European legislation, the police force and social media. The second refers to the sociopsychological phenomena that normalize hate speech, such as the bystander effect, the influence of politicians and the media and the fluid nature of hate speech. By acting on these obstacles, online hate speech may be reduced.
Keywords: detecting online hate, social control, European legislation, combating hate speech

THE PROPS PROJECT: INTERACTIVE NARRATIVES AS COUNTERPOINTS TO ONLINE HATE SPEECH IN VIDEO GAMES

Ana Filipa Martins
University of Algarve, Portugal
fcerolm@ualg.pt

Bruno Mendes da Silva
University of Algarve, Portugal
bsilva@ualg.pt

Alexandre Martins
University of Algarve, Portugal
acmartins@ualg.pt

Susana Costa
University of Algarve, Portugal
srsilva@ualg.pt

The following chapter discusses the research results from "PROPS – Interactive Narratives Propose Pluralistic Discourse" (2023-2024), a project focused on media education to address online hate speech, particularly in the context of online video games. The initiative aimed to develop a different approach to this issue through the creation of interactive counter-narratives, designed to motivate and engage educators, trainers, children, and young people. The term "PROPS" (slang for proper respect) encapsulates the project's ethos of fostering respect in digital spaces. PROPS began with a comprehensive review of existing literature on online hate speech, video games, and the use of interactive narratives as pedagogical tools. This was followed by surveys and focus groups conducted with students aged 10-18 to gather firsthand experiences and perspectives. The collected data was instrumental in the creation of six interactive narratives, which will serve as educational tools to foster reflection and discussion about online hate speech and its prevention, in educational settings.

Keywords: online hate speech, online video games, media literacy, interactive narratives

DEFINING HATE SPEECH: CONSTITUTIVE RHETORIC AND THE MEANING OF HATE ON SOCIAL MEDIA

Reed Van Schenck
/ IE University, Spain

The advent of social media has led to a global eruption of novel networks of communication. Unfortunately, some of these networks spread ideas that are anything but new. Hate speech has found a new home on social media, spreading to new audiences and sewing widespread consequences for the health of digital users and the societies in which they live. On the bright side, researchers have responded to this trend with urgent research from different disciplines, perspectives, methods, and regions attesting to the effects of hate speech on social media. This growing body of research stands as an inspiring testament to the global academic community's eagerness to stand against the spread of hateful ideas online.

However, among all the controversies that animate this literature, one simple yet profound debate stands out: how to define "hate speech." Scholars attest to the difficulty of defining of hate speech due to the diversity of actors at play: governments, platforms, companies, and users converge online with radically different understandings of the role of speech in society, the degree to which speech ought to be regulated, and thresholds for what makes a speech act hateful. To make matters more difficult, scholars from different fields and methods traffic in their own disciplinary assumptions: A legal

scholar's inquiry attends to different frames of reference than a scholar of cultural studies or computer science. Consequentially, researchers produce definitions that may appear similar yet functionally harbor profound differences. Should the reader approach the literature on hate speech and social media without paying attention to these definitions, confusion is inevitable.

To define hate speech is the first step to studying it. The definition that one assigns to "hate speech" alters each subsequent phase of research: from the methods they choose to collect and categorize hate speech, to the veracity of data analysis, to the means through which they recommend its curtailment. Should one give the same dataset to two scholars who operate under different definitions of hate speech, they will undoubtedly be returned with two very different sets of results. Therefore, it is of fundamental importance that scholars of hate speech on social media understand the different approaches taken to defining hate speech and how those definitions might be improved by way of multidisciplinary collaboration.

It is to this end that this chapter offers a critical rhetorician's perspective on the definition of hate speech. Despite the fact that communication scholars contribute a growing proportion of global research into hate speech on social media, scholars rarely consult rhetoric to understand the meaning of hate speech. To rectify this gap, this chapter presents a rhetorical additive to the definitional controversy. Through a review of the literature's many definitions of hate speech, I argue that scholars across fields should include hate speech's effect as constitutive rhetoric in their definitions. Constitution is a rhetorical form that traces how speakers use discourse to identify themselves and their audiences in common with or against another. Rooted at the intersection of rhetorical analysis and critical theories of identity and race, this constitutive definition of hate speech clarifies how hate speech enacts intolerable violence upon marginalized peoples. Even absent the direct incitement of violence, hate speech enforces exclusion from the grounds of humanity from which the speaker constitutes themselves. Constitutive rhetoric clarifies the material consequences of hate

speech while introducing readers to the growing body of rhetorical analysis of hate speech on social media.

This chapter proceeds in four parts. First, I review the field's three primary strategies for defining hate speech – legal, platform, and academic – and I situate my intervention within the academic perspective. Second, I offer a brief overview of the discipline of rhetoric, constitutive rhetoric, and the efforts made by rhetoricians to understand the persuasive function of hate speech. Third, I articulate how constitutive rhetoric can augment scholars' ability to identify and analyze the effects of hate speech beyond its immediate consequences. Fourth and finally, I articulate two unique benefits to the rhetorical definition: understanding unique modes of hate speech endemic to social media and recommending proactive regulatory strategies that can stymie its spread. This chapter concludes by offering practical recommendations for non-rhetorical scholars interested in using this definition to refine their own research.

Hate speech defined three ways: Legal, platform, and academic

Scholars have taken numerous approaches to defining hate speech in line with the purpose of their inquiry. These include legal definitions oriented toward identification and redress in the court of law, platform definitions which provide blueprints for content moderation on social media, and academic definitions which provide schemas to optimize research. Consequentially, the reader looking for a straightforward, universal definition of hate speech will emerge empty handed. This section offers a review of three main approaches taken by regulators and researchers to define hate speech online: legal, platform, and academic. By reviewing the affordances and limitations of each group of definitions for the scholarly researcher, I hope to establish the significance of the rhetorical contribution.

Scholars have utilized legal definitions of hate speech when studying online communities corresponding to national and international bodies of law. One of the clear benefits of the legal approach is that the resulting definition

draws from real precedent that can directly facilitate redress (Leets, 2001; Paz et al., 2020). Toward this end, many legal scholars consider eclectic definitions by selecting definitions from multiple governmental bodies and blending their components in order to make a case for their adoption. For example, in his influential defense of expanded anti-hate speech regulation, Jeremy Waldron (2012) defines hate speech as "the use of words which are deliberately abusive and/or insulting and/or threatening and/or demeaning directed at members of vulnerable minorities, calculated to stir up hatred against them" (Waldron, 2012: 8–9). He draws his definition from regulations found in Canada, Denmark, Germany, New Zealand, and the United Kingdom, five states with strong restrictions on hate speech. However, others note that the eclectic legal definition leaves room for misinterpretation and over- or under-regulation, perhaps undermining some of the benefits of this approach (Leiter, 2012). These definitions leave the threshold for establishing "abusive and/or insulting and/or threatening and/or demeaning" verbiage to judges and lawmakers bound by precedent.

While legal definitions remain crucial for facilitating specific policy change, their contextual baggage weakens their application in comparative and international studies. For example, many scholars cite a definition (or other derivatives) found in the European Union's Recommendation against Racism and Intolerance on combatting hate speech (Alkiviadou, 2019; Aswad, 2016; Whine, 2016). This definition is one of the broadest in the literature, covering:

> all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin (*Recommendation of the Committee of Ministers to Member States on "Hate Speech,"* 1997).

While this definition clearly facilitates regulation while covering a large breadth of harmful speech, it is inapplicable in many other contexts, such as in the United States where the prevailing interpretation of the Constitution prevents intent-driven prohibitions of expression. As a result, U.S. American researchers couch their definitions within local or state-level obscenity laws and tort violations, rendering the Commission's definition less relevant (Delgado & Stefancic, 2018: 108–112; Holling & Moon, 2021: 438–439). Consequentially, legal definitions face a troubling double-bind: while definitions minted from a single source struggle for relevance outside of their governing body of origin, eclectic definitions lose some of their meaning when taken out of context and blended with other fragments of definitions. Either way, their abilities to dictate redress, and to facilitate a research agenda, are limited.

Some scholars turn to platform-based definitions to understand how social media companies identify and regulate hate speech on their websites. These definitions offer researchers the most specificity for digital media studies because platforms are, in the words of Andrew Sellars (2016: 20), "perhaps the most active space in adjudicating definitions of hate speech." Compared to some governments, platforms enjoy relatively unrestricted freedom to regulate content on their own websites. Researchers who study specific platforms – Twitter being the most popular in the literature, trailed by YouTube, Facebook, Reddit, and others – often utilize the definition of "hate speech" presented in the platform's terms of service (Matamoros-Fernández & Farkas, 2021: 209–210; Ruwandika & Weerasinghe, 2018; Zhang & Luo, 2019). For example, scholars studying hate speech on Twitter often cite the platform's hateful conduct policy when categorizing content as hate speech (Iorliam et al., 2021). Before its rebrand under Elon Musk, Twitter's (2022) policy expanded to cover threats, calls for harm, references to violent events, incitement against marginalized people, slurs and tropes, hateful images, and other modes of speech which:

> promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories.

This approach benefits from its specificity, allowing the researcher the ability to comment more directly upon hate cultivated on certain platforms.

However, these definitions struggle in facilitating reflection upon platforms' shortcomings, rendering them insufficient to achieve a critical conception of online hate speech. By accepting a platform's own definition, we undercut our ability to evaluate the necessity and sufficiency of a platform's current regulatory approach (Konikoff, 2021). For example, suppose a researcher must identify how many Tweets out of a sample of 10,000 contain hate speech. If the researcher applies Twitter's definition of hate speech to find 1,000 offending posts, but an additional 200 contain euphemistic language that Twitter's definition cannot properly identify as hateful, then the researcher will miss an opportunity to recommend an improvement to the platform's hate speech policy. Furthermore, relying upon platform definitions makes it difficult to conduct research on extremist platforms, such as the white supremacist social network Gab. These websites often have few to no policies because they reject the notion that hate speech ought to be regulated at all. This is problematic because hateful content produced on these platforms can still reach mainstream social media audiences (Mathew et al., 2019). Platform-based definitions remain invaluable resources, but they ought to be supplemented by critical definitions to attend to the shortcomings of each platform.

Toward this end, many scholars prefer academic definitions that seek to establish far-reaching, universalizable conceptions of hate speech. Critical definitions of hate speech trace back to U.S. American legal scholar Richard Delgado (1982: 179) whose definition focuses on the perception and effects of hate speech. He establishes a criterion for proving racist hate speech

as language that is "intended to demean through reference to race… understood as intended to demean through reference to race, and that a reasonable person would recognize as a racial insult." This definition introduces the perception of a "reasonable person," a legal stand-in for the implied audience of the public, into the interpretation of hate speech. Later, Delgado and Jean Stefancic (2014: 320) applied this conception to online hate speech by classifying it within the category of online criminal behavior which "decrease trust, weakens social bonds, or erodes quality of life."

Academic definitions such as these are prescriptive, empowering scholars to define hate speech by way of what it *does* rather than what it *is* (Paz et al., 2020; Siegel, 2020; Woods & Ruscher, 2021). Delgado's and Stefancic's approach has spread widely across legal, communication, and media studies, informing lengthy taxonomies of hate-speech effects that enjoin definition. For example, Sellars (2016: 24–30) identifies eight common themes that cut across such definitions: (1) targeting a group or individual as a member of a group, (2) expressing hatred through the message, (3) causing verifiable harm, (4) possessing intent to harm (physically or otherwise), (5) incitement of bad actions beyond the speech act itself (physically or otherwise), (6) occurring in a public context with an audience and/or directing public hatred toward a person, (7) enabling violent response, and (8) lacking any redeeming purpose. Academic definitions are unrestricted by the limitations of any one governing body or platform. Therefore, they tend to interpret hate speech by its fullest social consequences and open up the possibility for reflexive criticism of current social media policies.

While the academic approach to definition offers the most affordances for scholars to assess online hate speech and its current regulation, this literature could be significantly improved by incorporating a perspective from critical rhetoric. Communication studies perspectives are underrepresented in the literature. Compared to legal and psychological studies, research from the discipline of communication is slight but increasing (Paz et al., 2020: 4–5). Critical perspectives from rhetoric and discourse analysis are also marginal. Surprisingly, despite the fact that so many scholars

owe their definition of hate speech to one of the founders of critical race theory (Delgado), fewer than a quarter of articles about social media and hate speech reference critical perspectives about identity or race, and fewer than 6% of quantitative studies on the subject engage this important field of inquiry (Matamoros-Fernández & Farkas, 2021: 212–213). Critical perspectives of rhetoric and race can improve the field's ability to craft comprehensive methods of analysis, irrespective of the researcher's home discipline (Waseem & Hovy, 2016). To set the foundation for how a rhetorical definition might facilitate more critical thought on hate speech, I now offer a brief overview of constitutive rhetoric.

Constitutive rhetoric: From persuasion to identification

Like hate speech, "rhetoric" is polysemous. Scholars understand rhetoric differently in correspondence with the conditions of communication that they observe in their disciplinary context. This chapter will be no different. The definition of rhetoric that I employ here does not encompass the field's full scope, but it does home in upon the insights that critical rhetoric can offer scholars of hate speech and social media. Therefore, this section outlines the development of a specific function of rhetoric that scholars of hate speech and social media would benefit from understanding: constitutive rhetoric, or the rhetoric of identity construction. I begin by briefly glossing the foundations of classical rhetorical inquiry before examining tropes of constitutive rhetoric at the intersection of rhetorical analysis and critical theory.

One of the most widely-cited classical definitions of rhetoric states: Rhetoric is persuasive discourse, or discourse that seeks to influence its audience (Black, 1978: 10–14). Rhetoric addresses the means through which a "speaker" delivers an argument to an "audience" of rational listeners in order to engender some change in their thought and/or action. This definition draws from the Greco-Roman schools of oratory, headlined by the likes of Aristotle, Quintilian, and Cicero, re-interpreted by the "neo-Aristotelian" tradition in the United States. Its traditional objects of study are speeches, particularly

those delivered within public fora, as well as literature. Using Aristotle's *Rhetoric* as a blueprint, centuries of Western rhetorical inquiry concerned themselves with evaluating the influential efficacy of discourse, or whether a speaker's form, style, and argument persuaded its audience. While oration dominated this genre of rhetorical inquiry, scholars also acknowledge rhetoric within written discourse and underwent criticisms of essays, novels, pamphlets, and other written compositions. Thus, rhetoric found homes among scholars of speech communication and literary composition.

With the twentieth century came the evolution of rhetoric beyond its Grecian cloister. The discipline faced identity crises from two developments. First, the emergence of new disciplines of humanistic inquiry – anthropology, psychoanalysis, sociology, and semiology, among others – contested rhetoric's exclusive grip upon the function of persuasion (Burke, 1951: 202–203). These disciplines challenged the foundational premise that audiences tend to *rationally* evaluate arguments. Critical scholars emerged to consider how social conditions, political circumstances, histories of oppression, and other factors subvert rational evaluation. This contention is greatly magnified by the second challenge to the discipline: The proliferation of mass media contested the duopoly of speaking and writing. As the ability to make and deliver arguments to mass audiences expanded beyond the illustrious orator or magnanimous author, the ability to critique persuasion expanded beyond speech and literature departments. Photographs, cartoons, and films persuade with as much effectiveness as spoken and written word (DeLuca, 2005: 89–93). Rhetoricians reckoned with the discipline's failure to pay due attention to the influence of technical and aesthetic components of communication – a fact made clear by social media.

As a result, midcentury rhetoricians invented "new rhetorics" that could attend to different forms of influence beyond persuasion through text. Kenneth Burke reconceived rhetoric's central function as identification rather than persuasion. Burke observed that each mode of persuasion canonized by Aristotle required a speaker to establish communality with their audience (Burke & Zappen, 2006: 335–336). Before one can evaluate an

argument, they must accept that the speaker shares their motive. For example, suppose a speaker wishes to persuade an audience to accept a certain national policy. To establish that their policy is in the best interests of their audience, they might identify themselves by way of the phrase, "My fellow citizens." This act of identification signals that, as citizens, speaker and audience share stakes in the advocacy because both stand to lose or gain. This establishes trust in the speaker's character as a precondition for persuasion. Audiences do not evaluate this act of identification consciously, as they might evaluate formal argument, but sub- and un-consciously, as a symbolic inducement of commonality that enables further evaluation. Observing this dynamic, Burke argued that identity is defined through social categories represented by symbols. That is to say, there exists no "unsocialized, pre-discursive, essential self" outside of rhetoric (Branaman, 1994: 445).

With the late twentieth-century introduction of rhetoric to critical and post-structuralist theories came the genre of constitutive rhetoric. Following Burke's turn to identification, constitutive rhetoric is a mode of persuasion that "calls its audience into being" (Charland, 1987: 134) As developed by Maurice Charland in his study of the rhetoric of Québécois nationalism, constitutive rhetoric expands upon Burke's insights by rejecting the existence of transhistorical "speakers" and "audiences," instead arguing that both are "constituted" through discourse in context (Charland, 1987: 133–135). Constitutive rhetoric invokes three key "narrative ideological effects" (Putman & Cole, 2020: 210–211). First, the rhetor constitutes a collective identity and positions both themselves and their audience within its purview. Second, the rhetor collapses any temporal, geographical, or other differences among persons, manufacturing feelings of commonality. Finally, the rhetor facilitates action through a call to act as the subject. These three processes ensure that subjects perceive themselves as belonging to and acting within a collective identity without assuming that the identity has been imposed upon them – it makes identification feel organic even though it is not, corroborating Burke's emphasis on the irrationality of persuasion.

Charland and other critical rhetoricians observe that all identities exist by tautology because it is through self-referential discourses that groups discover their commonalities. Let us return to the "My fellow citizens" example to make an important caveat: Constitutive rhetoric does not claim that social identities, such as the "citizen," are conjured out of thin air and flowery language alone. Clearly, there exist nation-states and real governing bodies through which states empower their citizens and disempower non-citizens. Rather, what it means to identify and act as a citizen is defined by discourse, and the appeal to "citizens" obtains persuasiveness through its symbolic functions. After all, to empower citizens requires that a government achieve recognition as sovereign by pre-existing nation-states, whose capacity to recognize a citizenry was similarly constituted through rhetorical negotiations with other sovereign states (Mills, 2014).

Rhetoric's contribution to the definition of hate speech emerges from a specific trope of constitutive rhetoric, called "identification by antithesis" (Goehring & Dionisopoulos, 2013). Borrowing from Burke (1973), identification by antithesis constitutes shared identity by virtue of a common opposition rather than a positive attribute. This occurs through the construction of two tropes to which the speaker opposes themselves: "generalized others" that lack specific markers of identity, such as the government, the media, and the wayward masses; and "specific others" marginalized by virtue of their identity, such as Jews and black people (Goehring & Dionisopoulos, 2013: 374; Stewart et al., 2012: 174–178). While generalized others usually refer to an un-marked social institution such as "government" or "society," specific others are singled out by virtue of some intrinsic trait, such as the "protected groups" including race, sex, gender, ethnicity, caste, religion, and many others.

Identification by antithesis tropically connects acts of hate speech with acts of terrorism. In their rhetorical analysis of the notorious white supremacist novel *The Turner Diaries* by William Pierce, Goehring and Dionisopoulos question, how could one obscure fiction book serve as inspiration for the

1995 Oklahoma City bombing, the 1999 London bombings, and the 1999 Columbine High School shooting? Their answer is that the novel utilized constitutive rhetoric to encourage its white readers to identify against marginalized populations and society itself. Identification against the generalized other of "corrupt society" allowed Pierce to forego the impossible task of reconciling his hateful ideology with the second narrative effect of constitutive rhetoric, which requires finding a common cause that intersects with intra-group diversity. Through this mode of hate speech, the majority of whites who do not foment animosity against other peoples are lumped into "corrupt society," excluded from the novel's constituted "white" identity due to their ideological difference. Identification by antithesis encourages readers to paper over the contradictions within their worldview, justifying violent acts to retaliate against society itself.

Constitutive rhetoric accepts that identification occurs when speakers mobilize discourse to form in-groups and out-groups. As such, it makes a necessary contribution to understanding why chauvinists invoke hate speech against those whom they perceive to be inferior. As we have observed, definitions of hate speech coined by governing bodies and platforms' terms of service exist primarily to facilitate regulation and redress. This is a critical task, but such definitions are insufficient for the distinct task of the academic: understanding how, why, and to what ends hate speech takes place. As Goehring and Dionisopoulos reveal, constitutive rhetoric can assist scholars of social media and hate speech by demonstrating how hate speech functions discursively: as a constitutive measure which conceals the intrinsically contradictory character of racist, antisemitic, misogynistic, and other supremacist ideologies, facilitating identification by denigrating the specific other. Put simply: Hate speech is not just an irrational explosion of vitriol. It is a tactic that seeks to create and enforce a hierarchical state of affairs through the rhetorical function of identification. In the next section, I will lay out this definition by way of the tropes of constitutive rhetoric.

Toward a constitutive definition of hate speech

If one has read this far into this chapter in search of a single sentence that suffices to define hate speech, then I am afraid that I will have to partially disappoint. I concur with Sellars (2016) that hate speech is most comprehensively defined by tracking the function of certain speech acts within their context, necessitating a schematic approach that attests to a series of effects attributable to hate speech. My purpose here is to add to, not to replace, these effects. I argue that rhetorical constitution is an effect of hate speech that elucidates its intrinsic harm, both to the targeted people and to the whole social system. I will demonstrate this point by situating constitution as a rhetorical function that differentiates hate speech from forms of protected expression. Thus, to keep my disappointment partial, my contribution can be summarized as such: Hate speech constitutes a protected group or person as a "specific other" inferior to and/or outside of the speaker's ideal humanity.

A functional definition of hate speech demands attention to every ramification of hate speech, not just those that are readily apparent to the observer. The harms of hate speech are conventionally divided into two general categories: consequential, referring to direct results of an act of hate speech (e.g. directly inciting an act of violence against a victim), and constitutive, pertaining to the subliminal effects of hate speech upon its victims. Regrettably, in some academic and regulatory contexts, only the consequential harms receive significant attention. Legal scholars Katharine Gelber and Luke McNamara (2016: 336) observe a litany of constitutive harms that create the conditions through which "consequential" violence becomes tolerable over time. These include "subordination, silencing, fear, victimization, emotional symptoms, restrictions on freedom, lowering of self-esteem, maintenance of power imbalances," and, most importantly, "undermining of human dignity." Understanding hate speech as constitutive rhetoric strengthens the significance that we assign these effects. While such effects may seem difficult to measure, their compounding effects upon the wellbeing of

marginalized populations merits further investigation. Mari Matsuda (2018) encourages scholars to account for these ramifications through critical, interdisciplinary inquiry that supplement legal and quantitative accounts with the lived experience of victims of hate speech, psychological and sociological accounts of its constitutive effects, and rhetorical studies of hate speech as a unique mode of discourse.

To understand the full ramifications of hate speech, it is necessary to look to how the speaker constructs their implied audiences on digital media. The implied audience, or "the second persona," suggests that any discourse is received not only by its direct addressee but also the implicit member of the public (Black, 1970: 111). Hate speech affects not only the direct target or the "specific other," but also the implied audience found among members of the public, or "generalized others," many of whom feel disturbed when they encounter hate speech. By defining themselves against both specific and generalized others, the hate speaker props up their own ethos or credibility as a rhetor. In this way, hate speech does not stop by harming its immediate audience, but as "the idiomatic token of an ideology" (Black, 1970: 115), it also reinforces the chauvinistic ideas that lend it credence within the public sphere. This clearly rings true on social media on which it is not uncommon to find posts circulate to audiences far beyond the intentions of the original poster. All posts imply the auditing of the hypothetical public viewer, no matter the size of one's current audience or accessibility of one's profile.

From a rhetorical perspective, hate speech is distinguished from protected modes of expression, such as social criticism, by its attempt to convince generalized others to re-join the hate speaker's ideal humanity by derogating specific others. Suppose someone were to post on Facebook, "The government is corrupt." This expression constitutes the government as a generalized other from which the speaker implicitly describes themselves as "not corrupt," and therefore a trustworthy critic, to the implied auditor. Nevertheless, it lacks a specific other and therefore neither victimizes a target nor calls a general other to adopt chauvinism. Therefore, it does not function as hate speech, meriting protection as free expression. Now,

imagine a similar post: "(((The government))) is corrupt." This post utilizes the "triple parentheses," a dog-whistle that signals antisemitic conspiracy within online white nationalist networks (Tuters & Hagen, 2020: 2228–2232). While the triple parentheses communicate naught but nonsense to most viewers, the reasonably-informed reader would immediately recognize that this post claims that the government is controlled by Jews which is why it is corrupt. With the mere addition of parenthetical signals, this post targets Jews as a specific other while still constituting both a generalized other ("the government" without parentheses) as well as the speaker's uncorrupt posture. The post targets Jews and attempts to convince general others, non-Jews, to do the same. Therefore, this post meets the rhetorical threshold for hate speech. This example demonstrates how rhetoric, attendant to the implicit meanings of discourses, is critical for understanding the evolution of hate speech on digital media.

Identification by antithesis helps scholars clarify how hate speech harms its victims and society by retrenching ideologies which position the in-group as the sole arbiter of what it means to be rightfully human. Hate speech constitutes the speaker (and the second personae who share the speaker's common oppositions) in a relationship of relative superiority over the others against which they identify themselves. According to Delgado and Stefancic, (2018: 113–120), even restrictive legal sources concur that hate speech gains much of its social impact by excluding its addressee from the vision of humanity that their speech constructs. A wealth of rhetorical and sociological scholarship attests to the power of identification by victimhood, which allows the speaker to establish human sympathy with the implied auditor while denying the same human recognition to the target of their speech (Bebout, 2020; Sharples & Blair, 2021). Unfortunately, this process encourages viewers, consciously or otherwise, to tolerate greater acts of dehumanization and violence against specific others.

These insights shine a light into the motivations of those who spread hate speech: not only for the thrill, but also to rhetorically enforce the truth of their ideas (Erjavec & Kovačič, 2012: 915). Indeed, "consequential" and

"constitutive" harms of hate speech are one in the same: rendered distinct for the purpose of neat analysis, but mutually reinforcing in practice. In the final section, I expand upon this insight by explaining how the constitutive definition may enhance the current state of scholarship about hate speech on social media.

Benefits of the constitutive perspective

Having worked through constitutive rhetoric's addendum to the definition of hate speech, I now demonstrate how this definition might ameliorate the study of hate speech on social media. I outline two general categories of benefit: understanding hate communication on social media and facilitating proactive regulation. Surely, this perspective cannot rectify all ongoing controversies in either area. Regardless, by forwarding the intrinsic significance of constitutive harms of hate speech, this definition fosters a more critical understanding of online hate speech and the necessity of proactive regulation at all relevant levels of governance.

The addition of constitutive rhetoric offers researchers more and better tools to understand different genres of hate speech that emerge from social media. One of the field's most pressing anxieties, held by quantitative and qualitative researchers alike, is how to evaluate the hatefulness of certain genres of online speech. Coded language, images, euphemisms, in-jokes, and memes are notoriously difficult for researchers to categorize as hate speech. For one, they are difficult to identify in data, even for deep-learning classification methods, because the full scope of injury can only be understood with contextual understanding gained through sustained engagement. The triple parentheses are exemplary. Qualitative scholars have made considerable progress in modifying their own methods to attend to vernacular hate speech on specific platforms, particularly on 4chan, an anonymous imageboard frequented by white supremacists, and within fringe communities on the popular social media and content aggregation platform Reddit (Rieger et al., 2021). While a theoretical definition is insufficient to resolve the technical problems which might impede detection, the constitutive addendum

should inspire researchers to design detection programs to acquire the "in-group knowledge" necessary to accurately classify forms of hate speech on social media.

Another barrier to identification of online hate speech is that it is difficult to quantify constitutive harms. Unlike inciting speech, whose harm is evident in the harmful act incited, constitutive effects are long-term, sometimes invisible, and often overlooked by unaffected observers. Legal and platform-based definitions, by ignoring constitutive effects, exacerbate these difficulties. The difficulty of representing constitutive harm hinders our capacity to understand various forms of online hate speech. For example, take the phrase "dindu nuffin," a racist meme that originated on 4chan but spread to mainstream social media in reaction to the Black Lives Matter movement in the United States. This term transliterates a caricature of black person saying, "I didn't do nothing." Critical and humanistic scholars recognize that "dindu nuffin" is a racial slur used to mock victims of police brutality against black people (Paul, 2021; Topinka, 2019). But beyond making a cruel joke at the expense of a victimized community, what does this phrase actually do?

This term is used by white supremacists on social media to exclude black people from humanity. Circulated after the murders of George Floyd and Breonna Taylor by police officers in the United States, "dindu nuffin" implies that black victims of police brutality deserved the violence that was inflicted upon them. By mocking the claim that these victims "didn't do anything," the users of this phrase instigate that Floyd and Taylor did do something to deserve their fate – by being black. The speaker rhetorically places themselves outside of and superior to black people, urging other non-black people to follow suit in thought and deed. Therefore, "dindu nuffin" should be classified as hate speech because it targets black people as a specific other and argues that they are less than human and consequentially deserving of police violence. The harm that such constitution causes is not felt by the deceased victims of the slander but rather by the black community that must face the emboldened assertion of their sub-humanity.

As such, the constitutive effects of hate speech are both internal – affecting the social, emotional, and physical health of oppressed communities – and external – reinforcing attitudes, ideas, behaviors, and practices which maintain unjust states of affairs online, offline, and in policy. These insights would be difficult to forge without making constitutive rhetoric an explicit part of hate speech's definition.

By including constitutive rhetoric in our understanding of hate speech, scholars can more adequately acknowledge all of its effects and recommend preventative measures on social media. Mainstream social media companies have increased their efforts to regulate hate speech, especially after researchers demonstrated how platforms facilitated the celebrity status of some extremist actors (Rogers, 2020). However, when determining whether a post contains hate speech, platforms often conflate the constitutive effects of hate speech with generalizable "offensive behavior." When paired with crude lexical detection methods, this causes regulators to neglect coded expressions like "dindu nuffin," to over-regulate non-hateful profanity, and to suspend non-offending accounts (Davidson et al., 2017). To be fair, an incomplete definition of hate speech does not explain all platform inconsistencies: the working conditions of human moderators, inefficient automated moderation systems, poor keyword databases, pressure from investors, and online celebrity status contribute to inconsistent platform governance.

Nevertheless, recognizing hate speech's constitutive nature should ameliorate moderation strategies in two notable ways: First, it clarifies that hate speech requires a specific other against which the speaker constitutes themselves, thus excluding from regulatory purview posts that interact with a hate speaker in the name of educating the public about the dangers of hate speech, or other non-hateful forms of free expression such as victimless banter. As such, the constitutive definition expands protections toward "counter-speech" which strengthens democratic practice on social media while simultaneously expanding the role of regulators. Second, constitutive rhetoric demands attention to context when discerning an act of hate speech, emphasizing the importance of human-mediated moderation. Just

because content lacks a hateful "keyword" such as a slur does not mean that it lacks the meaning and impact of hate speech in its context. By taking this intervention seriously, platforms might refine the tools at their disposal, and social media might render more useful tools for building a healthier online environment for all.

Conclusion

The definition of hate speech lies at the core of the legal, regulatory, and academic controversies surrounding the spread of discriminatory ideas online. While no single definition can suit every purpose, this chapter has identified shortcomings of current definitional approaches and sought to ameliorate some of them by defining hate speech as constitutive rhetoric. Such an addendum reintroduces critical identity perspectives to the academic community's understanding of hate speech, greatly improving our collective ability to understand its effects on marginalized communities, social networking media, and democratic societies. Furthermore, by calling explicit attention to hate speech's intrinsic constitutive effects, I have shown how this approach may facilitate improved understanding of the contextual nature of hate speech on social media, ameliorating research and regulation alike. By way of conclusion, I will outline three ways that non-rhetorical scholars might apply this definition in their own research.

First, scholars interested in improving our ability to detect hate speech on social media can use constitutive rhetoric as a flexible benchmark to determine what qualifies as hate speech and what does not, in concert with other attributes agreed upon in the literature. On the one hand, this definition expands our conception of "hate speech" by qualifying a variety of different signals through which social media interlocutors communicate the inferiority of others. On the other hand, this definition refines our approach by insisting that modes of expression which do not designate a specific other as inferior to the speaker should not be considered hate speech. By simultaneously expanding and restricting the definition of hate speech, this chapter sets forth an important challenge that researchers working with automatic

detection programs, and keyword-based detection in particular, should rise to meet. Such a task is daunting because it requires close attention to an expression's implicit and context-driven connotations. Fortunately, recent advancements indicate that refinement is possible and desirable (Mullah & Zainon, 2021).

Second, this definition emphasizes the importance of critical and humanistic inquiry to understand hate speech on social media, urging quantitative and qualitative scholars to consider critical approaches to identity when designing their experiments. Incorporating humanistic inquiry is crucial to fully understand and predict developments in the digital subcultures where many expressions of online hate speech gain initial purchase. This is illustrated by the predominance of critical race perspectives in the earliest studies of hate speech. The identities operationalized by hate speech are constructed through human discourse, not in programming environments. As such, it has been a regrettable development that critical communication perspectives have become sidelined over time. To resolve this problem, researchers need not develop encyclopedic understandings of rhetoric, race, or critical theory. Instead, they must reflect upon what their methods include as hate speech and what they exclude, attempting to situate those boundaries within the context of their object of study. Rather than seeking to eliminate all biases from one's understanding of hate speech, this approach requires scholars to think about hate speech as subjective rhetoric with objective effects, both qualities which ought to be rigorously measured.

Third and finally, the constitutive rhetorical approach calls scholars to understand hate speech's effects upon the whole public, not just upon the hate speaker's direct addressee. In line with the goal to consider hate speech's constitutive effects alongside its immediate consequences, researchers must recognize that communication does not occur in a vacuum isolated from the world in which it occurs. This insight is magnified when one attends to hyper-connectedness on digital media and the possibility for obscure posts to reach audiences that the original poster had not considered. With such an array of second personae available, it stands to reason

that hate speech on social media exacerbates violent social structures as much as it inflicts localized harm upon its immediate victim. By defining hate speech as constitutive rhetoric, scholars will emerge better equipped to name and analyze these important effects, even if and when they evade quantification. To do so will greatly improve our work's ability to describe the full impacts of hate speech. Yet most importantly of all, it will incline our research toward more effective strategies for regulation at all levels of governance, so that the international academic community may contribute bold proposals for the curtailment of hatred.

References

Alkiviadou, N. (2019). Hate speech on social media networks: Towards a regulatory framework? *Information & Communications Technology Law*, *28*(1), 19–35. https://doi.org/10.1080/13600834.2018.1494417

Aswad, E. (2016). The role of U.S. technology companies as enforcers of Europe's new Internet hate speech ban. *Columbia Human Rights Law Review Online*, *1*, 1–14. https://doi.org/10.2139/ssrn.2829175

Bebout, L. (2020). Weaponizing victimhood: Discourses of oppression and the maintenance of supremacy on the right. In *News on the Right* (pp. 64–83). Oxford University Press. https://doi.org/10.1093/oso/9780190913540.003.0004

Black, E. (1970). The second persona. *Quarterly Journal of Speech*, *56*(2), 109–119. https://doi.org/10.1080/00335637009382992

Black, E. (1978). *Rhetorical Criticism: A study in method*. Univ of Wisconsin Press.

Branaman, A. (1994). Reconsidering Kenneth Burke: His contributions to the identity controversy. *The Sociological Quarterly*, *35*(3), 443–455. https://doi.org/10.1111/j.1533-8525.1994.tb01738.x

Burke, K. (1951). Rhetoric—old and new. *The Journal of General Education*, *5*(3), 202–209.

Burke, K. (1973). The rhetorical situation. In L. Thayer (Ed.), *Communication: Ethical and moral issues*. Gordon and Breach Science Publishers.

Burke, K. & Zappen, J. P. (2006). On persuasion, identification, and dialectical symmetry. *Philosophy & Rhetoric*, *39*(4), 333–339.

Charland, M. (1987). Constitutive rhetoric: The case of the peuple québécois. *Quarterly Journal of Speech*, *73*(2), 133–150. https://doi.org/10.1080/00335638709383799

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), 512–515.

Delgado, R. (1982). Words that wound: A tort action for racial insults, epithets, and name-calling. *Harvard Civil Rights-Civil Liberties Law Review*, *17*(1), 133–182.

Delgado, R. & Stefancic, J. (2014). Hate speech in cyberspace. *Wake Forest Law Review*, *49*, 1–21.

Delgado, R. & Stefancic, J. (2018). *Must we defend Nazis?: Why the First Amendment should not protect hate speech and white supremacy* (Reprint edition). NYU Press.

DeLuca, K. M. (2005). *Image Politics: The New Rhetoric of Environmental Activism*. Routledge. https://doi.org/10.4324/9780203063088

Erjavec, K. & Kovačič, M. P. (2012). "You don't understand, this is a new war!" Analysis of hate speech in news web sites' comments. *Mass Communication and Society*, *15*(6), 899–920. https://doi.org/10.1080/15205436.2011.619679

Gelber, K. & McNamara, L. (2016). Evidencing the harms of hate speech. *Social Identities*, *22*(3), 324–341. https://doi.org/10.1080/13504630.2015.1128810

Goehring, C. & Dionisopoulos, G. N. (2013). Identification by antithesis: The Turner Diaries as constitutive rhetoric. *Southern Communication Journal*, *78*(5), 369–386. https://doi.org/10.1080/1041794X.2013.823456

Holling, M. A. & Moon, D. G. (2021). 20/20 in 2020?: Refractive vision, 45, and white supremacy. *Quarterly Journal of Speech*, *107*(4), 435–442. https://doi.org/10.1080/00335630.2021.1983195

Iorliam, A., Agber, S., Dzungwe, M. P., Kwaghtyo, D. K., & Bum, S. (2021). Comparative analysis of deep learning techniques for the classification of hate speech. *Nigerian Annals of Pure and Applied Sciences*, *4*(1), 103–108. https://doi.org/10.46912/napas.227

Konikoff, D. (2021). Gatekeepers of toxicity: Reconceptualizing Twitter's abuse and hate speech policies. *Policy & Internet*, *13*(4), 502–521. https://doi.org/10.1002/poi3.265

Leets, L. (2001). Responses to Internet hate sites: Is speech too free in cyberspace? *Communication Law and Policy*, *6*(2), 287–317. https://doi.org/10.1207/S15326926CLP0602_2

Leiter, B. (2012). The harm in hate speech [Review of the book *The harm in hate speech*, by J. Waldron]. *Notre Dame Philosophical Reviews,* https://ndpr.nd.edu/reviews/the-harm-in-hate-speech/

Matamoros-Fernández, A. & Farkas, J. (2021). Racism, hate speech, and social media: A systematic review and critique. *Television & New Media*, *22*(2), 205–224. https://doi.org/10.1177/1527476420982230

Mathew, B., Dutt, R., Goyal, P., & Mukherjee, A. (2019). Spread of hate speech in online social media. *Proceedings of the 10th ACM conference on web science* (pp. 173–182). https://doi.org/10.1145/3292522.3326034

Matsuda, M. J. (2018). Public response to racist speech: Considering the victim's story. In M. J. Matsuda, C. R. Lawrence III, R. Delgado, & K. W. Crenshaw (Eds.), *Words that wound: Critical race theory, assaultive speech, and the First Amendment*. Routledge.

Mills, R. E. (2014). The pirate and the sovereign: Negative identification and the constitutive rhetoric of the nation-state. *Rhetoric and Public Affairs*, *17*(1), 105–136. https://doi.org/10.14321/rhetpublaffa.17.1.0105

Mullah, N. S. & Zainon, W. M. N. W. (2021). Advances in machine learning algorithms for hate speech detection in social media: A review. *IEEE Access*, *9*(2021), 88364–88376. https://doi.org/10.1109/ACCESS.2021.3089515

Paul, J. (2021). "Because for us, as Europeans, it is only normal again when we are great again": Metapolitical whiteness and the normalization of white supremacist discourse in the wake of Trump. *Ethnic and Racial Studies*, *44*(13), 2328–2349. https://doi.org/10.1080/01419870.2021.1922730

Paz, M. A., Montero-Díaz, J., & Moreno-Delgado, A. (2020). Hate speech: A systematized review. *SAGE Open*, *10*(4), 1–12. https://doi.org/10.1177/2158244020973022

Putman, A. L. & Cole, K. L. (2020). All hail DNA: The constitutive rhetoric of AncestryDNA™ advertising. *Critical Studies in Media Communication*, *37*(3), 207–220. https://doi.org/10.1080/15295036.2020.1767796

Recommendation of the Committee of Ministers to Member States on "Hate Speech" (R (97) 20; pp. 106–108). (1997). Committee of Ministers of the Council of Europe. https://rm.coe.int/1680505d5b

Rieger, D., Kümpel, A. S., Wich, M., Kiening, T., & Groh, G. (2021). Assessing the extent and types of hate speech in fringe communities: A case study of alt-tight communities on 8chan, 4chan, and Reddit. *Social Media + Society*, *7*(4), 1–14. https://doi.org/10.1177/20563051211052906

Rogers, R. (2020). Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media. *European Journal of Communication*, *35*(3), 213–229. https://doi.org/10.1177/0267323120922066

Ruwandika, N. D. T. & Weerasinghe, A. R. (2018). Identification of hate speech in social media. *2018 18th international conference on advances in ICT for emerging regions*, 273–278. https://doi.org/10.1109/ICTER.2018.8615517

Sellars, A. (2016). Defining hate speech [Berkman Klein Center Research Publication No. 2016-20]. Boston University School of Law. https://papers.ssrn.com/abstract=2882244

Sharples, R. & Blair, K. (2021). Claiming 'anti-white racism' in Australia: Victimhood, identity, and privilege. *Journal of Sociology*, *57*(3), 559–576. https://doi.org/10.1177/1440783320934184

Stewart, C. J., Smith, C. A., & R. E. D., Jr. (2012). *Persuasion and social movements* (6th ed.). Waveland Press.

Topinka, R. (2019). "Back to a past that was futuristic": The Alt-Right and the uncanny form of racism. *B2o: An Online Journal*. https://www.boundary2.org/2019/10/robert-topinka-back-to-a-past-that-was-futuristic-the-alt-right-and-the-uncanny-form-of-racism/

Tuters, M. & Hagen, S. (2020). (((They))) rule: Memetic antagonism and nebulous othering on 4chan. *New Media & Society*, *22*(12), 2218–2237. https://doi.org/10.1177/1461444819888746

Twitter. (2022). *Twitter's policy on hateful conduct*. Retrieved August 24, 2022, from https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

Waldron, J. (2012). *The harm in hate speech*. Harvard University Press. https://doi.org/10.4159/harvard.9780674065086

Waseem, Z. & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. *Proceedings of NAACL-HLT 2016* (pp. 88–93).

Whine, M. (2016). National monitoring of hate crime in Europe: The case for a European level policy. In J. Schweppe & M. A. Walters (Eds.), *The globalization of hate: Internationalising hate crime?* Oxford University Press.

Zhang, Z. & Luo, L. (2019). Hate speech detection: A solved problem? The challenging case of long tail on Twitter. *Semantic Web*, *10*(5), 925–945. https://doi.org/10.3233/SW-180338

MONITORING HATE SPEECH AGAINST IMMIGRANTS IN SOCIAL MEDIA: A TAXONOMY AND A GUIDE TO DETECT IT

Berta Chulvi Ferriols
/ Universitat Politècnica de València and Universitat de València, Spain

Paolo Rosso
/ Universitat Politècnica de València

Karoline Fernandez De La Hoz Zeitler
/ Ministry of Inclusion, Social Security and Migrations of the Spanish Government, Spain

Introduction

As everybody knows, social media are being exploited as platforms for intolerance, mainly for the diffusion of hate speech. Some authors speak about a Hate Speech Epidemic that leads to political radicalization and deteriorates intergroup relations (Bilewicz & Soral, 2020). According to ECRI General Policy Recommendation No. 15, hate speech is based on the unjustified assumption that a person or a group of persons are superior to others; it incites acts of violence or discrimination thus undermining respect for minority groups and damaging social cohesion. This discrimination is based on a non-exhaustive list of personal characteristics or status including race, colour, language, religion or belief, nationality or national or ethnic origin as well as ancestry, age, disability, sex, gender, gender identity and sexual orientation.

Fighting the spread of hate speech in digital communication is a central concern for the United Nations[1], the Council of Europe[2] and to the European Commission[3]. These organisations have developed plans to contain its diffusion, and several national governments have put in place similar initiatives. The Spanish government is no exception and has promoted a protocol to combat illegal hate speech online. In March 2021, the Spanish Observatory on Racism and Xenophobia[4] (OBERAXE) published a protocol to combat illegal hate speech on media platforms. In addition, OBERAXE has a program to monitor the main social platforms daily to report the messages that express distinct expressions of hate against immigrants as they appear.

This chapter provides the results of a collaboration between the PRHLT Research Center[5] of the Universitat Politècnica de València and OBERAXE. The aim of this collaboration was twofold, first to elaborate a guide for monitoring hate speech that offers clear criteria to classify hate speech messages on social media and secondly, to develop an App to facilitate the task of monitoring hate speech on different social media platforms. The PRHLT Research Center has developed a line of research in Natural Language Processing (NLP) dedicated to automatically detecting hate speech which will also be briefly explained in this chapter.

In summary, in Section 1, we will present the aspects of hate speech that characterize the expression of prejudice in digital societies. In Section 2 we will revise the efforts made by computational social science to detect it automatically. In Section 3 we will give some insights on the performance of platforms against hate speech. In Section 4, we will clarify the concept from the point of view of a practitioner who must monitor hate speech messages against immigrants, and we will propose a taxonomy

1. https://www.un.org/en/hate-speech
2. https://www.coe.int/en/web/no-hate-campaign/coe-work-on-hate-speech
3. https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en
4. https://www.inclusion.gob.es/oberaxe/es/index.htm
5. https://www.prhlt.upv.es/

to classify them. In Section 5 we end with a practical guide for monitoring hate speech and with some insights for the future of monitoring this harmful communication.

Hate speech in social media: The new face of prejudice

Quantifying just how much hate speech circulates every day on social media is a difficult task. Estimations of percentages of hate messages circulating on social media are varied. Some authors report a low percentage of hate messages, around 1% of tweets posted in a given period (Pereira-Kohatsu et al., 2019) while other research estimates the number of hate messages to be between 10% and 15% of the total (Waseem, 2016). Recently, Carvalho et al. (2022) created a dataset with hate speech posts against people of African descent, Roma and LGBTQ+ Communities in Portugal collected from Twitter from the 1st of August 2018 to the 31st of October 2021. Their data allow us to estimate that nearly 18% of messages that mention these target groups contain offensive language or hate against these minorities.

Practically all of us have had contact with some hateful message and it is clear that hate speech is the new face of classical prejudice (Allport, 1954, Duckitt, 1994). As Allport explains in his seminal work, the verbal expression of prejudice, which he calls "antilocution", is the first step on a 5-point continuum: antilocutions, avoidance, discrimination, physical attack, and extermination. That is not an automatic process but is a series of steps. One phenomenon precedes the other. As Allport claims "it was Hitler's antilocution that led Germans to avoid their Jewish neighbours and erstwhile friends. This preparation made it easier to enact the Nürenberg laws of discrimination which, in turn, made the subsequent burning of synagogues and street attacks upon Jews seem natural. The final step in the macabre progression was the ovens at Auschwitz" (Allport, 1954: 15). Then, we can conclude that hate speech is not a new phenomenon but an old problem with a new nature.

The aspects that make hate speech a phenomenon with a new nature are related not to the content of the messages but to the communicational ecosystem in which it has been developed. This new communicational context has been defined by some authors as "digital societies" (Mossberger et al., 2008). These digital societies not only amplify classical prejudice but transform it, adding new features which include multiple senders outside the traditional elites (1), the anonymity of communication (2), and the disappearance of physical distances in everyday interaction among people who don't know one another (3).

Compared to mass communication societies, in digital societies, the democratisation of the ability to communicate with a huge audience is a fact: anyone has the possibility of creating a message for mass reception. In some sense, the concept of receivers has disappeared. Everyone in digital societies is required to act: to like, comment, post, retweet, etc (Pisani & Piotet, 2009). Recent research (Miranda et al., 2022) focusing on hate speech diffusion, shows that regarding the behavior of the audience, users prefer to place "likes" than to comment on posts or even share them with others. The interaction with the content is marked from a first -very basic- level. Moreover, the multiplication of the number of senders who are able to reach a massive audience without belonging to a social, cultural or political elite is crucial to understand the current spreading of hate speech.

The second characteristic of these digital societies is the intensive use of anonymous communication. Before the establishment of digital societies, the status of opinion leader was a privilege held by an elite group of people and only those with a recognisable identity could benefit from having access to a mass audience. The conjunction of anonymity and a massive audience in messages that circulate on social platforms has enormous consequences in terms of the dissolution of accountability. That is why in 2018, the Spanish Attorney General called for the regulation of anonymity on social media[6], and in October 2020 the French government made the same proposal[7].

6. https://www.eldiario.es/politica/fiscal-general-regular-anonimato-sociales_1_2241317.html
7. https://cadenaser.com/programa/2020/10/20/hora_25/1603219631_507670.html

An additional problem to the dissolution of accountability is the extension of the false consensus effect (Ross, et al. 1977): the illusion that minority opinions shared in a small community are consensual. There have always been political groups that have instrumentalised minorities for their own benefit, but their spokespersons were clearly identifiable. It was more or less evident that they were acting with the objective of gaining power. Nowadays, with anonymous users spreading hate speech, this ideological discourse that still serves a particular interest seems to be spontaneous and supported by ordinary citizens with no special interests. The idea that the minority is a threat to the majority group is spread by these repetitive messages and seems to justify prejudice. However, recent research maintains that the process is just the opposite: it is prejudice that gives rise to the sense of threat, not the other way round. When people feel prejudice toward a group, they can justify their prejudice by perceiving the group as threatening (Bahns, 2017, Perez et. al, 2022).

The third characteristic of these digital societies in relation to the increasing hate speech episodes is that they allow unknown people to interact without physical proximity and without any physical contact. The disappearance of physical distances as a condition for social relations makes it easier to connect radical individuals that would have been isolated some decades ago. This connection created between extremist minorities gives an appearance of normality to these manifestations of hate. Once geographical boundaries have disappeared as a condition for human interaction, it no longer matters if the in-group that legitimises hatred is located in the same neighbourhood or scattered all over the planet (Kaufman, 2015). Social media have created not only a new public sphere but also a new private life (Zafra, 2012). In social networks it is possible to develop a double life without reputational costs. Internet users can consume hateful content without fear of being the objects of social judgement. People only need to create an anonymous profile to access hateful content without leaving a trace.

Automatic hate speech identification

Given the enormous amount of user-generated content, the problem of identifying and, if possible, counteracting the spread of hate speech on the web and, in particular, on social media, is becoming a fundamental aspect of the fight against xenophobia. Hate speech identification has been studied with different strategies ranging from linguistic feature-based methods to machine learning techniques (Fortuna & Nunes 2018, Poletto et al., 2021). In most cases, the research makes use of traditional machine learning models, such as logistic regression (Waseem & Hovy, 2016). Some approaches have utilised external resources such as dictionaries and lexical repertoires. (MacAvaney et al., 2019). Recently, many systems are based on deep learning models, such as recurrent neural networks (Magalhaes, 2019) or BERT (Samghabadi et al., 2020).

Most approaches try to identify whether or not a text contains hate speech and a few works focus on the identification of hate speech at the user level. Mathew et al. (2019) analysed the profiles and networks of haters and non-haters by focusing on the dynamics of haters' dissemination and observed that content containing hate speech spreads more widely and more quickly than regular messages. ElSherief et al. (2018) compared users who spread hate speech on Twitter with those who receive their attacks based on their profile, activities and visibility on the networks. Their results suggest that users who disseminate hate content are more popular and that participating in the dissemination of hate speech can lead to greater visibility on social networks. Ribeiro et al. (2018) focused on profiling hate speech disseminators on Twitter using a methodology to obtain a graph from the full profiles of users and then investigated the difference between hate speech and non-hate speech spreaders in terms of activity patterns, words usage, and network structure. The authors observed that haters are tightly connected so they focused on exploiting the network of connections. Hagen *et al*. (2019) studied the use of emojis in white nationalist conversations on Twitter and observed the difference between pro and anti-stance.

In recent years, there have been a number of shared tasks for different languages (e.g., English, German, Hindi, Italian, Mexican-Spanish and Spanish) focusing on hate speech and issues related to online violence, reflecting the interest in addressing this important problem. For example, the first workshop on *Trolling, Aggression and Cyberbullying* (Kumar *et al*. 2018), which also included a task on the identification of aggression. The task about *Authorship and Aggressiveness Analysis* (MEX-A3T) (Carmona *et al*. 2018); the *Automatic Misogyny Identification* Task (Fersini *et al*., 2018a, 2018b, 2020, 2022); the one task of GermEval about *Identification of Offensive Language* (Wiegand *et al*. 2018; Struß *et al*. 2019); the *Hate Speech Detection tasks* (HaSpeeDe) in EVALITA (Bosco *et al*. 2018, Sanguinetti *et al*. 2020); OffenSeval (Zampieri *et al*. 2019, 2020) y *HatEval* (Basile *et al*. 2019) in SemEval; the task *HASOC* in FIRE about *HAte Speech and Offensive Content identification in Indo-European languages* (Modha *et al*. 2019, Mandl *et al*. 2021); and recently DETOXIS in IberLEF on *DEtection of TOXicity in comments In Spanish* about migrant people (Taulé *et al*. 2021) and DETEST, DETEction and classification of racial STereotypes in Spanish (Ariza-Casabona et al. 2022). Finally, in the PAN lab a shared work task on profiling hate speech spreaders on Twitter has been organised (Rangel *et al*. 2021).

Recently much effort has been made to improve the way research communities classify hate speech. For example, Kennedy et al. (2022) have developed the Gab Hate Corpus (GHC), consisting of 27,665 posts from the social network service gab.com, annotated for the presence of "hatebased rhetoric". Other studies on hate speech in social media include Müller and Schwarz (2020), which analyses the ways in which prejudice manifests itself in violence by means of a temporal analysis of anti-refugee activity on Facebook. An examination of the relationship between moral homogeneity in an online social network and the rate of posting hate speech has been carried out by Atari et al. (2022). Mathew et al. (2020) have also analysed the temporal and network structure of hate speech in an online social network.

Despite the efforts made in other languages, most of the research about hate speech has been conducted in English and is necessary to fill the gap

that exists in other languages, as hate speech against certain minorities is usually a local phenomenon. The spread of harmful stereotypes tends to be deeply rooted in a specific geographical location and historical context. For this reason hate speech monitoring carried out by different organisations at national level, as the case of OBERAXE, has a double effect: on the one hand, allows counteracting hate speech by forcing platforms to react to its denunciations, and, on the other hand, it allows accumulating very valuable textual information in native languages both for computational linguistics teams and for social science research.

The performance of platforms combating hate speech

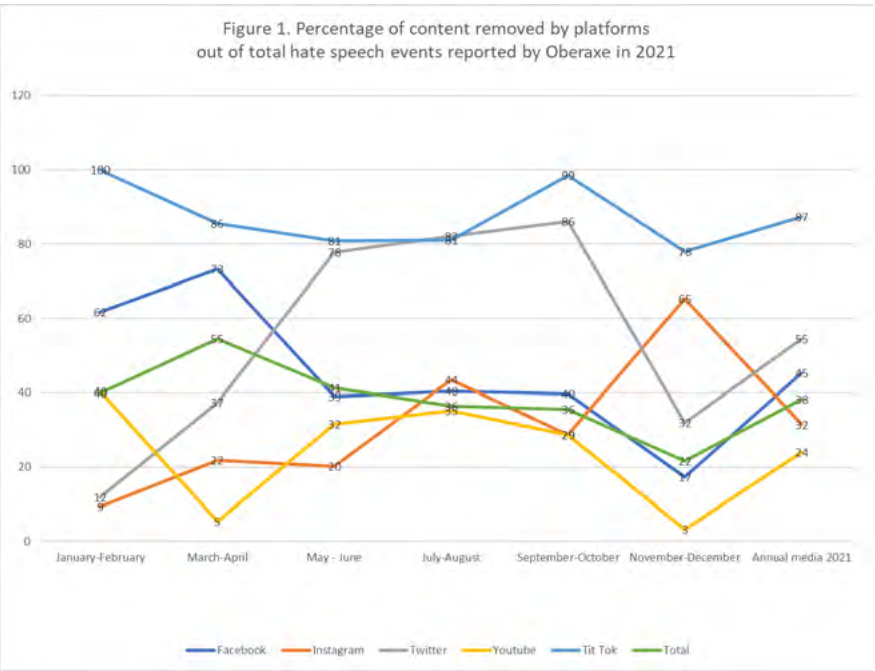
Platforms are no strangers to the problem of hate speech online spreading. Concerned about their public image and the responsibility they could bear for the circulation of such speech, the main platforms have issued their own rules to prevent and sanction hate speech, signing codes of conduct and collaboration protocols with public administrations. In May 2016, the European Commission[8] agreed with Facebook, Microsoft, Twitter and YouTube on a "Code of Conduct on Combating Illegal Online Hate Speech.". Rakuten, ViberSearch, Instagram, Snapchat, Dailymotion, Jeuxvideo, TikTok, and LinkedIn have also signed this Code of Conduct. The results of the sixth evaluation of the Code of Conduct on countering illegal hate speech online in 2021, show a mixed picture: social media platforms remove an average of 62.5% of flagged content. These results are lower than the average of 71% withdrawn in 2020.

In Spain, in July 2020, the main platforms adhered to the Protocol to Combat Illegal Hate Speech Online[9] promoted by OBERAXE. All platforms have a reporting mechanism that can be used by users to request the removal of content that violates the country's legal system or the platforms' terms of service. In addition, under the aforementioned protocol, some

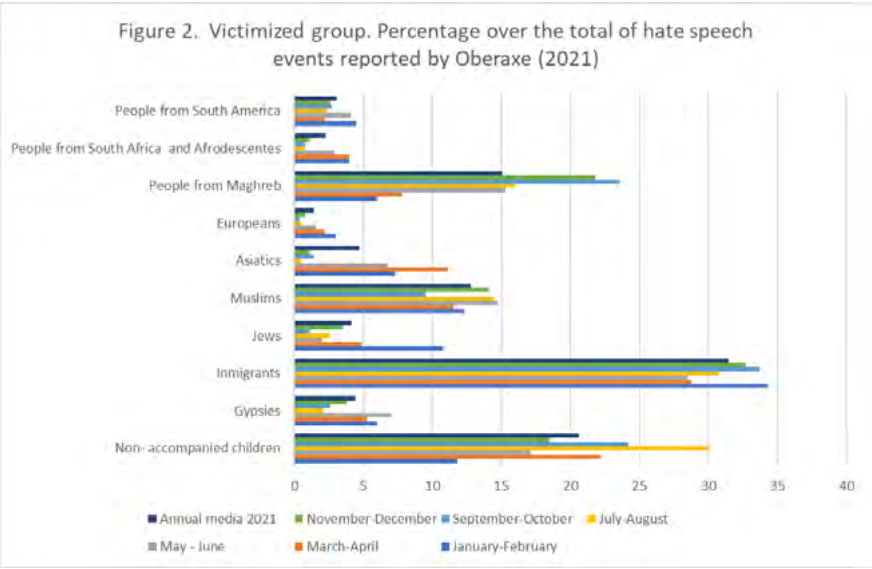
8. https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en
9. http://www.interior.gob.es/documents/642012/0/protocolo-discurso-odio/357cb9d2-e254-4303-a9bb-18b0027e4a42

organisations, including OBERAXE, are recognized as trusted flaggers and their requests receive preferential attention from the platforms.

As we have already mentioned, OBERAXE has a program to monitor daily the main social platforms to report messages that express hate against immigrants and other ethnic or national minorities. An analysis of the action taken by the platforms against the 3,378 hate speech events reported by OBERAXE in 2021 shows a removal rate of around 38% percent. This rate can vary considerably in a yearly period. In March and April 2021 it reached 55%, whereas in November and December of the same year it fell to 22%. Figure 1 shows a substantial disparity in the behaviour of the platforms in terms of the percentage of content removed in a year. Only Tik Tok shows homogeneous behaviour throughout the year with the highest rate of content removed. The erratic evolution of most of these platforms does not allow us to infer a process of progressive sensitization of these data hosting services.

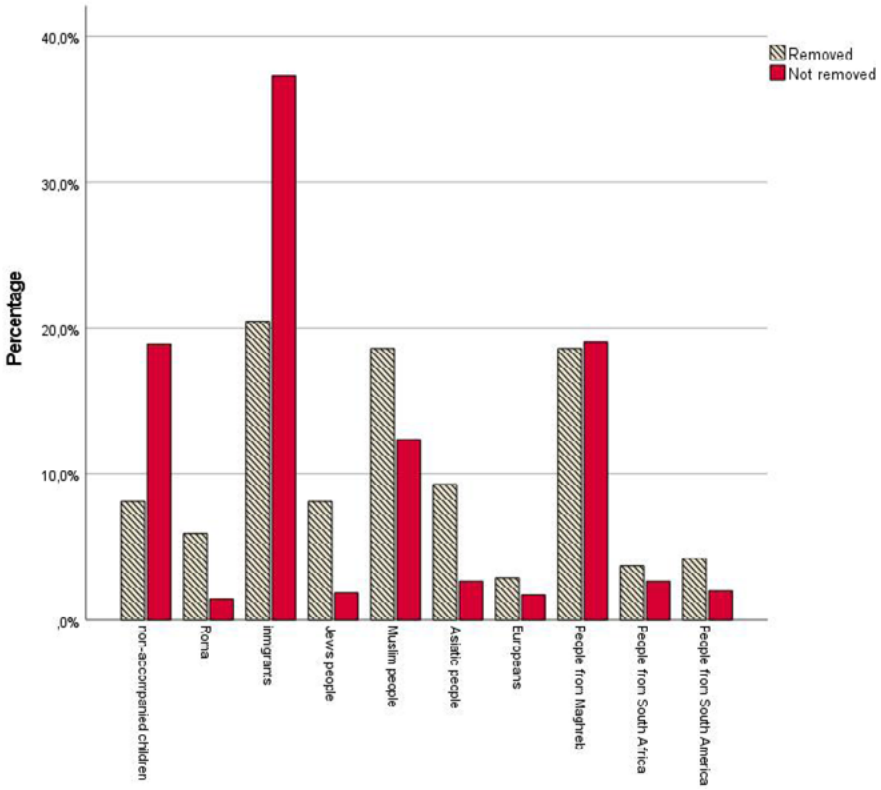


An analysis of the victimised groups in 3,378 hate speech posts shows that the social category of "immigrants", mentioned generically, receives the highest number of attacks (see Figure 2). As we can observe, some groups are more victimised in specific moments of the year. For example, Jews in January-February or people from Asia in March-April, possibly due to particular news items that focus on different minorities each time.



Further analysis of victimised groups and the content removed in 1,342 posts on three major platforms (Facebook, Twitter and Instagram) shows that platforms do not react in the same way to all victimized groups. Results confirm a statistically significant relation (Pearson$\chi 2$=148,948, df=9, $p<0.001$) between two variables, the group victimized and the fact of content being removed or not. As can be seen in Figure 3, hate speech posts have less probability of being removed if the victimized group is mentioned as "immigrant" generically, or as "non-accompanied children" (Menas, in Spanish). However, there is a greater probability that a hate speech post will be removed if the victimised group is referred as Jews, Roma, Asians, Muslims, or Latin Americans, in that order.

Figure 3. Platform's reaction to reported hate speech by group victimized



A new phenomenon related to the moderation of hate speech in mainstream platforms is the growth of alternative social platforms such as Gab and Parler, branded as 'free speech'. Touting their commitment to 'free speech' and 'no moderation' policy, Gab and Parler attract users that had been suspended from mainstream platforms (Israeli & Tsur, 2022)

A Taxonomy and a guide to monitoring hate speech in social media

Before developing a taxonomy of hate speech episodes, it is necessary to offer clear criteria on what hate speech is and what it is not. When we speak about hate speech in academic research everything seems to be more or

less clear. Controversy is centred on the ongoing debate between those who want to control hate speech and those who advocate for free speech (Gelber, 2002). The task of labelling and classifying concrete manifestations of hate speech is, however, much more difficult. A good orientation to address this task is to analyse each speech act by distinguishing three elements: who is being talked about, what is being said, and how it is being talked about.

Who is being talked about? A message that can be qualified as hate speech is one which is directed at a group of people or at an individual for being members of a regularly discriminated group. It is a speech act targeted to disadvantaged social groups in a harmful way (Jacobs & Potter, 1998). It is a message that refers to a group that has historically suffered, or currently suffers, a situation of discrimination, oppression or vulnerability, for example, minorities such as migrants, Roma, LGBTi+, Jews, people with disabilities, homeless people or religious minorities, etc. Hate speech is also that speech containing a sexist or misogynist component. This one has the effect of deepening the discrimination of women and it also has historical roots. The reason to limit the concept of hate speech to the acts against groups regularly discriminated is that a speaker who attacks a minority group is benefiting from an already existing discriminatory ideology that openly supports its exclusing intolerance towards certain groups. The pre-existence of such a discriminatory ideology is a cultural factor that generates an asymmetry between the aggressor and the minority benefiting the first one.

Consequently, any group can be presented as homogeneous in an argument. For example, when it is stated that "men are all the same", referring to "human males", we are dealing with a stereotype (Mackie & Hamilton, 1993), but not all groups have a history of discrimination or a reality of discrimination behind them. Only when the message refers to a minority that has historically suffered discrimination can be qualified as "hate speech". It is precisely for this reason that recently, the Spanish Supreme Court (ATS 10-01-2022) did not admit a complaint of the far-right party against a leftist Spanish politician who called them "open-faced Nazis" at a rally. The

Supreme Court points out that "it may be understood that such expressions are contrary to the due respect that should be given to the different political parties in electoral confrontation, but they do not constitute hate crime".

Being a persecuted minority is a factor closely linked to the historical moment and the context in which the speech act takes place. This is very clear when we talk about political or religious minorities. For example, Catholics are a religious minority in the United Kingdom and a majority in Spain. That is why any taxonomy or guide for monitoring hate speech is strongly linked to the socio-political and economic system in which the hate speech takes place. In any case, hate speech is always an attack to the fundamental rights of the persons to whom it is directed. Specifically, hate speech is an attack to the dignity of the person and contravenes the right to equality and non-discrimination that is at the basis of democratic states.

What is said? In its content, hate speech involves advocacy, promotion or instigation to hatred, humiliation or disparagement. It can include a variety of behaviours: threats, harassment, discrediting, disparagement or dissemination of negative stereotypes. It is characterized by the fact that rather than communicating an idea, its *raison d'être* lies in the effect it produces: that is, causing harm to a whole group (Kaufman, 2015). We speak of effect (and not motivation) because hate speech can be presented as a clearly voluntary act (an argument) or as an expression that might seem involuntary (an insult resulting from an argument). Regarding the criminal procedure, the discriminatory motivation is important, however, concerning the monitoring of hate speech in social networks, which aims at reducing the spread of racism and xenophobia online, the intention is less relevant than the fact itself. Regardless the speaker's intentions, the fact is that the message that spreads hateful content towards a particular minority is on the internet, accessible to anyone, and will remain there unless it is removed by the platform.

It is important to inquire about the effects of discourse when categorizing the content of a message as hate speech. To be operational, the diversity of effects that hate speech has can be arranged into two big categories: either

they pose a threat to the physical or psychological integrity of a person or group (incitement to violence) or they involve discrimination incompatible with the principle of equality enshrined in our legal system (incitement to discrimination). Therefore, speech that is dehumanising or seriously degrading is considered to be incitement to violence as it is the prelude, and the justification of a violent act.

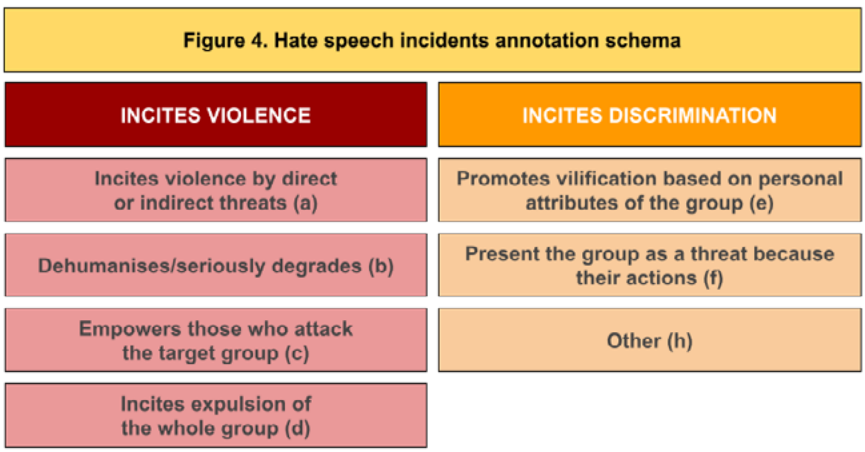
How is it expressed? Hate speech is a public manifestation, which is delivered in a physical or virtual public space. That is why it may involve incitement to actions of other people and it may reach recipients other than those to whom it was originally intended. When someone performs an act of communication in a virtual public space, he or she launches a message that may have both intended and unintended consequences. It will depend on the circumstances in which the audience receives each particular message. Human beings are responsible not only for their actions but also for the consequences of their actions, and incitement to hatred is considered a crime. It is in this sense that jurists speak of hate speech crimes as a paradigmatic example of "crimes of danger or climate", i.e. crimes that encourage, for example, the commission of other hate crimes that are not hate speech.

A hate message can be any communicative action, such as symbols (the swastika, for instance), images, videos, gestures, etc. Currently, the use of some emoticons has to be considered hate speech as far as they are equivalent to concepts and emotions expressed through words. IOS and Android originally support 845 emojis including options such as hearts/love symbols, stars, signs and animals.

In terms of textual resources, hate speech may use explicit aggressive language containing insults or threats ("these moors are pigs") or explicit but non-aggressive discriminatory language ("these immigrants do not deserve the medical care provided by the Spanish healthcare system") as recently has been shown by Sanchez et al. (2021) or ironic or sarcastic language as in the work of Frenda et al. (2022).

A taxonomy for hate speech

In Figure 4 a general binary classification of hate speech is proposed. The two principal categories distinguish between (1) hate that incites violence and (2) hate that incites discrimination. These two categories are consistent with the ones used by Kennedy et al. (2022), that distinguish between "calls for violence" and "assaults on human dignity". They are also supported by the two-class specification given by UNESCO for hate speech: (a) "expressions that advocate incitement to harm (particularly, discrimination, hostility or violence) based upon the target being identified with a certain social or demographic group" and (b) "expressions that foster a climate of prejudice and intolerance on the assumption that this may fuel targeted discrimination, hostility and violent acts" (Gagliardone et al., 2015: 10). Further subcategories are identified in each of these two large groups. Each of these subcategories is explained below with examples of hate messages from social media posts reported by OBERAXE in 2021.



Amongst hate speech that incites violence, we can distinguish four subcategories: inciting violence, dehumanisation, empowering of violent groups, and open calls to the expulsion of the minority. Some examples are provided below.

a. Hate speech that incites violence through direct or indirect threats. In this subcategory, we can find messages that attack people's physical integrity. The good to be protected is the humans being life. Some examples are:

- *"Boats (pateras in Spanish) that sink until you could use them as stepping stones".*
- *"Tie a block of concrete to his foot and let him swim back to his fucking country".*
- *"Do you want to go bang bang to the menas* (accompanied by a picture of a holstered revolver)

b. Hate speech that seriously dehumanizes and degrades its targets. A person's human nature is denied, the minority is said to resemble animals or waste. As the victims are dehumanized, their lives are not considered to have any value. For that reason this subcategory is considered included in the category of "hate speech that incites violence". The good to be protected is the human dignity attributable to anyone. Some messages taken from social media that fall into this category are the following:

- Accompanying a link to an article that reports on the finding of tortured cats in a cemetery, the social media user says: *"Mr Fernández Díaz, do us all a favour and get rid of this deluge of shit, for God's sake. Fucking South American and African immigration with their voodoo rituals and this awful satanic culture in Europe (y su puta madre que los parió)".*
- *"I think it's great that the boat full of African immigrants sank. That way they won't contaminate Europe any more by filling it with black muslims".*
- *"Let tanks roll over these fucking rats put here by these fuckers from Podemos and that dirty bitch Colau".*

c. Hate speech that empowers or praises those who attack a target group. These messages are intended to attack not only people's lives but also public order. These messages promote ideologies that endorse exclusing intolerance.

- *"Each day that goes by I feel more like this #Menas* (Unaccompanied Immigrant Minors) *Out"* (shown together with a picture of a shirtless young man with a swastika).

- *"Gipsies and Moros* (people from North Africa)*, races that deserve to be annihilated".*
- "Your duty as Amatsu's chosen one is to wipe out your enemies and that's what gypsies are. May the eternal glory of Amatsu guide your sacred sword and allow you to exterminate the unbelievers".

d. Hate speech that promotes the expulsion of a whole group. These are messages that promote the expulsion of a group from the country. Expulsion is considered to be an act of violence and therefore these messages are included in the category of hate speech that incites violence.

Among hate speech that incites discrimination, we distinguish two subcategories:

e. Hate speech that promotes vilification based on a group's personal attributes or based only on their belonging to the group. The good to be protected is the right to equality and non-discrimination. The messages included under this subcategory stigmatise a minority without mentioning a single aspect of their behaviour, only for being who they are. The following are some examples:

- "I don't understand why Muslims are allowed into Spain; don't forget that they're like cancer to western societies".
- "They want to replace us Europeans for a race that is worse and backward in every way, mixed-race and horrendous, really bad and bloodthirsty, this socio-communist scum... trying to brainwash everybody... lying".
- "Fucking Chinese and their fucking virus".
- "*When they are menas* (Unaccompanied Immigrant Minors) *they're delinquents and when they're ex menas* (adults) *they're fucking moros* (North Africans)".

f. Hate speech that presents a group as a threat because of its actions, and therefore describes a type of action that is collectively associated with the group. A particular behaviour that threatens the host society is attributed

to the group in question. The good to be protected is the right to equality and non-discrimination. Some examples are following:

> - "The difference is that this man integrated, the menas haven't come here to integrate, they've come here to do nothing and live a life of crime, and this scum, me at least, I don't want them in my country, they can call me whatever the fuck they want..."
> - "The menas have come here to rob, rape, kill gays and claim benefits".

The taxonomy we have suggested is in itself a scale of prejudice in descending order from greater to lesser gravity of consequences. For this reason, it is recommended that if a specific hate message falls into two subcategories, it should be classified according to the more serious one. Each research or monitoring team can decide to use the classification as multiple-choice questions or not.

Some insights for the future of monitoring hate speech and a guide

There are powerful reasons to monitor hate speech. Along the exposition to hate speech, people's sensitivity to hateful language diminishes. The more hate speech people observe in their environment, the less emotional arousal they tend to feel (Bilewicz & Soral, 2020). Moreover, frequent exposure to hate speech changes the dominant image of the outgroup targeted by such language. Being frequent targets of hate speech, minority group members become increasingly viewed as inferior to one's ingroup due to a system of justification mechanisms (Jost, 2019). Additionally, the monitoring activity can produce high-quality data to gain a better understanding of the phenomenon and acquire the means to prevent it. In order to meet both requirements, it is essential to consider the data collection process in all its complexity and for this purpose we offer a guide in seven steps.

As a starting recommendation when monitoring hate speech in social media, we suggest registering basic data such as the date and time of posting, the platform and the URL of the episode in question. As far as the characterization of a said episode we suggest collecting answers to the following questions:

Step 1. Is the victim an individual or a group?
This is a question to indicate whether a whole group is being attacked or a particular individual is under attack for belonging to a specific group. Hate incidents that refer to an individual can be the result of personal animosity or momentary resentment, whereas incidents where a group is named generically, and no specific references are made show a much clearer desire to stigmatise a minority.

Step 2. Who is the target group of hate speech? Which other groups are mentioned?
We suggest considering the possibility to name the group generically, with generic categories (immigrants, foreigners, etc.) and specifically (the Chinese, the Rumanians, blacks). In the same way, we consider taking the opportunity to take note of other groups being mentioned alongside (racialised minorities or victims of xenophobia) or whether other non-victimised groups are mentioned (for example, some political party).

Step 3. Is any sign of the sex/gender of the victim mentioned in the hate speech incident?
Try to identify double discrimination such as origin and gender.

Step 4. Does the post or comment refer to a prototypical episode?
The objective of this question is to see how the problem of immigration is framed in reference to a specific episode. In the Spanish context certain episodes, such as people climbing over the border fence of Ceuta, or the arrival of boats carrying immigrants into the country's shores can be defined as prototypical events. Other such episodes, in which hate speech shows up, might be events that affect public security, terrorism, pandemics, health issues or armed conflicts. In addition, other events in a country's life such as elections, actions taken by the government to protect more vulnerable sectors of the population or some other events related to the economy, for example, the publication of unemployment figures, could be added to this list. It is also useful to note whether the hate speech episode appears to be unconnected to a particular event and therefore represents an expression of profound animosity as no further justification is needed.

Step 5. What is the content of the hate message?
We suggest applying the taxonomy already described in section 4.

Step 6. What kind of language is used in the hate episode? Are images, videos or memes included?
We suggest a simple classification: (1) explicit aggressive language that includes insults and other aggressive expressions; (2) discriminatory non-aggressive language; (3) the use of humour or; (4) irony or sarcasm. Recently, Merlo (2022) has studied the degree of offensiveness in humour and she found that one of the elements that distinguish offensive humour from not offensive humour is a greater presence of social categories referring to ethnic minorities in the first.

Step 7. What is the platform's reaction?
If in addition to monitoring also reporting the hate content to the specific platform is done, we recommend recording the date that the content is reported and the platform's reaction (if the content is removed after 24 hours, 48 hours or a week of the notification). Some platforms will respond by explaining why they are not removing the content and when such a response occurs it is important to record it.

Last but not least, we need to take into account that although moderation is important, it does not fix the problem, as Lubin and Gilbert pointed out recently in the MIT Technology Review[10]. As the authors stand moderation can potentially work for harms directly caused by particular pieces of content, but this content also produces "structural" effects in society. Issues such as discrimination, worsening mental health and a decline in civic trust manifest themselves in many ways across the social media environment rather than through any concrete pieces of content. That is why it is necessary to have a deep understanding of the phenomenon, devote interdisciplinary efforts to analysing it, and propose comprehensive strategies to lead our societies on the path of harmony rather than hate.

10. https://www.technologyreview.com/2022/08/09/1057171/social-media-polluting-society-moderation-alone-wont-fix-the-problem/

Acknowledgements

The work of Berta Chulvi and Paolo Rosso was done in the framework of the research project: FAKEnHATE-PdC FAKE news and HATE speech (Grant PDC2022-133118-I00) funded by MCIN/AEI/10.13039/501100011033 and by European Union NextGenerationEU/PRTR.

References

Allport, G. (1954). *The nature of Prejudice* Addison-Wesley. Cambridge

Ariza-Casabona, A., Schmeisser-Nieto, W. S., Nofre, M., Taulé, M., Amigó, E., Chulvi, B., & Rosso, P. (2022). Overview of DETESTS at IberLEF 2022: DETEction and classification of racial STereotypes in Spanish. *Procesamiento del Lenguaje Natural*, 69, 217-228.

Atari, M., Davani, A. M., Kogon, D., Kennedy, B., Ani Saxena, N., Anderson, I., & Dehghani, M. (2022). Morally homogeneous networks and radicalism. *Social Psychological and Personality Science*, *13*(6), 999–1009. https://doi.org/10.1177/19485506211059329

Bahns, A. J. (2017). Threat as justification of prejudice. *Group Processes & Intergroup Relations*, *20*(1), 52–74. https://doi.org/10.1177/1368430215591042

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti,V., Rangel, F., Rosso, P., & Sanguinetti, M. (2019). SemEval-2019 Task 5: Multilingual detection of hate speech against immigrants and women in Twitter. *Proceedings of the 13th Int. Workshop on Semantic Evaluation (SemEval-2019)*, co-located with the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), 54-63.

Bilewicz, M. & Soral, W. (2020). Hate speech epidemic. The dynamic effects of derogatory language on intergroup relations and political radicalization. *Advances in Political Psychology*, *41*(1), 3-33. https://doi.org/10.1111/pops.12670

Bosco, C., Dell'Orletta, F., Poletto, F., Sanguinetti, M., & Tesconi, M. (2018). Overview of the EVALITA 2018 hate speech detection task. *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, CEUR-WS.org., 67-74.

Carmona, M. A., Guzmán-Falcón, E., Montes, M., Escalante, H. J., Villaseñor-Pineda, L., Reyes-Meza, V., & Rico-Sulayes, A. (2018). Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in Mexican Spanish tweets. *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages*, pp. 74-96. CEUR-WS.org. https://ceur-ws.org/Vol-2150/

Carvalho, P., Matos, B., Santos, R., Batista, F., & Ribeiro, R. (2022). Hate speech dynamics against African descent, Roma and LGBTQ+ communities in Portugal. *Proceedings of the 13th Conference on Language Resources and Evaluation* (LREC 2022), 2362–2370.

Duckitt, J. (1994). *The social psychology of prejudice*. Praeger.

ElSherief, M., Nilizadeh, S., Nguyen,D., Vigna, G., & Belding, E. (2018). Peer to peer hate: Hate speech instigators and their targets. *Proceedings of the International AAAI Conference on Web and Social Media,* 12.

Fersini, E., Rosso. P., & Anzovino, M. (2018a). Overview of the task on automatic misogyny identification at IberEval 2018. *Proceedings of 3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages*, 57–64.

Fersini, E., Nozza, D., & Rosso, P. (2018b). Overview of the evalita 2018 task on automatic misogyny identification (ami). *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian*.

Fersini, E., Nozza, D., & Rosso, P. (2020), «AMI @ EVALITA2020: Automatic misogyny identification". *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop*, Vol. 2765.

Fersini E., Gasparini F., Rizzi G., Saibene A., Chulvi B., Rosso P., Lees A., & Sorensen. J. (2022). SemEval-2022 Task 5: Multimedia automatic misogyny identification. *Proc. of the 16th Int. Workshop on Semantic Evaluation (SemEval-2022), co-located with NAACL-2022, Association for Computational Linguistics*, 533-549.

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

Frenda S., Cignarella A., Basile V., Bosco C., Patti V., & Rosso P. (2022). The unbearable hurtfulness of sarcasm. *Expert Systems with Applications (ESWA)*, 193. https://doi.org/10.1016/j.eswa.2021.116398

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering online hate speech*. Unesco Publishing.

Gelber, K. (2002). *Speaking back: The free speech versus hate speech debate*. John Benjamins. https://doi.org/10.1075/dapsac.1

Hagen, L., Falling, M., Lisnichenko, O., Elmadany, A. A., Mehta, P., Abdul-Mageed, M., Costakis, J., & Keller, E. T. (2019). Emoji use in Twitter white nationalism communication. *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing,* 201-205.

Israeli, A. & Tsur, O. (2022). Free speech or free hate speech? Analyzing the proliferation of hate speech in Parler. *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, 109-121.

Jacobs, J. & Potter, K. (1998). *Hate crimes: Criminal law and identity politics*. Oxford University Press.

Jost, J. T. (2019). A quarter century of system justification theory: Questions, answers, criticisms, and societal applications. *British Journal of Social Psychology*, 58, 263-314. https://doi.org/10.1111/bjso.12297

Kaufman, G. A. (2015) . *Odium Dicta. Libertad de expresión y protección de los grupos discriminados en internet*. Consejo Nacional para Prevenir la Discriminación. Conapred.

Kennedy, B., Atari, M., Davani, A. M., Yeh, L., Omrani, A., Kim, Y., Coombs, K., Havaldar, S., Portillo-Wightman, G., Gonzalez, E., Hoover, J., Azatian, A., Hussain, A., Lara, A., Cardenas, G., Omary, A., Park, C.,

Wang, X., Wijaya, C., & Dehghani, M. (2022). Introducing the Gab Hate Corpus: Defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation, 56*(1), 79–108. https://doi.org/10.1007/s10579-021-09569-x

Kumar, R., Ojha, A. k., Malmasi, S.. & Zampieri, M. (2018). Benchmarking aggression identification in social media. *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying,* 1–11. Association for Computational Linguistics.

Pereira-Kohatsu, J. C., Quijano-Sánchez, L., Liberatore, F., & Camacho-Collados, M. (2019). Detecting and monitoring hate speech in Twitter. *Sensors*, *19*(21), 4654. MDPI AG. http://dx.doi.org/10.3390/s19214654

Pérez, J. A., Ghosn, F., Chulvi, B., & Molpeceres, M. (2023). Does threat cause discrimination or does discrimination cause threat? *International Journal of Social Psychology, 38*(2), 279-303. https://doi.org/10.1080/02134748.2022.2158589

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2021). Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation*, 55, 477–523.

Mackie, D. M. & Hamilton, D. L. (1993). *Affect, cognition, and stereotyping. Interactive processes in group perception*. Academic Press.

MacAvaney, S., Hao-Ren, Y., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *Plos One*, *14*(8), 1-16.

Magalhaes, A. (2019). *Automating online hate speech detection: A survey of deep learning approaches*. [Master's thesis, School of Informatics, University of Edinburgh]. https://api.semanticscholar.org/CorpusID:237363061

Mandl, T., Modha, S., Shahi G. K., Jaiswal, A. K., Nandini, D., Patel, D., Majumder, P., & Schäfer, J. (2021). Overview of the HASOC track at FIRE 2020: Hate speech and offensive content identification in Indo-European Languages. *Notebook Papers of FIRE 2020, CEUR Workshop Proceedings*, Vol. 2826, 87-111.

Mathew, B., Dutt, R., Goyal, P., & Mukherjee, A. (2019). Spread of hate speech in online social media. *Proceedings of the 10th ACM conference on web science*, 173-182.

Mathew, B., Illendula, A., Saha, P., Sarkar, S., Goyal, P., & Mukherjee, A. (2020). Hate begets hate: Atemporal study of hate speech. *Proceedings of the ACM on Human-Computer Interaction - 4(CSCW2)*, 1–24.

Merlo, L. (2022). *When humour hurts: A computational linguistic approach*. [Final degree project, Universitat Politècnica de València]. http://hdl.handle.net/10251/188166

Miranda, S., Malini, F., Di Fátima, B., & Cruz-Silva, J. (2022). I love to hate! The racist hate speech in social media. *Proceedings of the European Conference on Social Media*, 9, 137-145. https://doi.org/10.34190/ecsm.9.1.311

Modha, S., Mandl, T., Majumder, P., & Patel, D. (2019). Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in Indo-European languages. *CEUR Workshop Proceedings,* 2517, 167-190.

Mossberger, K., Tolbert, C. J., & McNeal, R. S. (2008) *Digital citizenship: The internet, society, and participation*. The MIT Press.

Müller, K. & Schwarz, C. (2020). 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

Pisani, F. & Piotet, D. (2009). *La alquimia de las multitudes*. Paidós.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2021). Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation*, 55, 477–523.

Rangel, F., De-La-Peña-Sarracén, G. L., Chulvi, B., Fersini, E., & Rosso, P. (2021). Profiling hate speech spreaders on Twitter task at PAN 2021. In: G. Faggioli, N. Ferro, A. Joly, M. Maistro, & F. Piroi (Eds.), *CLEF 2021 Labs and Workshops, Notebook Papers*. CEUR-WS.org.

Ribeiro, M., Calais, P., Santos, Y., Almeida, V., & Meira Jr., W. (2018). Characterizing and Detecting Hateful Users on Twitter. Proceedings of the International AAAI Conference on Web and Social Media, 12(1). https://doi.org/10.1609/icwsm.v12i1.15057

Ross, L., Greene, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of experimental social psychology*, *13*(3), 279-301.

Sánchez-Junquera, J., Chulvi, B., Rosso, P., & Ponzetto, S. (2021). How do you speak about immigrants? Taxonomy and stereo immigrants dataset for identifying stereotypes about immigrants. *Applied Science*, *11*(8), 3610. https://doi.org/10.3390/app11083610

Samghabadi, N. S., Patwa, P., PYKL, S., Prerana, M., Amitava, D., & Solorio, T. (2020). Aggression and misogyny detection using BERT: A multi-task approach. *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, 126–131.

Sanguinetti, M., Comandini, G., di Nuovo, E., Frenda, S., Stranisci, M., Bosco, C., Caselli, T., Patti, V., & Russo, I. (2020). HaSpeeDe 2 @ EVALITA2020: Overview of the EVALITA 2020 hate speech detection task. *Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, CEUR Proceedings*, 2765, 1–9.

Struß, J. M., Siegel, M., Ruppenhofer, J., Wiegand, M., & Klenner, M. (2019). Overview of GermEval Task 2, shared task on the identification of offensive language. *Proceedings of GermEval-2019*, 352-363.

Taulé, M., Ariza, A., Nofre, M., Amigó, E., & Rosso, P. (2021). Overview of the detoxis task at iberlef-2021: Detection of toxicity in comments in Spanish. *Procesamiento del Lenguaje Natural*, 67, 209-221.

Waseem, Z. (2016). Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. *Proceedings of the First Workshop on NLP and Computational Social Science*, 138–142.

Waseem, Z. & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. *Proceedings of the NAACL Student Research Workshop*, 88–93.

Wiegand, M., Siegel, M., & Ruppenhofer, J. (2018). Overview of the GermEval 2018 shared task on the identification of offensive language. *Proceedings of GermEval 2018*, 1-10.

Zafra, R. (2012). *A connected room of one's own. (Cyber)space and (self) management of the self*. Fórcola.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). Semeval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). *Proceedings of SemEval-2019*, 75-86.

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., & Çöltekin, Ç. (2020). SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). *Proceedings of SemEval-2020*, 1425–1447.

TRIALS AND CHALLENGES MEASURING ONLINE HATE

Andre Oboler
/ La Trobe University, Australia

Introduction

In recent years there has been a noticeable shift in public attitudes towards internet regulation, particularly the regulation of online hate. The early internet promised a new world outside existing power structures, beyond the reach of governments, where a new social contract would emerge (Barlow, 1996). Regulation was seen as stifling innovation and retarding economic and social progress; disruption was the order of the day. As the internet has become an essential and established part of life, and leading internet companies have become some of the largest multinationals, public expectations have changed. The status quo is seen as a failure. Change is seen as needed, but the first step in securing improvements is understanding the current state of the problem. This requires measurement in a manner that is useful, reliable, and practical. This chapter focuses on efforts that have attempted to do this for hate speech, often specifically antisemitism, with a focus on the methods used and the measurements that were, or could have been, applied to support evaluation of the work.

The growing concern over online hate results from recognition that what starts online doesn't stay there. Part of the problem is the link between online hate and violent extremism. Radicalisation in online communities of hate has resulted in deadly attacks on multiple

occasions and the livestreaming of some attacks, such as in the case of the 2019 attack in Christchurch, have inspiring further attacks (Oboler et al., 2019). Predictions of the normalisation of hate in online communities leading to a normalisation of hate in society (Oboler, 2008a) have been demonstrated by, for example, the "Unite the Right" rally where hundreds of white nationalists marched confidently and proudly down the streets of Charlottesville (Elliot, 2022), the rise of QAnon, and the January 6th attack on the US Capitol. Online platforms responded with new policies, and widespread account closures, but this has been seen as insufficient to protect public safety (Timberg et al., 2021).

Reflecting societal concerns, the volume of scholarly work on the topic has been rising in recent years. Table 1 shows the number of publications on "Hate Speech" and "Online Hate" by two major publishers. Springer publishes across a wide range of disciplines from the social sciences, law, and history to computer science, while IEEE publishes across the engineering and computing disciplines.

Table 1: Publications on "hate speech" and "online hate"

	2022	2021	2020	2019	2018	2008-2017
"hate speech" in Springer	2068	1514	947	790	538	45
"hate speech" in IEEE	108	100	65	47	23	4
"online hate" in Springer	255	201	80	62	36	78
"online hate" IEEE	36	25	9	3	3	2

Online hate, however, is not new. An early incident in 1989 saw thousands of racist and misogynistic jokes shared by a computer at the University of Waterloo in Canada, to computers at Standard University in the United States. Stanford blocked access to prevent university resources being used to spread hate speech. Reflecting the attitudes at the time, protests followed and a professor of Computer Science circumvented the restriction ("Racist,

sexist jokes do not compute," 1989). Extremists' use of technology is also not new. They were already using Bulletin Board Services (BBS) for sharing messages and files and for internal communications in the 1980s and early 1990s (Potok, 2008). Anyone with a computer and modem could use the telephone network to connect and access their content if they knew the phone number (Edwards, 2016). Online hate also proliferated through USENET newsgroups before the internet, and then along with USENET itself became a part of the new internet. A neo-Nazis online activist called alt.politics.nationalism.white, alt.politics.white-power, alt.revolution.counter, and alt.skinheads an online "Aryan Resistance" (Kleim, 1995); Within 5 years these groups each averaged over a hundred messages per day (Mann et al., 2003).

Civil society groups dedicated to monitoring hate and extremism have been watching from the started. A report by the ADL (1995), for example, described alt.politics.white-power as one of the "traditional on-line racist haunts". When the first extremist website, Don Black's Stormfront, went live in March 1995 (Potok, 2008), at a time when there were fewer than 23,500 websites in existence (Gray, 1996), the Simon Wiesenthal Center reported on it the very next month (OSCE, 2008). As other hate sites followed, including one for the National Alliance, the most dangerous and well organised neo-Nazi organisation in the United States at that time (*National Alliance*), groups like the ADL kept track (ADL, 1995). When the rise of social media led to "antisemitism 2.0" (Oboler, 2008a) and more generally "hate 2.0" (Oboler, 2012) civil society organisations were caught largely off guard (Snyder, 2008). A working group at the Global Forum for Combating Antisemitism founds that among the challenges was the "lack of metrics" on how many items were reported to social media platforms, how many user reports were actioned, how effectively platforms responded to law enforcement requests, and how long the process took (Oboler & Matas, 2009).

Today, despite much work having been done, the need for research and reporting with clear, accurate and effective metrics remains. The rising public and scholarly interest in hate speech adds urgency. Public policy development would be better informed if based on data. The increasing shift to a regulatory approach for addressing online hate only becomes possible in a meaningful way if the hate itself is capable of being recognised and accurately measured. Exiting regulatory measures in the absence of effective metrics are crude in their measurement and wide in their tolerance.

The remainder of this chapter considers four approaches to mapping hate speech, the benefits and limitations of each approach, the ways in which each approach should be evaluated, and how they have been evaluated in past work. The four approaches are: demonstrating hate, counting hate, manually coding hate, and modelling hate. Work demonstrating hate uses qualitative analysis best evaluated through peer review. Work counting hate requires sufficient data to verify the accuracy of the count. Work manually coding hate requires both expertise and reliability, the use of inter-coder agreement rates provides a measure of reliability, but caution is needed to avoid getting a consistent (reliable) but wrong result due to a lack of knowledge. When it comes to models, confusion matrices, precision, recall, and F-score provide metrics to compared how well a model's classification of data matches what are assumed to be true values for the data. Using these metrics, some models will have greater precision, others will have greater recall, and some will be more balanced with a better F-score than others. Selecting a model that is fit for purpose requires not only the metrics, but a consideration of their impact in a particular use case. The chapter ends with a conclusion on the way the forward.

Demonstrating Hate

As Lord Kelvin once explained, "When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers,

your knowledge is of a meagre and unsatisfactory kind: it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science, whatever the matter may be." We want a deeper understanding, expression in numbers, but first must come the beginnings of knowledge.

Work demonstrating that a particular form of online hate exists in a particular environment represents the beginnings of the knowledge and can lead to effective measurement of the hate. Given the rapid pace of technological change, and the changing nature of online hate, new demonstrations are continually needed. This qualitative work can highlight online hate in places with no affordances for users to address it, demonstrate gaps in community standards, and highlight failures in policy implementation. Qualitative work demonstrating online hate is vital to informing stakeholders, including impacted communities, platforms, legislators, and regulators, of the gaps that need to be addressed.

Examples of work demonstrating hate include the article on the 1989 case of misogyny jokes ("Racist, sexist jokes do not compute," 1989), the Simon Wiesenthal Center's exposure of Stormfront in 1995 (OSCE, 2008), the ADL's 1995 report into the online recruitment practices of hate groups (ADL, 1995), and their 1996 report into websites belonging to thirteen significant promoters of online hate (Hoffman, 1996). Examples of early work demonstrating hate 2.0 focused on antisemitism, racism against Indigenous Australians, and Islamophobia (Oboler, 2008b, 2012, 2013a, 2013b, 2014). The approach remains relevant and the subject of ongoing work, for example at the Online Hate Prevention Institute in Australia where over 350 articles demonstrating and deconstructing various forms of online hate across a wide range of platforms have been published since 2012 (OHPI, 2023). Work demonstrating hate is also particularly important when new hate narratives emerge, or old narratives become significantly more common, both of which occurred in the context of anti-Asian hate during the

Covid-19 pandemic (Oboler, 2022). Demonstrations of hate help platforms, governments, and communities identify and respond to the problem.

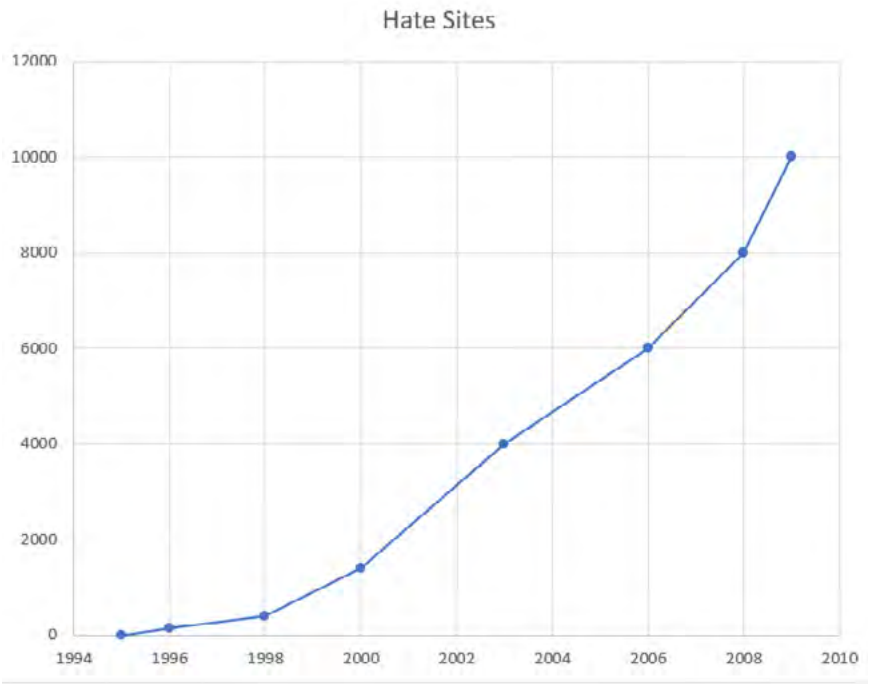
As the ADL concluded in the early days, "People of goodwill must continuously monitor the Internet, especially the World Wide Web, to counter messages of hate with information that challenges bigotry, exposes the bigots, and promotes tolerance, decency and truth" (Hoffman, 1996). We need to move beyond this into quantifying and measuring the hate, but demonstrations that aid identification are a necessary step.

Counting hate

The simplest empirical metric is the count of hate speech. Even an incomplete count is useful, demonstrating there are at least N cases of a particular problem. While individual demonstrations may be discussed and isolated incident then corrected manually, if the problem gains enough media attention, the creation of lists and counts raises questions about more systemic problems. It shifts the focus from users to the platforms and providers.

The proliferation of hate sites on the web led David Goldman, a reference librarian at Harvard Law School, to compile the in a list in 1995. Originally published as the *Guide to Hate Groups on the Internet* webpage at Harvard, it later became the *HateWatch* website ("Harvard Law School Librarian Discusses Cyberhate", 2001). The Magenta Foundation in the Netherlands established the *Complaints Bureau for Discrimination on the Internet* in 1997 and began collecting public reports (*History Magenta Foundation*, 2021). The Simon Wiesenthal Center started systematically monitoring online hate from 2018, counting the number of hate websites. Their 2008 report put the number of hate and terrorist websites at 8,000 (*iReport Online Terror + Hate: The First Decade*, 2008), while a report the following year graphed annual growth in online hate as summarised in Figure 1 (*Facebook, Youtube+: How Social Media Outlets Impact Digital Terrorism and Hate*, 2009). A Council of Europe report in October 2001 put the number of racist websites at 4,000 (Tallo, 2001), almost twice what the Simon Wiesenthal Center was finding at that time.

Figure 1: Rise of hate sites as monitored by the Simon Wiesenthal Center



Even for a simple metric like a count, the approach of expert staff finding hate sites, or verifying sites reported by the public, has become ineffective by 2008. The prolific growth of the Internet posed a challenge. From a little over 10,000 websites in early 1995, the web grew to about 180 million sites by early 2008 (Zakon, 2018). Even if the number of hate sites only grew in proportion to growth of the internet, the task of manually tracking and monitor them was becoming impractical. Using the Simon Wiesenthal Center's data as the baseline and assuming steady growth, then by 2022 there would have been between 78,000, and 180,000 sites. The real number is likely far higher as research into online hate in social media has demonstrated that online hate is growing faster than the internet itself (Matamoros-Fernández & Farkas, 2021; *Report: Online hate increasing against minorities, says expert*, 2021; Wendling, 2015). The initial estimates were also incomplete, as shown

by the vastly higher numbers found by the Council of Europe in 2001. While the sort of manual count done in the past is impractical, some count or way of identifying dedicated hate websites would clearly be useful for online safety. It is a task complicated by the efforts the most extreme sites take to evade bans by upstream providers and which can result in the same site returning under different domain names (Lavin, 2018).

The nature of Web 2.0 and social media led to large platforms which became microcosms of the internet itself, containing many online spaces catering to many communities. A smaller number of sites were so large there were fears of the way they might dominate the web (Berners-Lee, 2010). These sites were used for hate content, accounts, and spaces, and as they enveloped more of the online world, their share of the online hate increased.

Early work counting hate in social media was analogous to the approach used with websites. Online spaces dedicated to hate, such as Facebook pages, were compiled into lists. Early lists focused on spaces targeting First Nations Australians, Jews, and Muslims (Oboler, 2012, 2013a, 2013b). These lists appeared alongside work demonstrating the hate and were often cross referenced, with spaces of hate in the lists being included on the basis of examples of content they contained which were demonstrated to be hate in the same report. While many listed hate spaces eventually came down, the process often took months. The lists, related advocacy, and media coverage were essential to securing action by platforms. In the absence of such action, removal was the exception rather than the rule.

Coding hate

Coding involves the additional of meta data to classify identified items of hate speech according to a range of parameters such as: victim group (anti-Black racism, anti-Asian racism, antisemitism, Islamophobia, misogyny, homophobia, etc), jurisdiction reported, jurisdiction uploaded, platform, content type (e.g. post, video, image, comment), and type of hate narrative.

In work measuring hate speech the primary classification is usually based on the different narratives of hate expressed against the groups. The coding of hate according to the hate narrative being used is particularly important as a platform's effectiveness in responding to hate varies by narrative, even when the same group is targeted. A change in the prevalence of a particular narrative is also important to policy makers, educators, and civil society groups, as it may require a recalibration of educational efforts and campaigns.

The most striking example of the importance of the specific narrative occurs with Holocaust denial, a form of antisemitism identified in the original antisemitism 2.0 paper (Oboler, 2008a). It was demonstrated to be a form of hate narrative not being removed, even in countries where it was unlawful (Oboler, 2009). When Facebook did started to remove it or block it in these countries, they did so strictly based on legal requirements, and would not removing it in other jurisdictions (Oboler & Matas, 2013). This position was only reversed in 2020 (Bickert, 2020). The gap between the public's expectations about community standards, and the practice of major platforms led to significant interest in online hate narratives of Holocaust denial and related phenomena such as Holocaust distortion and Nazi glorification.

The value in coding hate narratives can also be seen in the "Measuring the Hate" report which demonstrated large variations between the handling of different narrative categories within each platform, as well as large variations between the platforms when examining the same category of antisemitism (Oboler, 2016). The report coded antisemitic content as traditional antisemitism, Israel related antisemitism, Holocaust denial, or incitement to violence and the data on removal rates is reproduced in Table 2. The data was collected through crowd sourcing with the custom built Fight Against Hate software (Oboler & Connelly, 2014) enabling the public to report and classify (code) content.

Table 2: Removal rates for antisemitism by category

	Traditional	Israel related	Holocaust denial	Violence	Platform avg.
Facebook	42%	27%	58%	75%	37%
Twitter	25%	20%	20%	14%	22%
YouTube	9%	4%	10%	30%	8%
Category avg.	21%	16%	22%	26%	

Coding schemes vary widely. A report by the World Jewish Congress (WJC) and Vigo Social Intelligence used five categories: Holocaust denial, dehumanisation, anti-Jewish hatred, antisemitic violence, and use of symbols (WJC, 2017) while a follow-up report focused just on Holocaust denial (WJC, 2018). An ADL report used six categories: Holocaust denial, classic antisemitic stereotypes, positive promotion of antisemitic people or publications/media, antisemitic conspiracy theories, slurs and epithets, and code words / antisemitic symbols ADL (2018). Work by Schwarz-Friesel (2018) on German language antisemitic content used three categories: classical antisemitism (as seen before 1945, post-Holocaust antisemitism (new forms arising after 1945 including those related to the Holocaust), and Israel-centred antisemitism (i.e. New Antisemitism). Ozalp et al. (2020) coded data on a binary basis as being "Antagonistic content related to Jewish identity". Chandra et al. (2021) used four categories of political, economic, religious, and racial antisemitism based on work by Brustein (2003) discussing pre-Holocaust antisemitism. To account for more recent forms of antisemitism they retrofitted new forms of antisemitism into these categories, for example, placing Holocaust denial under racial antisemitism and claims of Jews being more loyal to Israel under political antisemitism. Work by the *de-coding antisemitism project* codes content according to very specific antisemitic narratives such as "blood libel", "evil", "apartheid analogy", "denying Israel's right to exist", "terrorist state", and "Power/Influence" (Ascone et al., 2022). Ali and

Zannettou (2022) coded data as antisemitic or not, while Meta (2022) codes content only as hate speech or not, without a breakdown into victim groups, let alone hate narrative.

Agreed definitions, such as the Working Definition of Antisemitism from the International Holocaust Remembrance Alliance (IHRA, 2016) can provide a starting point for coding schemes. They can, however, be operationalised for coding in different ways. In 2018 the Australian Government's delegation to the International Holocaust Remembrance Alliance used the IHRA Working Definition of Antisemitism in combination with the IHRA Working Definition of Holocaust Denial and Distortion to create the coding schema for antisemitism with 4 major categories and a total of 26 more specific sub-categories, as shown in Table 2. This was implemented in the Fight Against Hate reporting software, with users now coding hate first by selecting a major category, then selecting from a list of related subcategories. Jikeli et al. (2019) coded content as antisemitic under the IHRA Working Definition of Antisemitism (IHRA, 2016) (yes or no) and separate coded the same content on whether it expressed negative sentiments to Jews, Judaism, or Israel (yes or no). While their work appears not to have recorded more detailed classifications, their approach also used a schema based on the IHRA definitions which could be used for detailed coding. It includes six major categories and 11 sub-categories within one of the major categories, as well as additional interpretative notes covering even more specific sub-categories such as alleged Jewish character traits, supposed Jewish physical stereotypes, antisemitic imagery, and alleged Jewish crimes. Chandra et al. (2021) used a reference work, Brustein (2003) on pre-Holocaust antisemitism, in place of a definition. When a check against the IHRA definition highlighted forms of antisemitism missing from the reference work (i.e. Holocaust era and post Holocaust forms of antisemitism), they tried to fit the missing types to the existing categories. An alternative approach, more in keeping with scholarship in the field, would have been to add additional categories and to understand them with reference to additional reference works, such as Rosenfeld (2015) on New Antisemitism.

Table 3: Australian IHRA Delegation's antisemitism taxonomy

Major category	Sub-categories
1. Holocaust related content	1.1 Denying the Holocaust 1.2 Accusing Jews or Israel of exaggerating the Holocaust 1.3 Blaming Jews for the Holocaust 1.4 Distort the facts of the Holocaust 1.5 Glorifying the Holocaust or suggesting it did not go far enough 1.6 Inappropriate comparisons with Nazis
2. Incitement to violence	2.1 Calling for, aiding, or justifying the killing or harming of Jews in the name of a radical ideology or an extremist view of religion. 2.2 Calling for harm to someone because they are Jewish 2.3 Calling for harm to Jewish people in general 2.4 Calling for harm to Jewish property 2.5 Calling for harm to someone believing they are Jewish 2.6 Calling for harm to non-Jews for supporting Jews or opposing antisemitism
3. Classic Antisemitism	3.1 Dehumanising Jews 3.2 Promoting the idea of a world Jewish conspiracy 3.3 Promoting the idea of Jews controlling the media 3.4 Promoting the idea of Jews controlling the economy 3.5 Promoting the idea of Jews controlling government or other societal institutions 3.6 Promoting traditional antisemitism such as blood libel and claims Jews killed Jesus 3.7 Holding Jews collectively responsible acts committed by individuals 3.8 Accusing Jews citizens of being disloyal to their country
4. Antisemitism related to Israel	4.1 Accusing Israel inventing or exaggerating the Holocaust 4.2 Denying Jewish people self-determination, e.g., by claiming Israel's existence is racist 4.3 Requiring a behavior from Israel not expected of other countries 4.4 Describing Israel or Israelis using antisemitic words or imagery (e.g., claims of Jews killing Jesus or blood libel) 4.5 Comparisons of Israeli policy to Nazism 4.6 Holding Jews collectively responsible for Israel's actions

The coding process relies upon, and represents, human judgement about the data, and that judgement is subject to errors and uncertainty. The approaches to minimizing errors generally assume the judgement of an

expert is more valuable than that of a non-expert, and that among people assumed to have similar expertise, a majority decision is likely to be correct more often than a minority. Taking a majority decision can reduce errors resulting from mistakes and fatigue, but it will also eliminate minority judgements from coders with additional knowledge or context relevant to the item being coded. Examining the past research shows the assumptions often being applied in practice. The crowd-sourcing approach in (Oboler, 2016) relied on non-experts who had identified hate and coded it, then a single expert who verified the data. Work by Schwarz-Friesel (2018) used two experts to code the data and operated on a conservative basis. Jikeli et al. (2019) used two trained non-experts to code the data. Ozalp et al. (2020) used a multi-stage process, four un-trained non-experts coded the data, and items coded as antisemitic by a majority were then reviewed by a trained non-expert which resulted in the trained non-expert rejecting the majority decision and re-coding 29% of the data. Chandra et al. (2021) used three trained non-expert coders, then discussed and reached a collective agreement on cases where the coders had disagreed. Ali and Zannettou (2022) in their work used a single person, one of the researchers, to code the data. Meta's sampling uses two coders with a third brought in as a tiebreaker when needed (Meta, 2022). As Ozalp et al. (2020) concluded, untrained people struggle with coding antisemitic hate speech, and as Jikeli et al. (2019) adds, coders need to be highly trained in the specific hate they are annotating, as well as knowledgeable about current events to account for context.

Given that human coding has a degree of uncertainty, the degree of that uncertainty (or its inverse "reliability") is an important research metric. It can be provided as an agreement rate measuring how often the same code is applied to the same data either by different coders (the inter-coder agreement rate) or by the same coder when the data is examined multiple times (the intra-coder agreement rate). Measuring and stating intercoder reliability is important when using multiple people to code the data (O'Connor & Joffe, 2020). There are multiple statistical approaches for calculating intercoder reliability in different circumstances and a range of common errors seen in

the published literature attempting to use intercoder reliability, including a common error of not providing it at all (Feng, 2014). The most intuitive approach, known as simple agreement, is the number of agreements divided by the number of decisions. This, however, doesn't take account of the fact people will also agree by random chance. There are various ways to adjust for this, for example, Cohen's kappa (κ) subtracts the degree of random chance from the simple agreement, while Krippendorff's alpha (α) divides the disagreement between coders by the expected disagreement if they were assigning codes randomly (Geisler & Swarts, 2019). Different approaches also have different limitations. As it can handle any number of coders, a variety of data types, and missing data, Krippendorff's alpha (α) is increasingly used as the preferred metric (Feng, 2014; Geisler & Swarts, 2019; Krippendorff, 2004).

In the work previously discussed, inter-coder agreement rates are seldom provided. Schwarz-Friesel (2018) does not include the inter-coder agreement rate between the two expert coder. Jikeli et al. (2019) also does not provide an agreement rate, but the work shows substantially different final counts for the classifications by different coders and states the differences were due to a lack of contextual knowledge, lapses in concentration, and different interpretations of the content. Ozalp et al. (2020) also does not provide an agreement rate but describes it as high, their coders only provided majority support for a coding a third of the time. Becker et al. (2021) states "intercoder reliability is calculated" but no measures are given in this or the subsequent three reports. Chandra et al. (2021) gives an inter-coder agreement score of 0.707 using Fleiss' Kappa, then coders discussed items with disagreement to reach a consensus. As a comparison, work by Saha et al. (2019) on hate speech more generally used two expert coders and found an inter-coder agreement rate of 0.8 using Cohen's κ. Becker et al. (2021) include neither the number of coders nor their inter-coder agreement. Facebook's transparency reporting does not provide an inter-coder agreement metric, despite a report they commissioned on the fitness for

purpose of their metrics recommending they "Share statistics on human reviewers' inter-rater reliability" (Bradford et al., 2019).

The lack of reliability metrics is not unique to work in online hate. It has been a systemic problem in content analysis research more broadly, despite experts in methodology highlighting that it is necessary (but not sufficient) for demonstrating research validity (Neuendorf, 2002; O'Connor & Joffe, 2020). Jikeli et al. (2019) highlights the need for expertise, not just consensus, noting that, "discussion between qualified annotators in which they explain the rationale for their classification is likely to result in better classification than using statistical measures across a larger number of (less qualified) annotators."

Sampling hate

Coding can be applied to a subset of data as a form of sampling, and the results used to extrapolate about the broader population. The accuracy of this approach depends on how representative the sample is of the broader data. This approach is used by ADL (2018) as well as in the Meta's transparency reports (Meta, 2023) as their methodology explains (Meta, 2022). The approach can be contrasted with coding all the content within a corpus, which becomes a census, as seen in the work of Schwarz-Friesel (2018) and the decoding antisemitism project (Ascone et al., 2022; Becker et al., 2021).

In sampling the key metric is the prevalence of hate, that is the rate at which hate speech occurs. This can be calculated by counting the number of hate items in a sample and dividing by the total number of items in the sample. This rate can then be multiplied by the size of the population to estimate the total number of hate items. This estimate will have some amount of random sampling error, which can be expressed for a given confidence interval by providing a margin of error using the usual statistical means. This margin of error is based on the assumption the items counted as hate truly are hate, and therefore it compounds the underlying uncertainty from the process of coding of the sample.

The use of crowd sourcing to gather a sample of online hate data can be seen as an approximation of a stochastic sampling technique. With no baseline to indicate the total volume of content consumed to generate the given number of reports, this approach can only give an indication of the relative prevalence of hate narratives, within or between platforms, it cannot estimate the prevalence in the broader population. It is only an approximation of a stochastic sampling as there are factors that will bias the data, such as differing numbers of observers (people engaging in reporting) on the different platforms, reporters focused on specific narratives, and even reporters that aren't focused on finding a particular hate narrative may be more likely to report some narratives as they are either more egregious or easier to identify. Oboler (2016) uses this approach, as summarised in Table 4, as well as tracking the relative rates of removal as previously shown in Table 2. The errors previously discussed are unquantifiable, however, the differences shown in both tables are so large as to indicate important variations within and between platforms.

Table 4: Categorisation of the data by platform and hate narrative from the measuring the hate report

	Traditional	Israel related	Holocaust denial	Violence
Facebook	447	237	43	16
Twitter	746	197	92	72
YouTube	834	264	106	27

The European Commission's monitoring exercise also uses a sampling technique. It collects data over a six-week period each year based on content reported to major platforms by a set of civil society and governmental organisations (Reynders, 2022). The 2022 monitoring exercise included 36 organisations from 21 countries that collectively reported 3534 social media items with 64% of them being removed within 24 hours. This represents a

decline in the 24 hour removal rate from 81% in 2021 and 90.4% in 2020 (Reynders, 2022). The data is a good sample for the average response times on online hate (in general) but is not representative of the relative volume of hate against each group as the different data collection entities focus on hate against different victim groups and applied different resourcing to their data collection. A further complication is that the collection methodology asked that only unlawful hate speech be reported, and whether something is unlawful may differs based on national legislation.

A sampling technique is also used in Meta's transparency reporting for Facebook and Instagram (Meta, 2023). The sample is of all content on the platform, not just reported content, and with a known sample size Meta is able to calculate the prevalence of hate speech. They define prevalence as the estimated number of views of hate speech on the platform, divided by the estimated number of total views of content on the platform and argue this "better reflects the effect on the community" than measuring the number of items of hate speech (Meta, 2022). The report they commissioned was critical of this approach recommending Meta "Report prevalence two ways: (1) Number of violating posts as a proportion of the total number of posts; (2) Number of views of violating posts as a proportion of all views" (Bradford et al., 2019). The prevalence of views is a good measure of the chance a person encounters violating content, under the clearly false assumption content is served randomly, however, the count of items is a better measure of a platforms effectiveness at removing hate speech. The prevalence of views will appear far lower than the prevalence of items as viral non-hate speech inflates it, while the same cannot happen for hate speech as soon after going viral it would likely trigger media attention leading to platform removal. The prevalence is graphed as a range corresponding to a confidence interval of 95%, which take account of the sampling error, though the underlying uncertainly in the judgement of those assessing the sample remains unknown.

Modelling hate

Modelling involves the use of an algorithm to classify whether online content is hate speech. It is an application of text filtering, an area that was first identified as a distinct topic of research at the Fourth Message Understanding Conference in 1992 (Lewis & Tong, 1992). A model may be created to identify hate speech at a generic level, or targeting a specific group, or using specific narratives. At its simplest, an algorithm might check the contents against a set of explicit rules created by experts. It might, for example, identify content as hate if any of the words in a designated list are present. Another approach is to use supervised machine learning, where a classification model is created by combining an artificial intelligence technique with a set of known data. In this approach a model is configured, used on the training data, then checked for accuracy (using the known values). The process is repeated multiple times with different configurations. The configuration that is most accurate in its classification decisions, that is the one that best fits the training data, is presented as the result. This model is then evaluated against a different set of known data to see how effective it is on data outside its training set. This may be done multiple times with different training data, or different settings, to determine the final model.

The advantage of modelling is that, once it has been created, it can be applied to a set of data to both identify individual items of hate speech, and to determine how many items of hate speech there are. Where there is access to all of the data posted to a particular platform, for example when it is the social media company itself running the model, this amounts to a census of the volume of hate speech on the platform. It may also lead to automated removal of the content, or to the content being given priority for human review. Where the model can only be applied to some of the data, for example where the model is being applied by a third party and the platform limits how much data the third party can access, or charges for access in a manner that makes larger volumes financially unviable, the available data can

be used as a sample to estimate the total number of items on the platform. The advantage of using a model over manual coding is one of speed and scale. The accessible sample of data will often be many times larger than it is practical to manually evaluate in a timely fashion. For social media companies, proactively identifying content using a model is more cost effective, and scalable, than human reviewers.

The limitation of modelling is that is it only an approximation of a reliable human decision, and depending on its accuracy it may or may not be fit for purpose. There are two possible mistakes a model can make: it could identify as hate speech items that are not hate speech, creating a false positive, or it could fail to identify items that are hate speech, creating a false negative. False positives in a system that automatically removes content or sanctions users undermines freedom of speech. False negatives allow hate speech to go undetected and undermine human dignity. The negative impacts of using artificial intelligence to make decisions can be mitigated in various ways, for example, by requiring human review before sanctions are applied, reviewing automated decisions after they are applied, providing an appeals process, providing other effective mechanisms for handling false negatives such as sufficient human reviewers to address user reports in a timely manner, or in the case of regulation increasing the regulatory tolerance. In some cases the decision is to abandon a model altogether, for example, concerns over false positives led Mark Zuckerberg to intervene blocking deployment of some mitigation approaches at Meta (Timberg et al., 2021).

Knowing how often a model is correct is a necessary but insufficient metric. To determine if a model is fit for a particular purpose information about the rate of false positives and false negatives is also needed. One way to present this data is in the form of a confusion matrix, a form of contingency table. A confusion matrix shows the number of correctly identified items, the number of correctly rejected items, as well as false positives and false negatives. Table 5 shown an example. From the confusion matrix precision and recall can be calculated.

Table 5: An example of a confusion matrix for a sample of 1000 items

		Actual values (based on human evaluation)	
		Hate Speech	Not Hate Speech
Predicted values (based on the model)	Hate Speech	100 (True Positives)	20 (False Positives)
	Not Hate Speech	180 (False Negatives)	700 (True Negatives)

Precision (P) measures the accuracy of the model in respect of items it classified as positives. It is the fraction of results identified as positive that are positives in truth. As the number of false positives drops towards zero, the precision increases towards 1.

$$P = \frac{True\ Positives}{True\ Positives\ + False\ Positives}$$

Recall (R) measures how effective the model is at finding all the available true positives, it is a measure of the model's coverage. As the number of false negatives (the number of positives the model failed to identify) heads towards zero, the recall increases towards 1.

$$R = \frac{True\ Positives}{True\ Positives\ + False\ Negatives}$$

There is a direct trade-off between precision and recall. In the most extreme case, if a model classed all the data as hate speech, it would have perfect recall, but very poor precision. With all the data classed as hate speech, there can be no false negatives but must be many false positives. The other extreme is a model that only selects the most direct and overt hate speech, getting very high precision, perhaps perfect precision if all the false positives are avoided, but the recall becomes low as the rest of the

hate speech is missed and becomes false negatives. Balancing precision and recall, the *F-score* has become a key metric for evaluating models for text filtering (Sasaki, 2007). It was proposed Lewis and Tong (1992) as a weighted mean of the two and defined as:

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

Altering the parameter allows the F metric to be adjusted to apply more weight to recall (if > 1), or to precision (if < 1). This allows models to be measured for their fitness towards a predefined balance between these two properties. The case without a weighting adjustment, where =1, gives the harmonic mean of precision and recall and is known as the F_1 metric. This is the most common form of the metric used in the research. In this case the equation can be simplified to:

$$F = \frac{2PR}{P + R}$$

Models with different sensitivities between precision and recall may be more appropriate for different tasks. A model with high recall (lower false negatives and higher false positives) may be appropriate when queuing items for manual review, while a model with high precision (lower false positives but higher false negatives) may be more appropriate for automated removal of content or application of sanctions. Regulation may also impact the sensitivity needed in a model. The average response time for a platform's staff to action user reports is a factor of how many items are reported divided by the number of staff. Meeting a fixed target, for example, removing the majority of reported content within 24 hours as required by the European Union, can either be done by hiring more staff, or by increasing the volume of content removed by artificial intelligence using models with higher recall

(and less precision). Conversely, regulations to protect freedom of speech may be implemented in a manner that penalises the use of artificial intelligence with low precision.

The selection of a sample of data to test the model on is also important. The F metric is sensitive to generality of the data examined, with precision dropping as generality increases (Lewis & Tong, 1992). This means a higher precision, and consequently a higher F value, when examining data that is rich in hate, such as from a community dedicated to hate, or newspaper articles likely to attract hate, than would apply in spaces with a lower rate of hate speech, such as general social media content. This also occurs if data is pre-filtered, for example, if models are run only on data that includes a predefined word or list of words. Pre-filtering means many possible true positives in the broader environment will have been excluded by the filter, and the stated precision and recall will be higher than the reality over the unfiltered data, inflating the F value.

An example of a simple model is the use of pattern matching to identify content as hate speech based on the presence or absence of words selected by experts. A report by the ADL (2016) used this approach to examine Twitter content from August 2015 to July 2016 and found 2,641,072 tweets containing language frequently used in antisemitic statements. The terms used included the hashtags #Jewish, #Israel, and #Holocaust, meaning the model would have many false positive from non-antisemitic tweets about Jewish events, Israel, and Holocaust remembrance and education. A sample examined manually was not representative of the data. Instead, tweets were reviewed if they contained the keywords and referenced specific journalists. This review resulted in 19,253 tweets being manually identified as antisemitic, but the total number of tweets manually reviewed is unknown. Without it, the precision can't be calculated. A later report estimated the number of antisemitic tweets between January 29, 2017 and January 28, 2018 and used more precise queries (ADL, 2018). It was run each week with a sample manually reviewed. Precision was calculated, then extrapolated back to the number of tweets identified to estimate the number of true

positives. Over the year a total of 55,000 tweets were reviewed and 4.2 million antisemitic tweets extrapolated. While a significant improvement, recall was still absent. To calculate recall a sample of the data from Twitter over the period would need to be manually coded, then the model applied. Only then could the false negatives (antisemitism not containing the keywords) be detected and measured and recall calculated so the query could be refined.

Ali and Zannettou (2022) focus on data from /pol/, an environment rich in hate speech. They further enrich the data by filtering based on selected terms to create a corpus. One term used to select data, "kike", was present in 67% of the corpus data. In a sample of 50 items containing kike, they manually coded 98% as antisemitic. Using 48 terms, each evaluated in this way, they determine that 93.83% of the content that used the terms was antisemitic. For a corpus of data selected using Islamophobic terms they determine 81% accuracy. This approach is similar to the ADL's, except that precision is calculated for each pattern and scaled according to its frequency, whereas the ADL appears to calculate precision wholistically. The approach of Ali and Zannettou highlights how overall precision could be improved, but recall reduced, if terms with less accuracy were dropped from their model.

The report by the World Jewish Congress and Vigo Social Intelligence used the social media listening tool Talkwalker in February 2017 to search for certain words or phrased appearing across tens of millions of items published during 2016 on a range of social media platforms and blogs (WJC, 2017). This is another implementation of the pattern matching model. A sample of 7,640 items (amounting to 2% of the identified data) was manually coded. The report does not provide a metric for precision, nor does it state that a precision metric was used to extrapolate the true number of antisemitic items based on the total number found to contain the target words on each platform, yet the data itself strong supports the idea this occurs. Across the data for the different platforms the number of items of *anti-Jewish hatred* is always a factor of 1580, the number of Holocaust denial items a factor of

140, and the number of dehumanisation items a factor of 250. Excluding the Twitter data, the number of symbols is always a factor of 770 and violence a factor of 155. A follow up report (WJC, 2018), focused on fewer categories of hate, over a shorter period of 28 days, but the resulting data still appears to have been scaled based on the sample.

Zannettou et al. (2020) used the data from /pol/ and the data from Gab to train the two-layer neural network word2vec (Mikolov et al., 2013) to create two "continuous bag-of words" models for antisemitism, one for each platform. They reduce the data into stemmed words (i.e. 'antisemitism' and 'antisemite' become 'antisemit'), then limit each corpus to stemmed words appearing at least 500 times. This resulted in 31,337 stem words on /pol/ and 20,115 on Gab. Word2vec is applied to each dataset, converting each stem word into a multi-dimensional vector. Each dimension represents an element of context and words that often appear together in a context are assigned values close together along that dimension. Rather than training the model on a sub-set of the data and then evaluating it on remaining data, Zannettou et al. train their model on all the data and provide no metrics on their model. Instead, they present a reduced version of the model graphically. The result shows 5 distinct groups (or "bags") of words associated with "jew", two of which represent antisemitic contexts. The first antisemitic bag they describe as containing words that present "Jews as a morally corrupt ethnicity" (i.e. negative stereotypes), while the second bag present Jews as "powerful geopolitical conspirators" (i.e. antisemitic conspiracy theories). The approach also generated three bags of non-antisemitic words related to Jewish ethnicity, Jewish mysticism, and Jewish theology. The paper suggests the approach can be useful to find new forms of coded hate speech.

Ozalp et al. (2020) tested four machine-learning methods, Decision Trees, Naïve Bayes, Support Vector Machine, and Fuzzy logic, each applied to the data using three approaches: Bag of Words, N-Grams, and Typed Dependencies. They note that absence of metrics related to the "accuracy of the content classification results" in past work and specifically the absence of information retrieval measures such as precision, recall, and F-measure,

which they provide for each combination of machine-learning method and approach in their work. Support Vector Machine using a Bag of Words approach gave the best results. Using a 10-fold test precision is given as 0.665, recall 0.531 (for antisemitic items), and the F Score as 0.590. While the paper considers this a good result, an artificial intelligence that is mislabelling content as antisemitic about a third of the time is not suitable for making automated decisions, and one that is missing about half the antisemitism is leaving substantial amount of antisemitism to be found by users, reported, and manually evaluated. The authors claim that with a 70:30 split (training on 70% of the data then testing on 30%) they get perfect precision and recall. This is highly unlikely, particularly given the comparison to the 10-fold test. It indicates the model was mistakenly trained and tested on essentially the same data, that is, the testing data was essentially duplicates of the training data.

In their fourth report (Ascone et al., 2022) the *decoding antisemitism* project introduced their first machine-learning models. They divided their large manually coded corpus into a training set and an evaluation set. Their model used a regression model to learn from the training set, and when applied to the evaluation is stated to have achieved an F_1 Score of 0.752. The specific values of precision and recall are not provided, however, the report states that details of the model will be published in a future report. The F_1 score is valuable for abstractly comparing models, but on its own provides insufficient information to determine whether one model would be better for a particular purpose than another. Their fifth report provides precision, recall, F_1 score, and sample size for two models used to identify antisemitic content (Chapelan et al., 2023). One model has greater precision, the other greater recall. The report explains the trade-off between recall and precision but uses F_1 rather than adjusting the F metric to give weight to either precision or recall. More concerning is that the precision, recall, and F_1 score are given for the class of content that is antisemitic *and* for the class of content that is not antisemitic. A weighted average of these two F_1 scores is then labelled "accuracy", a metric which is not meaningful.

Meta's metrics (Meta, 2023) are only reported as hate speech, not for specific forms of hate such as antisemitism. The metrics are the combined result of applying many different models. Meta reports on effectiveness by giving a "proactive rate", an approximation for recall under assumptions which favour would Meta and overstate effectiveness. The assumptions are that: all automated removals are correct, all rejections of user reports are correct, and all hate speech is either found by the models or reported by users. Under those conditions the proactive rate would be the same as recall. For the final quarter of 2022, 81.9% of removed hate speech was proactively identified, with the remaining 18.1% identified by user reports. This is the first quarter in which there was a significant drop in the rate of proactive identification, from 95.6% in the previous quarter. In the same period the number of hate items removed grew from 10.6 million to 11 million. The changes may be in part a result of Meta broadening its understanding of hate speech and accepting as valid a greater number of user reports, but it also indicated the absolute number of items identified by models has dropped. Either the hate speech is occurring outside patterns the models recognise, or the models have been refined to reduce the rate of false positives. There simply isn't enough data provided to know and this is despite a 2019 report Facebook commissioned that suggested company publish the "false positive, true positive, and false negative rates, as well as precision and recall" for automated decisions such as the detection of hate speech (Bradford et al., 2019).

Conclusion

This chapter explored the concept of measurement in relation to four approaches to mapping hate: demonstrating hate, counting hate, manually coding hate, and modelling hate. Each approach adds value but comes with limitations. Reviewing past work, it is clear much of the information needed to assess the quality of this work is often omitted. Where it is provided, the data used is often from sources with far higher occurrences of hate which result in the metrics appearing far better than they would when applied to more general online content. This is because the generality of the data impacts the metrics.

Efforts to map online hate occur in a context and for a purpose. Emerging forms of hate may go undetected until explained by qualitative work. A count may demonstrate a systemic problem. Manual coding can enable comparisons between the frequency of different types of hate. It can also create the data needed to train models. Proactive detection of hate speech at scale may only be possible using models.

The different approaches to mapping hate are evaluated in different ways. Work demonstrating hate is transparent and explicitly provides both data and a rationale for classification. This allows it to be reviewed. Work counting online hate may be compared or supported by examples that demonstrate the counted hate. The uncertainty in work coding hate can be measured using inter-coder agreement and comparisons to expert judgement. The ideal is inter-coder agreement based on the coding of multiple experts. Models seek to simulate expert judgement and can be compared to known expert decisions. The agreement between a model and expert judgement can be expressed using a confusion matrix. It can be summarised using precision, the fraction of items identified as hate that really were hate, and recall, the fraction of all actual hate items that were identified by the model. It can also be further summarised into a single number using the F-Score.

When models are used, the reason hate is being mapped will determine which metrics are more important. Models may increase precision by sacrificing recall, or vice versa. A model with higher precision may be needed when the result is autonomous action, such as the closure of accounts or removal of content. Models with higher recall may be needed to reduce the volume of hate that require reporting and the number of staff needed to review it in a timely fashion. The F_1 Score, which balances precision and recall, allows overall improvements to be measured, distinct from trade-offs between precision and recall.

Ultimately, mapping online hate in a manner fit for purpose requires a more consistent use of metrics. Descriptive work provides the beginning of knowledge, but deeper knowledge requires measurement and expression in meaningful numbers.

References

ADL. (1995). *Hate group recruitment on the InterNet*. http://nizkor.com/hweb/orgs/american/adl/recruitment/hgr-3.html

ADL. (2016). *Anti-semitic targeting of journalists during the 2016 presidential campaign*. https://www.adl.org/resources/press-release/adl-task-force-issues-report-detailing-widespread-anti-semitic-harassment

ADL. (2018). *Quantifying hate: A year of antisemitism on Twitter*. https://www.adl.org/resources/report/quantifying-hate-year-anti-semitism-twitter

Ali, M. & Zannettou, S. (2022). *Analyzing antisemitism and islamophobia using a lexicon-based approach*. ICWSM Workshops, Atlanta.

Ascone, L., Becker, M. J., Bolton, M., Chapelan, A., Krasni, J., Placzynta, K., Scheiber, M., Troschke, H., & Vincent, C. (2022). *Discourse Report 4* (Decoding antisemitism: An AI-driven study on hate speech and imagery online, Issue). https://decoding-antisemitism.eu/publications/fourth-discourse-report/

Barlow, J. P. (1996). *A declaration of the independence of cyberspace*. Electronic Frontier Foundation. https://www.eff.org/cyberspace-independence

Becker, M. J., Troschke, H., & Allington, D. (2021). *First Discourse Report* (Decoding antisemitism: An AI-driven study on hate speech and imagery online, Issue). https://decoding-antisemitism.eu/publications/first-discourse-report/

Berners-Lee, T. (2010). Long live the web. *Scientific American*, *303*(6), 80-85. http://www.jstor.org.ez.library.latrobe.edu.au/stable/26002308

Bickert, M. (2020). *Removing holocaust denial content*. Meta. https://about.fb.com/news/2020/10/removing-holocaust-denial-content/

Bradford, B., Grisel, F., Meares, T. L., Owens, E., Pineda, B. L., Shapiro, J. N., Tyler, T. R., & Peterman, D. E. (2019). *Report of the Facebook data transparency advisory group*. https://law.yale.edu/sites/default/files/area/center/justice/document/dtag_report_5.22.2019.pdf

Brustein, W. I. (2003). *Roots of hate: Anti-semitism in Europe before the Holocaust*. Cambridge University Press. https://doi.org/DOI: 10.1017/CBO9780511499425

Chandra, M., Pailla, D., Bhatia, H., Sanchawala, A., Gupta, M., Shrivastava, M., & Kumaraguru, P. (2021). *"Subverting the Jewtocracy": Online antisemitism detection using multimodal deep learning*. In 13th ACM Web Science Conference 2021, https://doi.org/10.1145/3447535.3462502

Chapelan, A., Ascone, L., Becker, M. J., Bolton, M., Haupeltshofer, P., Krasni, J., Krugel, A., Mihaljević, H., Placzynta, K., Pustet, M., Scheiber, M., Steffen, E., Troschke, H., Tschiskale, V., & Vincent, C. (2023). *Decoding antisemitism: An AI-driven study on hate speech and imagery online.* Discourse Report 5. T. U. Berlin.

Edwards, B. (2016, 5 November). The lost civilization of dial-up bulletin board systems. *The Atlantic*. https://www.theatlantic.com/technology/archive/2016/11/the-lost-civilization-of-dial-up-bulletin-board-systems/506465/

Elliot, D. (2022, 12 August). The Charlottesville rally 5 years later: 'It's what you're still trying to forget'. *NPR*. https://www.npr.org/2022/08/12/1116942725/the-charlottesville-rally-5-years-later-its-what-youre-still-trying-to-forget

Facebook, Youtube+: How Social Media Outlets Impact Digital Terrorism and Hate. (2009). Simon Wiesenthal Center. https://web.archive.org/web/20100331230627/www.wiesenthal.com/atf/cf/%7B54d385e6-f1b9-4e9f-8e94-890c3e6dd277%7D/LA-RELEASE_2.PDF

Feng, G. C. (2014). Intercoder reliability indices: disuse, misuse, and abuse. *Quality & Quantity, 48*(3), 1803-1815. https://doi.org/10.1007/s11135-013-9956-8

Geisler, C. & Swarts, J. (2019). *Coding streams of language: Techniques for the systematic coding of text, talk, and other verbal data*. University Press of Colorado. https://wac.colostate.edu/books/practice/codingstreams/

Gray, M. (1996). *Web growth summary*. http://www.mit.edu/people/mkgray/net/web-growth-summary.html

Harvard Law School Librarian Discusses Cyberhate. (2001). *Intelligence report*(Spring). https://www.splcenter.org/fighting-hate/intelligence-report/2001/harvard-law-school-librarian-discusses-cyberhate

History Magenta Foundation. (2021). *Magenta Foundation*. https://www.stichtingmagenta.nl/Archives.html

Hoffman, D. S. (1996). *Web of hate: Extremists exploit the Internet*. ADL. https://www.adl.org/sites/default/files/documents/assets/pdf/combating-hate/ADL-Report-1996-Web-of-Hate-Extremists-exploit-the-Internet.pdf

IHRA. (2016). *IHRA working definition of antisemitism*. https://www.holocaustremembrance.com/resources/working-definitions-charters/working-definition-antisemitism

iReport Online Terror + Hate: The First Decade. (2008). Simon Wiesenthal Center. https://www.csce.gov/sites/helsinkicommission.house.gov/files/Cooper%20Testimony.PDF

Jikeli, G., Cavar, D., & Miehling, D. (2019). Annotating Antisemitic online content. Towards an applicable definition of antisemitism. *arXiv preprint*. https://doi.org/10.5967/3r3m-na89

Kelvin, W. T. (1889). Electrical units of measurement (Speech to the Institution of Civil Engineers, London, 3 May 1883). In *Popular Lectures and Addresses*, 1, 80-81). Macmillan and Company.

Kleim, M. J. J. (1995). On tactics and strategy for USENET. In *alt.revisionism post Sep 20, 1995*.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). SAGE.

Lavin, T. (2018). The Neo-Nazis of the daily stormer wander the digital wilderness. *The New Yorker*. https://www.newyorker.com/tech/annals-of-technology/the-neo-nazis-of-the-daily-stormer-wander-the-digital-wilderness

Lewis, D. D. & Tong, R. M. (1992). *Text filtering in MUC-3 and MUC-4*. In Proceedings of the 4th conference on Message understanding. McLean, Virginia.

Mann, D., Sutton, M., & Tuffin, R. (2003). The evolution of hate: Social dynamics in white racist newsgroups. *Internet Journal of Criminology*. https://irep.ntu.ac.uk/id/eprint/10080

Matamoros-Fernández, A. & Farkas, J. (2021). Racism, hate speech, and social media: A systematic review and critique. *Television & New Media*, *22*(2), 205-224. https://doi.org/10.1177/1527476420982230

Meta. (2022, 18 November). *Prevalence*. https://transparency.fb.com/en-gb/policies/improving/prevalence-metric/

Meta. (2023). *Hate speech*. https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint*. https://arxiv.org/abs/1301.3781

National Alliance. Southern Poverty Law Center. https://www.splcenter.org/fighting-hate/extremist-files/group/national-alliance

Neuendorf, K. A. (2002). *The content analysis guidebook*. SAGE.

O'Connor, C. & Joffe, H. (2020). Intercoder reliability in qualitative research: Debates and practical guidelines. *International Journal of Qualitative Methods*, *19*. https://doi.org/10.1177/1609406919899220

Oboler, A. (2008a). Online antisemitism 2.0. 'Social antisemitism on the social web'. *Post-Holocaust and antisemitism*, *No 67*. https://jcpa.org/online-antisemitism-2-0-social-antisemitism-on-the-social-web/

Oboler, A. (2008b). The rise and fall of a Facebook hate group. *First Monday*, *13*(11). https://doi.org/10.5210/fm.v13i11.2254

Oboler, A. (2009). Facebook, Holocaust Denial, and Anti-Semitism 2.0. *Post-Holocaust and antisemitism*, Article No. 86. https://jcpa.org/article/facebook-holocaust-denial-and-anti-semitism-2-0/

Oboler, A. (2012). *Aboriginal memes and online hate*. Online Hate Prevention Institute. https://nla.gov.au/nla.obj-888711478/view

Oboler, A. (2013a). *Islamophobia on the internet: The growth of online hate targeting Muslims*. Online Hate Prevention Institute. https://nla.gov.au/nla.obj-1971792213/view

Oboler, A. (2013b). *Recognizing hate speech: Antisemitism on Facebook*. Online Hate Prevention Institute. http://ohpi.org.au/reports/IR13-1_Recognizing_hate_speech_antisemitism_on_Facebook.pdf

Oboler, A. (2014). *The antisemitic meme of the Jew*. Online Hate Prevention Institute. https://nla.gov.au/nla.obj-888640442/view

Oboler, A. (2016). *Measuring the hate: the state of antisemitism in social media*. Online Hate Prevention Institute. https://nla.gov.au/nla.obj-1971821446/view

Oboler, A. (2022). *Anti-Asian racism in Australian social media*. Online Hate Prevention Institute. https://nla.gov.au/nla.obj-3117746478

Oboler, A., Allington, W., & Scolyer-Gray, P. (2019). *Hate and violent extremism from an online sub-culture: The Yom Kippur terrorist attack in Halle, Germany*. Online Hate Prevention Institute. https://nla.gov.au/nla.obj-2286730824/view

Oboler, A. & Connelly, K. (2014). *Hate speech: A quality of service challenge* IEEE Conference on e-Learning, e-Management and e-Services (IC3e). Hawthorne, VIC, Australia.

Oboler, A. & Matas, D. (2009). *Report of the online antisemitism working group*. https://www.oboler.com/andre-oboler-and-david-matas-report-from-the-working-group-on-online-antisemitism-the-global-forum-to-combat-antisemitism/

Oboler, A. & Matas, D. (2013). *Online antisemitism: A systematic review of the problem, the response and the need for change*. Israeli Ministry of Foreign Affairs. https://www.jewishvirtuallibrary.org/jsource/anti-semitism/onlineantisem2013.pdf

OHPI. (2023, 22 Janaury). *Celebrating 11 years improving online safety*. https://ohpi.org.au/celebrating-11-years-improving-online-safety/

OSCE. (2008). *Hate in the Information Age: Hearing before the U.S Comm'n on Sec. & Cooperation in Europe (U.S. Helsinki Comm'n), 110th Cong. 1-9 (Remarks of Rabbi Cooper, Mark Potok, and Christopher Wolf)*. https://www.csce.gov/sites/helsinkicommission.house.gov/files/Hate%20in%20the%20Information%20Age_0.pdf

Ozalp, S., Williams, M. L., Burnap, P., Liu, H., & Mostafa, M. (2020). Antisemitism on Twitter: Collective efficacy and the role of community organisations in challenging online hate speech. *Social Media + Society*, *6*(2). https://doi.org/10.1177/2056305120916850

Potok, M. Prepared remarks to Hate in the Information Age: Hearing Before the U.S Comm'n on Sec. & Cooperation in Europe (U.S. Helsinki Comm'n), 110th Cong., (2008). https://www.csce.gov/sites/helsinkicommission.house.gov/files/Potok%20Testimony.pdf

Racist, sexist jokes do not compute. (1989). *On Campus With Women*, *19*(1). https://archive.org/details/sim_on-campus-with-women_summer-1989_19_1

Report: Online hate increasing against minorities, says expert. (2021). United Nations Office of the High Commissioner for Human Rights. https://www.ohchr.org/en/stories/2021/03/report-online-hate-increasing-against-minorities-says-expert

Reynders, D. (2022). Factsheet - 7th monitoring round of the Code of Conduct. In E. C. D.-G. f. J. a. Consumers (Ed.).

Rosenfeld, A. H. (Ed.). (2015). *Deciphering the new antisemitism*. Indiana University Press. http://www.jstor.org.ez.library.latrobe.edu.au/stable/j.ctt18crxz7.

Saha, K., Chandrasekharan, E., & De Choudhury, M. (2019). Prevalence and psychological effects of hateful speech in online college communities. *Proc ACM Web Sci Conf, 2019*, 255-264. https://doi.org/10.1145/3292522.3326032

Sasaki, Y. (2007). The truth of the F-measure. In.

Schwarz-Friesel, M. (2018). *Antisemitism 2.0 and the cyberculture of hate: Hostility towards Jews as a cultural constant and collective emotional value in the digital age (short version).* https://web.archive.org/web/20211108211557/https://www.linguistik.tu-berlin.de/fileadmin/fg72/Antisemitism_2.0_short_version_final.pdf

Snyder, T. (2008, 20 February). Anti-Semitism 2.0 going largely unchallenged. *The New York Jewish Week*. https://www.jta.org/2008/02/20/ny/anti-semitism-2-0-going-largely-unchallenged

Tallo, I. (2001). *Racism and xenophobia in cyberspace*. (Doc. 9263). https://assembly.coe.int/nw/xml/XRef/X2H-Xref-ViewHTML.asp?FileID=9540&lang=EN

Timberg, C., Dwoskin, E., & Albergotti, R. (2021, 22 October). Inside Facebook, Jan. 6 violence fueled anger, regret over missed warning signs. *Washington Post*. https://www.washingtonpost.com/technology/2021/10/22/jan-6-capitol-riot-facebook/

Wendling, M. (2015). 2015: The year that angry won the internet. *BBC News*. https://www.bbc.com/news/blogs-trending-35111707

WJC. (2017). *The rise of anti-semitism on social media: Summary of 2016*. World Jewish Congress and Vigo Social Intelligence. https://www.worldjewishcongress.org/download/RVsVZzRXTaZwO41YbzlWwg

WJC. (2018). *Anti-semitic symbols and holocaust denial in social media posts: January 2018*. World Jewish Congress and Vigo Social Intelligence. https://www.worldjewishcongress.org/en/news/holocaust-denial-and-anti-semitism-on-social-media-up-30-percent-in-january-2018-compared-to-2016-wjc-report-finds-2-3-2018

Zakon, R. H. (2018). *Hobbes' internet timeline 25*. https://www.zakon.org/robert/internet/timeline/

Zannettou, S., Finkelstein, J., Bradlyn, B., & Blackburn, J. (2020). A quantitative approach to understanding online antisemitism. In *Proceedings of the International AAAI Conference on Web and Social Media*, *14*(1), 786-797. https://doi.org/10.1609/icwsm.v14i1.7343

HARNESSING ARTIFICIAL INTELLIGENCE TO COMBAT ONLINE HATE: EXPLORING THE CHALLENGES AND OPPORTUNITIES OF LARGE LANGUAGE MODELS IN HATE SPEECH DETECTION

Tharindu Kumarage
/ Arizona State University, USA

Amrita Bhattacharjee
/ Arizona State University, USA

Joshua Garland
/ Arizona State University, USA

Introduction

Online social media platforms have become important channels of communication and sharing information, opinions, and connecting with other individuals and businesses. However, these platforms are also often used for hateful or toxic content, bullying and intimidation, etc. (Poletto et al., 2021). Given the scale of such platforms, hate speech and toxic content detection is a challenge and performing such detection manually is infeasible. This necessitates the use of automated detection systems Del Vigna et al. (2017); Schmidt and Wiegand (2017), which also is a challenge in practice due to the dynamic nature of hate speech. Hate speech can evolve with time, is highly subjective, and may be dependent on the context in which it is expressed MacAvaney et al. (2019).

With the advent of advanced large language models (LLMs), there is growing interest in leveraging these

models for content moderation. Specifically, using them to detect harmful and toxic content online by simply prompting the models. Several recent studies have examined the efficacy of GPT-3 Brown et al. (2020) and GPT-3.5 Huang, Kwak, and An (2023)[1] in detecting hate speech, encompassing both explicit and implicit forms. OpenAI has recently presented in-house experiments demonstrating GPT-4's OpenAI (2023) potential as a content moderator[2]. Similarly, the state-of-the-art open-source model, Llama 2 Touvron et al. (2023), has shown promise in hate speech detection. In this study, our objective is to thoroughly assess these claims and delve into the nuances behind the LLMs' ability to discern hate speech. To achieve this, first explore the space of LLMs as a detector or text classifier, with a focus on the task of hate speech detection. Then, we evaluate several candidate LLMs, spanning both open-source and proprietary models, and address the following research questions:

Q1: How robust are these LLMs in detecting hate speech? We will examine and compare multiple LLMs on various types of hate speech: both general and targeted towards specific minorities. We aim to determine if these LLMs primarily rely on specific keywords, such as profanities, for detection, or if they genuinely discern and characterize the hateful intent of the speech.

Q2: How do various prompting techniques influence the hate speech detection efficacy of LLMs? We will compare different prompting strategies, with varying degrees of complexity, to discern differences in how they affect the hate speech detection capabilities. Based on our findings, we will endeavor to provide insights into the specific elements and nuances of LLMs and best practices surrounding the use of LLMs for this particular task.

LLMs as text classifiers or annotators

Given the availability of several large language models, both open-source and proprietary (albeit via APIs), these technologies are increasingly being

1. ChatGPT and GPT-3.5 are used interchangeably here.
2. https://openai.com/blog/using-gpt-4-for-content-moderation

used in NLP applications such as text classification. Owing to the success of the more recent larger LLMs (such as Chat-GPT, GPT-4 OpenAI (2023), Llama 2 Touvron et al. (2023), etc.), researchers are actively exploring novel use-cases of such models in order to tackle issues such as generalization, data scarcity, etc. In this section we provide a brief overview on how language models (both pre-trained language models, and the more recent large language models) have been used in the task of text classification, first going over the general text classification task, before delving into hate speech specific classifiers.

General text classifier or annotator

In this section, we describe some works that have used language models for the general problem of text classification. We further divide this section into two categories: (i) the pre-LLM era, and (ii) the LLM era.

Pre-LLM Era

In the pre-LLM era, pre-trained language models (PLMs) such as BERT Devlin et al. (2018), RoBERTa Y. Liu et al. (2019), BART Lewis et al. (2019) etc. have been used extensively as language encoders. These PLMs are essentially transformer-based language models that are pre-trained on a large corpus of unlabeled text data (mostly webtext) and often fine-tuned on downstream task datasets to perform classification or detection. Given the extensive pre-training that these language models go through, PLMs are often used as general language encoders in a classification task, with additional classification layers or classification heads added to facilitate task-specific fine-tuning Howard and Ruder (2018); Arslan et al. (2021).

For example, authors in Kant et al. (2018) first pre-train and then fine-tune an encoder- decoder type language model on task specific data for the task of multi-dimensional sentiment classification and compare their method with a pre-trained ELMo Peters et al. (1802), which is then further fine-tuned on their tasks-specific dataset. BERT Devlin et al. (2018), which is

a bidirectional transformer-based language model, has shown impressive performance on many natural language understanding tasks. Authors in Sun et al. (2019) investigate the training regimes and different fine-tuning settings to understand how to get the most out of fine-tuning BERT for the task of text-classification. Through their experiments they advise that text classification using BERT can be improved via the following best practices: further pre-training on task-specific in-domain data, multitask fine-tuning rather than single task fine-tuning etc.

Given the smaller sizes of pre-trained language models as compared to more recent models like ChatGPT or Llama, these models have been used in several other text classification tasks, often with task-specific fine-tuning or in conjunction with other specialized architecture or training regimes Min et al. (2023). Examples of some tasks where such pre-trained language models have been used are toxic comment classification Zhao, Zhang, and Hopfgartner (2021), counter-speech detection Garland et al. (2020, 2022) text mining Zhang et al. (2021), sentiment classification Meng et al. (2020); Rathnayake et al. (2022), etc.

LLM Era

Given the impressive performance of newer LLMs such as ChatGPT and GPT-4 OpenAI (2023) on a variety of natural language tasks, that too in a zero-shot manner, researchers are evaluating the possibility of using such LLMs as annotators. This could potentially assuage data scarcity issues in tasks and thereby facilitate or improve training of better models. One recent work Gilardi, Alizadeh, and Kubli (2023) performed a systematic evaluation of the annotation capabilities of ChatGPT especially in comparison to annotations obtained from crowd workers on Amazon Mechanical Turk[3]. They evaluate the accuracy of ChatGPT and MTurk workers with annotations from trained annotators and show that ChatGPT out-performs the MTurk

3. https://www.mturk.com/

crowd workers, on a variety of content moderation tasks involving different four datasets of Tweets and news articles.

Another recent study Zhu et al. (2023) evaluated the capability of ChatGPT to reproduce human-generated labels on a set of five benchmark text datasets, on tasks such as stance detection, bot detection, sentiment analysis and hate speech detection. Results show that ChatGPT can replicate the human generated labels to a certain extent, achieving an accuracy of 0.609 across the five datasets, but is still far from being a perfect annotator. The authors also find varying performance of ChatGPT across different labels within one specific task. A similar observation has been made by authors in Bhattacharjee and Liu (2023) where ChatGPT was used to distinguish AI-generated text from human-written text, and an asymmetric performance across the two labels was identified. However, experiments demonstrate that GPT-4 has superior performance on the task. A similar work uses ChatGPT in automatic genre classification, where the task is to classify a given text into one of several genre categories such as News, Legal, Promotion, etc. The authors evaluate ChatGPT and compare its performance with a fine-tuned XLM-RoBERTa, and they test on both English and Slovenian language data. Interestingly, for the English split, ChatGPT performs better than the fine-tuned XLM-RoBERTa model, even without any labeled data, although the performance drops a bit for the Slovenian one.

Compared to all these works that demonstrate the potential for using LLMs and, in particular ChatGPT as an annotator, one interesting piece of work Reiss (2023) investigates the reliability of ChatGPT-derived annotations, and demonstrates that the annotations rely heavily on the temperature parameters and possibly other factors such as length of the text prompt and complexity of instructions.

Hate speech classifiers

In this section, we go over recent works that have used language models in a hate speech classification task, and we divide this section into two categories: (i) the pre-LLM era, and (ii) the LLM era.

Pre-LLM Era

Similar to the general classification, early applications of language models in hate speech detection employed pre-trained language models as rich embeddings or representations for the text. Since hate speech detection is often heavily dependent on language-specific words and phrases such as profanities, there have been many efforts in building hate speech classifiers for specific languages. Among methods that use pre-trained language models in the detection framework, some examples are Plaza-del Arco et al. (2021) for Spanish hate speech detection where they use both multilingual pre-trained LMs like mBERT and XLM Lample and Conneau (2019) as well as a Spanish version of BERT called BETO[4]. Authors in Pham et al. (2020) build a detector for Vietnamese hate speech by using a RoBERTa model, or in particular, a version trained for the Vietnamese language called PhoBERT Dat and Tuan (2020). Similar efforts involving detection using multilingual and monolingual versions of BERT or RoBERTa have also been done for Italian hate speech detection Lavergne et al. (2020), where alongside multilingual models, Italian versions such as AlBERTo, PoliBERT and UmBERTo have been used. Similar efforts for training language-specific hate speech detectors by fine-tuning different variants of the BERT family of models have been used in languages such as Marathi Velankar, Patil, and Joshi (2022), Polish Czapla et al. (2019).

Authors in Stappen, Brunn, and Schuller (2020) use frozen pre-trained language models as feature extractors in a framework for cross-lingual hate speech detection. Alongside comparing various framework designs for the task, authors also evaluate their proposed method in zero-shot and few-shot

4. https://github.com/dccuchile/beto

setting with substantial success. Another interesting work in multi-lingual hate speech detection uses a multi-channel BERT Sohn and Lee (2019), i.e., multiple language-specific pre-trained BERT models in parallel to facilitate transfer learning, The authors also experiment with adding additional classification signals by providing translated versions of the input to the classifier. Given that the lack of labeled data in low-resource languages is a major bottleneck in the development of hate speech detectors for these particular languages, Zia et al. (2022) proposed a framework that leverages labeled data from a high-resource language such as English and used a language model based teacher-student framework to perform transfer learning for hate speech detection on a target language, in the absence of target labels. To do this, they first fine-tune a multilingual language model on labeled training data from the source language. Then they use this model to generate pseudo-labels for samples from the target language, by simply predicting in a zero-shot manner. Finally, they use these pseudo-labels to fine-tune a monolingual pre-trained language model to perform hate speech detection on the target language without requiring any labels from the target.

LLM Era

Most of the works discussed above use pre-trained language models of parameter sizes in the range of a few hundred million. However, there is a growing trend towards developing and training larger language models, often with parameter sizes of a few hundred billion. Performance of language models on NLP tasks have shown huge improvements with increase in the scale of these models. These larger models, now often referred to as Large Language Models (LLMs) are trained on huge internet-scale data corpora. Due to their extensive pre-training, LLMs often demonstrate good performance on a variety of tasks even on a zero-shot manner. The standard mode of using these LLMs is via the task of text generation, whereby the user provides a text input as a 'prompt' to the LLM, and the LLM produces some text output conditioned on the input prompt.

Broadly, there are two categories of LLMs: base LLMs - that simply perform the task of next token prediction, essentially performing a text completion task; and instruction- tuned LLMs - where LLMs are specifically trained to follow instructions in the prompt. Instruction-tuned LLMs are useful for a variety of tasks. Examples of such instruction- tuned LLMs are ChatGPT, GPT-4, the Llama family of models, etc. An example of a base LLM is GPT-3 Brown et al. (2020) by OpenAI, with 175 billion parameters.

Authors in Chiu, Collins, and Alexander (2021) evaluate the performance of GPT- 3 Brown et al. (2020) on hate speech detection in a variety of settings, including zero-shot, one-shot (where a single example is provided in the prompt), few-shot (where a small number of samples are provided in the prompt as examples). The authors also evaluate the few- shot performance along with instructions in the prompt wherein a small instruction is also provided in the prompt, specifying what the possible labels are, such as 'sexist', 'racist' or 'neither'. Interestingly, the study finds that GPT-3 performs the best when prompted with- out instructions in a few-shot setting. In a similar direction, alongside experimenting with different prompt structures for this task, Han and Tang (2022) shows how increasing the number of labeled samples in the prompt in the few shot setting improves the performance of GPT-3.

Other recent prompt-based detection methods include Luo et al. (2023), where the authors propose a new category of the hate speech detection task: enforceable hate speech detection, where text content is classified as hate speech if it violates at least one legally enforceable definition of hate speech. For the detection method, the authors present various settings of prompt tuning on a RoBERTa-large model. Prompt-tuning is a new parameter- efficient fine-tuning method that enables fine-tuning of large language models in low- resource settings, by freezing the model weights and updating a small set of parameters instead. Del Arco, Nozza, and Hovy (2023) evaluates zero-shot hate speech detection by simply prompting instruction-tuned models FLAN-T5 Chung et al. (2022) and mT0 Muennighoff et al. (2022), and compare the performance with encoder-based language models such

as the BERT family of models. They perform the evaluation on an extensive collection of 8 benchmark datasets containing online hate speech. Their results show that the instruction-tuned models have superior performance.

Recently, the accessibility and ease of use of ChatGPT, along with its impressive performance has inspired a series of interesting exploratory efforts into using ChatGPT as a detector for many NLP tasks. Along this direction, authors in Huang et al. (2023) have experimented with ChatGPT to understand how well it can detect implicit hate speech in Tweets, and also whether it can provide explanations for the reasoning. Their experiments demonstrate that ChatGPT has the potential to be used for such subjective tasks such as implicit hate speech detection. Furthermore, ChatGPT generated explanations also appear to have more clarity than human-written explanations, although there was no significant difference in informativeness. ChatGPT has also been evaluated for language-specific hate speech detection in Portuguese Oliveira et al. (2023) and results show that even without any fine-tuning, ChatGPT performs well in the detection task.

Empirical analysis

In this section, we undertake several experiments utilizing representative LLMs to empirically assess their proficiency in identifying hate speech. Through these experiments, we address two primary research questions:

RQ1: How robust are LLMs in classifying hate speech?

RQ2: How do various prompting techniques influence the hate speech detection efficacy of LLMs?

Experiment design

In this subsection, we delve into the details of our experimental design, highlighting the critical decisions made to address the stated research questions. Paramount among these decisions were the choice of LLMs as the hate speech detector(classifier) and the dataset selection to rigorously assess the robustness of the chosen LLMs in detecting hate speech.

LLM selection

As mentioned in Section 2, numerous advanced LLMs are currently available, encompassing both open-source and proprietary options. From the open-source category, we opted for the Llama-2 model (7B parameters chat variant[5]) and the Falcon model (7B parameters chat variant[6]) based on their notable standings on the Open LLM Leaderboard hosted by Hugging Face[7] . For proprietary models, we selected GPT 3.5, which serves as a foundational benchmark for subsequent LLMs, including GPT 4. To enhance the reproducibility of our results, we utilized the specific version "gpt-3.5-turbo-0613." The experiments involving the chosen LLMs—Llama-2, Falcon, and GPT 3.5—were conducted between September 25th and October 8th, 2023.

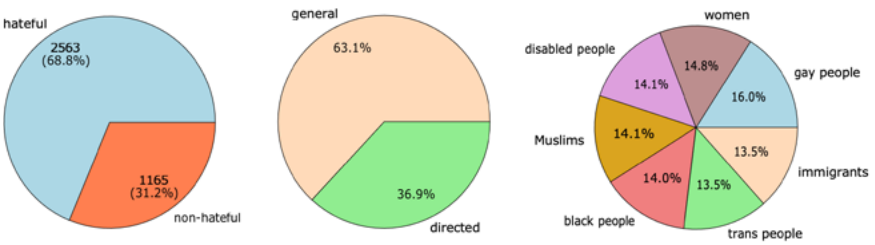
Dataset selection

The primary objective of this study is to assess the robustness and reliability of selected LLM classifiers in identifying real-world hate speech. To achieve this, our evaluation criteria encompass: i) determining the classifier's capability to recognize general hate speech concepts; ii) gauging its sensitivity to hate speech directed at specific groups, such as women or Muslims; and iii) ascertaining any biases the classifier might exhibit towards extraneous features, like swear words or group identifiers.

Consequently, we chose the HateCheck dataset Röttger et al. (2021) as the primary hate speech dataset to evaluate the LLM classifiers. This dataset is meticulously designed to serve as a diagnostic tool for hate speech detection systems, aligning perfectly with our evaluation criteria. It features annotations that categorize hate speech as either 'directed' or 'general' and identifies the target of the hate. Furthermore, the non-hateful examples within this dataset are constructed to probe potential biases in hate speech classifiers towards spurious correlations. Consider the example: "They

5. https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
6. https://huggingface.co/tiiuae/falcon-7b-instruct
7. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

are building a new dyke near the beach." The aim here is to discern if the classifier exhibits a bias towards the term "dyke" and misclassifies these examples as hate speech. For a comprehensive overview of the dataset statistics, refer to Figure 1.



(a) Label distribution (b) Directed vs general hate examples (c) Hate target distribution

Figure 1: HateCheck dataset statistics

Hate speech detection

The concluding aspect of our experimental design pertains to the methodology detailing our hate speech classification approach using the chosen LLMs. We employed the LLMs in a zero-shot fashion (i.e., without specific training on hate speech datasets), presenting them with examples from the HateCheck dataset and instructing each LLM to determine the hatefulness of the input text. Consequently, it is imperative to discuss both the *input prompt* utilized to query the LLM and the *labeling method* adopted to categorize the LLM's response and decide the final classification label.

Classification prompt

The input prompt is integral to LLM-based classifiers. Its primary objective is to concisely direct the LLM to analyze a provided text and determine the appropriate label, signifying whether the content is hateful or non-hateful. In our study, we incorporated the following prompt:

Prompt

Is the following text hateful or not? Just answer in Yes or No. Don't provide explanations.

text: {hate_speech}

Labeling method and caveats

The labeling method is employed to translate the text output of the LLM into binary class labels: 1 ('hate') and 0 ('non-hate'). When the LLM explicitly responds with 'Yes' or 'No', the label mapping process is straightforward. However, some scenarios necessitate a more nuanced approach to categorize the output:

- Caveat 1: Deviation from Instructions: LLMs occasionally diverge from the provided directives and offer explanations alongside the label. In these instances, we manually reviewed the diverse, unique outputs, determining the appropriate labels grounded in keywords like 'Yes,' 'hateful,' 'No,' and 'not hateful.'
- Caveat 2: Activation of LLM Guardrails: Certain examples within the HateCheck dataset activate the LLM's built-in guardrails, designed to identify and mitigate hateful or offensive content processing. When these guardrails are triggered, the LLM yields a message indicating the presence of hate or offensive language, leading us to categorize such instances as hate speech.

Experiment results

RQ1: LLM's hate classification performance

Table 1 displays the efficacy of selected LLMs in classifying hate speech, using data from the HateCheck dataset. The performance metrics, derived from direct prompt outcomes, reveal that both GPT-3.5 and Llama 2 exhibit

commendable efficiency, with accuracy and F1 scores ranging between 80-90%. This underscores their proficiency in identifying hate speech. GPT-3.5 outperforms the others, an expected outcome given it has benefited from numerous advanced iterations of Reinforcement Learning from Human Feedback (RLHF) (from November 2022 onwards), and it contains more parameters than the other LLMs we considered. In contrast, Llama 2, despite its smaller 7B parameter model, delivers a performance that nearly matches GPT-3.5. The Falcon model, however, demonstrates inferior classification, performing below the level of random guessing. This disparity in performance between Llama 2 and Falcon can be attributed to the specific tuning conducted to optimize their pre-trained versions for chat compatibility. Another potential explanation is that the Llama 2 authors deliberately retained toxic data during pre-training to enhance downstream task generalization Touvron et al. (2023), positioning it as a more adept hate speech classifier than the Falcon model.

Error analysis

We conducted an error analysis to delve into the challenges the existing LLMs face in identifying hate speech and to pinpoint specific contexts where these models struggle to discern hate speech effectively. For this examination, we utilized the directionality annotations and target annotations from the HateCheck dataset. Within the realm of directionality, we assessed the proportion of misclassified hate speech samples, distinguishing between errors in identifying directed hate speech and those in discerning general hate speech. As shown in Table 1, both Llama 2 and Falcon have equal error rates for directed and general hate speech, suggesting that these models possess comparable proficiency in detecting both types of hate speech. In contrast, GPT 3.5 exhibits a higher error rate for directed hate speech than general hate speech. Subsequently, we assessed the error rates of the LLMs concerning different hate targets. The objective of this segment was to ascertain which target-associated hate speech poses the most significant detection challenges for the LLMs. As demonstrated in Table 2, the error

rates for Llama 2 and Falcon regarding specific targets largely mirror the original distribution of these targets in the dataset. However, GPT 3.5 exhibits a disproportionately elevated error rate when identifying hate speech related to "women."

Performance attributed to spurious correlations rather than proper reasoning

It is crucial to examine whether the notable classification performance of LLMs can be attributed to spurious correlations, such as categorizing a text as hate speech based solely on the presence of swear words or group identifiers, rather than substantive reasoning. This consideration is facilitated by the non-hate examples included in the HateCheck dataset, which contains elements like swear words and group identifiers used in non-hateful con- texts. Evaluating the performance of LLMs in classifying these "non-hate" examples is essential to confirm their reliability as hate speech classifiers. As detailed in Table 1, although Llama 2 demonstrates impressive classification accuracy for "hate" content, its performance diminishes in identifying non-hateful content, suggesting a reliance on spurious correlations. Conversely, GPT 3.5 maintains robust performance in classifying both "hate" and "non-hate" content.

LLM	Hate Class			Non-Hate Class			Overall	
	P	R	F1	P	R	F1	Accuracy	AUROC
Falcon	0.69	0.43	0.53	0.3	0.56	0.4	0.47	0.49
Llama 2	0.80	**1.00**	0.89	**0.99**	0.46	0.63	0.83	0.73
GPT 3.5	**0.89**	0.98	**0.93**	0.93	**0.73**	**0.82**	**0.89**	**0.85**

Table 1: Hate classification results: Precision(P), Recall(R), F1-score(F1) values are recorded for both "Hate" and "Non-Hate" classes. Highest performance under each metric is in **bold**.

LLM	Directionality		Hate Target						
	Directed	**General**	**Women**	**Gay**	**Immigrants**	**Trans**	**Black**	**Muslims**	**Disabled**
Falcon	**53.7**	**59.0**	14.1	13.0	**14.3**	**13.1**	**15.4**	**15.6**	**14.6**
Llama 2	0.2	0.2	9.7	**15.6**	5.4	5.9	12.0	8.2	9.0
GPT 3.5	0.6	0.3	**47.6**	7.9	14.2	6.3	3.2	14.2	6.3

Table 2: Error analysis: error rate (%) under "directionality" and "hate-target". Highest error rate under each category is in **bold**.

We further investigated the specific types of spurious correlations influencing these LLMs using the functionality annotations of the HateCheck dataset. These annotations identify various categories of spurious correlations scenarios evident in non-hateful content, including "slur", "profanity", "negate hateful statements", "group identifiers", "countering of hate speech through quoting or referencing hate speech examples" and "abuse targeted at objects, individuals, and non-protected groups." As illustrated in Figure 2, Llama 2 exhibits more errors attributed to spurious correlations, further underlining its diminished performance in classifying the 'non-hate' category. Both Llama 2 and GPT 3.5 display heightened inaccuracies in distinguishing examples that counteract hate speech by referencing or quoting hate speech instances. This augmented error rate may be, in part, due to the labeling function, where specific counter-speech scenarios could trigger the LLM guardrails. As a result, the labeling function might mistakenly assume that the LLM's response to these examples implies a hate label. This underscores the significance of adequately addressing such scenarios when integrating LLMs into real-world hate speech detection frameworks.

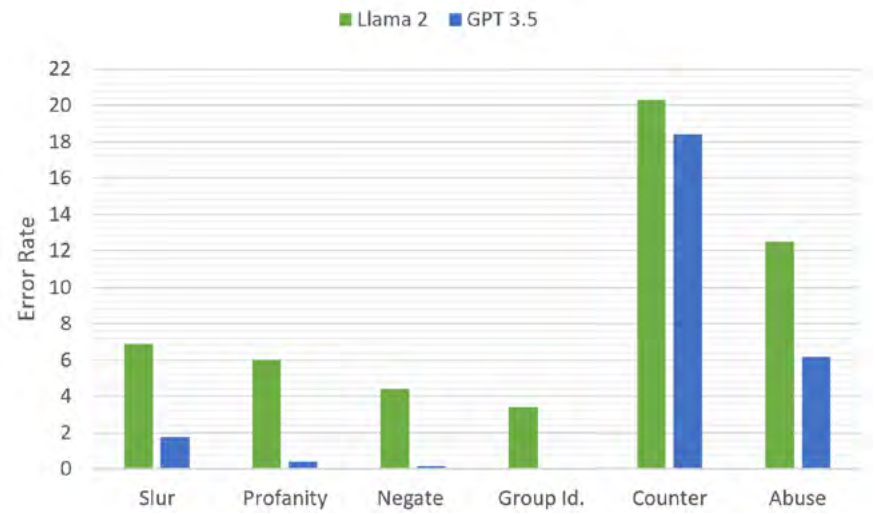


Figure 2: Error analysis on non-hate class

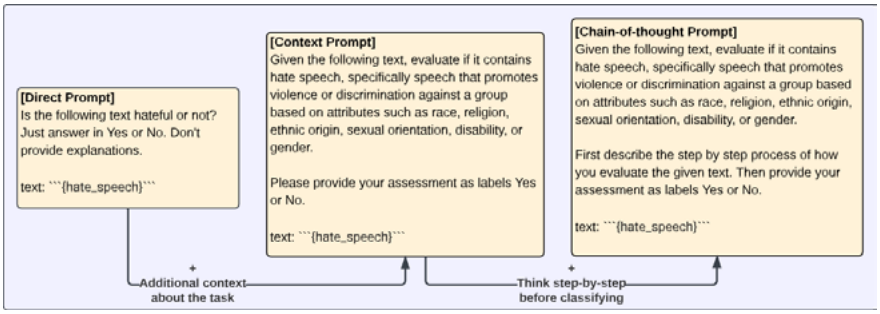


Figure 3: Prompt templates used for hate speech classification

RQ2: Effect of prompting

The input prompt plays an indispensable role in LLM-based classifiers. Generally, the efficacy of an LLM in classifying text is intrinsically tied to the quality of the input prompt. In light of this, we conducted an extended experiment involving the top-performing LLM, GPT 3.5, to explore the impact of various prompts on classification performance. As illustrated in Figure 3, we introduced two additional prompt types, referred to as *context prompt*, and *chain-of-thought(COT) prompt*.

Table 3 presents the classification results of GPT 3.5 using different prompts employed in our study. Intuitively, we anticipated the performance of the LLM classifier to improve as we transitioned through the prompts from left to right in Figure 3, particularly given the additional context and incorporation of the COT method. However, unexpectedly, the direct concise prompt yielded the most superior performance out of the three prompts. One potential rationale for this result is that an overly complex prompt, paired with the inherently intricate nature of hate speech detection, might obscure the LLM's understanding of the task rather than clarifying it. Another explanation aligns with recent findings on LLMs, suggesting that performance peaks when vital information is positioned at the beginning or end of the input context and diminishes substantially when models must retrieve relevant information from the middle of lengthy contexts N. F. Liu et al. (2023).

Discussion

In addressing the two research questions posed, our findings offer significant insights into the robustness and nuances of LLMs in hate speech classification.

Answering RQ1: LLM's robustness in classifying hate speech

For **RQ1**, the GPT-3.5 and Llama 2 models proved their robustness in classifying hate speech, boasting accuracy and F1 scores between 80-90%. Despite its fewer parameters, Llama 2 nearly matches the performance of GPT-3.5, although GPT-3.5 remains superior. We attribute this to its advanced RLHF iterations and larger parameter size. Falcon, conversely, demonstrated subpar performance, indicating its unsuitability for reliable hate speech classification. The error analysis further enriched our understanding. While Llama 2 and Falcon demonstrated equal proficiency in detecting directed and general hate speech, GPT-3.5 showed a higher error rate for directed hate speech. Additionally, it exhibited an increased error rate in identifying hate speech targeted at women, indicating potential areas for improvement in its training and calibration. Llama 2's diminished

performance in classifying non-hateful content hinted at its reliance on spurious correlations. Both Llama 2 and GPT-3.5 were challenged in scenarios involving the counteraction of hate speech through referencing or quoting hateful content, pinpointing a need to refine the LLMs' handling of such contexts.

Answering RQ2: Influence of prompting techniques

As for **RQ2**, the efficacy of LLMs is notably influenced by the employed prompting techniques. Contrary to our anticipation that more complex prompts (such as context and chain-of-thought prompts) would enhance classification performance, the direct concise prompts delivered best results. It suggests that simplicity and conciseness in prompts might facilitate clearer hate speech detection task comprehension for LLMs, leading to more accurate classifications.

Prompt	Hate Class			Non-Hate Class			Overall	
	P	R	F1	P	R	F1	Accuracy	AUROC
Direct	0.89	**0.98**	**0.93**	**0.93**	0.73	**0.82**	**0.89**	**0.85**
Context	**0.91**	0.85	0.88	0.71	**0.82**	0.76	0.84	0.83
COT	0.88	0.81	0.84	0.69	0.79	0.74	0.80	0.79

Table 3: GPT 3.5's hate classification results with different prompts: Precision(P), Recall(R), F1-score(F1) values are recorded for both "Hate" and "Non-Hate" classes. Highest performance under each metric is in **bold**.

Best practices and pro tips

Optimizing LLM performance

When utilizing LLMs as hate speech classifiers, certain practices can optimize their performance and reliability.

- **Select Appropriate LLMs**: GPT-3.5 and Llama 2 have shown notable efficacy; however, it's crucial to consider the specific needs and contexts

of the application. Evaluate multiple models to identify which offers the best balance of accuracy and computational efficiency.

- **Input Prompt**: Direct and concise prompts have been shown to be more effective. Avoid overly complex prompts that could potentially confuse the model or dilute the task's clarity. Experiment with various prompt designs to identify which yields optimal performance for the specific LLM and classification task.

- **Error Analysis**: Conduct detailed error analyses to identify specific areas where the LLM struggles, and consider this information when fine-tuning or selecting models for deployment.

- **Labeling Function**: The labeling function plays a pivotal role in the performance of LLMs in classification tasks. It's essential to optimize and test various labeling functions to ensure that they are accurate and reliable, avoiding misclassifications especially in complex scenarios like counter-speech.

Mitigating the influence of spurious correlations

The risk of LLMs relying on spurious correlations, as observed with Llama 2, underscores the necessity of specific strategies to mitigate such influences.

- **Balanced Fine-tuning**: Conduct additional fine-tuning of the LLM with balanced training data that includes diverse examples of hate speech and non-hate speech, reducing the model's reliance on specific words or phrases as indicators of hate speech.

- **Functionality Annotations**: Leverage functionality annotations to identify and analyze potential spurious correlations, enabling the refinement of the model's classification capabilities.

- **Real-world Testing**: Test the LLMs in real-world scenarios to assess their performance beyond controlled experiments. Adapt and refine the models continuously based on the emerging data and classification challenges.

Incorporating these insights and practices will be instrumental in enhancing the reliability, accuracy, and fairness of LLMs in hate speech classification, ensuring they are a valuable tool in combating online hate while preserving freedom of expression.

Conclusion

In our study, we provided a detailed look into the progression of language models for hate speech classification, from the days of pre-LLMs to the modern era of sophisticated LLMs like GPT. Earlier language models, often needed significant fine-tuning to work well, but new LLMs, like GPT-3.5 and Llama 2, have shown they can be effective at identifying some forms of hate speech right out of the box, even in zero and few shot settings.

We explored the capabilities of three LLMs, GPT-3.5, Llama 2 and Falcon, on the HateCheck dataset to gain deeper insights into their abilities and challenges in classifying hate speech. From our experiments, a few key points stood out: GPT-3.5 and Llama-2 were quite effective overall with accuracy levels between 80-90%, but Falcon lagged behind considerably. As we discussed, this may be an artifact of what data was used to train Falcon. When we looked into the nuances of hate speech, like understanding who the hate was directed at, all of these models faced challenges and their performance declined considerably. For instance, GPT 3.5 struggled particularly with recognizing hate directed towards women. We also found through experimentation that clear and straightforward prompts worked best, hinting that simplicity of classification instructions may be key for effective classification performance.

Hate speech classification remains a challenging area for many reasons, not just due to its nuanced nature but also the ethical concerns around data collection and especially labeling. LLMs, even in zero and few shot settings, present a potential exciting way forward. While they are promising, there is still much to understand and refine. We hope our findings and recommendations from this study offer a useful guide for those looking to delve

further into the capabilities of LLMs for managing online hate. Forging towards a safer, more inclusive digital landscape for everyone.

References

Arslan, Y., et al. (2021). A comparison of pre-trained language models for multi-class text classification in the financial domain. *Companion proceedings of the web conference 2021* (pp. 260–268).

Bhattacharjee, A. & Liu, H. (2023). Fighting fire with fire: Can chatgpt detect ai-generated text? *arXiv preprint arXiv:2308.01284*.

Brown, T. et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Chiu, K.-L., Collins, A., & Alexander, R. (2021). Detecting hate speech with gpt-3. *arXiv preprint arXiv:2103.12407*.

Chung, H. W. et al. (2022). Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Czapla, P. et al. (2019). Universal language model fine-tuning for polish hate speech detection. *Proceedings ofthePolEval2019Workshop*, 149.

Dat, N. & Tuan, N. (2020). Phobert: Pre-trained language models for vietnamese. *Findings of the Association for Computational Linguistics: EMNLP*, *2020*, 1037–1042.

Del Arco, F. M. P., Nozza, D., & Hovy, D. (2023). Respectful or toxic? Using zero-shot learning with language models to detect hate speech. *The 7th workshop on online abuse and harms (woah)* (pp. 60–68).

Del Vigna, F. et al. (2017). Hate me, hate me not: Hate speech detection on facebook. *Proceedings of the first italian conference on cybersecurity (itasec17)* (pp. 86–95).

Devlin, J. et al. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Garland, J. et al. (2020, November). Countering hate on social media: Large scale classi- fication of hate and counter speech. *Proceedings of the fourth workshop on online abuse and harms* (pp. 102–112). Association for Computational Linguistics.

Garland, J. et al. (2022). Impact and dynamics of hate and counter speech online. *EPJ Data Science*, *11*(1), 3.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

Han, L. & Tang, H. (2022). Designing of prompts for hate speech recognition with in- context learning. *2022 International Conference on Computational Science and Computational Intelligence (csci)* (pp. 319–320).

Howard, J. & Ruder, S. (2018). Universal language model fine-tuning for text classifica- tion. *arXiv preprint arXiv:1801.06146*.

Huang, F., Kwak, H., & An, J. (2023). Is chatgpt better than human annotators? Potential and limitations of chatgpt in explaining implicit hate speech. *arXiv preprint arXiv:2302.07736*.

Kant, N. et al. (2018). Practical text classification with large pre-trained language models. *arXiv preprint arXiv:1812.01207*.

Lample, G. & Conneau, A. (2019). Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Lavergne, E. et al. (2020). Thenorth@ haspeede 2: Bert-based language model fine-tuning for italian hate speech detection. *7th evaluation campaign of natural language processing and speech tools for italian. final workshop, evalita* (Vol. 2765).

Lewis, M. et al. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Liu, N. F. et al. (2023). Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Liu, Y. et al. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Luo, C. F. et al. (2023). Towards legally enforceable hate speech detection for public forums. *arXiv preprint arXiv:2305.13677*.

MacAvaney, S. et al. (2019). Hate speech detection: Challenges and solutions. *PloS One*, *14*(8), e0221152.

Meng, Y. et al. (2020). Text classification using label names only: A language model self-training approach. *arXiv preprint arXiv:2010.07245*.

Min, B. et al. (2023). Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, *56*(2), 1–40.

Muennighoff, N. et al. (2022). Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Oliveira, A. S., et al. (2023). How good is chatgpt for detecting hate speech in portuguese? *Anais do xiv simpo´sio brasileiro de tecnologia da informac¸a˜o e da linguagem humana* (pp. 94–103).

OpenAI, R. (2023). Gpt-4 technical report. *arXiv*, 2303–08774.

Peters, M. E. et al. (1802). Deep contextualized word representations. corr abs/1802.05365 (2018). *arXiv preprint arXiv:1802.05365*.

Pham, Q. H. et al. (2020). From universal language model to downstream task: Improving roberta-based vietnamese hate speech detection. *2020 12th International Conference on Knowledge and Systems Engineering (kse)* (pp. 37–42).

Plaza-del Arco, F. M. et al. (2021). Comparing pre-trained language models for spanish hate speech detection. *Expert Systems with Applications*, *166*, 114120.

Poletto, F. et al. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, *55*, 477–523.

Rathnayake, H. et al. (2022). Adapter-based fine-tuning of pre-trained multilingual lan- guage models for code-mixed and code-switched text classification. *Knowledge and Information Systems*, *64*(7), 1937–1966.

Reiss, M. V. (2023). Testing the reliability of chatgpt for text annotation and classification: A cautionary remark. *arXiv preprint arXiv:2304.11085*.

Röttger, P. et al. (2021, August). HateCheck: Functional tests for hate speech detection models. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long papers)* (pp. 41–58). Association for Computational Linguistics. https://aclanthology.org/2021.acl-long.4 doi: 10.18653/v1/2021.acl-long.4

Schmidt, A. & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. *Proceedings of the fifth international workshop on natural language processing for social media* (pp. 1–10).

Sohn, H. & Lee, H. (2019). Mc-bert4hate: Hate speech detection using multi-channel bert for different languages and translations. *2019 International Conference on Data Mining Workshops (icdmw)* (pp. 551–559).

Stappen, L., Brunn, F. & Schuller, B. (2020). Cross-lingual zero-and few-shot hate speech detection utilising frozen transformer language models and axel. *arXiv preprint arXiv:2004.13850*.

Sun, C. et al. (2019). How to fine-tune bert for text classification? *Chinese Computational Linguistics: 18th China National Conference, ccl 2019, Kunming, China, October 18–20, 2019, Proceedings 18* (pp. 194–206).

Touvron, H. et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Velankar, A., Patil, H., & Joshi, R. (2022). Mono vs multilingual bert for hate speech detection and text classification: A case study in marathi. *Iapr workshop on artificial neural networks in pattern recognition* (pp. 121–128).

Zhang, T. et al. (2021). Smedbert: A knowledge-enhanced pre-trained language model with structured semantics for medical text mining. *arXiv preprint arXiv:2108.08983*.

Zhao, Z., Zhang, Z., & Hopfgartner, F. (2021). A comparative study of using pre-trained language models for toxic comment classification. *Companion proceedings of the web conference 2021* (pp. 500–507).

Zhu, Y. et al. (2023). Can chatgpt reproduce human-generated labels? a study of social computing tasks. *arXiv preprint arXiv:2304.10145*.

Zia, H. B. et al. (2022). Improving zero-shot cross-lingual hate speech detection with pseudo-label fine-tuning of transformer language models. *Proceedings of the in- ternational aaai conference on web and social media* (Vol. 16, pp. 1435–1439).

MAPPING THE HATE SPEECH ON TWITTER: POLITICAL ATTACKS ON JOURNALIST PATRÍCIA CAMPOS MELLO

Fábio Malini
/ Federal University of Espírito Santo, Brazil

Jéssica do Nascimento Oliveira
/ Federal University of Espírito Santo, Brazil

Gabriel Herkenhoff Coelho Moura
/ Federal University of Espírito Santo, Brazil

Introduction

In the process of constructing discursive productions during a social interaction, the image of each individual can be preserved or attacked. The digital sphere has been fertile ground for this type of reflection due to the fact that social media platforms, each with their own particularities, are marked by the connection between profiles, a connection marked by reverence or repulsion, by *likes* or *dislikes*. In such a context, there must be a particular look when it comes to the representations of women, considering that the historical social construction of the image of women as an *absolute other*, in the expression of Simone de Beauvoir (1967 [1949]), continues to stimulate asymmetrical relationships that result in attitudes that deny women's dignity and, in more serious cases, violent attacks. It is noticeable, therefore, that the formation of some categories are not intertwined in the discursive material that is exposed, but constructed in the interpretative manifestations of the digital. We mean, people shape and develop these

categories through their own interpretations, perceptions, and attitudes when consuming and interacting with online content. These interpretations can be influenced by gender stereotypes, prejudices, ideologies and other subjective influences. Ultimately, categories related to gender, especially those that affect women, are not simply an objective representation of reality, but are influenced by people's interpretation and are often subject to distortions, prejudices and stereotypes that are amplified in the digital environment.

We understand that hate speech can present particularities that aim not only to demoralize, but also to build a negative image of a specific person. Considering the journalism-society interface, we note that attitudes and expressions of violence remain close to journalistic practice. According to the annual report of the National Federation of Journalists, 2020 was the most violent year for Brazilian journalists since the beginning of the historical series of records of attacks on press freedom made by the Federation, which began in the 1990s.[1] According to the same report, "the explosion of cases is associated with the systematic action of the President of the Republic, Jair Bolsonaro, to discredit the press and with the action of his supporters against media outlets and journalists" (FENAJ, 2020: 6). In this sense, this research focuses on the attacks suffered by journalist Patrícia Campos Mello, from Folha de São Paulo, motivated by Hans River's testimony to the Joint Parliamentary Commission of Inquiry (in Portuguese, Comissão Parlamentar Mista de Inquérito ou CPMI) on Fake News.

Hans, a former employee of *Yacows*, a digital marketing company that worked on the campaign of Jair Messias Bolsonaro during the 2018 presidential election, accused the reporter of offering him sex in exchange for information. Bolsonaro, president at that time and target of the investigation led by Campos Mello, used the deponent's speech and, making the

1. On this topic, see: https://www.abraji.org.br/abraji-aponta-que-mulheres-jornalistas-foram-vitimas-de-mais-da-metade-das-agressoes-no-meio-digital. Last access: 11/12/2022.

statement "She wanted a scoop. She wanted to scoop the scoop at any price against me", he stimulated numerous sexist comments and sexual insinuations against the journalist on social media.

Based on this episode, we aimed, through the analysis of comments (tweets and retweets) on the Twitter platform and prior observation of the material collected with the help of the Ford software developed by the Image and Cyberculture Studies Laboratory (Labic) at the Federal University of Espírito Santo (Ufes), identify words and expressions that characterizes hate speech directed against women in the digital environment. After filtering the content of the most shared tweets in the database, we labeled the linguistic material identified using the network discourse perspectives method (Malini, 2015). The team of coders grouped these words/expressions into categories depending on the context in which they were used and on their characteristics. Then, this supervised database was applied to all the rest of the retweets, taking topic modeling and machine learning techniques for classification algorithms as a starting point. It was also possible to highlight some evidence that contributes to delineating the concept of hate speech and to discuss how this speech is operated against women.

The present work is therefore in the intersection between the fields of Linguistics, Communication Studies and Data Science in order to reflect on an important topic in contemporary society: the profusion of hate speech on social media platforms. In our case, it involves discussing violent speech acts against a woman, journalist Patrícia Campos Mello, on Twitter. The methodological contribution of this work is to apply the perspectivist method of Social Network Analysis (Malini, 2015), combined with Digital Discourse Analysis (Paveau, 2017/2021), to reflect on topics in the field of Linguistics. Furthermore, this work contributes to the disclosure of power relations created through language and dominant ideologies in situations of violence against women in the digital environment.

Methodology

On a general level, the methodologies used in this work were bibliographical review, with the aim of understanding the field of insertion of this work, and descriptive-exploratory analysis, that is, the presentation of and deepening into our research object. From a more specific point of view, the methodological procedures used to construct the analysis material for this study are divided into three stages: 1) creating the dataset, 2) delimiting the *corpus*, and 3) modeling topics.

Creating the dataset

During the period from February 7 to 15, 2020, 418,891 posts were extracted from Twitter, directly from the platform's Search API (an acronym for Application Programming Interface, is a programming interface that allows the development of applications linked to a specific platform. The API establishes a set of standards that allows developers to access part of the internal structure of a platform), using a script in Python language, which, in addition to collecting data on Twitter, also generates statistical, textual and relational files (these for plotting in graph visualization software).

The choice of terms for data collection in the API was based on the following keyword (all of them in Portuguese) classes:

> (class 1) the target of hateful offenses, in this case, journalist Patrícia Campos Mello. The search terms were "patricia Campos Melo", Camposmello, "Patricia Campos Mello", "Campos Mello".
>
> (class 2) the author of the statement against Patrícia, namely: the businessman Hans River. The search terms were hans+patricia[2], hans+journalista.

2. According to Twitter API standards, the use of the plus symbol (+) allows the detection of posts that contain the queried terms in any position in a message. Thus, if one uses the query 'rio+janeiro', the API will return tweets that contain expressions such as "Rio de Janeiro" or "rio, de janeiro a janeiro".

(class 3) the lexicons of the sexist offense made by River in the National Congress. The search terms were: hans+sexo [hans+sex], foda+furo+-fake [fuck+scoop+fake], meretrizes+furo+busca [whore+scoop+search], xerecard+folha [pussycard+folha], foda+sao+paulo [fuck+sao+paulo], ela+dar+furo [she+give+scoop], patriciaxerecard, metodo+foia [method+foia], jornalista+folha+falsa+ultrapassou [journalist+folha+-fake+surpassed].

We call the first two classes of queries "multiverse-words" (Malini, 2020), as they are terms and expressions that encompass a vast array of topics and positions, making it impossible for them to be named as belonging to just one social group. They are otherwise cemented in the everyday life itself, and therefore almost always required to mediate participation in the public discursive arena of networks in the topic that is this work object, by continually evoking and notifying new situations that demand debates, announcements, statements and denunciations. Furthermore, this group of words presents marks that may reveal more segmented relationships between the actors, considering that they are not linked exclusively to an identity or preference (Malini, 2020).

All research, whose analysis is based on data science techniques, depends on understanding a multiverse that escapes an identity corpus, typical of hashtags, to expand the voices and approaches to the topic of the research on social media.

At the other end, the terms in class 3 refer to "attack-words", coined in order to reproduce the feeling of hatred of a social group and to give meaning to a narrative that aims to subjugate and humiliate others in the digital public arena. In the view of Paveau (2021), the elements linked to cyberviolence, taking into account a linguistic typology and enunciative organization of flaming (hostile interaction between users of digital social media), are marked by direct addresses in the second person. Furthermore, the *techno-linguistic* issue is of a socio-discursive nature, i.e., it takes into account the parameters of acceptability of speeches in a digital environment and the role

of impostors in the elaboration of speeches. It is also pragmatic, as it seeks to verify the effects of violent speeches in this environment. In the case studied, the control group, the flamers, is formed by supporters of President Jair Bolsonaro, responsible for spreading (through retweets and replies) the hashtag #xerecard [pussycard], which associated, in a defamatory way, the Folha de São Paulo journalist with a prostitute looking for information.

During the 2018 electoral dispute for the presidency of the Republic, the journalist published a report on the mass triggering of messages made on WhatsApp to benefit certain political groups. On February 18, 2020, the then-president of Brazil, Jair Messias Bolsonaro, insulted Patrícia Campos Mello during an interview in front of the Palácio da Alvorada (Official residence of the President of Brazil).

Bolsonaro, in front of a group of journalists, relied on the testimony given to CPMI (Joint Parliamentary Commission of Inquiry (CPMI) of the National Congress whose purpose is to investigate cyber attacks that undermine democracy and public debate, in addition to the use of fake profiles to influence the results of the 2018 elections) on Fake News by Hans River do Rio Nascimento, a former employee of *Yacows*, one of the digital marketing companies responsible for bulk messaging services via app of messages. The deponent accused Campos Mello of offering him sex in exchange for information to compose the report: Hans River said that she "wanted a certain type of material in exchange for sex" (2020: n.p.).

In his speech during the interview, Bolsonaro alludes to River's accusation and inserts a sexual insinuation: "She wanted a scoop. She wanted to scoop the scoop at any price against me" (Bolsonaro, 2020: n.p.). In the journalistic context, the word "scoop" (in Portuguese, "furo", whose literal translation would be "hole") is a jargon used to name information published exclusively in a media outlet. However, when using the expression, the president emphasized the double meaning of the word, making "scoop" (i.e., "furo") refer to the female sexual organ. The speech, intended to offend and delegitimize the journalist's reputation, resulted in a series of attacks against her on Twitter.

The expected effect of attacking the journalist was, above all, to construct a political scenario that was capable of discrediting the content of the article written by her, in a way that would question the legitimacy of the content and, thus, benefit Jair Bolsonaro before public opinion.

In misogynistic practices, one of the ways to discredit the work done by a woman is to associate her with sexual conduct seen as inconvenient by society. As a form of attack and humiliation, Hans insinuated that the journalist offered him sex as payment in exchange for information to write the story, producing thus the image of a morally corrupt woman.

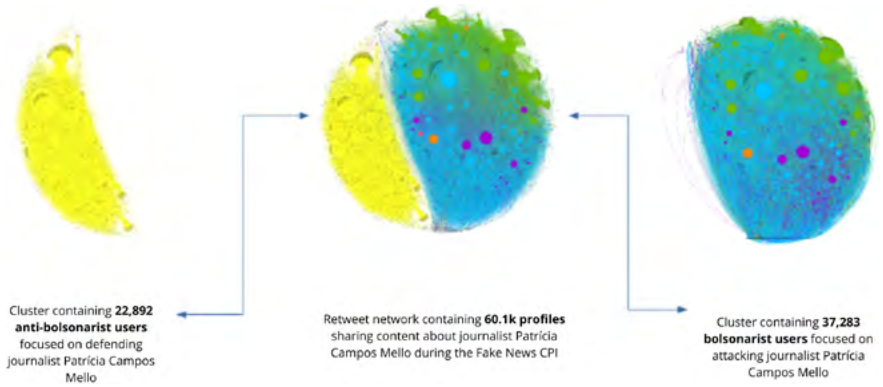
When analyzing a mobilization of violence against others on Twitter, a possible task for the researcher is to map the semantic indexing strategies that create and encourage users to act like online gangs. Every account that verbally abuses another on Twitter presents itself, in fact, as a network of accounts, acting collectively in the name of a practice of power that seeks to inferiorize and silence others.

Corpus

To detach the discursive material of this Bolsonarism control group, we had to filter the dataset using a clustering technique of the profiles. To do this, with the help of the software Ford (developed by the Internet and Data Science Laboratory to refine big datasets), we separated all tweets classified as retweets, drawing a matrix containing two columns, source and target, corresponding, respectively, to the profile that replicates the post (source) and the author whose message is replicated (target). This matrix was transformed into a file with the extension *.gdf, which, when plotted in the Gephi[3] software, allows the visualization of a graph, whose representation is made up of points (profiles) and lines (retweets), as can be seen in Figure 1.

3. Free open-source software for visualizing, analyzing and manipulating networks and graphs.

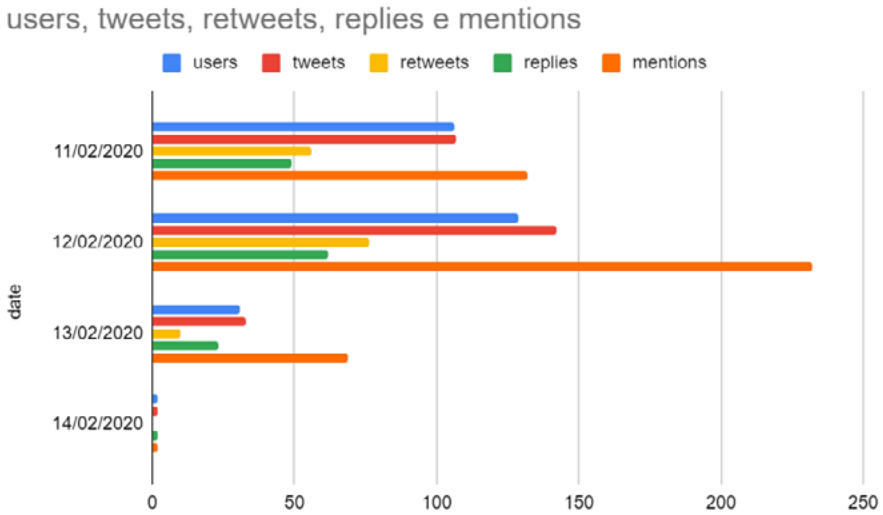
Figure 1: Clusterization process of anti-Bolsonarist and Bolsonarist profiles



Source: authors.

After detecting the Bolsonarist network (Figure 1), we filtered the posts shared by this group of users, totalizing 217,101 retweets replicated by 37,283 profiles. The frequency of retweets was 35,375.4 retweets per day between February 7th and 15th, 2020. This database was exported to a text file (csv), called "tweets.csv". The document is made up of 217,101 lines (each one corresponding to one retweet) and 52 columns (corresponding to the metadata extracted from the posts). As seen in Figure 2, the daily action was concentrated from February 11th to 13th, 2020 (Hans River gave a statement to the CPMI on Fake News in Congress on February 11th). On February 12th alone, 120,152 tweets were generated (of these 91,522 retweets), 25,514 users participated in the action and 2,890 different hashtags were created.

Figura 2: Discursive viral overload: frequency of daily activities carried out by the Bolsonarist cluster (users, tweets, retweets, replies and mentions)



Source: authors.

For this study, the following metadata from retweets were used: 1) "rt_ text", which brings together the content of the message reposted by the Bolsonarist group (this metadata had allowed messages to be labeled based on a list of categories that express the type of violence speech addressed to Campo Mello); 2) "link", which specifies the electronic address of the link replicated by the retweeted message, if any (this metadata had allowed the identification of the pro-Bolsonarism news ecosystem that echoed narratives against the Folha de São Paulo journalist); 3) "from_username", which refers to the name of the Twitter account that originated the retweet (useful information for detecting Bolsonarism influencers who disseminated and fueled violence against the journalist); 4) "time", which concerns when the message was published, in the format day, month, year, hour, minute and seconds (which had allowed the analysis of changes in discourse against Patrícia as the days passed).

Topic modeling and machine learning techniques

After filtering the content of Bolsonarist posts (rt_text), the 1,000 most shared tweets in the database were labeled by a team of human coders. After coding, each labeling was supervised by the study's main investigator, in order to determine whether the common meanings of the messages were labeled in different ways by the researchers. There were 6 labels classified as discursive violence intensified by Bolsonaro trolls, they are:

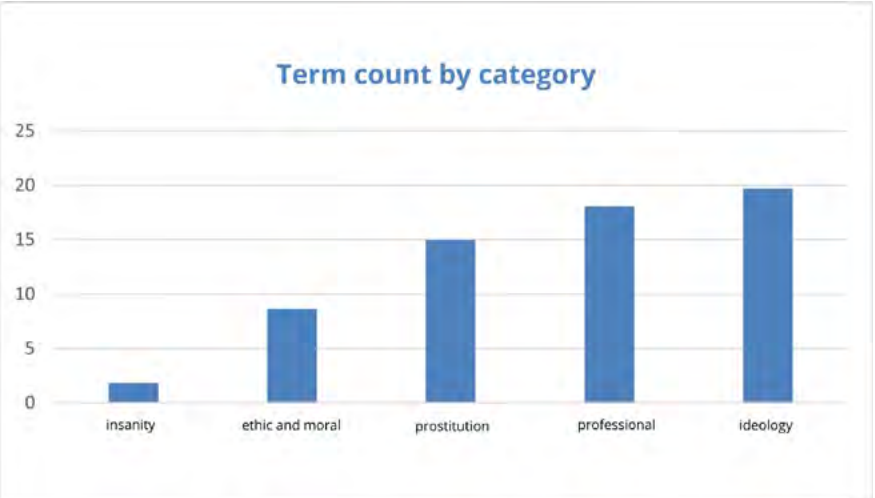
(C1) Discrediting due to ideology;
(C2) Discrediting due to ethics/morals;
(C3) Discrediting the profession;
(C4) Sexual insinuation;
(C5) Insinuation of insanity;
(C6) Accusation of racism.

The machine learning process, carried out by feeding the system with human labels, represents a critical step in content analysis projects, especially when dealing with significant volumes of data and the need for accurate categorization. In this context, machine learning classification algorithms play a central role in transforming human labels into processable knowledge.

Once these initial 1,000 tweets were labeled and the consistency of the labels was checked, the machine learning classification algorithms came into play. These algorithms were trained to learn to classify through the labels provided by human coders. They analyzed the linguistic patterns, contexts and characteristics of the labeled posts, seeking to understand how different categories of discursive violence were manifested in the messages.

Finally, we arrive at the results that can now be viewed through a graph that represents the count of terms by category. This graph is a comprehensive representation of the results obtained from the application of these advanced discursive analysis techniques in a digital environment.

Figure 3: Chart representing term count by category



Source: authors.

The Term Count by Category chart provides a detailed look at trends and patterns identified across different categories of discursive violence. It allows us to understand how the various forms of discursive violence are manifested in Bolsonarist messages on Twitter. The chart shows how many times key terms or linguistic patterns associated with each category of discursive violence were identified in the analyzed posts. This allows us to identify which categories were most frequent and which specific terms or patterns were prevalent in each of them.

Exploring the conversations and discourses present on the internet has become an increasingly feasible task, due to technological and methodological advances combined with the new developments of data science and netnography in general. However, analyzing big data from online sources, that is, large sets of unstructured data on the internet, can easily lead us to distorted and partial conclusions. We are fully aware of the limitations of this analysis method. The lack of historical context and the emphasis on the "now" or even the trap of searching for correlations between data are tendencies to avoid (Bollier, 2010: 18-19).

Boyd and Crawford (2011) created a list of provocations and questions about the use of big data in academia. Among the points mentioned, we highlight the fact that the insertion of automatized processes into research can change the definition of knowledge, that is, computational tools can change the reality of what they measure (Boyd & Crawford, 2011: 3). Our proposal does not involve delivering data to an algorithm and reproducing a statistical result. We will always assume that a result from big data analysis is subject to our interpretation. At issue is not the opposition between theory versus data, but how the latter helps to test theories and improve them (Bollier, 2010: 7). On the other hand, we also do not see these big data analysis methodologies as an exclusive area of data science, but as an expanding transdisciplinary resource:

> The use of digital technologies also contributes to the understanding of complex social and cultural phenomena, which go beyond merely numerical and statistical presentation. The increasing use of digital technologies therefore contributes to a different conceptualization of science (...). It is, therefore, a conceptualization that blurs the boundaries between disciplines, and their characteristic methodological processes, moving towards an increasingly transdisciplinary field. (Goveia & Carreira, 2013: 58).

Twitter data was collected using the Labic-Ford tool. Ford, programmed in *Python*[4], is a set of data collection and analysis tools that composes a single interface (*wrapper script*). The Ford was used in two important steps in obtaining the data we analyze.

Firstly, it collected the data through communication with the Twitter API. An API, or Application Programming Interface, is a computer structure for integrating systems, in this case Twitter and Ford. Thus, data can be shared securely and quickly, even when both sides (technologies) do not share the same programming language. It is through APIs that social media

4. Python is a programming language commonly used in web applications, data science and artificial intelligence.

sites limit and provide certain information. Different social media, different openings in the APIs. In the case of Twitter, 46 fields are provided. Not strictly speaking, APIs encompass: a physicality in terms of the bodily landscape of infrastructure and technology, through the economic logic at work (i.e. business models, ownership, licensing of APIs); functions and services (i.e. access to data); user practices (i.e. ways of working, playing and collaborating); discursive formations (i.e. statements, knowledge, ideas); rules and standards (i.e. design principles, terms of service, technical standards); as well as social imaginaries and desires (Bucher, 2018).

The second important step to be carried out by Ford is data parsing. This phase produces a statistical survey and transforms the results into files, so they can be analyzed and imported by other software. These statistical listings are great starting points for data analysis. Knowing the frequency of publication on a topic, which users were most active or which words were used most often (or the exact opposite) can become valuable indicators for identifying themes, points of view, the presence of influence in a network, among other elements. This process is preceded by a filtering that removes stop words. That is, prepositions, articles or conjunctions that, if not removed, would pollute the results. These stop words are removed so that the focus falls on other words that have meaning and give the context of the object to be analyzed. Skipping this step would result in terms like "a" or "e" receiving the most prominence in any statistical calculation of mentions in Portuguese. Finally, having files with the information collected means the possibility of importing them into other programs and, thus, continuing deepening the analysis.

We enter, at this point, the third stage of the process: analysis of results. Since we are interested in narratives that attack or try to diminish journalist Patrícia Campos Mello, we isolated the mentions present in the clusters that reproduce negative discourses. This way, our new database was formed by 37,283 tweets. We then started a new data structuring, in which we categorized terms and expressions based on samples of 300 tweets. Each

researcher categorized these mentions separately, thus preventing one researcher from influencing others in the process. After categorization, we compared the different terminologies used for the names of the categories, as well as the terms and expressions included in each of them.

Verbal violence and cyberviolence in the digital arena: Some considerations on the attack strategies

The term cyberviolence, as Paveau (2021) points out, has been used internationally and recurrently, given the rapid pace that technological innovations have imposed on linguistic-social dynamics. This scenario therefore presents the need to reflect on the ways of creating attacks on subjects' "faces" (according to Linguistics vocabulary) on social media such as Twitter, for example. In this sense, this work is inserted in the perspective of digital discourse analysis, proposing an observation of the technopragmatic effects of these linguistic constructions.

Studies of pragmatics outside a virtual world indicate that relationships between interlocutors result from a socially implicit contract regarding the preservation or threats to faces for the construction and maintenance of interactions in a social coexistence. However, the techno-discursive environment, according to Paveau (2021), brings new parameters that need to be considered to understand the discursive phenomena on the network and, consequently, on the elaboration of this new kind of social interaction. If pragmatic studies already conceived the idea that subjects occupy, in any communicative situations, a place of vulnerability in relation to the construction of a social image, this becomes intensely present when we consider anonymity-pseudoanonymity, the effect of absence, the cockpit effect, the displacement of the power relationship, the inseparability and virality of interactional processes (Paveau, 2021; Richardson-Self, 2021; Di Fátima, 2023).

Goffman's (1959) idea of the "face" as a positive social frame that individuals claim during interactions needs to be considered from a new perspective in

this scenario in which there is anonymity-pseudoanonymity during interactions, since this *techno-characteristic* calls into question a new dynamism in face-threatening acts proposed by Brown and Levinson (1987). Linguistic production is here associated with social production mediated by the dynamics of comments, responses, retweets and, therefore, available for expansion in a dimension that does not occur during face-to-face interaction.

> Online, the holder of discursive power is the one who has technological, computational and digital know-how, the publishing, disseminating, indexing and sharing practices. Recovering the strength of pseudoanonymity, the absence effect and the cockpit effect, the digital speaker displaces the traditional power relationship, dominating the technopragmatic effects of digital discourses (Paveau, 2021: 71).

In this way, the subjects' faces, when attacked by processes such as flaming war, are not restricted to the interior or surface of what the person chooses to expose, but are elaborated in the interpretation of the events manifested in the digital environment. Thus, even if, when making a post, there is an interest in consciously drawing a certain face of oneself, there is no way to control the construction of it for others, since one cannot control the interpretation made, nor the effect of virality. This is due to the virtual space configuring a free *locus*, as suggested by Seara (2021).

Given the previous reflection on the complexities of online speech and its impacts on individuals, it is pertinent to provide a brief approach to the definition of hate speech. According to Di Fátima (2023), there is no universally accepted definition for this type of speech, and its characterization is a reason for intellectual controversy, as it encompasses different forms of expression.

> In general, hate speech is an attack on a person or group, usually targeting members of a social minority. Thus, it can be classified as sexist, racist, xenophobic, ageist, fatphobic, or homophobic, among others.

> Haters direct their attacks, for example, against women, Black people, immigrants, seniors, disabled people, and the LGBTQ+ community (Di Fátima, 2023: 11).

The Twitter platform, when it emerged in 2006, proposed a unique form of messaging. However, after ten years, it presents several possibilities characterized by the platform's own lexicon, often including neologic terms and some signs (Paveau, 2021). These new forms create a *technolinguistic grammar* of the network itself, as its evolutionary nature presents new scriptural possibilities. The main genre that circulates within this environment is the *tweet*, understood as a *complex plurisemiotic statement*.

In socio-interactional processes, therefore, the act of language is an exposure of oneself, which normally claims, as we mentioned, a positive image. In the process of posting on the enunciator's own profile, we may suppose the attempt to enhance one's own image in a certain environment to the appreciation of others who access that space. However, at the same time, it is necessary to understand that "social networks are like virtual spaces or virtual squares (in the sense of the Roman forum) where relationships are developed, shared and modified in an infinite number of connections" (Seara, 2021: 289). As a result, there is no direct and objective control of the relationship of others with the image that is desired to be built. The modes of interaction (reply, commented retweet, sharing) that interlocutors make in publications can appear either in the sense of accepting this claim or attacking it.

Paveau (2021) then proposes that we understand how techno-discursive responses to discursive cyberviolence are organized. In the scope of this work, we are essentially interested in the relationships of *flame wars*, *shitstorms* and *tweetclashes*, which show that the attack carried out on social media is not only in the discursive field, but also linked to the image. This means that in relation to journalist Patrícia Campos Mello, for example, the hate speech directed at her refers to an attempt not only to demoralize her discursive constructions, but rather to construct a negative image of her as

a woman and professional: "One imagine that the person who utters hate speech does so to exercise sovereign power, to do what he or she says when he or she says it." (Butler, 2021: 35).

According to Brown and Levinson (1999/1978), marking an opposition, in a certain way, implies emulating the interlocutor and consequently leads to the construction of an action that in itself hurts only the negative face of the other. Threats to faces can be interpreted as a violent act. In this way, we seek to mitigate the opposition, that is, we try mild and subtle ways of countering the other's reasons. However, this caution does not eliminate the existence of explicit acts of violence, especially those where individuals intend to disqualify and insult another person. Thus, configuring a scenario of verbal violence.

According to Amossy (2014), violence towards the opponent is constituted by their disqualification, this mechanism being one of the strategies in the discourse of polemic. Following this premise, the theoretical framework that deals with verbal violence is extremely relevant to understand the argumentative strategies that were used by the opposing group, in this context of analysis, the Bolsonarist group, to attack and delegitimize the image of the journalist.

For Bousfield (2008), acts of verbal violence are intentional. This means that whoever attacks one of his peers does so with a specific objective, which may be the disqualification of the individual himself, as well as depreciating his arguments, in a way that nullifies their validity. Culpeper (2008) also argues that the use of a verbal utterance, of a violent nature, carries with it the intention of attacking.

From this point of view, there is an intentionality in the attacks and they seek to cause harm. This can be observed from acts whose objectives are to threaten the interlocutor's positive face, such as criticizing, insulting, disapproving, and also from acts that threaten the interlocutor's negative face, such as threatening the interlocutor's freedom of action, direct questions

without showing courtesy, indiscreet questions, unsolicited advice, orders, demanding previous favors, among others (cf. Saito & Nascimento, n.d., 4).

In addition to this, Locher and Watts (2008) emphasize that attitudes that go beyond social norms must be negatively evaluated; in other words, they perceive them as insulting and aggressive. In controversies, as already mentioned, aggression gains materiality through the disqualification of the other. In these cases, one of the strategies is the use of pejorative qualifiers which, according to Kerbrat-Orecchioni (1997 [1980]: 89) "concerns nouns or adjectives used to qualify an individual or a group in a derogatory way".

In this regard, we will verify in the *corpus* under analysis, the presence of pejorative statements and expressions used by the attack group with the aim of disqualifying Patrícia Campos Mello under different aspects, such as her attitude as a woman, her professional competence and, even, her lucidity. Furthermore, we will analyze how these pejorative expressions, understood here as verbal violence and cyberviolence, were coined either to disparage the journalist or to maintain an argumentative strategy that, by sustaining the ongoing controversy, would politically favor the opposing group, by polarization and disqualification of the other.

Our work highlights the construction of patterns of attacks targeting a journalist who was openly opposed to the government of President Jair Messias Bolsonaro. Previous research has shown that the Bolsonaro government was characterized by the continuous attack against communicators, including female journalists, and media (Cowley Forner; Muñoz Gallego, 2022; Capoano; Silva; Prates, 2023). Given this, we propose to observe and compile linguistic-discursive strategies that are configured as cyberviolence and group them into categories depending on the context in which they were carried out. To do this, we will analyze how these strategies take place in the discursive materiality of the interaction resources present on Twitter.

Results e analysis

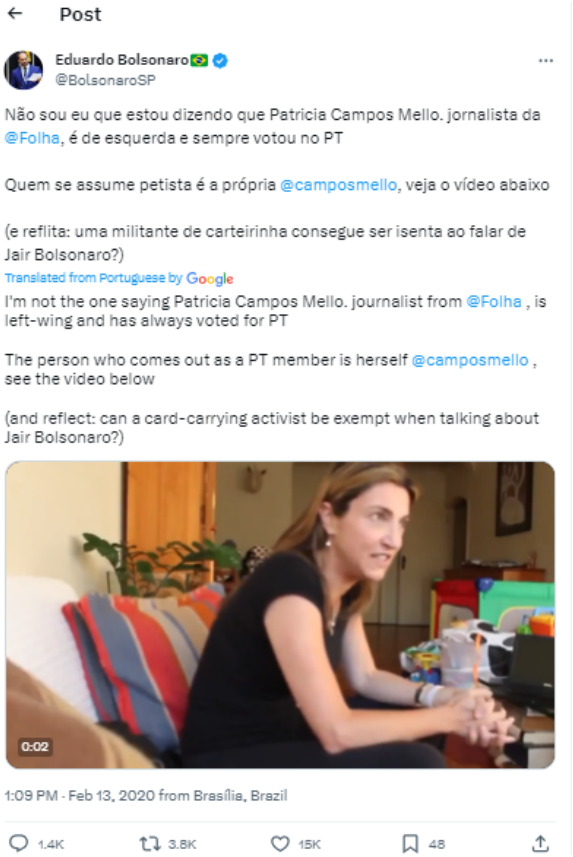
There are several possibilities for analysis when dealing with discourse structures, especially when dealing with material extracted from the social media under study. However, taking into account the prior observation of the data to be discussed, we have listed some strategies that best meet the purpose of this research. Therefore, we opted for an analysis from a qualitative and qualitative-interpretative perspective, starting from a macro observation of the categories established as cyberviolence, as main structures and, subsequently, an investigation of a micro nature, in order to understand the discursive strategies that confirm such structures.

Starting from the analysis of global semantics, understood as a more recurrent subject that, therefore, appears in other structures of discourse, it was possible to coin categories capable of more incisively encompassing socially shared interpretations. Thus, this general matrix of significations and meanings is the basis of the discourse that is assimilated during interactions.

To understand this matrix, we assume that the global semantics must specify in terms the meanings of the parts that compose it. Therefore, in order to convey the global meaning of the 6 categorical macrostructures found in the analysis corpus, we constructed identifying concepts that will be specified below.

(C1) Discrediting due to ideology: in this first category (C1), the words or expressions are linked to political gender violence, that is, the sentences pronounced during the attacks make reference to ideological issues or the journalist's political preference. The attack group uses terms such as: "jornalista+PTista [journalist+PTist]", "militante+PTista [militant+PTist]", "militante+maliciosa [militant+malicious]", "esquerdopata [left-winger]", among others. This case can be seen in the tweet below:

Figure 4: Tweet from @BolsonaroSP representing the category (C1) Descredibilization due to ideology



Source: https://twitter.com/BolsonaroSP/status/1227988094193930240.

The former president's son and congressist, Eduardo Bolsonaro, uses pejorative terms and negative insinuations to question the impartiality of journalist Patrícia Campos Mello based on her supposed political inclinations, which is a form of attack on her professional integrity.

(C2) Discrediting due to ethics/morals: taking into account the historical and social construction of a patriarchal society in which women's ethics were associated with their moral values that determine how their social behavior should be, associating explicitly the conduct of the female figure with

a sexual episode, becomes taboo. Based on the fact that the reporter had her image associated by the President of the Republic with a circumstance of "exchange of sexual favors", the attack group constructed speeches to discredit the complaint against Bolsonaro in the published article, since the professional was "morally corrupt". In this second category (C2) it is possible to find expressions such as: "perdeu+credibilidade [lost+credibility]", "sem+rePutação [without reputation]", "desqualificada [disqualified]", "mentirosa+sem vergonha [liar+shameless]", etc. In the text below, taken from a tweet, we can see a clear example of how this happens:

> Tweet from @PorTiMeu_BR representing the category (C2) Discrediting due to ethics/moral
>
> "@AndreiaSadi @camposmello Cowardice is what Madame FAKE News does daily, she tries to assassinate the reputation of people who just want to work for the country. The rotten game you played will be unmasked and we will move on. Woman? Woman has shame on her face, she doesn't lend herself to the ridicule that you lend yourselves to"[5]
>
> Source: https://www.twitter.com/PorTiMeu_BR/status/1227555611115687936

The use of the term "madame" is pejorative, suggesting a condescending attitude towards the journalist. The allegation of spreading "FAKE News" is a serious accusation, implying that the person is deliberately spreading false information.

(C3) Discrediting the professional: verbal violence was a resource widely used to demoralize and discredit female communicators throughout the electoral period. Using similar resources from C1 and C2, in the third category (C3) Bolsonaro's supporters used pejorative terms, also linked to Campos Mello's ideology and morals, in order to belittle the quality of

5. The Portuguese version is: "@AndreiaSadi @camposmello Covardia é o que a madame FAKE News faz diariamente,tenta assassinar a reputação de gente que só quer trabalhar pelo país. O jogo podre que vcs fizeram será desmascarado e vamos pra cima. Mulher?Mulher tem vergonha na cara não se presta a esse ridículo que vcs se prestam"

his work. To attack her, they use expressions such as: "militante+profissional [militant+professional]", "pseudo+jornalista [pseudo+journalist]", "Pseudojornalista [Pseudojournalist]", "jornaleira [papergirl]", "sujeitinha+desqualificada [little bloke+disqualified]", etc. The text below shows this:

> Tweet from @hans_sincero representing the category (C3) Discrediting the professional
>
> "Folha's presumption is so enormous that it believes it is capable of saying that Hans is lying. They do not subject the material presented by Patrícia Papergirl to expertise and competent authorities and act as if they were the owners of the truth. Abuse of power against a poor black man. #hansTHEMYTH https://t.co/PhaCJTgz4K"6
>
> Source: https://www.twitter.com/hans_sincero/status/1227702556631191552.

The use of the term "papergirl" is an attempt to disqualify Patrícia Campos Mello's profession as a journalist. The term is used here in a derogatory manner, suggesting that she is not a serious and trustworthy journalist, but rather someone who would be involved in questionable practices in journalism.

(C4) Sexual insinuation: in misogynistic practices, one of the ways to discredit the work done by a woman is to associate her with sexual conduct seen as inappropriate by society. As a form of attack and humiliation, Jair Bolsonaro insinuated that the journalist offered sex to Hans as payment in exchange for information, building the image of a morally corrupt woman. However, the comments made on social media against the journalist consist of statements that demonstrate not only hate speech, but also a political bias. The political discourse is constructed seeking to discredit journalistic discourse, that is, through the deconstruction of the image

6. The Portuguese version is: "A soberba da Folha é tão gigante, que ela se julga capaz de afirmar que Hans mente. Não sujeitam o material apresentado pela Patrícia Jornaleira à perícia e autoridades competentes e agem como se donos da verdade fossem. Abuso de poder contra um negro pobre. #hansOMITO https://t.co/PhaCJTgz4K"

of the journalist as a broadcaster of true reality, in this case, labeling her as a disseminator of *fake news*. In this category (C4), the main expressions were: “oferecer+xerecard [offering+pussycard]”, “furo+vagabunda [scoop+slut]”, “jornalista+vagabunda [journalist+slut]”, “Xota News [Pussy News]”, “prostituta+troca+informação [prostitute+exchange+information”, among others.

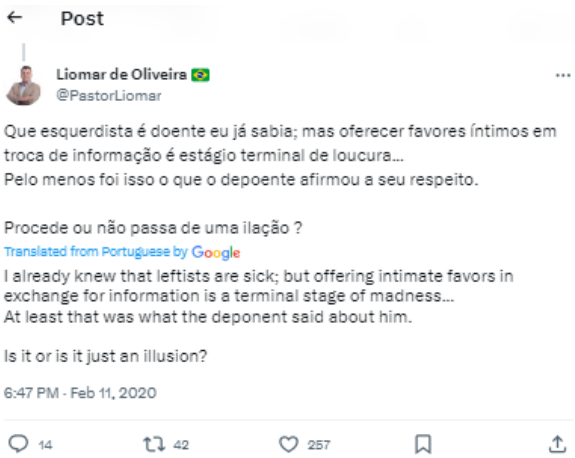
Figure 5 - Tweet from @carivaldomelo representing the category (C4) Sexual insinuation

Source: https://twitter.com/carivaldomelo/status/1227689529928187904

The expression "pussycard" refers to a situation in which a woman uses sex in exchange for some favor or benefit. The term is a neologism formed by the combination of the words "pussy", which is a nickname for the female genitals, and "Mastercard", a famous brand of credit and debit cards. In this context, it insinuates that journalist Patrícia Campos Mello offered sexual favors in exchange for information.

(C5) Insinuation of insanity: violence against women, in the case under analysis, is not something particular, it is influenced by the social, cultural and historical environment. Another sexist discourse strategy used with the aim of questioning the reliability of a woman is to accuse her of being insane. In this sense, in the fifth category (C5), the attack group used words such as: "histérica [hysterical]", "esquizofrênica [schizophrenic]", "estágio+terminal+loucura [terminal+stage+madness]", etc. We can see an example in the message below:

Figure 6 - Tweet from @PastorLiomar representing the category (C5) Insinuation of insanity



Source: https://twitter.com/PastorLiomar/status/1227348408635142145

The expression "terminal stage of madness" used in this context is a figure of speech loaded with a negative and hyperbolic connotation. The author

of the tweet is using this expression in an exaggerated way to describe the behavior he alleges that journalist Patrícia Campos Mello adopted.

(C6) Accusation of racism: the group attacking Patrícia Campos Mello also sent messages on Twitter accusing her of racism. Users took into account the difference between the social classes of the journalist and the deponent, Hans River, in addition to the color and professional position held by both. Expressions such as "patricinha rica [rich preppy]", "rica+opressora [rich+oppressor]", "patricinha+branca+elite [preppy+white+elite]", etc. were the most used. An example:

Figure 7 - Tweet from @JornalNoAtaque representing the category (C6) Accusation of racism

Source: https://twitter.com/JornalNoAtaque/status/1228316789932806144.

The term “preppy” is often used pejoratively to describe a woman who is seen as spoiled, rich, superficial and concerned with appearances and social status. The expression “white elite” refers to the idea that Patrícía Campos Mello belongs to a privileged social class and is of white ethnic origin. The use of this term suggests a criticism of her social privilege.

That said, the results of the analysis of the categories of discursive violence highlight the extent of the hate speech and discredit that journalist Patrícia Campos Mello faced on social media. These categories not only reveal the specific strategies used by perpetrators, but also shed light on the multifaceted nature of online discursive violence.

It is important to stress that these attacks are not limited to simple defamation, but also reflect the perpetuation of gender stereotypes. The journalist’s association with sexual, political and racial stereotypes reveals the complexity of the dynamics of hate speech on social media and how these dynamics can be used to achieve multiple objectives.

Furthermore, qualitative and qualitative-interpretative analysis allows us to deepen our understanding of the discursive strategies employed by attackers. By examining the specific terms, expressions, and contexts in which these categories of discursive violence emerge, we can identify the underlying narratives and ideological discourses that fuel them.

Final remarks

Nowadays, the possibility of public expression has reached new levels. In such a context, it is remarkable the importance of studying discursive constructions in social interactions, particularly in the digital environment of social media. On these platforms, relationships between individuals can be marked by both reverence and repulsion, expressed through *likes* and *dislikes*.

In our analysis, we observed the persistence of gender stereotypes and prejudices that affect online interactions. The historical construction of women

as “the absolute other”, as expressed by Simone de Beauvoir, continues to influence social relations and often results in attitudes that challenge the dignity of women, which has its worst expression in the violent attacks.

This study focused on analyzing the attacks directed at journalist Patrícia Campos Mello, from Folha de São Paulo, after Hans River’s testimony to the Joint Parliamentary Commission of Inquiry (CPMI) on Fake News. Hans accused the journalist of offering sex in exchange for information, triggering a series of sexist and sexist comments on social media.

Our analysis was based on the identification of words and expressions that characterizes hate speech against women in the digital environment, using text analysis and machine learning methods. Furthermore, we highlight the intersection between Linguistics, Social Communication and Data Science in approaching this relevant topic in contemporary society.

The categories of discursive violence (C1 to C6) not only exposed the specific strategies used by attackers, but also presented the complexity of the dynamics of online hate speech. These attacks were not limited to simple defamation, but also reflected the perpetuation of gender stereotypes. Furthermore, the journalist’s association with sexual, political and racial stereotypes showed how the dynamics of power and prejudice intertwine on social media.

At a time when hate speech is proliferating on social media, this study highlights the importance of promoting critical discussion and seeking solutions to mitigate this problem, ensuring a safer and more respectful online environment for all users. Journalist Patrícia Campos Mello is an example of the real and harmful consequences of this type of attack, which not only affects individuals, but also the quality of public debate and press freedom. Therefore, it is essential to continue investigating and combating this phenomenon in all its forms.

References

ABRAJI. (2022). *Abraji points out that women journalists were victims of more than half of the attacks in the digital media*. https://www.abraji.org.br/abraji-aponta-que-mulheres-jornalistas-foram-vitimas-de-mais-da-metade-das-agressoes-no-meio-digital.

Amossy, R. (2014). *Apologie de la polémique*. Presses Universitaires de France.

Beauvoir, S. (1967). *The Second Sex II: The lived experience* (S. Milliet, Trans.). Difusão Européia do Livro.

Bollier, D. (2010). *The promise and peril of Big Data* (1st ed.). The Aspen Institute.

BolsonaroSP [@BolsonaroSP]. (2020, February 13). *Descredibilization due to ideology (C1)* [Tweet]. Twitter. https://twitter.com/BolsonaroSP/status/1227988094193930240

Bousfield, D. (2008). *Impoliteness in interaction*. John Benjamins.

Boyd, D. & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Brown, P. & Levinson, S. (1987). *Politeness: Some universals in language usage*. Cambridge University Press.

Bucher, T. & Helmond, A. (2018). *The affordances of social media platforms*. In J. Burgess, A. Marwick, & T. Poell (Eds.), The SAGE handbook of social media (pp. 233-253). Sage. http://dx.doi.org/10.4135/9781473984066.n14

Butler, J. (2021). *Hate speech: A politics of the performative* (R. Fabbri Viscardi, Trans.). Editora Unesp.

carivaldomelo [@carivaldomelo]. (2020, February 13). *Sexual insinuation (C4)* [Tweet]. Twitter. https://twitter.com/carivaldomelo/status/1227689529928187904

Capoano, E., Sousa, V., & Prates, V. (2023). Circulation systems, emotions, and presentism: Three views on hate speech discourse from attacks on journalists in Brazil. In B. Di Fátima (Ed.), *Hate speech on social media: A global approach* (pp. 159-184). LabCom Books & EdiPUCE.

Cowley Forner, O. M. & Muñoz Gallego, A. (2022). The hate speech against female journalists in the government of Jair Bolsonaro. *Tematica*, *18*(12), 126-141. https://periodicos.ufpb.br/index.php/tematica/article/view/64819.

Culpeper, J. (2011). *Impoliteness: Using Language to Cause Offence*. Cambridge and New York: Cambridge University Press.

Di Fátima, B. (2023). *Hate speech on social media: A global approach* (Communication Books, Vol. 7). LabCom Books & EdiPUCE.

FENAJ. (2020). *Violence against journalists and freedom of the press in Brazil: 2020 report*. https://fenaj.org.br/wp-content/uploads/2021/01/relatorio_fenaj_2020.pdf

Forner, O. M. C. & Gallego, A. M. (2022). The hate speech against women journalist in the government of Jair Bolsonaro. *Revista Temática: Cultura e Sociedade*, XVIII(12). http://periodicos.ufpb.br/index.php/tematica/index

Goffman, E. (1956). *The presentation of self in everyday life* (M. C. S. Raposo, Trans.). Vozes.

Goveia, F. G. & Carreira, L. S. (2013). Data research and the issue of image abundance: Relations between science and art. *Icone*, *15*(1). https://doi.org/10.34176/icone.v15i1.230707.

hans_sincero [@hans_sincero]. (2020, February 13). *Discrediting the professional (C3)* [Tweet]. Twitter. https://twitter.com/hans_sincero/status/1227702556631191552

JornalNoAtaque [@JornalNoAtaque]. (2020, February 15). *Accusation of racism (C6)* [Tweet]. Twitter. https://twitter.com/JornalNoAtaque/status/1228316789932806144

Kerbrat-Orecchioni, C. (1997). *L'énonciation*. Armand Colin.

Locher, M. A. & Watts, R. J. (2008). *Relational work and impoliteness: Negotiating norms of linguistic behavior*. In D. Bousfield & M. A. Locher (Eds.), Impoliteness in language: Studies on its interplay with power in theory and practice (pp. 77-99). Mouton de Gruyter.

Malini, F. (2015). A perspectivist method for social network analysis: Mapping topologies and temporalities in the network. *Presented at the XXV Annual Meeting of Compós*, Federal University of Goiás, Goiânia. https://www.labic.net/wp-content/uploads/2016/06/compos_Malini_2016.pdf

Malini, F. (2020). *The word and the "things": How to create your list of terms for data collection on social media*. https://fabiomalini.medium.com/a-palavra-e-as-coisas-como-montar-a-sua-lista-de-termos-para-coleta-de-dados-em-redes-sociais-39ed3648ea4

PastorLiomar [@PastorLiomar]. (2020, February 12). *Insinuation of insanity (C5)* [Tweet]. Twitter. https://twitter.com/PastorLiomar/status/1227348408635142145

Paveau, M. (2021). *Digital discourse analysis: Dictionary of forms and practices* (J. L. Costa & R. L. Baronas, Eds.). 1st ed. Pontes Editores.

PorTiMeu_BR [@PorTiMeu_BR]. (2020, February 12). *Discrediting due to ethics/moral (C2)* [Tweet]. Twitter. https://twitter.com/PorTiMeu_BR/status/1227555611115687936

Richardson-Self, L. (2021). *Hate speech against women online: Concepts and countermeasures*. Rowman & Littlefield.

Saito, C. L. N. & Nascimento, E. L. (n.d.). *Face preservation and politeness: A seduction game in face-to-face interactions*. http://www.diritto.it/archivio/1/20656.pdf

Seara, I. R. (2021). Dizzying connections: Verbal violence in 'comments' on social media. *Calidoscópio*, *19*(3), 385-397. https://doi.org/10.4013/cld.2021.193.07

OBSTACLES TO DETECTING AND SUPPRESSING ONLINE HATE SPEECH

Mariana Magalhães
/ University of Porto, Portugal

Sara Alves
/ University of Porto, Portugal

Márcia Bernardo
/ University of Porto, Portugal

Decoding the language of hate

Hate speech is a phenomenon that, both due to its increasing frequency and its social impacts, has been attracting more and more public attention. Despite the absence of an official or legal definition of hate speech, the Recommendation of the Committee of Ministers of the Council of Europe (1997) defines it as "all forms of expression that propagate, incite, promote or justify racial hatred, xenophobia, anti-Semitism and other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism or ethnocentrism, discrimination and hostility against minorities, migrants and people of migrant origin". This scourge can be conceived as a verbal aggression directed at a group or individual, based on an ideology anchored in negative stereotypes about its target groups, perceived as less capable, meritorious, and worthy of respect (Keen & Georgescu, 2016; Ruwandika & Weerasinghe, 2018). It is traditionally anchored in attributes indicative of vulnerable social groups (e.g., gender, ethnic origin, nationality, ability, religion, sexual orientation) and is

often associated with attitudes of demonization, dehumanization, and social exclusion (United Nations, 2019). Thus, hate speech is considered a strategy for maintaining and reinforcing the hierarchical social system, in which its targets are described as a threat to society and the status quo, against which the majority group must defend itself (Weber, 2009).

When considering the expression of hate in the online context, the European Union Agenda for Fundamental Rights (FRA, 2023) found that hate expressions could be found in five different forms: incitement to violence, denigration, offensive language, negative stereotyping and, finally, other content. Incitement to violence, discrimination or hatred only includes speech that specifically calls for action. Denigration translates into targeted attacks on the capacity, character or reputation of a person or group, based on their membership in a particular social group. Offensive language is present in situations marked by obscene, hurtful and derogatory language, which is highly dependent on the context. Negative stereotyping is based on disseminating negative traits and characteristics assigned to a social group and its individual members. The fifth and last category comprises hateful content that does not fit into the other categories, representing a residual class that includes support for hateful ideologies or Holocaust denial. Nevertheless, the categorization of hate speech into one of these categories remains subjective and not mutually exclusive.

This violent social phenomenon can have a negative impact on several levels, either for the individual, the group, or their society. Victimization by hate speech (as well as exposure to this verbal aggression when directed at a group to which we belong) has the potential to impact one's mental health, both in the short and long term (Leets, 2022), namely by promoting increased feelings of insecurity, anguish, revolt, fear and shame, and decreased self-esteem (Keen & Georgescu, 2016; Sarmento, 2016), which, ultimately, can lead to harmful behaviors, such as suicide (Keen & Georgescu, 2016). Moreover, considering that hate speech is directed at socially vulnerable groups, the collective experience of hate speech increases feelings of insecurity and social exclusion (Mullen & Rice, 2003), leading to a decrease

in social trust (Näsi et al., 2015) and in the collective confidence (Boeckmann & Liew, 2002), as well as increased social isolation, with the aim of preventing future incidents (Gelber & McNamara, 2015). In the case of the majority communities (the main sources of hate speech), the frequent witnessing of such incidents, especially when sanctions are not applied to the aggressors, normalizes the hatred inherent in the speech and, consequently, deteriorates relations between the various communities (Leets, 2001), maintaining unbalanced power relations between groups (Gelber, 2017). Witnessing hate speech may even encourage discriminatory attitudes, episodes of physical violence (Keen & Georgescu, 2016), and related crimes (Aluru et al., 2020). Online hate speech, in particular, is also often a predictor of offline violence perpetration (Müller & Schwarz, 2021).

Defining hate speech is a challenge yet to be overcome. The identification of language excerpts as hate speech is made difficult by several factors, from the absence of a common definition to the possibility of this aggression being manifested in a more subtle and camouflaged way, using statements that can, in an initial analysis, be characterized as rational or normal (Weber, 2009).

The nuances of hate: Unraveling its complexity online

Social networks and other digital platforms have revolutionized the ways of communicating and interacting available until then, quickly becoming the tools of choice for this purpose (ElSherief et al., 2018; Latour et al., 2017). This has led to positive and negative outcomes, of which the dissemination of hate speech is one of the most worrisome.

Communication mediated by technological equipment presents peculiarities that influence interaction and promote the previously mentioned propagation of hate speech. In the online context, there is greater behavioral disinhibition, leading people to display attitudes they would not in person (Keen & Georgescu, 2016; Seixas et al., 2016; Suler, 2004). Additionally, several studies demonstrate that, in these circumstances, people exhibit more

deceptive behavior, lower levels of empathy, and greater moral disengagement regarding their conduct (Seixas et al., 2016; Suler, 2004). This could be enhanced by, among other factors, the fact that communication mediated by a screen often prevents access to the receiver's immediate reaction to the message sent (Suler, 2004). The absence of feedback and associated physical signs leads to a lack of information about the receiver's emotional state, which can promote conflict (Seixas et al., 2016).

Another peculiarity of online interactions is that the people involved may choose to keep their identity anonymous or use pseudonyms to hide their identity (ElSherief et al., 2018). In this scenario, people perceive the behavior they adopt online and their personal identity as separate aspects, meaning they feel less vulnerable, exposed, and liable to be punished for their actions. For these reasons, one can witness the alternation of roles between victim and aggressor (Tarouca & Pires, 2016).

The online context is fed by an immeasurable amount of content, which can be easily replicated and shared, potentially reaching a wide audience at a considerable speed (Seixas et al., 2016). Unlike in face-to-face situations, online aggressions can become viral quickly (Seixas et al., 2016).

For all the above-mentioned reasons, the online context can be described as especially challenging. Thus, detection and intervention mechanisms should take into account its peculiarities in order to promote success.

Hate speech detection

Detecting and formally recording hate speech incidents is a top priority for national governments and supranational organizations, such as the European Union (EU). This recording allows the characterization of the phenomenon, concerning its prevalence, both general and disaggregated by target groups, and the different contexts in which these incidents occur. It would also allow us to understand the evolution of the content of hate speech, since there has been a transformation in the form of expression of hate towards an increasing subtlety, particularly online (Siegel, 2020). Using

this information, it is possible to outline effective strategies to reduce hate speech incidents to the greatest extent, which are adapted and differentiated according to the most frequent target groups, contexts, and content; their implementation would also send a message to the most vulnerable communities that this phenomenon is taken seriously by national authorities, and to the majority groups that this type of behavior will not be tolerated (FRA, 2018; cf. Crandall et al., 2002). This information will also support the need to create appropriate responses to support victims. However, in 2018, only 19 EU Member States made data on reported hate crimes public, and the types of information they provided about these cases varied between these countries (for example, the type of prejudice associated, the type of crime committed, the population groups at greatest risk of victimization, and levels of satisfaction with the police response; FRA, 2018).

Existing detection mechanisms

Detecting and formally recording hate speech incidents is the responsibility of the police force. Therefore, for a hate speech incident to be officially counted, the victim(s) must file a report with the police, who, in turn, must record the incident as bias-motivated. However, an analysis of the recording mechanisms for hate crimes (which includes hate speech) of EU Member States conducted by the FRA found a large discrepancy between the analyzed countries regarding the way this registration is carried out (FRA, 2018). In some countries, the flagging of a crime as a hate crime, either on general crime registration forms or specific forms for hate crimes, is not possible, nor is the police force provided with a list of bias indicators necessary to identify prejudiced motivations underlying a crime, either at the time of reporting or throughout the investigation process. Particularly in Portugal, the police did not report having a list of bias indicators, and forms for recording crimes did not incorporate options for flagging the incident as bias-motivated. The police also did not receive any guidance or training in recording hate crimes (FRA, 2018).

Civil society organizations (CSOs) dedicated to promoting human rights could also play a fundamental role in detecting hate crimes, including hate speech (FRA, 2018). Cooperation between CSOs and the police may occur at several levels, including through the exchange of data and information related to hate incidents, the creation of working groups, and the co-development of guidelines on bias indicators. Many countries have developed cooperative relationships between CSOs and police forces. In Portugal, there is no information about these relationships (FRA, 2018).

In the online context, hate speech detection is carried out by the platform (e.g., a social network) where this speech occurs. In the event of a user report, there is a moderation process where the reported content is analyzed in light of the platform's terms of service and, consequently, deleted or retained. Platforms can, therefore, be distinguished by the tolerance they demonstrate towards hate speech, with less tolerant platforms (such as YouTube, Facebook, and Instagram), where it occurs infrequently, and more tolerant ones (such as Gab, Telegram, and 4Chan), where it ends up becoming the most frequent type of speech (Mathew et al., 2020). Furthermore, the platforms make use of hate speech detection algorithms, automatically eliminating content identified as such. The IT strategies developed so far for detecting hate speech have approached the problem as one of classification, seeking to classify the analyzed content as being, or not, hate speech or as falling into one or more types of hate speech, according to the underlying prejudice (Fortuna & Nunes, 2018). However, there are still limitations to all the developed strategies, meaning that this classification cannot be carried out with complete precision.

The gap between occurred and reported incidents of hate speech

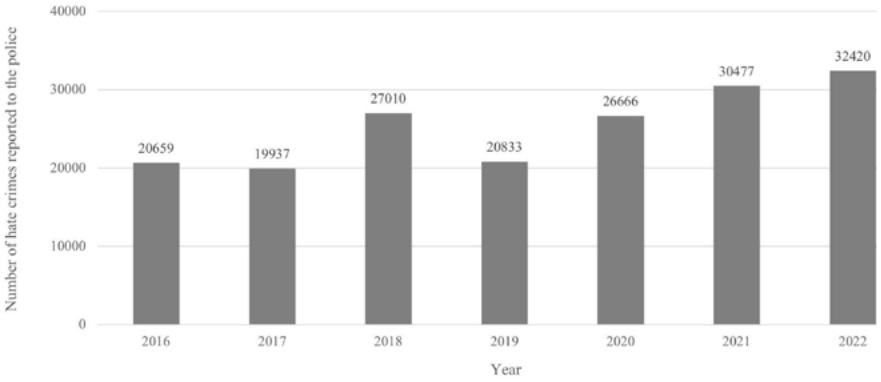
The variability in methods of recording hate crimes raises a great concern among EU and national authorities related to the invisibility of hate crimes. Indeed, a crime will not be included in a country's official hate crime data, nor will it be investigated and punished as such if there is no possibility of

marking it as bias-motivated. However, this variability is only part of the problem concerning the invisibility of hate crimes.

According to a FRA report (2021), which aggregates data from four surveys conducted with various minority groups, reporting rates vary between 6% and 19% for bias-motivated harassment. Thus, there appears to be a gap between hate crimes that occur and those effectively reported. It is also important to highlight that different groups report at different frequencies, suggesting different social dynamics between these groups and the societies in which they live, that will affect their motivation to make an official report. For example, reporting rates of bias-motivated harassment are higher among Jewish people and lower among Roma and Travellers.

When analyzing the official data made available by European countries, collected and centralized by the OSCE Office for Democratic Institutions and Human Rights (OSCE/ODIHR), we notice a slight increase in the number of hate crimes recorded by the police between 2016 and 2022 (OSCE/ODIHR, 2023; see Figure 1). Although not all countries disaggregate their data according to the type of prejudice that motivated the hate crime, there is a predominance of racist and xenophobic motivations in hate crimes reported in 2022 (OSCE/ODIHR, 2023; see Figure 2). Additionally, data disaggregated by type of crime is even more scarce, which makes it very difficult to monitor the evolution of the prevalence of this type of hate incident. It is important to highlight that, when referring to an increase in the number of hate crimes recorded by the police, this may reflect an actual increase in the frequency of occurrence of hate crimes, an improvement in the methods of recording hate crimes, or a greater predisposition of the population for reporting hate incidents. In the online context, hate speech seems to be increasing (e.g., Kim & Kesari, 2021), especially on social networks with reduced moderation (Mathew et al., 2020). In line with this, there is evidence of high rates of self-reported exposure to hate speech on social media (e.g., UK Safer Internet Centre, 2016).

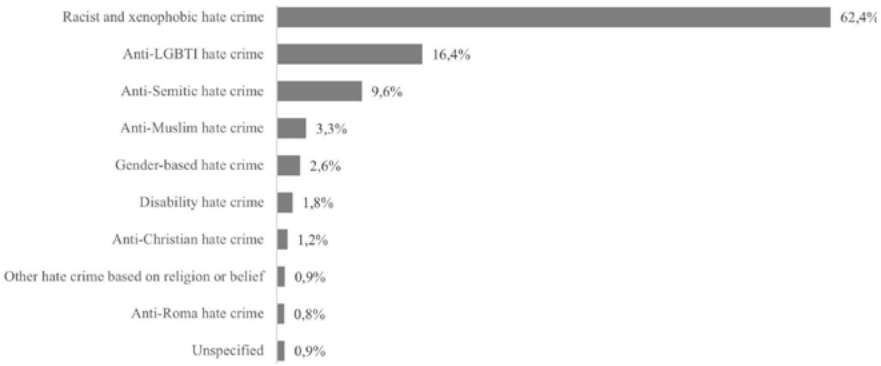
Figure 1: Number of Hate Crimes Reported to the Police from 2016 to 2022



Note: Data from OSCE/ODIHR, 2023, from EU countries that have publicly reported hate crime data.

Considering that combating hate speech will be more effective when its prevalence and characteristics are well known, and that this knowledge results from reporting these cases and registering them as hate speech, it is necessary to identify the obstacles to obtaining this information so that strategies to reduce hate speech can be developed. In this sense, we highlight the limitations of existing social control mechanisms and the psychosocial phenomena that limit reporting and normalize hate speech, such as the bystander effect, the impact of political speeches and the media, and the contextual nature of hate speech.

Figure 2: Proportion of bias-motivations in 2022



Note: Data from OSCE/ODIHR, 2023, from EU countries that have publicly reported disaggregated hate crime data.

Limitations of the existing social control mechanisms

The monitorization and suppression of hate speech is the responsibility of various social control mechanisms, namely the police, the judicial system, and social media platforms. However, several limitations to all of these mechanisms contribute to the perception that reporting hate incidents will not have any consequences for offenders (but may, however, have consequences for victims), reducing, consequently, the intention to report (FRA, 2021). These limitations will be addressed in more detail below.

Freedom of expression versus hate speech

Potentially, the greatest obstacle for suppressing hate speech is the limitations it imposes on individuals' freedom of expression, one of the fundamental human rights according to the Universal Declaration of Human Rights (United Nations, 1948). Indeed, attempts to suppress hate speech could be understood as putting into conflict the right to freedom of expression and other rights, such as the right to non-discriminatory treatment and a free and safe life.

There are several perspectives regarding this subject (Tontodimamma et al., 2021). On the one hand, defenders of free speech argue that the prohibition of any type of speech, even when it disrespects, offends, or creates discomfort, opposes the principle of content neutrality, which defends the non-restriction of expressions based on its content (Brettschneider, 2013). On the other hand, some authors defend the need to find a balance in order to guarantee both freedom of expression and the protection of vulnerable minorities, who deserve to be treated as human beings and members of the community at the same level as majority groups (Cohen-Almagor, 2019).

This debate has consequences on different levels. At the legislative level, the difficulty in finding the limit on the type of speech that should not be tolerated leads to the implementation of laws that vary in the scope of speech that can be judicially punishable (Bleich, 2011). A greater belief in the right to freedom of expression has been shown to predict a lower perception of

hate speech incidents as hate crimes (Roussos & Dovidio, 2018), which may impact reporting intentions. On social networks, the defense of freedom of expression is the flagship of platforms with reduced content moderation, allowing the exponential dissemination of hate speech (Mathew et al., 2020). However, the application of restrictions on freedom of expression based on the protection of vulnerable groups seems to become more accepted after their implementation (Bleich, 2011), which allows the continuation of the work carried out so far.

Hate speech in European and Portuguese legislation

In the EU, the European Convention on Human Rights (Council of Europe, 1950) is the first instrument to bind some of the rights set out in the Universal Declaration of Human Rights. In this convention (ECHR), restrictions can be put on freedom of speech as long as they are "necessary in a democratic society", which includes restrictions made "for the protection of the reputation or rights of others". The ECHR also includes the prohibition of discrimination based on "the sex, race, color, language, religion, political or other opinions, national or social origin, membership of a national minority, wealth, birth or any other situation". Though this provides the legal basis for the introduction of national legislation relating specifically to hate crimes, the ECHR does not require their implementation (OSCE/ODHIR, 2009).

In view of this limitation, the Council Framework Decision 2008/913/JHA (Framework Decision on the fight through criminal law against certain forms and manifestations of racism and xenophobia) was developed, which determines the conditions for the criminalization of hate speech and hate crime in the EU Member States. In it, hate speech is defined as "public incitement to violence or hatred directed against a group of persons or a member of such a group defined on the basis of race, colour, descent, religion or belief, or national or ethnic origin", as well as "publicly condoning, denying or grossly trivialising crimes of genocide, crimes against humanity and war crimes as defined in the Statute of the International Criminal Court

(Articles 6, 7 and 8) and crimes defined in Article 6 of the Charter of the International Military Tribunal, when the conduct is carried out in a manner likely to incite violence or hatred against such a group or a member of such a group" (Council of Europe, 2008). The application of this Framework Decision was analyzed in 2014, and it was found that there were variations between Member States in the application of its different provisions (European Commission, 2014). It is also important to mention the European Code of Conduct on Countering Illegal Hate Speech Online, initiated in 2016 by the European Commission, and signed by several technology companies over the last few years, which requires the removal within 24 hours of hate speech that infringes the platforms' terms of service and European law, after reporting by users. However, an examination by the European Commission revealed that platforms that signed the Code of Conduct analyze only 64.4% of reports within the 24 hours agreed in the Code of Conduct and remove only 63.6% of hate speech flagged by users (European Commission, 2022).

Indeed, the different national understandings about what should be considered hate speech (versus freedom of expression) largely explain the divergences in EU Member States' legislation regarding hate speech (European Commission, 2014). In Portugal, hate speech has a limited legal presence in the Penal Code; it can only be prosecuted as the crimes of defamation or incitement to hatred and violence (Article 240), which requires the speech to be publicly disclosed and able to be disseminated.

It appears that there are several limitations not only in European but also in Portuguese legislation, with a distinction being made between "hard" hate speech, which includes forms of hate speech punished by law, and "soft" hate speech, which is legal, despite being a form of discrimination and intolerance (Baider et al., 2017). In different national contexts, what can be considered hard and soft hate speech varies, and this legislative variation (both inside and outside the EU) makes it even more complex to combat hate speech in the online context, characterized by its borderlessness (Alkiviadou, 2017).

The role of police forces

A hate crime cannot be judged as such without previously flagging it as having a bias-motivation. The police forces are responsible for this registration, so it is necessary to ensure they do so. Apart from the limitations of the methodologies available to conduct this registration, studies conducted by the FRA demonstrate that part of the reasons for victims' decision not to report a hate incident is related to a lack of trust in the police and low satisfaction with previous experiences with this institution (FRA, 2021).

In this sense, internalized prejudice can lead police officers to not register a hate incident as having a bias motivation, even if there are methods at their disposal to do so, because they are unable to discern this motivation or do not believe in it when explicitly indicated by the victims (FRA, 2021). A survey conducted with justice professionals found that two in five professionals consider it fairly or very likely that police officers share the prejudices of offenders (FRA, 2016a). These agents have even been identified as the aggressors (FRA, 2016b). Among those victimized by the police, the majority (63%) did not make a report. In Portugal, a recent report revealed the existence of a Facebook group in which racist and xenophobic comments were made by more than 600 police officers (Pena et al., 2022). Thus, trust in the police is reduced, simultaneously with the victims' motivation to report the hate incidents they experienced, which contributes to their invisibility (FRA, 2021).

Reduced control on social media

The high prevalence of hate speech in the online context highlights the need for social media platforms to moderate their content. As already mentioned, this moderation is highly dependent on user reports and the detection of hate speech content through algorithms. However, there are limitations associated with this moderation.

Regarding algorithms, hate speech's complexity and constant evolution make it difficult to detect and eliminate (Siegel, 2020). Indeed, only the most

flagrant forms of hate speech are easily detected by algorithms (Fortuna & Nunes, 2018). Human moderation suffers from the same limitation in that decisions regarding what does or does not fall into the category of hate speech can vary depending on the moderators' social identity(ies) and beliefs (Wojatzki et al., 2018). It is also necessary for algorithms to be able to detect hate speech in the different languages in which it is written, something that has not received sufficient financial investment (Laub, 2019). The large amount of content makes this moderation process even more demanding which, accompanied by the lack of human resources available for this task, leads to the non-detection and removal of much content that could be considered hate speech (Laub, 2019). A report by the FRA (2023) revealed that, despite platform moderation efforts, about 53% of manually analyzed posts were classified as containing hate, with 55% of these posts containing hate based on protected characteristics. These findings highlight that, even though moderation systems function to some extent, many posts considered online hate speech still go unnoticed, as moderation tools fail to properly identify content constituting online hate speech.

Some platforms may also hesitate to remove content, considering that they cannot be legally penalized for the users' speech and that removing content could demotivate the use of the platform by offenders and those who follow them (Banks, 2010). Indeed, there is little transparency in the way platforms carry out content moderation (Laub, 2019; Siegel, 2020), and, in some cases, the actions taken by platforms have been criticized for appearing to go against the values defended by them (Ray, 2019). This can be demotivating for users who disagree with the prejudiced content they are exposed to, inhibiting them from reporting it because they believe nothing will be done due to their reporting (Jubany & Roiha, 2016).

Sociopsychological phenomena normalizing hate speech

Strengthening the current social control mechanisms to motivate reporting and demotivate prejudiced actions is undeniably important. Nonetheless, it is possible to point out some socio-psychological phenomena that work

against them by normalizing hate speech and, thus, fueling the perception that it is not serious enough to be reported (FRA, 2021).

The bystander effect

Added to the complexity of evaluating and combating hatred on online platforms is the influence of social phenomena in emergency intervention. In this context, there is a theoretical model that acquires special relevance: the bystander effect (Latané & Darley, 1970). According to the model, the presence of third parties inhibits interventions in emergency situations, even when recognized as harmful or dangerous, due to the assumption that someone else will take the initiative to help (Obermaier et al., 2023). In this regard, Latané and Darley (1970) identified three underlying psychological processes behind the bystander effect and spectators' passivity in emergency cases. These phenomena include diffusion of responsibility, where the presence of more people reduces the sense of individual responsibility to act; apprehensive evaluation, which reflects the fear of being judged or negatively interpreted when intervening; and pluralistic ignorance, occurring when people base their actions on the visible reactions of others in uncertain situations. The interaction between these three psychological processes often results in a lack of intervention (Nickerson et al., 2014). Specifically, when it comes to hate speech occurring online, this reality becomes evident in a report by the UK Safer Internet Centre (2016) that shows that the majority of youths (82%) had witnessed online hate, with most of them (53%) choosing not to take any action regarding such content. Furthermore, 58% of young individuals who witnessed online hate admitted they wouldn't be able to tell whether it crosses legal boundaries. Also, almost half (45%) of these witnesses expressed concerns about speaking up, fearing becoming targets themselves.

The actions of witnesses in these incidents become crucial as they can both contribute to mitigating or perpetuating these harmful behaviors. Supporting the victim or confronting the aggressor can discourage such behaviors (Rudnicki et al., 2022), while witness passivity can inadvertently

legitimize or perpetuate hate speech since silence or lack of intervention can be interpreted as consent, normalization, or even tacit acceptance of these harmful attitudes. Thus, constructive intervention regarding online hate speech (e.g., defending the victim, reporting the incident, confronting the aggressor, or supporting the victim) can deter its normalization, support affected groups, and foster a safer, more inclusive online environment for all (Rudnicki et al., 2022). Furthermore, counterspeech (e.g., presenting facts, pointing out logical inconsistencies in hateful discourses, targeting the perpetrators, supporting the victims, disseminating neutral messages, or flooding a discussion with unrelated content) plays a pivotal role in discouraging hate speech by providing direct and constructive responses aimed at interrupting, challenging, and reducing the spread of these harmful messages (Garland et al., 2022).

The Bystander Intervention Model outlines three key steps necessary for action: firstly, recognizing an emergency situation and acknowledging the need for help; secondly, feeling a personal obligation to intervene; and finally, accepting this responsibility and taking action (Latané & Darley, 1970). Indeed, recent studies based on the Bystander Intervention Model suggest that perceiving hate speech as threatening or harmful enhances individuals' readiness to take action. This heightened perception escalates their sense of personal responsibility to take a stand (Leonhard et al., 2018). Hence, it is crucial for individuals to recognize the severity of the situation and assume responsibility for action.

Media and political influence

Political speeches and the media play a crucial role in amplifying and normalizing hate speech. These opinion makers validate and legitimize these attitudes and behaviors by framing the existence of minority groups as an imminent threat (Pereira et al., 2010), and as malevolent agents aiming to subjugate the ingroup (Esses et al., 2013). This representation not only contributes to growing rejection and hostility toward these groups, but also morally validates, normalizes, and legitimizes hostile and violent

hate-directed attitudes toward them (Obaidi et al., 2018). These narratives contribute to an ingroup centrality by presenting the ingroup objectives as being threatened by the mere existence of outgroups (Hogg & Adelman, 2013) while also perpetuating the idea that the ingroup is morally and ideologically superior to the outgroup (Woitzel & Koch, 2022), which can thus justify the morality of acting in a hostile and violent manner towards these groups.

These speeches may be particularly effective in times of uncertainty (e.g., economic crises and political instability). Feelings of identity uncertainty, i.e., uncertainty about how to act as an ingroup member and how the ingroup relates to other groups, may rise, which triggers an urgent need for clear and unequivocal information about the group's prototype, leading to a pursuit of leadership that provides clarity regarding social identity (Hogg, 2018). In their discourses, leaders may exploit this uncertainty, presenting themselves as the only ones capable of taking decisive actions to protect the ingroup (Hogg et al., 2010). These autocratic leaders and the extremist groups they usually belong to can effectively reduce uncertainty as they offer a sense of security and stability, thus becoming an attractive option for those who are highly uncertain (Hogg, 2018). Identification with these groups may foster the adoption of aggressive non-normative behavior, including hate speech, by increasing one's susceptibility to radical and exclusionary attitudes (Gøtzsche-Astrup et al., 2020).

With the online context emerging as the primary source of information for many people, especially in the case of social networks, these distorted perspectives of the world are further exacerbated within "echo chambers" – spaces where groups reaffirm and strengthen their ideologies and opinions without being exposed to alternative views, thus reinforcing their pre-existing convictions (Goel et al., 2023). These spaces facilitate the viral dissemination of unverified statements perpetuating and encouraging hatred. This rapid spread of often unchecked information feeds distorted and harmful narratives, increasing division and promoting hostile and violent attitudes toward minority groups (FRA, 2016c). This propagation is driven

by the circulation of misinformation, often stemming from political speeches and exacerbated by media channels (FRA, 2016c). In this regard, a report by FRA (2014) regarding hate speech and crimes against Jews effectively reflects this minority group's perception, with 75% of participants considering antisemitic comments on the internet as a fairly or very big current problem in their country, and 73% consider that it has increased in the last five years. Additionally, 59% of participants consider antisemitic comments in the media as a fairly or very big current problem in their country, while 44% of participants consider the same for political speeches and debates. Indeed, these discourses perpetuate and normalize hate speech, making it challenging to adopt effective measures to combat it, as it is perceived as legitimate behavior that does not require reporting or correction (Harel et al., 2020).

Fluid and contextual nature of hate speech

Hate speech exhibits a fluid, adaptable, and heavily context-dependent nature. For instance, during the pandemic, there was an alarming increase in expressions of hatred directed toward Asians (Kim & Kesari, 2021) and individuals of Asian descent (Haft & Zhou, 2021), revealing the ability of these discourses to quickly transform and adapt to current circumstances. As such, the COVID-19 pandemic context was conducive to spreading stereotypes, conspiracy theories, and incite xenophobia arising from the perception of threat and the unfair attribution of blame for the spread of the SARS-CoV-2 virus to these groups (Kim & Kesari, 2021). More recently, a wave of hatred stemming from the Israel-Palestine conflict has increased expressions of Islamophobia and Antisemitism (CAIR, 2023; ADL, 2023).

Moreover, identifying hate speech becomes challenging when it is indirectly conveyed through contextual nuances and coded expressions. This challenge is particularly evident in online communication, where language becomes highly specialized. For example, specific terms understood solely by certain groups are used as concealed racial or ethnic insults. Additionally, deliberate misspelling tactics often evade detection by online

systems designed to identify hate speech (Siegel, 2020). Moreover, the use of more subtle language complicates its classification as hate speech since it may not be readily recognized as such (Parekh et al., 2012). These practices may complicate identification for average online users and impede the recognition of hate speech within the reporting systems of online platforms (Siegel, 2020), as these systems commonly rely on predetermined word lists, making it challenging to identify offensive content that employs unconventional codes or terms (Parekh et al., 2012).

Conclusion

In summary, online hate speech is a complex and multifaceted phenomenon. The advancement of online platforms and social networks has intensified the rapid spread of hate, often evading moderation systems and being aggravated by the negative and hostile portrayal of minority groups made by the media and political speeches. Its ability to adapt to current contexts, such as the pandemic and geopolitical conflicts, exposes challenges in its effective detection. Additionally, the lack of consensus on the legal definition of hate speech and the difficulty in detecting and measuring it poses significant challenges in its prevention and combat. Therefore, a joint effort between platforms, politicians, media, and society becomes imperative to implement more assertive policies regarding this social phenomenon. Raising awareness of the effects of hate and stimulating collective responsibility in denouncing these manifestations is crucial. Empowering media literacy and critical content analysis is essential for confronting hate-inciting speeches, discouraging intolerance, and promoting an inclusive and respectful environment for all.

References

ADL. (2023). *ADL reports unprecedented rise in antisemitic incidents post-oct. 7.* https://www.cair.com/press_releases/cair-reports-sharp-increase-in-complaints-reported-bias-incidents-since-107/

Alkiviadou, N. (2017). Regulating hate speech in the EU. In S. Assimakopoulos, F. H. Baider, & S. Millar (Eds.), *Online hate speech in the European Union: A discourse-analytic perspective* (pp. 6-10). SpringerOpen.

Aluru, S. S., Mathew, B., Saha, P., & Mukherjee, A. (2020). *Deep learning models for multilingual hate speech detection*. arXiv. https://doi.org/10.48550/arXiv.2004.06465

Baider, F. H., Assimakopoulos, S., & Millar, S. (2017). Hate speech in the EU and the C.O.N.T.A.C.T. Project. In S. Assimakopoulos, F. H. Baider, & S. Millar (Eds.), *Online hate speech in the European Union: A discourse-analytic perspective* (pp. 1-6). SpringerOpen.

Bleich, E. (2011). The rise of hate speech and hate crime laws in liberal democracies. *Journal of Ethnic and Migration Studies*, *37*(6), 917-934. https://doi.org/10.1080/1369183X.2011.576195

Boeckmann, R. J. & Liew, J. (2002). Hate speech: Asian American students' justice judgments and psychological responses. *Journal of Social Issues*, *58*(2), 363-381. https://doi.org/10.1111/1540-4560.00265

Brettschneider, C. (2013). Value democracy as the basis for viewpoint neutrality: A theory of free speech and its implications for the state speech and limited public forum doctrines. *Northwestern University Law Review, 107*, 603–646.

CAIR. (2023). *CAIR reports sharp increase in complaints, reported bias incidents since 10/7.* https://www.cair.com/press_releases/cair-reports-sharp-increase-in-complaints-reported-bias-incidents-since-107/

Cohen-Almagor, R. (2019). Racism and hate speech: A critique of Scanlon's contractual theory. *First Amendment Studies, 53*(1–2), 41–66. http://doi.org/10.1080/21689 725.2019.1601579

Council of Europe. (1950). *European Convention on Human Rights*. https://www.echr.coe.int/documents/d/echr/convention_eng

Council of Europe. (2008). *Council Framework Decision 2008/913/JHA - Framework Decision on combating certain forms and expressions of racism and xenophobia by means of criminal law*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=LEGISSUM%3Al33178

Crandall, C. S., Eshleman, A., & O'Brien, L. T. (2002). Social norms and the expression and suppression of prejudice: The struggle for internalization. *Journal of Personality and Social Psychology, 82*(3), 359–378. https://doi.org/10.1037/0022-3514.82.3.359

ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., & Belding, E. (2018). Peer to peer hate: Hate speech instigators and their targets. *12th International AAAI Conference on Web and Social Media, ICWSM 2018*, Icwsm, 52–61.

Esses, V. M., Medianu, S., & Lawson, A. S. (2013). Uncertainty, threat, and the role of the media in promoting the dehumanization of immigrants and refugees. *Journal of Social Issues, 69*(3), 518–536. https://doi.org/10.1111/josi.12027

European Commission. (2014). *Report from the Commission to the European Parliament and the Council on the implementation of Council Framework Decision 2008/913/JHA on combating certain forms and expressions of racism and xenophobia by means of criminal law*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52014DC0027

European Commission. (2016). *The EU Code of conduct on countering illegal hate speech online*. https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

European Commission. (2022). *Countering illegal hate speech online: 7th evaluation of the Code of Conduct*. https://commission.europa.eu/system/files/2022-12/Factsheet%20%207th%20monitoring%20round%20of%20the%20Code%20of%20Conduct.pdf

Fortuna, P. & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR), 51*(4), 1-30. https://doi.org/10.1145/3232676

FRA. (2014). *Discrimination and hate crime against Jews in EU Member States: Experiences and perceptions of antisemitism*. Publications Office of the European Union.

FRA. (2016a). *Ensuring justice for hate crime victims: Professional perspectives*. Publications Office of the European Union.

FRA. (2016b). *Second European Union Minorities and Discrimination Survey*. Publications Office of the European Union.

FRA. (2016c). *Incitement in media content and political discourse in EU Member States*. Publications Office of the European Union.

FRA. (2018). *Hate crime recording and data collection practice across the EU*. Publications Office of the European Union.

FRA. (2021). *Encouraging hate crime reporting - The role of law enforcement and other authorities*. Publications Office of the European Union.

FRA. (2023). *Online content moderation - Current challenges in detecting hate speech*. Publications Office of the European Union.

Garland, J., Ghazi-Zahedi, K., Young, J. G., Hébert-Dufresne, L., & Galesic, M. (2022). Impact and dynamics of hate and counter speech online. *EPJ data science*, *11*(1), 1-24. https://doi.org/10.1140/epjds/s13688-021-00314-6

Gelber, K. (2017). Hate speech-definitions & empirical evidence. *Constitutional Commentary*, *32*, 619-629.

Gelber, K. & McNamara, L. (2016). Evidencing the harms of hate speech. *Social Identities*, *22*(3), 324-341. https://doi.org/10.1080/13504630.2015.1128810

Goel, V., Sahnan, D., Dutta, S., Bandhakavi, A., & Chakraborty, T. (2023). Hatemongers ride on echo chambers to escalate hate speech diffusion. *PNAS nexus*, *2*(3), 1-10. https://doi.org/10.1093/pnasnexus/pgad041

Gøtzsche-Astrup, O., Van den Bos, K., & Hogg, M. A. (2020). Radicalization and violent extremism: Perspectives from research on group processes and intergroup relations. *Group Processes & Intergroup Relations*, *23*(8), 1127-1136. https://doi.org/10.1177/136843022097031

Haft, S. L. & Zhou, Q. (2021). An outbreak of xenophobia: Perceived discrimination and anxiety in Chinese American college students before and during the COVID-19 pandemic. *International Journal of Psychology*, *56*(4), 522-531. https://doi.org/10.1002/ijop.12740

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

Hogg, M. A. (2018). Self-uncertainty, leadership preference, and communication of social identity. *Atlantic Journal of Communication*, *26*(2), 111-121. https://doi.org/10.1080/15456870.2018.1432619

Hogg, M. A., Meehan, C., & Farquharson, J. (2010). The solace of radicalism: Self-uncertainty and group identification in the face of threat. *Journal of Experimental Social Psychology*, *46*(6), 1061–1066. https://doi.org/10.1016/j.jesp.2010.05.005

Hogg, M. A. & Adelman, J. (2013). Uncertainty–identity theory: Extreme groups, radical behavior, and authoritarian leadership. *Journal of Social Issues*, *69*(3), 436–454. https://doi.org/10.1111/josi.12023

Jubany, O. & Roiha, M. (2016). *Backgrounds, experiences and responses to online hate speech: A comparative cross-country analysis*. https://www.rcmedia-freedom.eu/Publications/Reports/Backgrounds-Experiences-and-Responses-to-Online-Hate-Speech-AComparative-Cross-Country-Analysis

Keen, E. & Georgescu, M. (2016). *Manual para o combate contra o discurso de ódio online através da educação para os direitos humanos - Edição revista 2016 [Handbook for combating online hate speech through human rights education - 2016 revised edition]*. Council of Europe.

Kim, J. Y. & Kesari, A. (2021). Misinformation and hate speech: The case of anti-Asian hate speech during the COVID-19 pandemic. *Journal of Online Trust and Safety*, *1*(1), 1-14. https://doi.org/10.54501/jots.v1i1.13

Latané, B. & Darley, J. M. (1970). *The unresponsive bystander: Why doesn't he help?* Appleton-Century Crofts.

Latour, A., Perger, N., Salaj, R., Tocchi, C., & Viejo, P. (2017). *ALTERNATIVAS - Agir contra o discurso de* ódio *através de contranarrativa*s *[ALTERNATIVES - Acting against hate speech through counter-narratives]*. Council of Europe.

Laub, Z. (2019, June 7). Hate speech on social media: Global comparisons. *Council on Foreign Relations Backgrounder*. www.cfr.org/backgrounder/hate-speech-socialmedia-global-comparisons

Leets, L. (2001). Explaining perceptions of racist speech. *Communication Research*, *28*(5), 676-706. https://doi.org/10.1177/009365001028005005

Leets, L. (2002). Experiencing hate speech: Perceptions and responses to anti-semitism and antigay speech. *Journal of Social Issues*, *58*(2), 341-361. https://doi.org/10.1111/1540-4560.00264

Leonhard, L., Rueß, C., Obermaier, M., & Reinemann, C. (2018). Perceiving threat and feeling responsible. How severity of hate speech, number of bystanders, and prior reactions of others affect bystanders' intention to counterargue against hate speech on Facebook. *Studies in Communication and Media*, *7*(4), 555-579. https://doi.org/10.5771/2192-4007-2018-4-555

Mathew, B., Illendula, A., Saha, P., Sarkar, S., Goyal, P., & Mukherjee, A. (2020). Hate begets hate: A temporal study of hate speech. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 1-24. https://doi.org/10.1145/3415163

Mullen, B. & Rice, D. R. (2003). Ethnophaulisms and exclusion: The behavioral consequences of cognitive representation of ethnic immigrant groups. *Personality and Social Psychology Bulletin, 29*, 1056–1067. https://doi.org/10.1177/0146167203254505

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

Näsi, M., Räsänen, P., Hawdon, J., Holkeri, E., & Oksanen, A. (2015). Exposure to online hate material and social trust among Finnish youth. *Information Technology & People*, *28*(3), 607-622. https://doi.org/10.1108/ITP-09-2014-0198

Nickerson, A. B., Aloe, A. M., Livingston, J. A., & Feeley, T. H. (2014). Measurement of the bystander intervention model for bullying and sexual harassment. *Journal of Adolescence*, *37*(4), 391–400. https://doi.org/10.1016/j.adolescence.2014.03.003

Obaidi, M., Thomsen, L., & Bergh, R. (2018). "They think we are a threat to their culture": Meta-cultural threat fuels willingness and endorsement of extremist violence against the cultural outgroup. *International Journal of Conflict and Violence*, *12*(12), 1-13. https://doi.org/10.4119/UNIBI/ijcv.647

Obermaier, M., Schmuck, D., & Saleem, M. (2023). I'll be there for you? Effects of Islamophobic online hate speech and counter speech on Muslim in-group bystanders' intention to intervene. *New Media & Society*, *25*(9), 2339–2358. https://doi.org/10.1177/14614448211017527

OSCE/ODIHR. (2009). *Hate crimes law: A practical guide*. Organization for Security and Co-operation in Europe.

OSCE/ODIHR. (2023). *Hate crime data*. https://hatecrime.osce.org/hate-crime-data

Parekh, B. (2012). Is there a case for banning hate speech? In M. Herz & P. Molnar (Eds.), *The Content and Context of Hate Speech: Rethinking Regulation and Responses* (pp. 37–56). Cambridge University Press.

Pena, P., Teles, F., & Coelho, P. (2022, November 16). Quase 600 membros das forças de segurança usam as redes sociais para violar a lei [Almost 600 members of the security forces use social media to violate the law]. *Público*. https://www.publico.pt/2022/11/16/sociedade/noticia/quase-600-membros-forcas-seguranca-usam-redes-sociais-violar-lei-2027932

Pereira, C., Vala, J., & Costa-Lopes, R. (2010). From prejudice to discrimination: The legitimizing role of perceived threat in discrimination against immigrants. *European Journal of Social Psychology*, *40*(7), 1231-1250. https://doi.org/10.1002/ejsp.718

Ray, J. (2019, June 5). During Pride Month, YouTube shows its true colors. *Columbia Journalism Review*. https://www.cjr.org/analysis/carlos-maza-youtube-steven-crowder.php

Roussos, G. & Dovidio, J. F. (2018). Hate speech is in the eye of the beholder: The influence of racial attitudes and freedom of speech

beliefs on perceptions of racially motivated threats of violence. *Social Psychological and Personality Science*, *9*(2), 176-185. https://doi.org/10.1177/1948550617748728

Rudnicki, K., Vandebosch, H., Voué, P., & Poels, K. (2023). Systematic review of determinants and consequences of bystander interventions in online hate and cyberbullying among adults. *Behaviour & Information Technology*, *42*(5), 527-544. https://doi.org/10.1080/0144929X.2022.2027013

Ruwandika, N. & Weerasinghe, A. (2018). Identification of hate speech in social media. 1*8th International Conference on Advances in ICT for Emerging Regions (ICTer)*. https://doi.org/10.1109/icter.2018.8615517

Sarmento, D. (2016). A liberdade de expressão e o problema do "hate speech". S*apere Aude*, *6*(12), 755. https://doi.org/10.5752/p.2177-6342.2015v6n12p755

Seixas, S. R. P. M. M., Fernandes, L., & Morais, T. (2016). Bullying e cyberbullying em idade escolar [Bullying and cyberbullying at school age]. *Revista de Psicologia da Criança e do Adolescente, 1*(7), 205-210.

Siegel, A. (2020). Online hate speech. In J. Tucker & N. Persily (Eds.), *Social media and democracy: The state of the field* (pp. 56-88). Cambridge University Press.

Steering Committee on Anti-Discrimination, Diversity and Inclusion (CDADI). (2023). *Study on preventing and combating hate speech in times of crisis*. Council of Europe Publishing.

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology and Behavior*, *7*(3), 321–326. https://doi.org/10.1089/1094931041291295

Tontodimamma, A., Nissi, E., Sarra, A., & Fontanella, L. (2021). Thirty years of research into hate speech: Topics of interest and their evolution. *Scientometrics*, *126*, 157-179. https://doi.org/10.1007/s11192-020-03737-6

UK Safer Internet Centre. (2016). *Creating a better internet for all: Young people's experiences of online empowerment + online hate*. https://childnetsic.s3.amazonaws.com/ufiles/SID2016/Creating%20a%20Better%20Internet%20for%20All.pdf

United Nations. (1948). *Universal Declaration of Human Rights*.

United Nations. (2019). *United Nations strategy and plan of action on hate speech*. https://www.un.org/en/genocideprevention/documents/advising-and-mobilizing/Action_plan_on_hate_speech_EN.pdf

Weber, A. (2009). *Manual on hate speech*. Council of Europe Publishing.

Woitzel, J. & Koch, A. (2023). Ideological prejudice is stronger in ideological extremists (vs. moderates). *Group Processes & Intergroup Relations*, *26*(8), 1685-1705. https://doi.org/10.1177/13684302221135083

Wojatzki, M., Horsmann, T., Gold, D., & Zesch, T. (2018). Do women perceive hate differently: Examining the relationship between hate speech, gender, and agreement judgments. In A. Barbaresi, H. Biber, F. Neubarth, & R. Osswald (Eds.), *Proceedings of the 14th Conference on Natural Language Processing (Konvens 2018)*. Austrian Academy of Sciences Press.

THE PROPS PROJECT: INTERACTIVE NARRATIVES AS COUNTERPOINTS TO ONLINE HATE SPEECH IN VIDEO GAMES

Ana Filipa Martins
/ University of Algarve, Portugal

Bruno Mendes da Silva
/ University of Algarve, Portugal

Alexandre Martins
/ University of Algarve, Portugal

Susana Costa
/ University of Algarve, Portugal

Video games and online hate speech

Although online hate speech (OHS) is a widespread phenomenon across all virtual media, it is also important to examine and understand how it manifests in the extensive and growing field of video games, specifically online multiplayer games. In 2023, there were 1.1 billion online video gamers globally (Clement, 2023). The act of playing online video games and being a part of a gaming community has become an indispensable element for many individuals. Besides the excitement of competition, online gaming offers the chance for different social interactions. For many online gamers, these forms of experience are both unique and rewarding, enabling players to stay connected and form new bonds. Such positive features can have a meaningful impact in the digital and physical lives of human beings (Costa et al., 2023a; Kwak & Blackburn, 2014), having the capacity to regulate behaviors. They can provide certain skills

in problem-solving, verbal cognitive performance, and conflict resolution (Malik, 2008). Multiplayer games help individuals build digital communities with shared conducts and values, where they work toward collective and common objectives (Rivera-Vargas & Mino-Puigcercos, 2018).

But despite the clear and important benefits of online gaming, much like major sports events, the sentiments of competition that derive from these contexts, often create moments of frustration, anger and tension, which can lead to aggressive verbal expressions, reactions which are sometimes perceived as frequent and acceptable (Breuer, 2017; Uyheng & Carley, 2021). During online gameplay, chat interactions are commonplace, which can range from compliments to ironic commentaries, or from insults to discrimination, harassment and attacks based on personal and social traits - real or perceived.

Previous studies confirm that factors such as anonymity or the lack of consequences might encourage toxic discourses, a form of alleviating such negative feelings (Soral et al., 2018; Breuer, 2017). These behaviors, however, can produce physical and psychological harmful effects, both on victims and perpetrators. On top of that, the continued exposure to OHS might result in desensitization and lead to an increased apathy for the victims and a sense of conformity regarding prejudiced attitudes (Costa et al., 2023b; Uyheng & Carley, 2021; Costa et al., 2020; Soral et al., 2018; Breuer, 2017).

In online multiplayer matches, players frequently engage in complex and dynamic interactions, which can happen through unmoderated voice conversations, increasing the probability of conflicts and toxic language use. Veteran gamers may also harass others who show a lack of experience, by deceiving, sabotaging, or engaging in acts which may spoil their overall enjoyment of the game. Besides the issue of competence, or lack of it, when speaking about the motives or targets of hate speech, women and minorities are frequent targets, partly due to underrepresentation in the narratives of video games (Williams et al., 2009; Fragoso et al., 2017). Therefore, these groups can be more exposed to rejection and vulnerable to OHS (Silva & Martins, 2024).

For some time now, studies have focused their concerns on the fact that video games, especially the ones characterized as violent, might have a pernicious influence on the minds of individuals (Maher, 2016). However, the interest may not be, for example, a young person fictionally shooting another, but instead what is being said when that action takes place, i.e, shooting while spouting racist, xenophobic, misogynistic, and homophobic slurs. Because OHS acts on undermining the dignity of others, it becomes important to better understand the dynamics of this (online) issue and to seek solutions to effectively address it.

Regulating and containing online hate speech in video games

Even though scientific studies have been increasingly focusing on hate speech, there is still a lot that remains unknown regarding its pervasiveness, motives and repercussions across multiple digital platforms (Siegel, 2020). Moreover, it is also important to examine further the effectiveness of practical methods to limit hate speech and the collateral consequences of such interventions. Despite existing laws explicitly prohibiting certain forms of hate speech, how these policies are or should be applied in digital spaces is still a matter of serious and continuous debate. The Council of Europe (CoE) (2022) asserts that measures targeting hate speech must always be appropriate and proportional to the severity of the respective instance. While some manifestations require a response from criminal, civil or administrative law, others may require non-legal reactions, such as education or awareness initiatives (Silva & Martins, 2024).

When it comes to the gaming industry and companies, their policies are typically focused on restricting or removing users. In both video games and video game platforms there are protocols capable of detecting forbidden words or messages using artificial intelligence (AI) and machine learning. The relationship between AI and hate speech is a complex and multifaceted matter. It can be used as a tool to identify, monitor, and combat OHS, as algorithms can be trained to detect linguistic patterns associated with this form of speech on different platforms, enabling a quicker and more effective response from

companies and online moderators (Alkiviadou, 2022). For example, the company Activision has sought to combat toxic voice chat in Call of Duty with the ToxMod tool developed by Modulate, which is able to identify discriminatory content and acts of harassment in real time (Acres, 2023).

Although current technologies have significantly evolved in detecting harmful text using AI, there are still limitations. Difficulties in understanding the context and the speaker's intention can pose risks to freedom of expression, access to information, and equality (Alkiviadou, 2022). Automated mechanisms trained to detect offensive speech may exhibit biased datasets, making them incapable of identifying the nuances of language (Alkiviadou, 2022; Finck, 2019). There are also concerns about the use of AI in the propagation and amplification of hate speech. Content recommendation algorithms can, either intentionally or inadvertently, promote harmful content by highlighting sensationalist or extremist messages. Because of its limitations, the CoE has proposed a series of recommendations to protect human rights regarding the use of AI (Council of Europe, 2019). To ensure that content moderation is done in a way that protects human rights and public discourse it is important to consider a balance between automation and human oversight when moderating hate speech.

Besides moderation and restriction, there is also the notion of "don't feed the troll", an expression that is sometimes brought to this discussion. When confronted with offensive messages, indifference can sometimes be the best or only form of reaction. This principle acknowledges that the offenders who practice OHS are not just aiming to cause harm but are also seeking a response to magnify the "problem" they are promoting (Costa et al., 2023a; Titley et al., 2014). In this perspective, ignoring and not reacting to offensive messages might be the most suitable response.

Limiting OHS lies in balancing different fundamental rights, such as freedom of expression and the rights to equality, inclusion, and protection. The issue lies in the way we handle different fundamental rights, such as freedom of expression and the rights to equality, inclusion, and protection. This

also presents a significant challenge in the realm of online video games. Strategies like ignoring or reporting and removing hateful messages are valid options to curb this problem, although they might not be sufficient. On the one hand, content removal presents itself as a powerful tool that can be over or misused by certain agents, potentially removing people from the discourse that, despite being offensive, may have not incite violence or intentionally inflict emotional distress. On the other, the indifference strategy might hinder the scrutiny and discussion about the causes and motives behind OHS (Latour, 2017). Therefore, it is crucial for people and institutions to explore alternative approaches.

Media and digital literacy

Fostering a healthy and safe gaming culture can also require diverse pedagogical approaches. As a symptom of deep societal issues, (online) hate speech cannot be adequately addressed solely through monitoring, control, and censorship. Pedagogical interventions are equally essential (Council of Europe, 2021). Current younger generations are extensively using digital devices, the Internet, social media, and video games for information-seeking, knowledge acquisition, communication, socialization, as well as for entertainment, creative expression, and collaboration. Media literacy could and should be used to transform these digital realities into opportunities for learning and skill development (Liu, 2020; Shaffer et al., 2005).

Digital literacy becomes a crucial approach for empowering young people, equipping them to recognize and develop resilience against hate speech. The CoE's recommendations underscore the significance of ensuring that children, young adults, and educators use ICTs effectively. The CoE (2021) also elucidates the concept of digital citizenship, emphasizing that it entails:

- using technology safely, ethically, and responsibly, along with possessing the skills to engage positively, critically, and competently in the digital environment;

- engaging positively in the creation, work, sharing, socialization, research, communication, and learning in a constantly evolving society influenced by digital technologies;
- knowing how to wisely enjoy the different forms of entertainment that technologies allow, as well as balancing one's exposure to media appropriately to avoid excessive or inappropriate use.

Concerning video games, the CoE emphasizes that digital citizenship education strives to promote a positive and mindful gaming culture for future generations. By enhancing comprehension about video games' economic models, structures, languages, risks, and opportunities it is possible to cultivate societies that better appreciate the inherent value of this medium. This knowledge can also be useful to enhance the overall quality of content produced, while mitigating potential risks and issues, as it happens with other forms of media.

Digital citizenship education aims to empower individuals to practice informed and conscious citizenship by understanding key concepts such as freedom of expression and social and civic responsibility, while also enhancing people's resilience against extremist messages, misinformation, and hate speech (Council of Europe, 2016). The effectiveness of digital literacy hinges on individuals' roles in relation to OHS; whether as victims, bystanders, propagators, or offenders (Latour, 2017). Digital literacy initiatives, promoted through gaming culture and other audiovisual and digital media, should strive to promote democratic values and digital citizenship, fostering positive behaviors that mitigate this problem (Silva & Martins, 2024).

Interactive narratives as pedagogical tools

Playing video games can be understood as a learning process that engages players almost unconsciously. To progress, they present several challenges and obstacles that must be overcome. Some games are quite complex, including intricate instructions, controls, and gameplay. During the act of play, children become deeply immersed in problem-solving while learning the

game's internal mechanics. They are unafraid to make mistakes, persisting through the toughest parts of the game. Through tutorials subtly embedded in the system, children learn all the required commands simply by playing (Malik, 2008).

In this context, there are objects functionally designed for educational purposes. "Serious games" (Laamarti et al., 2014) like *Human Resource Machine* (2016), *Father and Son* (2017), *A Gender Story* (2018), and *Bury me, my love!* (2017) represent a pedagogical approach that can be very useful in teaching complex skills and fostering deeper understanding of specific topics. These games often provide immersive experiences that promote critical thinking, empathy, and awareness of social issues. Alternatively, the use of "entertainment video games,", e.g., Triple-A Games, also proves suitable in capturing students' attention, while motivating them to learn. This method is based on a type of tangential learning (Council of Europe, 2021), which suggests that some people will independently begin a learning process if parents, teachers, or game designers introduce topics in an engaging and stimulating context.

They can also address serious themes or current events, encouraging individuals to think critically. This is a process that involves reflection, discernment, analysis, evaluation, and responsible action, which can, in some cases, help to dismantle stereotypes and oppose prejudices, providing valuable opportunities for classroom discussions. Issues such as ethics, morality, empathy, racism, legal matters, gender and LGBT representation, violence, current events, and other sensitive or controversial subjects can be explored through video games, either by playing them, showing excerpts, or discussing the games. Overall, video games have the potential to be important educational resources, motivating young people to acquire specific skills by fostering critical thinking, cooperation, and interaction, while also stimulating the development of physical and emotional skills through immersive narratives, puzzles, and logical or deductive problem-solving (Laamarti et al., 2014).

Similarly, other types of interactive narratives have emerged as promising tools for both entertainment and educational purposes (Si & Marsella, 2014).

Narratives have historically been conveyed through various media, such as films, books, and oral storytelling. With the rapid advancement of computing technologies, another form of media has become prominent: interactive narratives and films. Here the spectator becomes an active participant in the story, choosing the paths of the events. They offer unique opportunities to exercise cultural and social skills, combining the pedagogical potential of narratives with an active learning experience (Silva & Martins, 2024).

By engaging the audience and establishing direct connections between actions and outcomes, users are encouraged to explore alternative story paths and invest more time in learning. Interactivity and the power of agency foster new and stronger motivations for learning (Park & Kim, 2008). Transforming audiovisual content, such as fictional and non-fictional narratives, into interactive formats creates new opportunities for actively involving the public in social causes. This medium allows viewers to become active participants, encouraging broader dialogue in a digital environment and promoting greater pluralism, tolerance, and engagement (Wintonick, 2013). Interactive narratives can effectively raise awareness and mobilize citizens around critical social issues, such as online hate speech (OHS).

Finally, the advancement of gaming and entertainment technologies promotes the development of new e-Learning methods. One example in this category are pedagogical itineraries (PIs), which are valuable tools for teachers and educators. They enable children and young people to explore subjects in innovative ways. One of its significant features is the connection they establish between the analog and digital spaces or objects. PIs offer a diverse group of hands-on activities aimed at digital education, allowing students to engage more deeply and enhance their diverse skills (European Commission, 2020).

A notable project in this field was Play Your Role (European Commission, 2020), funded by the European Commission's Rights, Equality, and Citizenship Programme (2014-2020). This initiative involved the collaboration of various European partners with diverse pedagogical experiences

and cultural contexts. They used innovative tools to develop 15 PIs, each addressing OHS from different perspectives. The goals of the project were to enhance video games and gamification processes as tools to reinforce positive behaviors in adolescents; and to raise awareness and understanding of OHS, xenophobia, and racism, promoting empathy and critical thinking.

PIs commonly employ gaming technology and design principles to educate students in an engaging and playful manner within the classroom. The use of games and other interactive tools and narratives for educational purposes is widely supported by previous studies, which confirm their effectiveness in teaching (Martín-SanJosé et al., 2014; Taran, 2005; Albert & Mori, 2001). This approach has proven effective in creating software that motivates and engages users more effectively during the learning process (Silva & Martins, 2024).

PROPS Project

Due to the potential of using interactive tools and narratives as pedagogical resources, and their relevance to talk about societal issues such as OHS, the "PROPS - Interactive Narratives Propose Pluralist Speech" project was born. It is an initiative funded by the Portuguese Foundation for Science and Technology, developed by the University of Algarve, University of Beira Interior, Santarém Polytechnic University and Universidade Aberta. PROPS is a project focused on media education, aimed at curbing hate speech in online video games. The objective is to address this problem in online gaming by developing multiple interactive digital narratives designed to attract, motivate, and engage educators, teachers, children, and young people in reflecting on and discussing OHS. PROPS is composed by a team of researchers and artists, with extensive experience in the fields of education, media literacy and digital media-arts, and in the creation and production of interactive films, video games, and pedagogical itineraries based on gamification.

One of the goals of this project is to enhance understanding of the complexities inherent in gaming communities. To achieve this, the project's team

conducted a comprehensive study using a variety of data collection tools and techniques, incorporating both quantitative and qualitative data processing methods. This approach allowed the project's group to gain a deeper understanding into this topic and field of study. By employing surveys and focus groups as data collection tools, the team was able to gather perspectives from young gamers regarding their experiences with offensive messages, toxic online environments, and game conduct norms.

The analysis of the data collected from students aged 10 to 18, from schools in the Algarve region of southern Portugal, revealed what hate speech means to them, how it has impacted them, and what potential responses they envision to address this phenomenon. These studies, along with a substantial corpus of scientific research on (online) hate speech in video games, have been fundamental in clarifying the complex dynamics of this issue within the gaming landscape, as well as in the development of six interactive narratives aimed at raising awareness about OHS in these settings.

Survey and focus groups

The first stage of this initiative involved the creation and dissemination of a survey in three schools in the Algarve region, southern Portugal (Costa et al., 2024). The survey sought to gain comprehensive insights into the personal experiences and perspectives of individuals aged 10 to 18 regarding OHS in video games and social platforms related to games. Besides the presentation of the empirical results, the goal was also to engage in critical dialogue about the wider implications of these findings. By doing so, it aimed to contribute to the ongoing conversation about online safety and digital citizenship.

The research focused on key aspects of the phenomenon, including (1) the extent of young people's exposure to OHS; (2) the most common types of OHS encountered; (3) the games and platforms where OHS incidents were most frequently observed; (4) the reactions and responses to such content; and (5) the occurrence of OHS during gameplays.

To understand more deeply about the viewpoints, behaviors and motives surrounding OHS in gaming communities, the team also conducted three focus groups held in multiple schools in the Algarve (Costa et al., in press). Nineteen students, between 12 and 18, from four different nationalities participated in the study. Employing this focus group approach granted the project team with direct access to participants' opinions, perspectives and experiences. The script used during each session included questions that sought to get students' first-hand accounts with OHS encountered in video games and gaming platforms, as well as their reactions to those experiences, or to their overall perspectives on the matter.

Both studies were conducted in accordance with the guidelines set forth by the Directorate General for Education (Direção-Geral da Educação), regarding their application in schools. These guidelines ensured compliance with privacy, security, protection, and confidentiality standards in respect to the collection and processing of personal data. The participation of the students was voluntary, and they were given clear information about the aim and the objectives of these studies and project.

The survey results provided direct reports of young people's exposure to OHS during video gaming. Nonetheless, they also indicate that many players do not perceive themselves as being under threat in these situations. Participants demonstrated a clear awareness that hate speech is prevalent in online video games, yet they also tended to minimize its impact. This juxtaposition of perspectives may suggest a certain level of acceptance of OHS as an inevitable or unproblematic aspect of these gaming experiences.

The survey also uncovered that a significant trigger for OHS was often linked to victims' inexperience and insufficient gaming skills, as observed by victims, bystanders and offenders. Additionally, both the survey and focus groups highlighted numerous instances of hate speech targeting individuals based on factors such as gender, ethnicity, nationality, physical appearance, sexual orientation, and religion. Furthermore, both studies indicated that the most popular games were more susceptible to hateful

discourse. Focus group discussions also showed that OHS was more likely to occur in less regulated or unregulated environments, or in highly competitive settings. Participants cited anonymity, frustration, and entertainment as potential reasons behind such behavior.

Regarding emotional reactions, the data reveals paradoxical results, where individuals who admit disliking toxic interactions also perceive them as somewhat acceptable. Nevertheless, a significant number of participants across both studies reported experiencing various negative emotions: in the focus groups feelings of insecurity and fear were commonly mentioned. OHS can detrimentally affect victims' self-esteem or be utilized to promote extremist ideologies.

Certain testimonies and opinions also highlighted a connection between online and offline behaviors. Students noted that concerns expressed online can spill over into their physical lives, illustrating the offline impact of OHS to some extent. As for participants' behavioral responses to OHS, the survey data paradoxically showed that it was both common to report and ignore incidents, while many respondents lacked awareness of the consequences of reporting hate speech, underscoring a absence of transparency from gaming companies regarding what are the possible outcomes of engaging in toxic behavior.

Finally, participants in the focus groups proposed various actions to address the issue, with some advocating for stricter regulations and sanctions, such as permanent bans for offenders to deter the spread of such behaviors. Others had different opinions, suggesting using educational tools to empower individuals to recognize and combat hate speech.

While hate speech extends beyond video games and online spaces, these findings underscore the importance of digital literacy initiatives and pedagogical resources in fostering safer and more inclusive digital gaming environments. Achieving this requires a comprehensive approach involving players, students, teachers, and academia. The results from the survey and focus groups provided valuable insights for the PROPS team into the

motivations, triggers, targets, and responses to OHS in video games, aiding in the development of relevant counter-narratives to raise awareness about this critical issue.

The six counter-narratives

The data collected from the survey and focus groups helped to consolidate the production of six different interactive narratives produced by a multidisciplinary team of nine researchers, from the University of Algarve, University of Beira Interior and Santarém Polytechnic University. The goal was to render these narratives - two video games, an interactive comic book, an interactive audiovisual essay, and two pedagogical itineraries - into educational tools to be used in the school and classroom contexts, directed at students between the ages of 10 and 18, to reflect and discuss different themes related to OHS. All the narratives will be available on the project's official website (CIAC, 2024).

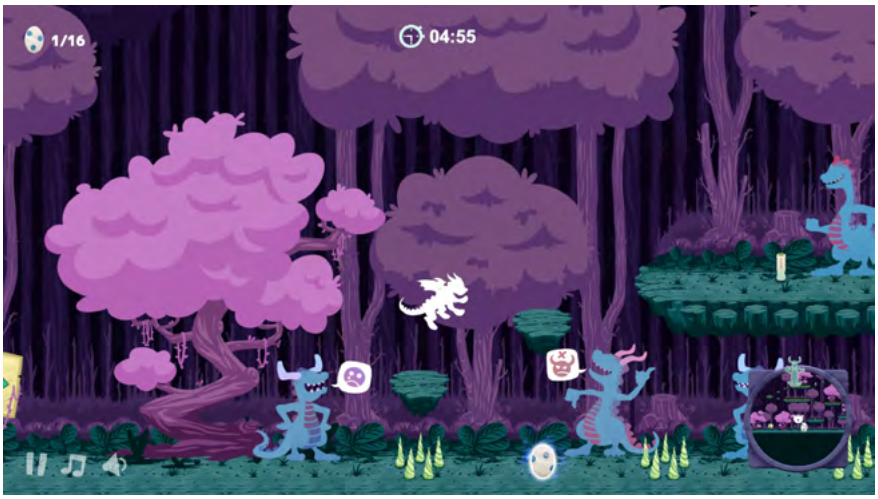
a. *Unbully*

Unbully is a computer game compatible with MacOS, Windows, and Linux platforms, requiring a display monitor, mouse, and keyboard. It is a single-player platform and exploration game where players control the dragon Tales. The game was designed for ages 10 to 12 with a playtime of 15 to 30 minutes.

Set in the Black Forest, the game features various elements: traps, poisonous slugs, friendly and oppressive dragons, and collectible items (Figure 1). The main objective is to locate the sacred dragon statue and to gather dragon eggs. Collecting these eggs grants the player the ability to defeat enemies, essential for navigating the entire game map and finishing the game. The secondary characters have a dual role: they influence gameplay and challenge the player while also being part of the narrative. The oppressive dragons mock and harass the main character. By collecting eggs, friendly dragons appear to create a positive and supportive atmosphere throughout the player's journey.

Unbully aims to counter OHS by promoting enjoyment in challenging scenarios, reflecting on toxic attitudes towards underperformance, and demonstrating growth through positive behaviors.

Figure 1: *Unbully*. Game's world with oppressive dragons mocking the main character



b. IN[The Hate Booth]

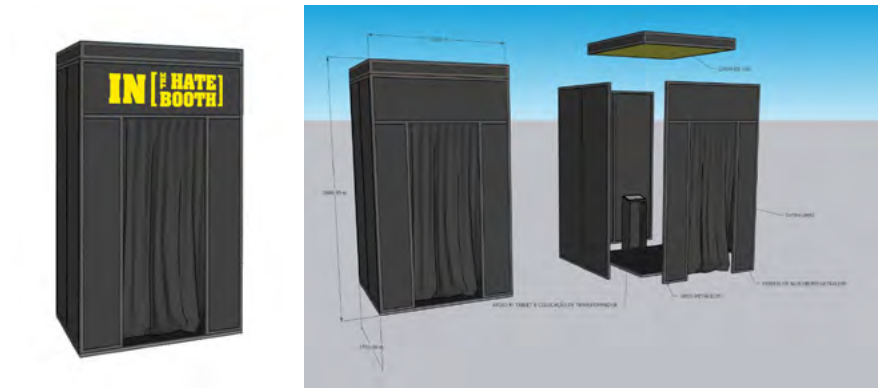
The installation-game *IN[The Hate Booth]* is composed of two dimensions: a physical light booth and a virtual interactive fiction game that leads users on a journey of discovery. This game reflects toxic disinhibition in cyberspace, where trolls and bots operate within its monitored yet unregulated environment.

The physical element of this installation, i.e., the booth, is an immersive room limited by three panels and a curtain, which invites users to step into its luminous environment (Figures 2 & 3). Inside, there is a pulpit that holds a 13-inch screen with an announcement stating that a blog will be shut down by its authors: Hazuka and Gotcha. Through comments, users can access the blog's post-mortem: a collection of pages from the inception of the webpage until its shutdown.

This takes users on a journey of exploration, uncovering the reasons behind the website's shutdown, leading them to discover a series of archived messages that gradually reveal the continuous and exponential increase of hate comments among the former webpage followers. This game allows the players to comment on posts, continuing the narrative and involving them in actively understanding and addressing OHS.

IN[The Hate Booth] aims to achieve four main effects: a) immerse the user in the experience; b) induce feelings of confusion and discomfort; c) refer to the concepts of stage and role, reminding users of their rights and responsibilities in both digital and physical dimensions; and d) symbolize the processes of remote communication inherent to digital media, which contribute to the emergence of OHS and toxic disinhibition.

Figures 2 & 3: *IN[The Hate Booth]*. Physical structure of the installation-game



c. *THE UPDATE*

THE UPDATE is an interactive comic book, designed to illustrate how children and young people are increasingly exposed to hateful comments (Figures 4 & 5). In this story, *The Best Game in the World* has a new update which brings new conduct rules. The story follows Leo, a video gamer who engages in OHS and must reflect on his behavior after losing the privilege of

playing with his friends. It also explores the concept of an excessive code of conduct where all players engaging in hate speech end up isolated.

The comic includes both an e-book/print-ready version and an interactive digital/WordPress version. The use of interactivity aims to engage readers in constructing the narrative, allowing them to make decisions that steer the story in different directions. *THE UPDATE* encourages reflection on the impact that OHS has on the characters and the gaming environment.

The data collected from the focus groups and surveys showed that a significant number of children and young people have already encountered hate speech. Noteworthy examples include prejudiced insults based on ethnicity or gender, with the primary trigger being a lack of gaming experience. Additionally, some participants showed acceptance of this phenomenon by ignoring or tolerating negative comments.

Figures 4 & 5: *THE UPDATE*. Example of a page and cover of the interactive comic book

Based on these insights, the comic medium was chosen to address this theme for the following reasons: a) it is a pedagogical tool capable of explaining concepts clearly and educationally; b) visual imagery aids in understanding concepts and makes the story more engaging; and c) comics often use *Grawlixes*, *Nittles*, or *Quimps* - terms coined by Mort Walker in his 1980's *The Lexicon of Comicana* -, typographic symbols that represent obscenities, which simplifies the depiction of hate speech in a manner suitable for the target audience. This comic aims for children and young people to recognize that some things are not acceptable, whether directed at a friend or an online stranger, while fostering critical thinking about the toxic behaviors encountered on digital platforms.

d. *G.G.*

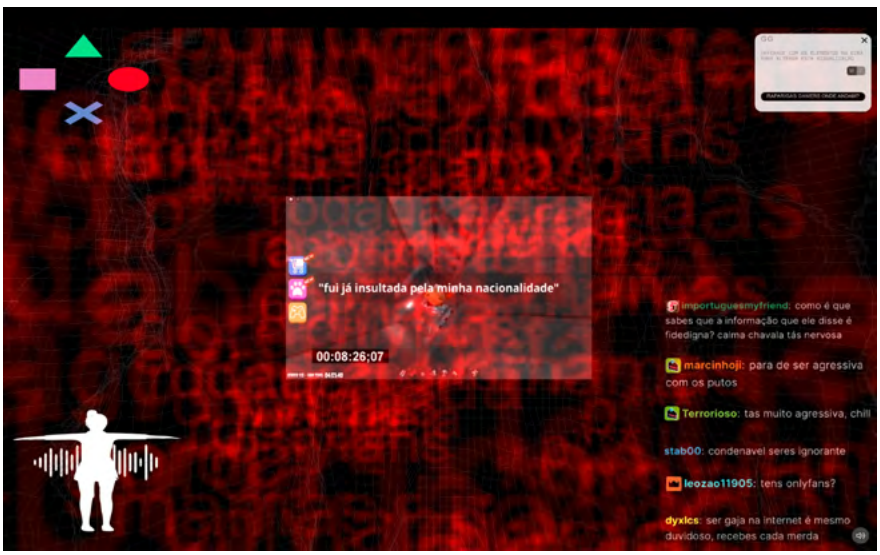
G.G. ("Good Game" or "Gamer Girl") is an interactive audiovisual narrative that allows users to play with various written, aural, and visual objects (Figure 6). From chat messages, gameplay footage, and sound recordings of Portuguese streamers, as well as comments and memes from gaming community forums, this digital collage aims to explore the broader gaming landscape, addressing the significance and consequences of good and bad performance in video games, as well as aspects of the female experience in these environments.

The survey and focus groups conducted within the PROPS project helped identify the most played video games among students, the streamers they follow most frequently, and common targets or triggers for insults among players. Together with content taken from online gaming communities and forums, the information gathered helped to consolidate the selection of the themes for this audiovisual narrative: a) "the impact of performance in online video games" (the survey identified player performance as the primary trigger for OHS); and b) "male/female experiences in video games" (gender issues were frequently reported as motives for hate speech, with multiple first-hand accounts from the focus groups, of young girls talking about the challenges of being a female player in online games). Using these key ideas,

a narrative was crafted using real examples, primarily from Portuguese gaming personalities and communities.

G.G. was produced on cables.gl and can be played from any browser, requiring only a computer with internet access. The goal was to create a user-friendly application that utilizes minimal and easily accessible materials. This app is designed for classroom use by students aged 15 to 18 within the Portuguese school system. The initiative aims to develop an educational interactive experience to help teachers and students reflect on and discuss topics related to online video games. This educational resource can prompt discussions on issues such as hate speech and its prevalence in gaming spaces, gender differences in online multiplayer game experiences, and the potential consequences of toxic gaming environments. This tool intended to help students: a) relate the examples presented in the narrative to their own experiences; b) analyze and discuss digital content; c) use digital content critically, effectively, and safely; and d) develop skills related to inclusivity and respect in online contexts.

Figure 6: *G.G.*. Screen of the interactive audiovisual narrative with various image and textual elements

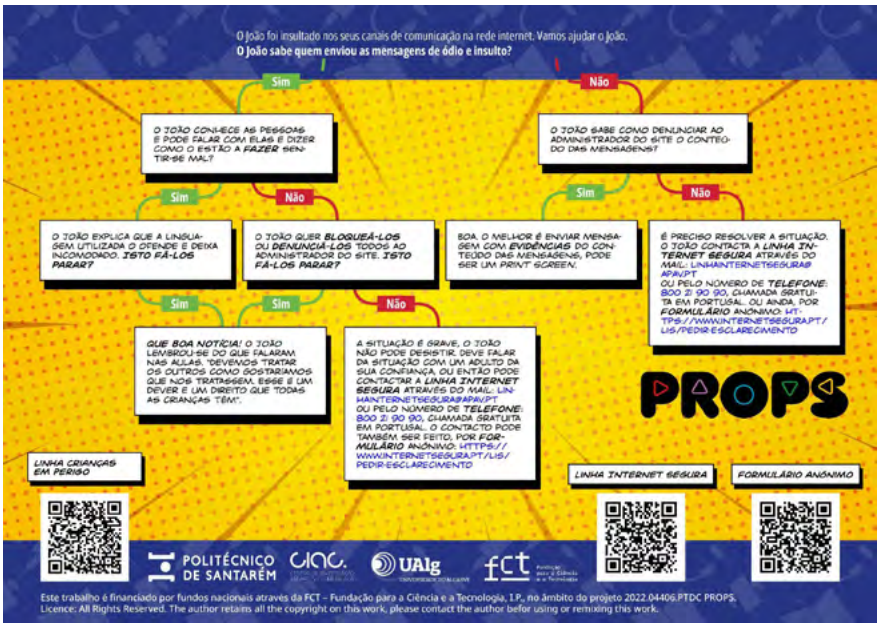


e. *Polígrafo, a Verdade e o Outro!*

Polígrafo, a Verdade e o Outro! [*Polygraph, the Truth and the Other!*] is a pedagogical itinerary (PI) that employs a personalized and varied learning approach to address the topic of OHS, tailored to each student's experiences and needs. This experience creates space for individualized learning adapted to personal and group goals, based on personal experiences or interests.

Though primarily focused on Citizenship and ICTs, for students aged 10 to 14, this PI can be implemented across various subjects, such as Portuguese classes, aligning with the goals of teachers or the needs and motivations of students.

Figure 7: *Polígrafo, a Verdade e o Outro!*. Decision tree where students must help João deal with insulting messages



In the PI's narrative, participants embark on a journey to seek justice and truth, summoned to undertake a special mission as "Defenders of Digital Well-being." Their task is to combat the "Digital Trolls," evil creatures

spreading false and hateful messages via social media. This sets the stage for students, who must help the city of "Ciberlândia" confront these threats and address online disharmony.

The adventure begins with an independent, home-based, activity guided by a script that directs students to watch a video on hate speech provided by the Portuguese Institute for Sport and Youth. To deepen their understanding, they are encouraged to explore additional online educational resources, including the Portuguese Safe Internet Center, the No Hate Speech Movement website, and the podcast ZigZaga na Net. Subsequently, in the classroom, students first engage with the narrative, where they must assist the main character, João, in making decisions to counter the hate speech infiltrating the game city he oversees (Figure 7). How will they help João restore truth in Ciberlândia and combat the fears, threats, and online discord they are facing?

f. *Better, not Best*

Better, Not Best is an interactive narrative and a pedagogical itinerary aimed at countering ethically reprehensible behaviors in online entertainment, targeting young people aged 15 to 18. The primary goals are to promote respect for others, foster healthy competition, and encourage empathy.

In this interactive narrative, students choose one of four inexperienced non-player characters (NPCs) as their teammate (Figure 8). They must continuously make decisions based on the teammate's performance, where they can opt to assist, teach, guide, reprimand, or criticize. Their choices can be rewarded or penalized in two different categories: "empathy towards the teammate" and "team performance." Depending on the paths they take, and the empathy shown towards the NPCs, students will be presented with one of four possible outcomes: "Perfect!", "Applause!", "Caution!", and "Danger!".

The narrative is composed of ten scenes, each presenting participants with three options that influence future routes, final scores, and outcomes. The narrative emphasizes coordination within the team and the importance of

choices made to achieve desired goals. The core message is that winning should not come at any cost, and the ends do not justify the means.

The main pedagogical and ethical aim is to demonstrate the importance of accommodating diverse personalities, contributions, motivations, perspectives, strategies, and skill levels to achieve a common goal. The key lesson is that true success requires acceptance and inclusivity, encapsulated in the message: being better is more important than being the best.

Figure 8: *Better, not Best*. Panel where students can choose an NPC as their teammate

Final considerations

Addressing OHS requires a multifaceted strategy that considers its complexity and its connection to broader societal issues. Effective regulation, both nationally and internationally, is crucial to mitigate its harmful effects on individuals and groups of people. Despite this, finding the balance between protecting people from harm and upholding freedom of expression in online spaces is a significant challenge.

Even though online video games promote connectivity and new social bonds, they often become hotspots for verbal abuse, discrimination, and harassment, which can profoundly affect players' virtual and real lives. The practices governing multiplayer games can enable abuse, conflict, and extremist rhetoric, with women and minorities frequently becoming prime targets. This highlights the need to address systemic issues such as underrepresentation and harmful stereotypes in video games and gaming communities, and to develop innovative strategies to effectively combat OHS.

There is ongoing debate on the best methods to address hate speech in online video games. Strategies include reporting and removing offensive content, educating users, using counter-speech, or ignoring hate speech. Each approach has its potential downsides, such as the risk of misusing reporting systems, which could silence individuals who did not directly incite violence. Consequently, lawmakers, institutions, and gaming companies must explore alternative, effective methods to curtail this issue. Creating safer and more welcoming gaming communities, hinges on promoting inclusivity. Combating discrimination and fostering mutual respect through collective action is essential to maintaining the enjoyment and social benefits of online gaming. Effectively managing OHS in video games demands a holistic approach, incorporating technological solutions, community moderation, and digital citizenship education (Silva & Martins, 2024).

Digital literacy initiatives are essential for empowering individuals to identify and combat hate speech effectively. Media literacy programs that promote critical thinking and empathy can guide users in navigating online spaces responsibly and ethically. Interactive storytelling, video games, and pedagogical itineraries present unique opportunities to engage students in experiences that enhance cultural understanding and empathy. Integrating these tools into educational curricula allows educators to help students explore complex topics interactively, fostering creativity, critical thinking, and encouraging participation in positive, inclusive virtual communities.

Acknowledgements:

This work is supported by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., in the framework of the project 2022.04406. PTDC PROPS.

References

Acres, T. (2023, September 1). Call Of Duty using AI to listen out for hate speech during online matches. *Sky News*. https://news.sky.com/story/call-of-duty-using-ai-to-listen-out-for-hate-speech-during-online-matches-12952063

Albert, D. & Mori, T. (2001). Contributions of cognitive psychology to the future of e-learning. *Bulletin of Graduate School of Education, Hiroshima University, Part I: Learning and curriculum development* (50), 25-34. http://doi.org/10.15027/18093

Alkiviadou, N. (2022). Artificial Intelligence and online hate speech moderation. *SUR, 19*(32), 101-112.

Breuer, J. (2017). Hate speech in online games. In K. Kaspar, L. Gräßer, & A. Riffi (Eds.) *Online hate speech, Perspektiven auf eine neue form des Hasses* (pp. 107-112). Kopaed.

Clement, J. (2023). *Online gaming - statistics & facts*. Statista. https://www.statista.com/topics/1551/online-gaming/#topicOverview

CIAC (2024). *PROPS – Interactive Narratives Propose Pluralist Speech*. Research Center for Arts and Communication. https://props.ciac.pt/en

Costa, S., Mendes da Silva, B., Martins, A. F., & Martins, A. (in press). Exploring online hate speech in gaming communities: Insights from focus groups with young players. *Communication Studies, 39*.

Costa, S., Mendes da Silva, B., Martins, A. F., & Martins, A. (2024). Perceptions about online hate speech in games and gaming communities: Results from a survey in Portugal. *Journal of Cyberspace Studies*, *8*(1), 145-176. https://doi.org/10.22059/jcss.2024.95910

Costa, S., Tavares, M., Bidarra, J., & Silva, B. M. (2023a). IN[The Hate Booth]: A gamified installation to counteract hate speech. In A. L. Brooks (Ed.), *ArtsIT, interactivity and game creation* (pp. 161–173). Springer. https://doi.org/10.1007/978-3-031-28993-4_12

Costa, S., Tavares, M., Bidarra, J., & Silva, B. M. (2023b). The enredo game-installation: A proposal to counter hate speech online. In N. Martins & D. Brandão (Eds.), *Advances in design and digital communication III* (pp. 307–320). Springer. https://doi.org/10.1007/978-3-031-20364-0_27

Costa, S., Tavares, M., Silva, B. M., Isca, B., & Cerol, F. (2020). Hate speech in video games and in online gaming communities – A state of art. *Revista Comunicando, 9*(1), 261-278.

Council of Europe (2022). *Combating hate speech*. Council of Europe.

Council of Europe (2021). *Educating for a Video Game culture - A map for teachers and parents*. Council of Europe.

Council of Europe (2019). *Unboxing Artificial Intelligence: 10 steps to protect Human Rights.* Council of Europe.

Council of Europe (2016). *Developing media literacy and critical thinking through education and training - Council conclusions (30 May 2016)*. Council of Europe.

European Commission (2020). *Play your role: The Project Toolkit*. European Commission.

Finck, M. (2019). *Artificial Intelligence tools and online hate speech.* Centre on Regulation in Europe. https://cerre.eu/wp-content/uploads/2020/05/CERRE_Hate-Speech-and-AI_IssuePaper.pdf

Fragoso, S., Recuero, R., & Caetano, M. (2017). Violência de gênero entre gamers brasileiros: Um estudo exploratório no Facebook. *Lumina, 11*(1), 1-21. https://doi.org/10.34019/1981-4070.2017.v11.21367

Kwak, H. & Blackburn, J. (2014) Linguistic analysis of toxic behavior in an online video game. In L. Aiello & D. McFarland (Eds.), *6th International Conference on Social Informatics* (pp. 209–217). Springer. https://doi.org/10.1007/978-3-319-15168-7_26

Laamarti, F., Eid, M., & Saddik, A. (2014). An overview of serious games. *International Journal of Computer Games Technology, 2014*(3), 1-15. https://doi.org/10.1155/2014/358152

Latour, A., Perger, N., Salaj, R., Tocchi, C., & Otero, P. V. (2017). *WE CAN! Taking action against hate speech through counter and alternative narratives*. Council of Europe. https://rm.coe.int/wecan-eng-final-23052017-web/168071ba08

Liu, H. (2020). Early adolescents' perceptions and attitudes towards gender representations in video games. *Journal of Media Literacy Education, 12*(2), 28-40. https://doi.org/10.23860/JMLE-2020-12-2-3

Maher, B. (2016). Can a video game company tame toxic behaviour? *NATURE, 531*, 568-571. https://doi.org/10.1038/531568a

Malik, K. (2008). Impact of computer and video games on the development of children. In V. Godara (Ed.), *Risk assessment and management in pervasive computing* (pp. 343–351). IGI Global. https://doi.org/10.4018/978-1-60566-220-6.ch019

Martín-SanJosé, J., Juan, M., Gil-Gómez, J., & Rando, N. (2014). Flexible learning itinerary vs. linear learning itinerary. *Science of Computer Programming, 88*(1), 3-21. https://doi.org/10.1016/j.scico.2013.12.009

Park, S. & Kim, K. (2008). The use of pedagogical agent as a tool to improve learning interest: Based on the distinction between individual interest and situational interest. In R. McFerrin, R. Weber, R. Carlsen, & D. A. Willis (Eds.), *Proceedings of the Society for Information Technology and Teacher Education International Conference* (pp. 2777–2781). AACE.

Rivera-Vargas, P. & Miño-Puigcercós, R. (2018). Los jóvenes y las comunidades virtuales. Nuevas maneras de aprendizaje y de participación social en la sociedad digital. *Páginas De Educación, 11*(1), 67-82. https://doi.org/10.22235/pe.v11i1.1554

Shaffer, D. W., Squire, K. D., Halverson, R., & Gee J. P. (2005). Video games and the future of learning. *Phi Delta Kappan, 87*(2), 104-111. http://doi.org/10.1177/003172170508700205

Si, M., Marsella, S. C., & Pynadath, D. V. (2005, May 6). *THESPIAN: An architecture for interactive pedagogical drama* [Paper presentation]. Conference on Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology (AIED '05), Amsterdam, The Netherlands. https://dl.acm.org/doi/abs/10.5555/1562524.1562605

Silva, Mendes da, & Martins, A. (Orgs.) (2024). Interactive Narratives Propose Pluralist Speech Results of a project aimed at researching and countering online hate speech in video games. CIAC Edições. https://doi.org/10.34623/ggqf-0g28

Siegel, A. A. (2020). Online Hate speech. In N. Persily & J. A. Tucker (Eds.), S*ocial media and democracy: The state of the field, prospects for reform* (pp. 56-88). Cambridge University Press.

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

Taran, C. (2005, July 5-8). *Motivation techniques in eLearning* [Paper presentation]. International Conference on Advanced Learning Technologies (ICALT), Kaohsiung, Taiwan. https://doi.org/10.1109/ICALT.2005.206

Titley, G., Keen, E., & Földi, L. (2014). *Starting points for combating hate speech online*. Council of Europe.

Uyheng, J. & Carley, K. M. (2021). Characterizing network dynamics of online hate communities around the COVID-19 pandemic. *Applied Network Science*, *6*(1), 1–21. https://doi.org/10.1007/s41109-021-00362-x

Williams, D., Martins, N., Consalvo, M., & Ivory, J. (2009). The virtual census: Representation of gender, race and age in video games. *New Media Society, 11*(5), 815-834. https://doi.org/10.1177/1461444809105354

Wintonick, P. (2013). *Doc the world: My personal manifesto and Credo*. POV Magazine.

EDITOR

Branco Di Fátima is a non-fiction writer with a PhD in Communication Sciences from the University Institute of Lisbon (ISCTE). He wrote the book *Dias de Tormenta* (Geração Editorial, 2019) and edited the collection *Hate Speech on Social Media: A Global Approach* (LabCom Books, EdiPUCE, 2023). He has published more than 90 scientific works and participated in 11 research projects funded by national and international organizations. His current research focuses on the pathologies and dysfunctions of democracy, journalism studies, online hate speech, and social network analysis. He is currently a contracted researcher at LabCom – University of Beira Interior (UBI) in Portugal.

AUTHORS

Alexandre Martins (UAlg/CIAC) is a Ph.D. student in Digital Media-arts (UAlg/UAb), Master in Heritage, Arts and Cultural Tourism (ESE-IPP, 2020) and has a degree in Foreign Languages and Cultures (ESE-IPP 2018). He has pre- viously collaborated with Direção Regional de Cultura do Norte (DRCN), providing support in the revision and edition of the monographic collection "Património a Norte" and with Cine-Clube de Avanca in the organization of its documentary archive. He is a researcher at Centro de Investigação em Artes e Comunicação (CIAC), where he develops studies in digital arts and audiovisual communication.

Amrita Bhattacharjee is a 4th year Computer Science Ph.D. student at the School of Computing and Augmented Intelligence at Arizona State University. She is advised by Dr. Huan Liu. Broadly the goal of her research is data and resource efficient representation learning for NLP applications. With this broad research goal in mind, her work primarily focuses on: (i) leveraging large language models (LLMs) for human-intensive tasks in a machine learning pipeline, (ii) robust detection of AI-generated content, and (iii) NLP applications in domain adaptation and domain generalization. She earned her B.Tech in Computer Science and Engineering from Heritage Institute of Technology, Kolkata, India, where she worked on social network analysis on developer social networks on large-scale open-source software systems.

Ana Filipa Martins (UAlg/CIAC) is an adjunct professor at the University of the Algarve's School of Education and Communication, Portugal, where she lectures in the degree course in Communication Sciences and the master's program in Communication and Digital Media. She holds a PhD in Communication and is a researcher at CIAC – Research Centre for Arts and Communication. She has participated in various research projects in the fields of media production and media literacy, as a researcher, local coordinator and Co-IR. She has also promoted and coordinated various projects in collaboration with news organizations and other entities.

Andre Oboler is an Honorary Associate in the Law School at La Trobe University and CEO of the Online Hate Prevention Institute. He has served as an adviser to governments, companies, charities, and individuals. He is an expert member of the Australian government's delegation to the International Holocaust Remembrance Alliance and chairs the IEEE's Global Public Policy Committee. He is a former Vice President of the IEEE Computer Society. He holds a PhD in Computer Science from Lancaster University (UK) and a law degree, LLM (Juris Doctor) and B. Comp. Sci. (Hons) from Monash University (Australia).

Berta Chulvi Ferriols holds a PhD in Social Psychology from the Universitat de València and a degree in Journalism from the Universitat Politècnica de València. She holds a postdoc research position in the Pattern Recognition and Human Language Technology (PRHLT) Research Center at the Universitat Politècnica de València. She is also assistant professor at the Social Psychology Department at Universitat de València. Her research focuses on hate speech detection, disinformation, stereotypes against minority groups and social minority influence. She is a member of the EC IBERIFIER project on monitoring the threats of disinformation.

Bruno Mendes da Silva (UAlg/CIAC) holds a Habilitation and Postdoctoral degree in Communication, Culture and Arts at the University of Algarve (UAlg). He is the President of the Technical-Scientific Council and Director of the Department of Communication at the School of Education and Communication (ESEC) of UAlg and Vice-coordinator of the Research Center in Arts and Communication (CIAC). He has been an invited speaker in Portugal, Spain, France, Italy, Tunisia, Mozambique, Brazil, Colombia, Sri Lanka, Thailand, Vietnam, South Korea and China. He is Director of Rotura Journal (Scopus) and Coordinator of Metared Portugal's Educational Technologies Working Group.

Fábio Malini is professor in the Department of Communication Studies at the Federal University of Espírito Santo (UFES). He holds a PhD in Communication and Culture from the Federal University of Rio de Janeiro (UFRJ) and completed a post-doctorate at King's College London, United Kingdom. He is the coordinator of the Internet and Data Science Laboratory (LABIC) at UFES.

Gabriel Herkenhoff Coelho Moura is Assistant Researcher at the Internet and Data Science Laboratory (LABIC) at UFES. He holds a PhD in Philosophy from the Federal University of Paraná (UFPR) and is a post-doctoral researcher affiliated with the Federal University of São Paulo (UNIFESP).

Jéssica do Nascimento Oliveira has a Master's degree in Linguistic Studies from the Federal University of Espírito Santo (PPGEL/UFES).

Joshua Garland holds a Ph.D. in Computer Science and an M.S. in Applied Mathematics from the University of Colorado at Boulder. He currently serves as the Interim Director and Associate Research Professor at Arizona State University's Center on Narrative, Disinformation, and Strategic Influence (NDSI). Before this, he was an Omidyar and Applied Complexity Fellow at the Santa Fe Institute. Dr. Garland's areas of interest focus on a wide variety of complex systems and applications, such as the climate, ecology, politics, dynamical systems, and many more. However, his primary focus at NDSI is online human social dynamics, where Joshua combines social theory, machine learning, time series analysis and natural language processing to understand the fascinating intersections of AI, global politics, social media, narratives, disinformation and strategic influence operations.

Karoline Fernandez De La Hoz Zeitler holds a PhD in Medicine from the Universitat de Barcelona. She did an MSc in Epidemiology at the London School of Hygiene and Tropical Medicine. Since April 2015, she has been the Director of the Spanish Observatory on Racism and Xenophobia (OBERAXE) of the State Secretariat for Migrations at the Ministry of Inclusion, Social Security and Migrations. She is the Chair of the Council of Europe's Committee on Intercultural Inclusion. She is the Spanish representative at the European Agency for Fundamental Rights (FRA), at the Office for Democratic Institutions and Human Rights (ODIHR), and a member of the HighLevel Group for Hate Crime and Hate Speech of the European Commission, and also a member of the group of experts on intercultural integration (ADI_INT) of the Council of Europe.

Márcia Bernardo possesses a Master's degree in Psychology and is currently a doctoral candidate in Psychology at the Faculty of Psychology and Educational Sciences, University of Porto. She is an active member of the Social Psychology Laboratory at the University of Porto and has been involved in several projects since 2021, addressing pivotal societal issues like

online hate speech and populism. Her primary research interests revolve around civic and political participation, extremist social movements, intergroup conflicts, populism, and discrimination against immigrants.

Mariana Magalhães has a PhD in Psychology from the Complutense University of Madrid, and is an active member of the Social Psychology Laboratory at the University of Porto and of the Violence and Crime Permanent Observatory at the University Fernando Pessoa. Since 2021, she has been involved in different research projects (e.g., EducHate – Detect, Combat, Prevent, Educate), addressing in her research interests and work topics such as violence in intimate relationships, bullying, cyberbullying and hate speech.

Paolo Rosso is Full Professor at the Universitat Politècnica de València, where he is also a member of the Pattern Recognition and Human Language Technology (PRHLT) Research Center. His research is mainly focused on the detection of harmful information in social media, both fake news and hate speech, also when hate speech is conveyed in an implicit form via stereotypes or figurative language. Currently, he is the PI of several research projects: XAI-DisInfodemics on eXplainable AI for disinformation and conspiracy detection during Infodemics, FairTransNLP on Fairness and Transparency for equitable NLP applications in social media; and a member of MARTINI project on profiling and detecting malicious actors in online social networks through AI, and of the EC EDMO hub IBERIFIER-Plus on monitoring the threats of disinformation.

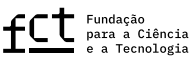
Reed Van Schenck is Assistant Professor of Communication at IE University and has a PhD in Communication from the University of Pittsburgh. His research focuses on white supremacist networks and digital platform governance in the United States, using critical frames from rhetoric, media studies, and cultural studies. His work appears in venues like Communication and Critical/Cultural Studies and Media, Culture & Society.

Sara Alves is currently a doctoral candidate in Psychology, affiliated with the Psychology Center of the University of Porto. She is also a member of the Social Psychology Laboratory of the University of Porto, assisting in the research project VigilHate (VIGILANT CITIZENS AGAINST HATE: How to counter bystander apathy and increase citizens' commitment against online hate speech?). Her research interests revolve around attitudes towards immigrants, perception of threat and nationalism.

Susana Costa (UAlg/CIAC) is a Ph.D. student in digital media art at the University of Algarve (UAlg) and Universidade Aberta (UAb). She is a collaborator at the Research Center for Arts and Communication (CIAC). Her research has been published in peer-reviewed journals, and she has presented her scientific work at national and international conferences. As a researcher, Susana focuses on the intersection of education, arts, and technology. Her current research delves into the study of hate speech manifestations and effects in games and online communities of young gamers. She proposes art and game-based approaches to effectively address this pervasive problem.

Tharindu Kumarage is a Computer Science Ph.D. student at Arizona State University, where he is advised by Prof. Huan Liu. His research intersects Natural Language Processing for Social Good (NLP4SG) and AI-LLM safety. Tharindu has actively worked and published papers on domain-generalized hate speech detection, stress detection in social media, and forensics of AI-generated text, showcasing his expertise in large language models. Prior to his Ph.D. studies, Tharindu was a Consultant Research Engineer for Digital Mobility Solutions Lanka (PVT) Ltd. He earned his BSc in Computer Science and Engineering from the University of Moratuwa in Sri Lanka.

DOI FCT - LABCOM
https://doi.org/10.54499/UIDB/00661/2020



This is the third book in the **Online Hate Speech Trilogy.** It focuses on presenting methods for detecting, analysing, and combating toxic language on the Internet. Alongside the legal dilemmas born from a desire to punish hate speech disseminators, identifying online hate speech is one of the biggest challenges in the field of studies on violent narratives and virtual attacks. The authors analyse the challenges of identifying violent narratives through automation, the advantages of manually coding social media posts, and the opportunities offered by AI in this field of research.

LabCom Books
www.labcom.ubi.pt

Editorial Universidad Icesi
www.icesi.edu.co/editorial