ONLINE HATE SPEECH TRILOGY – VOL II

LEGAL CHALLENGES AND POLITICAL STRATEGIES IN THE POST-TRUTH ERA

BRANCO DI FÁTIMA [ED]

LABCOM COMUNICAÇÃO & ARTES

Editorial Universidad Icesi

ONLINE HATE SPEECH TRILOGY — VOL II

LEGAL CHALLENGES AND POLITICAL STRATEGIES IN THE POST-TRUTH ERA

BRANCO DI FÁTIMA [ED]

LABCOM COMUNICAÇÃO & ARTES

Editorial Universidad Icesi

Technical Specification

Title
Legal Challenges and Political Strategies in the Post-Truth Era — Online Hate Speech Trilogy - vol II

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
978-989-9229-08-2 (print)
978-989-9229-09-9 (pdf)

Legal Deposit
538470/24

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

VIOLENT NARRATIVES, LEGAL CHALLENGES AND POLITICAL STRATEGIES

Branco Di Fátima
/ LabCom – University of Beira Interior

This is the second book of the **Online Hate Speech Trilogy**. The work focuses on the legal challenges of combating toxic language and retaliating against those who spread hate on the Internet. Although the need for fighting violent narratives appears evident, given the role of hate in eroding trust and fragmenting the social fabric, there are many sensitive layers to the matter.

The debate is controversial because, in some cases, laws designed to combat hate speech have been used to punish political dissidents and individuals who challenge prevailing norms (Munoriyarwa, 2023; Chekol, 2023). On the other hand, opponents of these laws advocate for unrestricted unconditional freedom, which is also difficult to defend. A path of compromise needs to be defined. The fight against hate speech must not violate other rights, such as freedom of the press or freedom of expression, but must protect society, especially its most vulnerable groups.

Since there is no universally accepted definition of hate speech, identifying violent narratives and measuring their impact is not an easy task either (Müller & Schwarz, 2021). Generally, hate speech can be understood as a verbal or non-verbal attack on an individual or group, usually a social minority. However, as its roots lie in the values of a particular culture (Matamoros-Fernández & Farkas, 2021), defining what constitutes a hateful attack is dependent on those same cultural codes.

Toxic language has also become a political strategy in the post-truth era, characterised by decision-making that is based more on emotional impulses than verifiable facts (Fischer, 2021). This has occurred partly due to the popularisation of digital technologies such as social media platforms and smartphones, and also because of the way society is structured. In other words, hate speech is also shaped by the collective values of the community and the power struggles that permeate it (Di Fátima, 2023).

This book brings together chapters written by 14 authors from 9 universities, examining hate speech within their unique socio-cultural contexts. They achieve this by employing both traditional and digital methods, utilizing quantitative and qualitative data gathered from diverse digital platforms such as websites, instant messaging apps, and social media.

The authors analyse the deep origins of hate speech and its manifestations online. They highlight the weaknesses of platform self-regulation, the European Union's legal approach to combating online hate, the use of toxic language as a political weapon in Latin America, and the risks it poses to peace in Africa. In addition, the authors examine how biases from media outlets can be amplified on platforms such as Facebook, X, and YouTube, creating social divisions.

Although it is not a new phenomenon, hate speech has become increasingly complex on the Internet. It is omnipresent, interactive, and multimedia in nature (Di Fátima, 2023). Haters hide behind the anonymity provided by digital technologies and find online support for their violent ideas (Amores et al., 2021). These issues have become more prominent in the last decade, largely due to their strong correlation with populism, disinformation, and the rise of the new anti-system far-right.

This book addresses the legal challenges of combating toxic language while safeguarding freedom of expression in an era when hate speech has become a political strategy. Volumes 1 and 3 of the **Online Hate Speech Trilogy** explore the close links between disinformation, polarization, and virtual attacks. They also examine cutting-edge techniques for identifying violent

narratives and developing counter-narratives to mitigate hate speech. 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

Amores, J. J., Blanco-Herrero, D., Sánchez-Holgado, P. & Frías-Vázquez, M. (2021). Detectando el odio ideológico en Twitter: Desarrollo y evaluación de un detector de discurso de odio por ideología política en tuits en español. *Cuadernos.info*, 49(2021), 98-124. https://doi.org/10.7764/cdi.49.27817

Chekol, M. A. (2023). Ethiopian socio-political contexts for hate speech. In Di Fátima, B. (Ed.), *Hate speech on social media: A global approach* (pp. 227-254). LabCom Books & EdiPUCE.

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

Fischer, F. (2021). *Truth and post-truth in public policy: Interpreting the arguments*. Cambridge University Press.

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

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

Munoriyarwa, A. (2023). Mapping social media hate speech regulations in Southern Africa: A regional comparative analysis. In Di Fátima, B. (Ed.), *Hate speech on social media: A global approach* (pp. 203-226). LabCom Books & EdiPUCE.

CALLING FOR CONSEQUENCES: HOW TO MOTIVATE SOCIAL MEDIA COMPANIES TO BETTER MODERATE HATE SPEECH

Caitlin Ring Carlson
Seattle University, USA
carlso42@seattleu.edu

The volume of hate speech in our digital media environment is staggering. Although social media companies have enacted complex self-regulatory measures to minimize its spread and thus its impact, hate speech continues to proliferate, particularly on social media. Therefore, a change is needed. This essay argues that self-regulation is insufficient, particularly given the of profitability hateful and divisive content for publicly traded social media platforms like Meta or X. Instead, what's needed are substantial legal incentives or consequences. Requiring social media platforms to earn their shield from legal liability by complying with the Santa Clara principles or another existing framework for best practices in content moderation is one approach. Countries could also consider offering tax breaks for compliance with transparency reporting or content moderation spending or staffing. Conversely, it may be worth exploring an approach like Germany's which fines social media companies for failing to comply with existing hate speech laws. Regardless of the specific approach taken, the time to act is now.

Keywords: hate speech, social media, content moderation, media policy, media regulation

Hate speech on social media in the electoral year context in Brazil

Rubens Beçak
University of Sao Paulo, Brazil
prof.becak@usp.br

Kaleo Dornaika Guaraty
University of Sao Paulo, Brazil
kaleodornaika@gmail.com

Tiago Augustini de Lima
University of Sao Paulo, Brazil
tiagoaugustini@gmail.com

The social media are a reality for the Brazilian partisan political marketing strategy, at least since the 2018 general elections. Not surprisingly, Brazil is a major consumer of digital platforms, whether for communication or for use and performance in the marketplace. With these two data in mind and analyzing the concepts and definitions of hate speech, above all, that which is made up as speech and not as speech-act, according to J. Waldron, it appears that, in electoral years in Brazil, there is a significant increase in hate speech complaints on social media. In the first half of 2022 alone, there was a 650,0% increase in reports of hate speech compared to the same semester in 2021. Hate speech used as an electoral campaign (and in the campaign) undermines the voter's free conviction, in addition to giving voice to intolerant speeches to authoritarian candidates, which can lead to corrosion of the democratic system.

Keywords: hate speech, Brazilian elections, digital platforms, social media

POLITICAL ENGAGEMENT AND AGGRESSIVE USE OF SOCIAL NETWORKS. PRESIDENTIAL CAMPAIGNS IN A HIGHLY POLARIZED ELECTORAL SCENARIO

Adolfo A. Abadía
Universidad Icesi, Colombia
aaabadia@icesi.edu.co

Luciana C. Manfredi
Universidad Icesi, Colombia
lcmanfredi@icesi.edu.co

Juana L. Rodriguez
Universidad Icesi, Colombia
juanalrp1@gmail.com

This chapter analyzes the impact of Twitter® on the 2018 presidential elections in Colombia, in a context of high political polarization due to the negotiation process, plebiscite, and beginning of implementation of the peace agreements with the FARC. This study examines the link between aggressive messages on social networks and the intention to vote for candidates, intending to understand how political communication on this social network can propose a possible explanation for the results of the first round of the elections. During the 90-day campaign period for the first round of the presidential elections, candidates were classified as either Aggressors or Targets based on their aggressive messages on Twitter®. In general, Gustavo Petro's aggressive approach generated greater public debate and engagement during his candidacy. However, his voting intention did not significantly increase between March and May 2018. On the other hand, Iván Duque's less aggressive strategy, which mainly targeted aggressive messages, resulted in a notable growth in electoral support during the same period.

Keywords: political engagement, aggressive message, social network, presidential campaigns, elections in Colombia 2018

THE EUROPEAN LEGAL APPROACH TO FIGHT HATE SPEECH ON SOCIAL MEDIA

Ana Gascón Marcén
University of Zaragoza, Spain
angascon@unizar.es

This chapter studies the legal mechanisms developed by two of the main European organisations to fight hate speech online. Regarding the Council of Europe, the analysis covers its soft law, the Protocol to the Budapest Convention and the case-law of the European Court of Human Rights. Regarding the European Union, it assesses the E-Commerce Directive, the Audiovisual Media Service Directive, the Code of Conduct and the Digital Services Act. They delineate a roadmap for States and social media to fight hate speech while respecting freedom of expression although they have some weaknesses and are not always enforced.

Keywords: hate speech, Council of Europe, article 10 ECHR, European Court of Human Rights, EU Law

HATE POSTINGS ON SOCIAL MEDIA AND PEACE IMPERATIVES IN NIGERIA

Nosa Owens-Ibie
Caleb University, Nigeria
nosa.owens-ibie@calebuniversity.edu.ng

Eric Msughter Aondover
Caleb University, Nigeria
eric.aondover@calebuniversity.edu.ng

The potential and capacities of social media to influence and impact socioeconomic and political changes in contemporary society, is established in literature. Nigerians as active users of various social media networks continue to share their thoughts on domestic and other issues as reflection of the freedom of access. This chapter analyses the problems and patterns of hate speech on social media in Nigeria, based on the security implications and threats to peace which the spread of such hate speech could and does trigger. It shows through the tracking of posts that social media users are divided in their opinions along ethnic, regional and religious lines, and discusses relevant legislation, the impact of hate speech on press freedom and free speech, the spread of hate speech on social media, and the impacts of hate speech on peaceful coexistence. It shows a relationship between social media and the incitement of violence, and underscores the importance of addressing the trend of using social media as trigger for crises, given the pervasiveness of social media, and its features which enables people to read content anonymously and respond with disparaging remarks that mock or insult the ethnic, political, regional, and religious affiliations of other groups in the nation's diverse population. The need for a clear policy framework to foster balance between the right to free expression on social media, and the demands of peaceful coexistence, is canvassed.

Keywords: hate speech, social media, peace imperative and posting, Nigeria

THE POLITICAL USE OF HATE SPEECH THROUGH SOCIAL MEDIA IN BRAZIL

Joelma Galvão de Lemos
Federal University of Sergipe, Brazil
joelmalemos@outlook.com

Daniel Menezes Coelho
Federal University of Sergipe, Brazil
daniel7377@gmail.com

In Brazil, social media is used not only for socialization and communication, but also as a tool employed by some political groups to encourage and spread a specific kind of online participation: the "hate speech". Based on a psychoanalytic standpoint, this paper presents and analyzes a selection of speeches from 2016 (the year of the coup d'etat in Brazil) and 2018 (the subsequent election year). It becomes evident, throughout our discussion, that the logic that organizes social media platforms and algorithms has contributed to the amplification of hate speech in Brazil. This happens because social media creates online bubbles which isolate users from the very "other" with whom they should learn to co-exist, thus limiting the possibility of dialogue between different groups. Therefore, we believe that it is important to contemplate the functioning, transparency, and regulation of these social media platforms.

Keywords: hate speech, social media, politics, psychoanalysis

FREEDOM OF THE PRESS OR HATE SPEECH? REGULATING MEDIA OUTLETS IN THE POST-TRUTH ERA

Branco Di Fátima
University of Beira Interior (UBI), Portugal
brancodifatima@gmail.com

Marco López-Paredes
Pontifical Catholic University of Ecuador (PUCE), Ecuador
mvlopez@puce.edu.ec

This chapter delves into the complex dynamics of regulating media outlets in the post-truth era, focusing on the challenges posed by the proliferation of online hate speech. These issues are particularly pronounced in highly polarized societies, where media outlets often blend opinionated narratives with factual news, sometimes neglecting the ethical principles of journalism. Historically, the emphasis on media freedom has sometimes compromised other rights, facilitating the spread of hate speech. The chapter also uses Brazil as a case study, highlighting a country without a regulatory agency for media outlets and significantly impacted by political polarization over the last decade. This example illustrates how the absence of regulation can endanger liberal democracies and the safety of marginalized social groups, who are more vulnerable to online attacks.

Keywords: media regulation, hate speech, legal frameworks, media outlets, post-truth era

CALLING FOR CONSEQUENCES: HOW TO MOTIVATE SOCIAL MEDIA COMPANIES TO BETTER MODERATE HATE SPEECH

Caitlin Ring Carlson
/ Seattle University, USA

Introduction

Despite the increasingly complex framework for content moderation that has emerged in the past decade, hate speech on social media remains a problem. While the exact volume is difficult to capture, transparency reports from Facebook, Twitter, and other social media organizations indicate that there are millions of posts featuring hateful content circulating at any given time. A recent transparency report from Facebook shows that the company took action on 18 million posts containing hate speech during a three-month period (Facebook, 2023). However, documents released by Facebook whistleblower Frances Haugen in 2021 indicate that this number likely represents only 3-5 percent of the total amount of hate speech on Facebook (Allyn, 2021).

The impact of hate speech on society is far-reaching. Hate speech undermines collective autonomy by making people fear for their well-being (Leets, 2002). In extreme cases, hate speech campaigns levied via social media have led to offline violence, discrimination, and persecution. For example, a Reuters investigation done in conjunction with the Human Rights Center and the U.C. Berkley School of Law found over 1,000 posts calling the Rohingya and other Muslims in Myanmar

dogs, maggots, and rapists (Stecklow, 2018). Hate speech on Facebook has turned Myanmar into a hotbed for extremism, coinciding with religious riots across the country between Buddhists and Muslims.

On an individual level, exposure to hate speech generates similar short- and long-term effects as other kinds of traumatic experiences (Leets, 2002) and can lead to increased stress levels, as well as symptoms of depression (Wypych & Bilewicz, 2022). In addition to the negative consequences it has on victims, hate speech desensitizes people and can increase feelings of prejudice and dehumanization toward those targeted (Soral, Bilewicz & Winiweski, 2018).

Recognizing the harm caused by hate speech on social media, international bodies, individual governments, and social media organizations have undertaken various actions to address the issue. The United Nations has issued a Strategy and Plan of Action on Hate Speech, detailing its commitment to addressing the issue. Germany enacted Netz DG, a law requiring social media companies to promptly remove illegal hate speech or face substantial fines. Social media organizations like Facebook, YouTube, and Twitter, have developed extensive community standards prohibiting certain forms of hateful content on their sites. Artificial intelligence, platform users, and human content moderators are all used to identify and help remove hate speech from these sites.

In spite of these efforts, the problem remains. This has led to extensive debates among elected officials, scholars, and media activists about what should be done to address the issue. Together, members of these groups have developed several viable solutions to best deal with the problem of hate speech while respecting social media users' freedom of expression. However, I argue in this essay that consequences are missing from many of these approaches. With few exceptions, most of the proposals for how to best moderate hate speech on social media are inherently flawed because they lack incentives for compliance and punishments for failing to comply. These proposals do not account for the fact that several of the most popular

social media platforms are run by publicly traded corporations operating under a shareholder primacy approach, which prioritizes profits above all else. If the solutions proposed to tackle online hate speech remain voluntary, we can only expect platforms to comply with those that align with their financial goals.

To rectify this issue, we must first understand its origins. This essay will begin by offering a snapshot of existing approaches to regulating hate speech on social media. Next, I'll demonstrate how incorporating incentives or consequences into the existing framework would increase compliance and minimize the volume of hate speech that appears on social media.

Regulatory overview

This section will outline how various practices and policies come together to regulate hate speech on social media. I will begin by outlining how social media platforms define and moderate hate speech on their sites. Next, I will discuss platforms' legal liability in various countries. Then, the UN guidance on combating hate speech on social media will be examined, as will the individual laws of various countries. Finally, I will present the joint efforts from governments, social media companies, activists, and academics that have emerged to provide recommendations and best practices for addressing hate speech on social media.

Social Media Content Moderation

The current approach to dealing with the problem of hate speech on social media is complex and at times, convoluted. There are recommendations from international bodies, individual states' laws, voluntary partnerships between governments and platforms, and social media companies' community standards and identification and removal processes. Globally, hate speech is considered illegal in some but not all countries. Notably, in the United States, home to Meta, which owns Facebook, the world's largest social media platform, hate speech is protected by the First Amendment.

Hate speech, by its very nature, is subjective and, therefore, difficult to define. Legal definitions are often different from the definitions used by social media platforms in their content moderation efforts. Moreover, platforms themselves each have their own unique definitions for the term. On Facebook, hate speech is defined as "anything that directly attacks people based on what are known as their 'protected characteristics,' such as race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease" (Facebook, 2023). For YouTube, hate speech is content that promotes violence or hatred toward groups based on a list of 13 specific identity characteristics (YouTube, 2023). Similarly, TikTok prohibits content that attacks, threatens, incites violence against, or otherwise dehumanizes an individual or a group on the basis of 12 protected attributes (TikTok, 2023).

As private virtual spaces, social media companies are free to regulate content on their sites in any way they wish. To access the platform, users must agree to the terms of service, which often require them to adhere to the platform's community standards (Gillespie, 2018). Automatic detection using artificial intelligence and community flagging are both used to identify content that violates the platform's hate speech policy. A combination of AI technology and human content moderators then decide if the content should be removed and whether action should be taken on the account (Gillespie, 2018).

In addition to removal, platforms can also minimize the impact of content featuring hate speech without removing it entirely. Reducing algorithmic amplification of hate speech, shadow banning, which limits the ability for posts from certain accounts to be viewed, and warnings prior to posting can all be used to minimize the spread of hate speech (Land & Hamilton, 2020). Platforms can also use counter-speech and provide additional information alongside problematic posts to educate users. Some platforms, such as YouTube, may remove sharing tools for posts containing hate speech. Requiring users to use their real names and identities is another tool platforms have available to reduce hate speech (Land & Hamilton, 2020).

Despite the varied approaches available to platforms, often, the best solution is to remove hate speech entirely. If done well, this means that content that incites hatred toward individuals and groups based on their fixed identity characteristics is removed from the public sphere. When done poorly, this process can lead to the removal of expression that does not actually violate the platform's community standards for hate speech.

Platforms Protected from Legal Liability

Users in the United States and elsewhere disagree about whether the content moderation process is fair or biased. People and elected officials on the political right often feel their content is being wrongly targeted and removed. At the same time, those on the left argue that platforms are not doing enough to eliminate hate speech from their sites. Critics have also raised questions about the extent to which the platform's algorithms are being manipulated to privilege extreme content, which keeps people on the site longer and ultimately increases advertising revenue (Vaidhyanathan, 2018).

Perhaps most importantly, though, platforms' approach to hate speech regulation highlights the tension inherent in allowing private corporations to serve as arbiters of free speech. In countries where hate speech is legal, social media companies decide where the line between protected and unprotected free expression should be. In most countries where hate speech is illegal, platforms are not liable for the illegal content posted to their sites (Germany is an exception to this rule). This is true in the United States, where Section 230 of the Communications Decency Act protects computer services such as social media platforms from legal liability for illegal content users post on their sites (Communications Decency Act, 1996).

This approach is largely supported by international bodies such as the United Nations.

In 2017, the UN Special Rapporteur on Freedom of Expression adopted a Joint Declaration on Freedom of Expression, and "Fake News," Disinformation, and Propaganda (Organization for Security and Co-Operation in Europe,

2017). The Declaration strongly restated the position that intermediaries, like web-hosting platforms or social media sites, should never be held liable for content posted by third parties unless they specifically intervene in that content or refuse a court order to remove it. Since then, the Special Rapporteur on Freedom of Expression has called on social media companies to align their content moderation practices with UN Human Rights Standards, which he argues favor free expression over censorship (Kaye, 2019).

UN Guidance

The United Nations advocates for prohibiting advocacy of racial or religious hatred but encourages member states to do so in a way that minimizes suppression of speech. Section 20(2) of the International Covenant on Civil and Political Rights (ICCPR) states that "advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law" (United Nations Treaty, 1966). However, states must show that the harm of discrimination cannot be lessened by means other than the suppression of speech, such as the use of educational initiatives. Section 4 of the International Convention on the Elimination of All Forms of Racial Discrimination (ICERD), which requires signatories to condemn all propaganda and organizations based on ideas of racial or ethnic superiority, while still giving due regard to the right to freedom of thought, religion, opinion, expression, peaceful assembly and association (United Nations Treaty, 1965). Article 19 of the ICCPR holds that everyone has the right to hold opinions without interference. Restrictions on speech must be legitimate, legal, and necessary. It also says that everyone shall have the right to freedom of expression. Hate speech restrictions mandated under the ICCPR and ICERD are required to comply with these limits.

Recognizing its role in offering states guidance on this issue, in 2019 the UN launched its Strategy and Plan of Action on Hate Speech, which commits the UN to monitor and analyze hate speech, support victims, convene relevant

actors, advocate, educate, and support member states in developing policies for countering hate speech (United Nations, 2019).

Since then, the Special Rapporteur on Free Expression, David Kaye, as well as notable academics such as Nadine Strossen (2021) and Evelyn Mary Aswad (2018) have called on platforms to align their content moderation practices with the UN framework, which they argue requires that restrictions on speech be minimal to comply with Article 19. Special Rapporteur Kaye says that platforms should combat hateful attitudes with education, counter-speech, and other tools such as deamplification, demonetization, reporting, and training (Kaye, 2019).

Council of Europe & European Union

In the early 2000s, the Council of Europe's European Commission Against Racism and Intolerance (ECRI) issued its Convention on Cybercrime, which included a separate, Additional Protocol on Internet Hate Speech. This Protocol called for an update to countries' offline laws to include prohibitions for online content that "advocates, promotes or incites hatred, discrimination or violence, against any individual or group of individuals, based on race, color, descent or national or ethnic origin, as well as religion" (Council of Europe, 2003). In 2016, the European Commission adopted a proposal to amend its Audiovisual Media Services Directive to enhance the effectiveness of the legal regulation of hate speech on social media by prohibiting the transfer of material that incites violence or hatred directed against a group of people of member of a group defined by reference to sex, race, color, religion, descent, national or ethnic origin. Notably, there are no protections for incitement to hatred based on sexual orientation or gender identity.

The European Council has chosen to relieve ISPs of most responsibility for hateful or incendiary material published on their sites. The Protocol limits the liability of third-party intermediaries, making only the individuals posting the illegal content liable. Notably, the Protocol leaves room for countries to adopt an expansive definition of intent, meaning that third parties may be

held accountable if they receive notification of racist or xenophobic expression on their platform and fail to remove it.

The European Union also has laws prohibiting hate speech in online contexts. The Framework Decision on Combating Certain Forms and Expressions of Racism and Xenophobia by Means of Criminal Law requires member states to sanction "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, color, descent, religion or belief, or national or ethnic origin" (2008).

The EU's E-Commerce Directive (2000) provides the legal framework for ISPs and other third-party intermediaries' responsibilities regarding hate speech on their online platforms. Under this Directive, ISPs do not have a duty to monitor conduct and are governed by the laws of the member state in which they are established.

Individual State Laws

Many countries have established laws prohibiting online hate speech, including hate speech posted to social media platforms. However, with the exception of Germany, most countries do not hold social media platforms liable for the illegal hate speech posted by users. Germany's Network Enforcement Act (Netz DG) requires social media platforms to remove illegal hate speech within 24 hours of receiving a user notification. If it is unclear whether the Content violates the country's Criminal Code regarding public incitement to hatred, the social media organization then has seven days to decide (Network Enforcement Act, 2017). Social media organizations that receive more than 100 complaints per year are also required to produce bi-annual reports that include the number of incoming complaints about unlawful content, the number of complaints for which an external body was consulted, and complaints in the reporting period that resulted in the deletion or blocking of the content in question (Cannan, 2022). Social media companies must also designate an in-country agent to manage compliance with the law. Failure to adhere to these requirements can result in fines of up to € 50 million.

To address a rise in right-wing violence, in 2020 Germany passed an updated version of NetzDG that requires additional information from the platforms regarding the accounts actioned under the law. While this new version did strengthen the appeals process for users who feel their content was wrongly removed, it also requires social media organizations to provide Germany's Federal Criminal Police Office with IP addresses, login information, and passwords for accounts that post illegal hate speech (Cannan, 2022). Facebook (Meta), YouTube, Twitter, and TikTok have filed suit claiming that the revised version of Netz DG, which requires platforms to share user's personal data with federal police officers, represents a violation of privacy law (Goujard, 2022).

While only Germany requires platforms to comply with laws regarding hate speech, many other countries recognize that existing laws prohibiting hate speech do apply to social media and will punish individual users for violations. Canada, for example, recently proposed an amendment to its Criminal Code against hate speech and its Human Rights Act that would apply to individuals who publish on the internet – including on social media, on personal websites, and in blog posts, as well as the operators of websites. If a person was found guilty of hate speech that personally identified a victim, they could be fined C$20,000. Notably, though, social media companies were excluded from potential liability in this proposal (Ljunggren, 2021).

In many instances, these laws have been manipulated to silence political dissent. Scholars have noted that these laws are "likely to be enforced in ways that further entrench dominant political and societal groups and that further disempower marginalized individuals and groups" (Strossen & Lukianoff, 2021: 1).

In the United States, neither users nor platforms are prohibited from using or allowing hate speech on social media. However, because they are private entities, platforms can choose to create community standards that prohibit hate speech, which users are required to adhere to. As private entities, social media companies can curtail more speech than the government permits under the First Amendment. Dissatisfied with this approach, some states have

begun to develop their own laws regarding content moderation. Florida and Texas have established regulations that would require platforms to carry certain speech. Those laws are being challenged and the cases are slated to be taken up by the Supreme Court in 2024 (NetChoice v. Paxton (2022); NetChoice v. Attorney General, State of Florida (2022)). Focusing instead on transparency, New York and California have enacted laws requiring large social media platforms to report what their hate speech policies and removal practices are to the states. Those laws are also being challenged in court.

Joint Efforts at Regulation

In addition to laws in individual countries and international guidance, joint regulatory efforts between social media organizations and states have emerged. Most notable among these is the Code of Conduct, which is an agreement between the European Commission and Facebook, Microsoft, YouTube, and Twitter to remove hate speech that violates community standards. The Code of Conduct, signed in 2016, recognizes the critical role these companies play in protecting free speech while also addressing the problem of illegal hate speech as defined by the Framework Decision on Racism and Xenophobia. Since its creation, several other IT companies have joined the Code, including Instagram, Google+, Snapchat, Dailymotion, Jeuxvideo.com, and TikTok.

To date, there have been six distinct review periods, each lasting approximately 6-8 weeks, measuring the response time and rates for hate speech takedowns. In 2016, Facebook, YouTube, and Twitter removed 28.2% of reported hate speech. In 2021, the removal rate was 62%. The average of notifications reviewed within 24 hours was 81%, however, that represented an almost 10% decrease compared to 2020 (European Commission, 2021).

The major criticism of the Code offered by scholars, such as Natalie Alkiviadou (2019), is that it relies entirely on social media users to come across, identify, and report hate speech instead of placing the burden of detection on the social media companies themselves.

The Santa Clara Principles on Transparency and Accountability in Content Moderation are another voluntary agreement that 12 major companies, including Apple, Facebook (Meta), Google, Reddit, Twitter, and GitHub, have all voluntarily agreed to adhere to. The Santa Clara Principles, now in their second iteration, were co-created by a coalition of academics, advocates, and organizations. There are five foundational principles: Human rights and due process, understandable rules and policies, cultural competence, state involvement in content moderation, and integrity and explainability (Santa Clara Principles, 2022).

The first principle of human rights and due process says that social media companies should ensure that human rights and due process considerations are integrated at all stages of their content moderation processes. It also asks them only to use automated processes to identify or remove content when there is high confidence in the quality and accuracy of that automated detection. In addition, it says that companies should provide users with clear and accessible methods of obtaining support regarding content and account actions.

The second principle asks companies to publish clear and precise rules regarding when an action will be taken regarding a user's content or account. The third principle recognizes the importance of having content moderators that understand the language, culture, and social and political context of the posts they are moderating. The fourth principle asks companies to recognize that state involvement in content moderation processes can threaten users' rights. Finally, the fifth principle calls on companies to ensure that their content moderation systems are accurate and effective and do not discriminate.

In addition to the core principles, this agreement also includes three operational principles. The first, "numbers," asks companies to transparently report items such as the amount of content and accounts actioned, the number of appeals, the percentage of appeals that result in removal, and more. The second operational principle, "notice," requires companies to provide

notice to users whose content is removed or their account suspended. Finally, the third operational principle, "appeal," covers social media and other web service companies' obligation to make the explanation, review, and appeal process available to users.

Lastly, the Santa Clara Principles include two specific recommendations for governments and other state actors. The first, "removing barriers to company transparency," asks governments to make clear that companies are "not prohibited from publishing information detailing requests or demands for content or account removal or enforcement which come from state actors" (Santa Clara Principles, 2022). The second recommendation for state actors is to promote government transparency by reporting their own involvement in content moderation decisions, demands for data, or requests for accounts to be actioned.

According to its creators, the Santa Clara Principles are not intended to be a template for regulation. Instead, they are designed to support companies' efforts to respect human rights and enhance their accountability.

Compliance requires incentives & consequences

The snapshot provided here of the regulatory framework that currently applies to hate speech on social media is by no means exhaustive. Instead, my goal has been to provide examples of each of the various approaches different stakeholders are adopting to combat hate speech on social media. As this overview suggests, many efforts are underway – by international bodies, advocacy groups, and social media organizations – to make headway in how hate speech is identified in virtual spaces and how platforms respond.

However, despite the recommendations, policies, and even laws in place, the problem persists. Why? I believe most of these approaches fail to hold platforms accountable and lack any real consequences or incentives for compliance. The majority of solutions offered fail to acknowledge the reality that social media companies are for-profit entities whose primary goal is to generate profits. Many of the world's largest social media platforms,

such as Facebook, Instagram, and YouTube, are owned by companies like Meta or Alphabet, Inc., which are publicly traded on the US stock market. This means that the Board of Directors for each of these companies has a fiduciary responsibility to shareholders to maximize profits.

Corporations today operate by a model of corporate governance called, "shareholder primacy." According to this theory, the primary purpose of a corporation is to generate returns for shareholders (Paladino & Karlsson, 2019). Therefore, the singular goal driving each decision-making process is maximizing shareholder value. While organizations may consider the needs of other stakeholders, which for social media companies would include users, advertisers, employees, governments, and society at large, the duty to shareholders looms large over considerations made to serve these other stakeholder groups. Critics of this approach assert that corporate rights should include societal responsibilities. However, the reality is that in the current US corporate environment, corporations are not required to serve the public interest (Paladino & Karlsson, 2019).

Nowhere is this fact more evident than in the information released by Facebook whistleblower Frances Haugen. In October 2021, Haugen, who was part of Facebook's Civic Integrity Department, testified to the US Congress about how the company's engagement-based formula helps sensational content, such as posts that feature rage, hate or misinformation, gain traction. The cache of internal Facebook documents Haugen released confirmed what scholars have long suspected, that Facebook's algorithms are designed to feed people extreme viewpoints to keep them on the platform longer (Allyn, 2021). This engagement and attention is then sold to advertisers, which generates substantial profits for the organization. In 2021, Facebook's parent company, Meta, generated $117 Billion in revenue. According to the World Bank (2022), that amount is higher than the Gross National Income of at least 136 countries.

Scholars, activists, and elected officials have created thoughtful regulations and recommendations to thread the needle between protecting free

expression and human dignity. However, without consequences to ensure compliance, I believe social media companies will continue to make decisions based on what is best for their bottom line.

Moreover, the lack of transparency in the content moderation process often means that users, activists, and others have little insight into what is happening. As the documents released by Haugen indicate, even Facebook's "transparency report" fails to capture the true scope of the problem. Like many social media companies, Facebook reports on the pieces of content actioned, not on the total amount of content on the site that includes hate speech.

So, what do we do? To begin, any question about platforms' content moderation practices and the role of government in regulating or providing oversight of those practices must consider the right to free expression and human dignity. Governments can establish incentives and consequences to ensure that platforms comply with the best practices included in the various approaches laid out in this chapter, including individual state laws, the UN recommendations for content moderation of hate speech, and the Santa Clara Principles.

Utilizing Incentives

One solution scholars have proposed would require social media platforms to earn their shield from legal liability. As mentioned previously, in the United States, Section 230 prohibits third-party intermediaries, such as social media platforms, from being held liable for illegal content posted to their sites by users. Legal scholars Danielle Citron and Benjamin Wittes (2017) have suggested that Section 230 be revised to provide platforms with legal liability only if they show that their response to unlawful uses of their services has been reasonable.

This approach could be expanded to use in other countries, particularly those where hate speech is illegal. Rather than simply awarding immunity to platforms, governments could develop legislation that required social

media platforms to engage in certain activities to earn immunity. States could mandate that platforms ensure that their hate speech removal policies are clear and available in multiple languages. They could also require social media organizations to act quickly to remove illegal hate speech from their platforms and provide transparency reports about the content that's being evaluated and, in some cases, removed.

Unlike the German approach under Netz DG, incentivizing platforms to earn their legal immunity would combat the problem of online hate speech while at the same time avoiding some of the concerns raised regarding the removal of lawful expression. One of the primary critiques of Netz DG is that social media organizations will often remove more content than is necessary to meet the requirements of the law and avoid fines. Here, platforms could maintain their immunity provided they engaged in certain regulatory activities outlined by the government (Cannan, 2022).

Another potential incentive governments could offer social media companies might be in the form of tax breaks. Voluntary compliance with established regulations could be rewarded through a decrease in the amount of corporate taxes social media companies are required to pay. This approach could use the US environmental regulations as a model. In the United States, tax breaks are provided to individuals and organizations for environmentally responsible actions such as purchasing an electric vehicle or constructing an energy-efficient building. Similarly, governments could offer tax breaks to social media companies for producing transparency reports, hiring content moderators that speak the languages of the country, and other "best practices" outlined in the Santa Clara Principles.

Establishing Consequences

While incentives are one way to ensure compliance, consequences for failing to adhere to certain content moderation practices provide another alternative for combatting hate speech that appears on social media. As discussed earlier, the United Nations has called on social media companies to

consider human rights in each stage of their content moderation practices. UN Special Reporter David Kaye has suggested that social media platforms align their approach to content moderation to adhere to UN Free speech law. This would mean speech restrictions prohibiting the incitement of hatred must conform to the requirements laid out by ICCPR 19(3), which states that governments demonstrate the legitimacy, legality, and necessity of any laws restricting free expression.

Kaye's proposed approach, which is supported by Strossen (2021), represents a viable pathway for navigating content moderation of hate speech. However, without consequences in place, social media companies have no reason to comply. One consequence that governments could establish for failure to comply would be fines. This is similar to the penalties used in other industries. For example, a chemical company that pollutes a local river may be fined for failing to follow established environmental regulations. So too could a social media platform be fined for failing to follow the guidelines laid out by the United Nations regarding how to limit the spread of hate speech and protecting users' free expression.

The Santa Clara Principles, while not draft legislation, also represent best practices for content moderation. Created by activists, nonprofit organizations, and scholars, the Santa Clara Principles call on platforms to respect basic human rights, establish clear, understandable rules, dedicate resources to ensure that ai, algorithms, and human content moderators have necessary cultural competencies to evaluate speech in the appropriate context, protect users privacy from government interference, and be accountable for their decisions. One way to ensure compliance with these ideas would be for governments to clearly lay out what is expected of social media organizations regarding transparency, reporting, and cultural competency and be prepared to levy fines against those platforms that fail to comply. For example, a country could require that social media companies publish their policies in all the languages spoken in the country and hire a significant number of human content moderators who speak those languages. Fines could then be levied for failure to comply.

An approach like this would require countries to dedicate significant resources to oversight. Like the energy industry, government officials would likely work closely with social media companies to ensure compliance. This too borrows one of the most useful ideas from Netz DG, which mandates that social media companies designate an in-country agent to manage compliance with the law.

Conclusion

Whether through incentives like earning a shield from legal liability, tax breaks, or consequences such as fines, efforts to regulate social media companies' content management of hate speech must have teeth. It is unreasonable to expect for-profit companies operating under a framework of shareholder primacy to act against their own financial best interests. Therefore, we need to find financial incentives or consequences to motivate compliance. This is the only way to ensure that these organizations remove content that otherwise might be quite lucrative for them. If we want social media companies to take meaningful actions such as hiring more content moderators, publishing community standards in multiple languages, and ensuring the cultural competence of automatic detection algorithms, then we must motivate them using financial incentives or consequences. Without these in place, social media companies will remain free to pick and choose which, if any, of the best practices for minimizing hate speech they will adhere to.

References

Alkiviadou, N. (2019). Hate speech on social media networks: Towards a regulatory framework? *Information and Communications Technology Law, 28*(1), *19-35.*

Allyn, B. (2021, October 5). Here are 4 key points from the facebook whistleblower's testimony on Capitol Hill. *NPR*. https://www.npr.org/2021/10/05/1043377310/facebook-whistleblower-frances-haugen-congress

Aswad, E. M. (2018). The future of freedom of expression online. *Duke Law & Technology Review, 17,* 26-70.

Canaan, I. (2022). Netzdg and the german precedent for authoritarian learning. *Columbia Journal of European Law*, *28*(1), 101-133.

Citron D. K. & Wittes, B. (2017). The internet will not break: Denying bad samaritans § 230 immunity. *Fordham Law Review*, *86*(2), 401-423. https://ir.lawnet.fordham.edu/flr/vol86/iss2/3

Communications Decency Act 1996, 47 US § 230(c)(2).

Community Guidelines. (2023). *TikTok*. Retrieved Oct. 10, 2023, from https://www.tiktok.com/community-guidelines?lang=en#38

Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law.

Council of Europe. (2003). *Additional Protocol to the Convention on cybercrime, concerning the criminalization of acts of a racist and xenophobic nature committed through computer systems*, opened for signature January 28, 2003, ETS 189 (entered into force March 1, 2006).

Directive 2010/13/EU of the European Parliament and of the Council of 10 March 10, 2010 on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive) OJ L 95, 15.4.2010, p. 1–24.

Directive 2000/31/EC of the European Parliament and of the Council of 17 July 2000 on Certain Legal Aspects of Information Society Services, in Particular Electronic Commerce, in the Internal Market ('Directive on electronic commerce).

European Union (2021, October 7). *EU Code of Conduct against illegal hate speech online: results remain positive but progress slows down* [Press release]. https://ec.europa.eu/commission/presscorner/detail/en/ip_21_5082

Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken[Netzwerkdurchsetzungsgesetz] [hereinafter NetzDG], Jun. 30, 2017, BUNDESGESETZBLATT, Teil I [BGBLI] at 3352, Nr. 61 (Ger.), available at https://www.gesetze-im-internet.de/englisch_stgb/englisch_stgb.html

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Goujard, C. (2022, Feb. 2). Big Tech Takes on Germany. *Politico.* https://www.politico.eu/article/big-tech-takes-on-germany-over-demands-to-forward-illegal-content-to-federal-police/

Hate Speech Policy. (2023). *YouTube*. Retreived Oct. 10, 2023, from https://support.google.com/youtube/answer/2801939?hl=en

Hate Speech Transparency Report. (2023). *Facebook*. Retrieved October 10, 2023, from https://transparency.fb.com/data/community-standards-enforcement/hate-speech/facebook/

International Convention on the Elimination of All Forms of Racial Discrimination, opened for signatures Dec. 21, 1965 S. Exec. Doc. C, 95-2 (1978); S. Treaty Doc. 95-18; 660 U.N.T.S. 195, 212 (1967).

International Covenant on Civil and Political Rights, opened for signatures Dec. 16, 1966 999 UNTS 171; S. Exec. Doc. E, 95-2 (1978); S. Treaty Doc. 95-20; 6 I.L.M. 368 (1967).

Kaye, D. (2019). *Promotion and protection of the right to freedom of opinion and expression*. i 28, UN Doc. A/74/486.

Land, M. K. & Hamilton, R.J. (2020). Beyond takedown: Expanding the toolkit for responding to online hate. In P. Dojcinovic (Ed.), *Propaganda, war crimes trials and international law: From cognition to criminality* (143-157). Routledge.

Leets, L. (2002). Experiencing hate speech: Perceptions and responses to anti-Semitism and anti-gay speech. *Journal of Social Issues, 58*(2), 350-358. https://doi.org/10.1111/1540-4560.00264

Ljunggren, D. (2021, June 23). Canada unveils plans to make online hate speech a crime. *Reuters.* https://www.reuters.com/world/americas/canada-unveils-plans-make-online-hate-speech-crime-2021-06-23/

NetChoice, LLC v. Attorney General, State of Florida, 34 F.4th 1196 (11th Cir. 2022).

NetChoice, LLC v. Paxton, 49 F.4th 439 (5th Cir. 2022). Organization for Security and Cooperation in Europe [OSCE], *Joint Declaration on Freedom of Expression and "Fake News," Disinformation and Propaganda*. (March 3, 2017). https://www.osce.org/fom/302796.

Palladino, L. & Karlsson, K. (2019, February 11). *Towards accountable capitalism: Remaking corporate law through stakeholder governance.* Harvard Law School Forum on Corporate Governance. https://corpgov.law.harvard.edu/2019/02/11/towards-accountable-capitalism-remaking-corporate-law-through-stakeholder-governance/

Santa Clara Principles 2.0. (2022). Santa Clara Principles. Retrieved October 12, 2022, from https://santaclaraprinciples.org/.

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

Stecklow, S. (2018, August 15). Why Facebook is losing the war on hate speech in Myanmar. *Reuters.* https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/

Strossen, N. (2021). United Nations free speech standards as the global benchmark for online platforms' hate speech policies. *Michigan State International Law Review, 29*(2), 307-361.

Strossen, N. & Lukianoff, G. (2021, September 30). *Hate speech laws backfire: Part 3 of answers to bad arguments against free speech*. The Fire. https://www.thefire.org/hate-speech-laws-backfire-part-3-of-answers-to-bad-arguments-against-free-speech-from-nadine-strossen/

United Nations. (2019). *Strategy and plan of action on hate speech*. https://www.un.org/en/hate-speech/un-strategy-and-plan-of-action-on-hate-speech

Vaidhyanathan, S. (2018). *Antisocial media: How Facebook disconnects us and undermines democracy.* Oxford University Press.

World Bank. (2021). *World Development Indicators: Size of the economy.* http://wdi.worldbank.org/table/WV.1

Wypych, M. & Bilewicz, M. (2022). Psychological toll of hate speech: The role of acculturation stress in the effects of exposure to ethnic slurs on mental health among Ukrainian immigrants in Poland. *Cultural Diversity and Ethnic Minority Psychology,* advance online publication. https://doi.org/10.1037/cdp0000522

HATE SPEECH ON SOCIAL MEDIA IN THE ELECTORAL YEAR CONTEXT IN BRAZIL

Rubens Beçak
/ University of Sao Paulo, Brazil

Kaleo Dornaika Guaraty
/ University of Sao Paulo, Brazil

Tiago Augustini de Lima
/ University of Sao Paulo, Brazil

Introduction

> What I look for in speech is the response of the other. What constitutes me as a subject is my question. To make myself recognized by the other, I only say what was with a view to what will be. To find him I call him by a name he must assume or refuse to answer me (Lacan, 1998: 301)

The analysis of hate speech in societies is not a new phenomenon, neither is it exclusive to Brazil. Both Marilena Chauí (2000) and Lilia Moritz Schwarcz (2019) attest that a peaceful and cordial Brazil (taking Sergio Buarque de Holanda's concept out of the context of passionality) never existed. The Brazilian historical logic of colonialism, slavery and racism, bossiness, patrimonialism, corruption, violence, social inequality and intolerance are not the fruits of contemporaneity. Hate speech had been used in our history as a matter of political propaganda and, currently, the biggest problem related to hate speech is its speed of dissemination due to the expressive use of social networks for communication and expression of thoughts.

The speed of dissemination of content on the internet in Brazil is immeasurable, and, according to a survey carried out by Statista, of internet use in the country (with more than 167 million users), 93% is used for instant messaging, this occurs also because the social media whatsapp has Brazil as the second largest country in number of users in the world, around 146 million users. The exclusive use of social media in the country accounts for 81% of all internet use.

Through this, it is easy to see that hate speech can reach a large part of the Brazilian population. Therefore, victims or vulnerable groups such as, for example, stigmatized populations such as the black population, women, the LGBTQI+ population, quilombolas, riverside dwellers, the elderly, asians, etc., do not have the same space to defend themselves, thus, hate speech has a greater impact when used in social medias, due to the fact that there is an increase in spread, dissemination and possibility of contact with a greater number of people than in the social environment.

In this way, the objective of this paper is to demonstrate the best way to conceptualize hate speech or, at least, as a starting point for its hermeneutics to, later, demonstrate how hate speech occurs in electoral years, especially in Brazil, through the analysis of data already collected by SaferNet, ComunicaQueMuda and the Violence Dossier. Finally, we analyze whether there is a correlation between the increase in hate speech on the social media in an election year according to party political objectives as a form of political strategy.

Thus, to achieve the proposed objectives for the work, we use the dialectical method, so that we have a juxtapositive analysis in addition to the different positions and concepts about hate speech. The analytical method of the data presented was used. In this way, it is concluded that we did not aim to exhaust the concept of hate speech, it was only used to conceptualize and analyze the data presented that were compared with the increase in cases of hate speech on the internet as a result of the global pandemic of Covid-19.

The concept of the hate speech

Caitlin Ring Carlson (2021) asserts that there is no society, culture or any media that is exempt, protected or even influenced by hate speech. However, the analysis of the definition of hate speech for the proposed theme needs more specificity.

The concept or rapid compression for hate speech has always been present in society, as Carlson guides, however, there are notable differences related to the agent that propagates it, the means used, which the offense is targeted and for what purpose it is perpetrated. Taking as an example the case of nazism, especially from 1933 to 1945, when Hitler was Chancellor and, later, Führer of Germany, the persecution of gypsies, of black people, homosexuals; and especially Jews, etc., the hate speech was realized and legitimized as state policy.

Starting from this point and already understanding our society as a digital society, the biggest dangers related to hate speech are: (i) its speed of spread on the internet and (ii) the danger that represents the breadth of this hatred when hate clusters take to itself the propagation of the offense.

As the topic does not have an exact concept, as it depends on the historical construction of each nation, the definition we will use for hate speech will be with the objective of demonstrating the brazilian experience, especially in an election year.

In this way, Andrew Altman, conceptualizes hate speech through a perspective of moral and physical superiority in relation to the other persons. Based on this assumption, subordination is a key factor for sexist, homophobic, racist and anti-religious content to be perpetuated, because

> Treating persons as moral subordinates means treating them in a way that takes their interests to be intrinsically less important, and their lives inherently less valuable, than the interests and lives of those who belong to some reference group. There are many ways of treating people as moral subordinates that are natural as opposed to conventional: the

> status of these acts as acts of subordination depends solely on universal principles of morality and not on the conventions of a given society. Slavery and genocide, for example, treat people as having inferior moral standing simply in virtue of the affront of such practices to universal moral principles.

Altman (1993) also emphasizes that this type of language that treats the other person as a moral subordinate is called speech-act and, for that, they need regulation and prohibition, because

> In general, what are needed are rules that prohibit speech that (a) employs slurs and epithets, conventionally used to subordinate persons on account of their race, gender, religion, ethnicity, or sexual preference, (b) is addressed to particular persons, and (c) is expressed with the intention of degrading such persons on account of their race, gender, religion, ethnicity, or sexual preference.

Therefore, moral subordination in relation to the other subject can only be verified when analyzing the concrete case, conditioning that both Charles Lawrence III (1990) and Mari Matsuda (1990), do not defend.

Lawrence III (1990) and Matsuda (1989) start from the interpretation that offenses that have a racist content, gender, religion, ethnicity, sexual preferences already constitute moral subordination, because they are conduct and not discourse.

Charles Lawrence III (1990) understands hate speech as offenses, slander and defamation in relation to race, gender, religion, nationality, ethnicity and sexuality and that, therefore, starting from the fact that freedom of expression is not a right in the abstract sense, as it victims may suffer to gain access to the labor market. Thus, equality among all citizens will only be achieved in its materiality if hate speech is not carried out and it is for this reason that the author emphasizes freedom of expression as the end of the master and slave dichotomy, in the Hegelian sense.[1]

1. "*Most importantly, we must continue this discussion. It must be a discussion in which the victims of*

Mari Matsuda (1989) states that hate speech is directed to race, religion, gender, sexual preference and, therefore, subordination needs to be analyzed in each of these hypotheses due to the different forms and biases that constitute each of these offenses. It is for this reason that Matsuda, when analyzing US jurisprudence, demonstrates that freedom of expression is not considered an absolute right for the courts.

Hate speech, therefore, can generate harm to the offended. Psychological and psychic problems, increased cardiorespiratory frequency, hypertension and the development of diseases such as depression. Heinz Häfner[2] (1968) argued that people who suffered persecution and discrimination based on ethnicity developed, after a certain period of time, chronic anxiety and depression, a range of neuroses, personality disorders, compulsive obsessions, etc.

However, as mentioned elsewhere, hate speech, which has a speech-act basis, is understood as action. Nevertheless, what is more common and what the legal currents defend today is hate speech understood as speech and not as action.

The theory of hate speech understood as speech gained notoriety with the work The harm of hate speech, by Jeremy Waldron (2012), because when analyzing the limits of freedom of expression, "publications which express deep disrespect, hatred and vilification for the member of minority groups" there is a barrier that needs to be looked at more carefully so that no citizen has their dignity offended.

In this way, Waldron (2012), when criticizing the limits of freedom of expression conceptualized by US jurisprudence, understood as incitement

racist speech are heard. We must be as attentive to the achievement of the constitutional ideal of equality as we are to the ideal of untrammeled expression. There can be no true free speech where there are still masters and slaves." LAWRENCE III, Charles R. Frontiers of Legal thought the New First Amendment: If he hollers let him go: regulating racist speech on campus. *In Duke Law Journal*, jun., 1990.

2. HÄFNER, Heinz. Psychological disturbances following prolonged persecution. *In Social Psychiatry*, *3*(3), 1968: 81. Available on: https://link.springer.com/epdf/10.1007/BF00577832?sharing_token=sXhxic3dKNASfAZkmvRR3ve4RwlQNchNByi7wbcMAY6ZvMpKQkjIG7l5VzghcEc7EZas8yycjQ4CkYddCyrGqV4BwqjMny_oL3mBOboXXCWNA9bhrXDli8XLcWWedURAMriW8zO1f3BhC40sxSfuhQ%3D%3D.

to probable violence, but present danger, clarifies that hate speech causes damage, even if its consequences are not reached. The author states that hate speech with biases of nationality and ethnicity that correlate the Arab population to terrorist groups does not cause any direct action to this group, however, the speech itself represents harm, as it allows the questioning of its dignity, pushing it away. of the community.

Thus, Jeremy Waldron (2012: 35-37) develops at the beginning of his critical work the very concept of hate speech. The word hate cannot be understood as just a feeling, as it would imply establishing the difference between feeling disgust for a certain thing or something and feeling hate. Thus, law enforcers and legislators must understand that it is not about the feeling of hate, but with the result that hate generates to target groups, minorities are always vulnerable.

The word discourse or speech also needs special attention for regulators, legislators and law enforcers. This is because the word emitted orally has the ability to hurt and this is obvious, however, vulnerable groups are attacked, for the most part, by written speeches and allocated in public environments. Thus, the medium in which hate speech is distributed, whether with posters on the streets of cities, articles in newspapers, publications, pictures, or even on the internet, is what refers to speech, because the assimilation of society in contact with hate speech is what generates the moral subordination of these minorities and their exclusion from social life[3].

Anna Laura M. Fadel (2018: 56), also based on Waldron's concepts, she understands that

> Hate speech [...] makes it difficult for individuals belonging to minority groups to live their lives in a dignified way. In a well-ordered society, legislation is not eliminated from its basic structure, as the coercive power

3. "The restriction on hate speech that I am interested in are not restricttioons on thinling; tjey are restrctions on more tangible forms of message. The issue is publication and the harm done to individuals and groups through the disfiguring our social environment by visible, public and semipermant announcements to the effect that in the opinion of one group in the comunity, perhaps tha majority members of another group are not worthy of equal citizenship". W.J. (2012). *The Harm in Hate Speech*. Boston: Harvard Press, p. 39.

> of the State is necessary to maintain stability and social cooperation. It would be a kind of guarantee that individuals expect to be treated equally, both by the State and by other citizens. [...] Furthermore, a well-ordered society must cultivate and cherish a sense of security and a guarantee of equal consideration.

Therefore, Waldron (2012: 108-110) remember that the objectives that legislations that combat hate speech must adhere to are the protection of individuals regarding their dignity, this because, the protection of the offense is related to the protection of the impacts that the effects of the offense generate on feelings (disgust, displeasure, embarrassment) of individuals, as for the protection of dignity and reputation in public spaces, it is the inclusion of individuals in the social environment in a decent way, without diminishing the elementary status of these human beings.

The hate speech on the internet

In Brazil, according to data collected by ANATEL (National Telecommunications Agency), 98.2% of Brazilians have a mobile telephone network and, of this amount, 3G technology reaches 99.3% of Brazilians, in a total of 5,301 cities; 4G technology reaches 94.4% of the population in 4,122 cities. The states with the highest percentage coverage of 4G technology are, respectively, the Federal District, with 99.64% of residents using this type of technology; São Paulo came second with 98.52% and Rio de Janeiro, with 98.27% of residents. Even the states that have the lowest coverage of 4G technology, there are more than 70% of residents with coverage of this type of technology, for example, the State of Pará, with 73.69%; Piauí, with 71.84% and Maranhão, with 71.47% of the dwellers.

As for the use of social media, Brazil is the second largest country with WhatsApp users, which, in the year 2021, has about 146.8 million users [4], behind only India, which has about 400 million users.

4. Data extratected from Statista.

Moreover, of users of the Facebook platform, Brazil has about 148.57 million users in 2021. As for Instagram, in July 2021, the country had about 110 million users, only behind India, with 180 million and the USA, with 170 millions of users. The latest social media, the Chinese TikTok, has 4.72 million users in Brazil. Twitter has, in 2021, in Brazil, about 17.46 million users.[5]

In this way, the digital environment does not have a clear distinction between what is private and what is public and, according to Perrone and Pftscher (2016: 148), this characteristic makes the internet a paradoxical medium in which the "private is validated by public exposure." The characteristic of the internet being the environment that takes any type of information to a large number of users, also brings the problem of the radicalization of violence, this is how the authors describe that:

> The immateriality of the internet has generated the false impression that it does not produce damage and there is a questioning of the material effectiveness of the processes it triggers. It is possible to observe a denial of the symbolic processes present in the flows of images and words that circulate on the network because, faced mainly with their destructive materiality, it is argued that they are just innocuous narratives, a private opinion, a word without harm.

The internet these days, together with the use a lot of social media, is the main public environment for the spread of hate speech, based on the concepts of Jeremy Waldron. Second, Santos and Silva (2016), hate speech practiced on the internet has the harm of assimilation and speed of dissemination as we said above, thus, this language used to offend and incite violence to minority and vulnerable social strata, recognized as difficulty in recognizing the difference, however, seems to generate social gain for the issuers of these discourses in the digital environment. Network algorithms that are based on engagement favor the offending agent's gain and, with that, their visibility increases and their popularity grows.

5. All this data was extracted from the Statista platform in 2022, regarding users of WhatsApp, Instagram, TikTok, and "X," formerly known as Twitter.

The argument of the psychoanalytic study of language in relation to the one who utters hate speech responds, at least in the field of psychoanalysis, but that we can use it to the Law to, above all, understand the agent's intent, what the aggressor seeks, so, acoording to Perrone and Pftscher (2016: 150-151):

> The one who utters the hate speech, not only builds himself, but also becomes an instrument of the Other. How does he confirm that the Other exists and is his instrument? Confirmation comes from the reaction provoked in the victim. Aggression addresses the victim's traumatic core, around which he organizes his identity. Words hurt when they leave the victim mute, unable to reflect or act [...] The victim's pain validates what the aggressor seeks. The Other has to exist, to sustain the perverse phantom of serving the Other's enjoyment. The invention of the dangerous Other (blacks, Jews, gays, etc.) acts as a significant core of love, capable of conjuring up disparate elements and giving a clear and coherent meaning. Thus, the dangerous Other gives consistency.

And this characteristic of the subject who utters the hate speech was well explained by Butler (2021), as the agent only replicates the speeches of the community in which he is inserted and, further, "whoever utters the hate speech is responsible for the way in which it is is repeated, for reinforcing this type of discourse, for re-establishing contexts of hatred and injury."

This pattern of hate speech, therefore, is transferred to the digital environment – the offenses against race, gender, religion, nationality, ethnicity and sexuality observed by the first theorists of hate speech, are also repeated on the internet, therefore, minority and vulnerable groups that Jeremy Waldon cites in his work, are also the same groups that most receive this discursive language on the internet.

In Brazil, the SaferNet (2022)[6], which receives complaints from all over the country related to cyber crimes, has already accounted for more than 2.5

6. SaferNet Brazil is a private civil association, with nationwide presence, nonprofit, non-economic, and without any political, partisan, religious, or racial affiliation. They are the creators and maintain-

million complaints of hate crime on the internet since its creation in 2005. According to the platform, 59.7% of these victims of hate speech are people black, 67% are women. In 2021, SaferNet accounted for 44,131 complaints: LGBTphobia accounted for 12.11% of all complaints; misogyny with the mark of 18.52%; racism with 12.60% and what stands out is the expressive mark of neo-Nazism, with 32.80%.

Hate speech on the internet in election year

In Brazil, the use of social networks in the 2018 elections really had a high level of use by candidates in the election, as Jairo Nicolau (2020) demonstrates, because from the post-redemocratization elections to the 2016 municipal elections, any candidate needed (i) large sum of money for proper campaign funding; (ii) reasonable time for electoral propaganda on television and radio and (iii) strong political support in the states of the Federation.

In addition to this electoral campaign structure, the 2018 election was the first general election to apply the electoral reforms provided for in Law 13,165/2015[7], such as, for example, the reduction of the official period of electoral campaign and the reduction of the deadline for affiliation to a party to run for office. Also, it was the first time that the fund created exclusively for electoral campaigns, introduced by Law 13,488/2017, had been used.

Along these lines, and recalling the party structure used in the 2016 US elections, the use of social networks for the 2018 Brazilian election campaign gained very high proportions. The legal regulation of social networks in the electoral context of that year was still incipient. In this way, politicians and their campaign structures turned to acting on the networks, with immense movement and organization of partisan ideological groups linked to the, until then, candidate Jair Bolsonaro.

ers of the National Cybercrime Reporting Center, operated in partnership with the Public Prosecutor's Offices and the Secretariat of Human Rights of the Presidency of the Republic (SDH) to strengthen actions against cybercrimes targeting Human Rights.

7. BRASIL. *Lei nº 13.165, de 29 de setembro de 2015.* Disponível em:<http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2015/lei/l13165.htm>. Acesso em: 03 de set. de 2021.

According Guaraty (2020), in 2018:

> Brazilian access to the internet already corresponded to 64.7% of the population, with social networks being the most accessed sites, as well as the massive use of mobile phone conversation applications. The use of political marketing tools by social networks, although incipient, is marked by the decentralized way in which it reaches the population, even though it can be directed by target audience selection tools. The 2016 American elections already signaled the relevance and dangers of electoral use of social networks, with abusive sharing practices using robots having been identified to expand the delivery of content and the dissemination of fake news.

The amount of misinformation generated, especially in the 2018 election campaign, is immeasurable due to the amount of shares not only on Twitter, but also on WhatsApp and Facebook. The advantage of WhatsApp is the possibility of sharing a huge amount of messages to a large mass of users, as there is the possibility of creating numerous groups with up to 256 (two hundred and fifty-six) people each.

This mobilization on social networks in an electoral environment is also reflected in hate speech, rallying feelings of inclusion and exclusion in the name of partisan ideological manifestation. With this, Kaleo D. Guaraty (2020), states that there is an amplification of a supposed moral superiority in relation to the other in an electoral environment that favors prejudice and political hatred against the oponente:

> In addition to the programmatic content discussed in terms of disagreement or non-acceptance, polarization engenders forms of personal identity with stereotypes, prejudices, exaggerations and radicalism. The debate no longer revolves around ideas, but the feelings of repulsion towards the other, fueling fear and hatred.

Through their algorithms, social networks favor the formation of echo chambers, that is, a large number of users engaging in certain political or controversial content that causes a moral debate. According Guaraty (2020:

25), this is identified by the platform as extremely relevant and, thus, other content similar to this is shown to users individually and, therefore, these echo chambers strengthen the identification with that large number of users.

This movement can be described by mass psychology as identification, in this way, the set of individuals forming this mass generates in the unconscious of the whole the image that each one is a "super strong" individual "among a bunch of equal companions".

However, whenever there are mass behaviors, Sigmund Freud (2011) teaches that there are:

> Atrophy of the conscious individual personality, the orientation of thoughts and feelings in the same directions, the predominance of affectivity and of the unconscious psyche, the tendency to the immediate execution of the purposes that arise, all this corresponds to a state of regression to a primitive psychic activity. [...] Even today, the individuals of the mass lack the illusion of being loved equally and justly by the leader, but the leader does not need to love anyone else, he is allowed to be of a lordly nature, absolutely narcissistic, but self-assured and independent. [...] the restless and compulsive character of the formation of the mass, evinced in its phenomena of suggestion, may then justly be traced back to its origin from the primeval horde. The mass leader continues to be the feared primordial father, the mass still wants to be dominated with unrestricted force, it has an extreme craving for authority, a thirst for submission. The primeval father is the ideal of the mass, which dominates the ego in place of the ego ideal.

That said, it is understood that, in an election year, there is an increase in cases of hate speech complaints in Brazil. According to data from SaferNet, the 2022 elections will be the third election year in which there is an increase in reports of hate crimes compared to non-election years. In the first half of 2022 alone, SaferNet received 23,947 complaints, which represents an increase of 67.5% compared to the same period last year.

Figure 1: Reports of hate speech to SaferNet increase during election years

Hate Speech	2017	2018 (General Election)	Increase in 2018	2019	2020 (Municipal Election)
Apology for Crimes Against Life	10611	27713	161,17%	8182	11852
Homophobia	2592	4244	63,73%	2752	5293
Misogyny	961	16717	**1639,50%**	7112	12698
Neo-nazism	1172	4244	**262,10%**	1071	9004
Racism	6166	8336	35,10%	4310	10684
Xenophobia	1395	9703	595,50%	978	2066
Religious Intolerance	1459	1084	-25,70%	1413	1321
Total numbers of reports	24356	72041	195,78%	25818	52918

Source: SaferNet: Available on: https://new.safernet.org.br/content/crimes-de-odio-tem-crescimento-de-ate-650-no-primeiro-semestre-de-2022.

Figure 2: Reports of hate speech to SaferNet increase during election years

Hate Speech	2021	1º Sem 2021	1º Sem 2022 (General Election)	Increase in the 1º Sem 2022 (General Election)
Apology for Crimes Against Life	7390	2374	3573	50,50%
Homophobia	5347	3206	4733	47,60%
Misogyny	8174	5593	7096	26,80%
Neo-nazism	14476	578	1273	**120,2 %**
Racism	6888	1807	2237	23,7 %
Xenophobia	1097	358	2222	**520,60%**
Religious Intolerance	759	373	2813	**654,10%**
Total numbers of reports	44131	14289	23947	67,50%

Source: SaferNet: Available on: https://new.safernet.org.br/content/crimes-de-odio-tem-crescimento-de-ate-650-no-primeiro-semestre-de-2022.

In analysis of this data and according to the platform, until the end of June 2022, xenophobia (520%) and crimes related to religious intolerance (654%) were the hate crimes that grew the most in relation to the year of 2022. 2021. In 2020, reports of neo-Nazism increased by 740.7% compared to 2021, and religious intolerance and xenophobia increased by 148% and 111%, respectively, compared to 2019.

In addition, this dizzying increase in intolerance that hate speech generates on social networks, presented in 2022, was already seen since 2016.

Comunica Que Muda[8], together with the Nova/SB agency, through the use of a word tracking software, Torabit, mapped, from April to June 2016, 393,284 (three hundred and ninety-three thousand, two hundred and eighty-four) mentions in 10 researched topics: politics, misogyny, homophobia, disability, racism, appearance, age/generation, social class, religiosity and xenophobia, 84% of the total mentions were negative. The highest percentages of negative mentions were racism, with 97.6%, and politics, with 97.4%.

The survey also shows that political intolerance received 273,752 mentions, while misogyny, which ranks second as the bias that most received intolerant speeches, received 79,484 mentions, 69% and 20.21%, respectively.

Hate speech is not a new concept and it arises especially with social networks, it also does not have a possible legal concept and, therefore, even with its measurement in social media, it is difficult to regulate. One of the problems for this conclusion is that the standards to recognize and protect the conduct are not common among the countries that try to regulate it.

Hate speech in Brazil can represent offenses against constitutional precepts, especially in the electoral sphere, such as, for example, the manifestation of

8. Comunica Que Muda It is a Brazilian communication agency that deals with sensitive issues for the entire Brazilian population with the goal of raising awareness and providing guidance on topics such as internet intolerance, suicide, among others. The essay generated the data through analyses of posts on Facebook, Twitter, Instagram, blogs, or websites. Every time one of the ten keywords appeared on any of these platforms, Torabit would collect it. Comunica Que Muda. *Dossiê da Intolerância*. (2016). Recuperado de: https://abcpublica.org.br/wp-content/uploads/2016/08/dossie_intolerancia.pdf.

thought, with regard to prejudice against people's origin, race, sex, gender, color, age and any other forms of discrimination, discriminatory offense to fundamental rights and freedoms, to anything that offends the foundation of human dignity. There is also infraconstitutional regulation provided for in the Elections Law, especially arts. 57-A to 57-J, provisions that regulate electoral propaganda on the internet, the Electoral Code (Law 4,737/65) and the jurisprudence of the STF and TSE.

However, the lack of specific regulation makes accountability difficult when hate speech is carried out in a virtual environment. It is, therefore, up to digital platforms to remove content that violates the conditions of use.

Diogo Rais and Camila Tsuzuki (2021), recall that, according to the project "victims of violence", during the 2018 election campaign, it gathered reported cases of victims of aggression motivated by political intolerance in Brazil since August 15, 2018. 88 cases, including assaults, murders, vandalism and threats, against women, LGBTQs and political opponents in more than 18 states in the country.

Also in 2018, the Brazilian Association of Investigative Journalism (Abraji) recorded 156 cases of violence against journalists and communicators who were in a political, electoral and/or partisan context. Of this amount, 85 of them took place in the virtual environment (on the internet) and 71 cases took place physically[9].

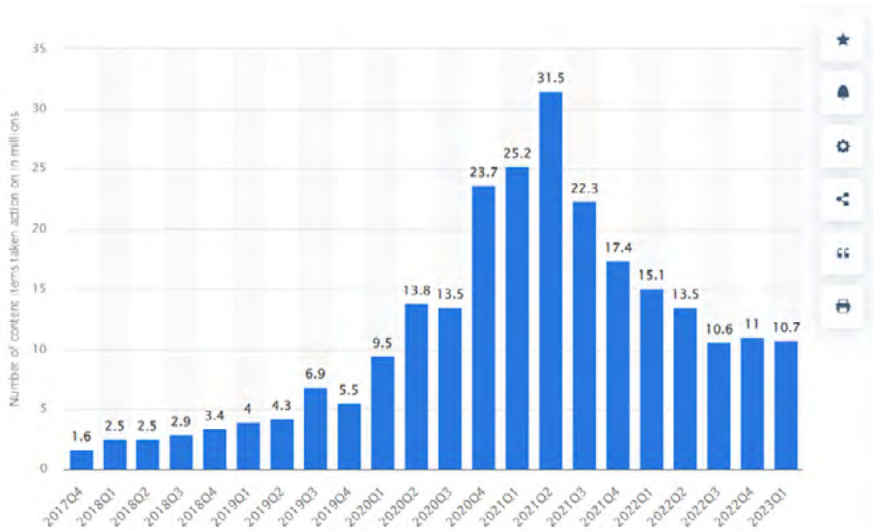
In the context of the increase in cases of hate speech during elections in Brazil, there was an increase in these cases when the world pandemic of covid-19 began in March 2020. This is what Dietch the Label demonstrates, which carried out searches on the internet in the period from 2019 to 2021, in the US and UK, obtained surprising results: (i) there was an average of

9. "The data collected by Abraji's monitoring in 2018 was mentioned in the annual report of the non-governmental organization Human Rights Watch, released on January 17, 2019, when discussing Brazil. On January 1, 2019, two journalists from different media outlets were targeted on social media after expressing dissatisfaction with the working conditions of the press during President Jair Bolsonaro's inauguration." Available on: https://www.abraji.org.br/noticias/abraji-registra-156-casos-de-agressoes-a-jornalistas-em-contexto-politico-eleitoral-em-2018.

new hate posts about race or ethnicity every 1.7 seconds; (ii) there has been an increase in hatred against the Asian population by 1,662% since the beginning of the pandemic; (iii) There was a 28% increase in hate speech based on racism; (iv) gender-biased hate speech rose 14%; homophobia with an increase of 85%[10].

This data can be seen by the significant increase in the removal of content classified as hate speech by the main social media. For example, worldwide, Facebook, as shown in Image 2, has removed 192.6 million offensive posts since 2020 to date.

Figure 3: Facebook hate speech removal *(in millions)*

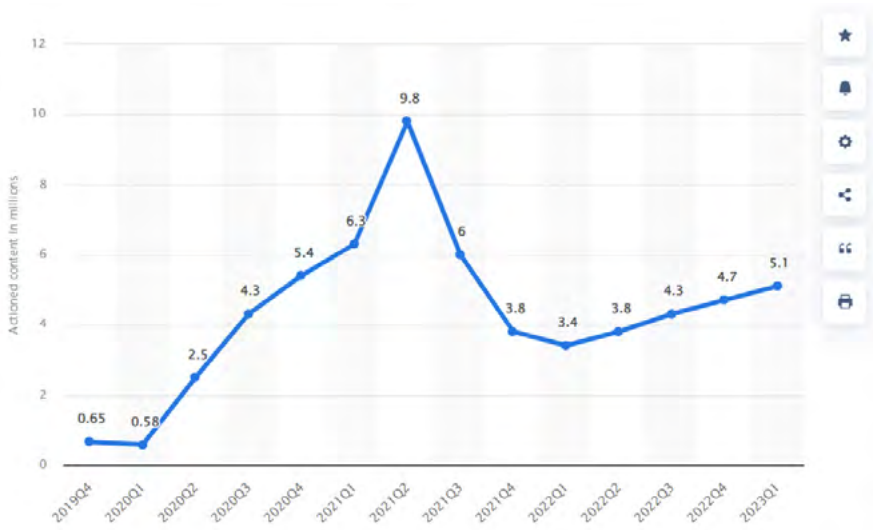


The increase in the removal of hate speech by Facebook in the period of the covid-19 pandemic.
Source: Retrieved from:https://www.statista.com/statistics/1013804/facebook-hate-speech-content-deletion-quarter/

This same curvature can be verified when analyzing the removals of hate speech on the social network Instagram in the world.

10. Available on: https://www.brandwatch.com/wp-content/uploads/2021/11/Uncovered_Online_Hate_Speech_DTLxBW.pdf .

Figure 4: Removal of hate speech content by Instagram (*in millions*)



Source: Available on: https://www.statista.com/statistics/1275933/global-actioned-hate-speech-content-instagram/.

From the analysis of these data, we can infer that hate speech on the internet in the context of the covid-19 pandemic also increased in Brazil, which was exacerbated with the municipal elections of 2020 and also the general elections of 2022, as already demonstrates SaferNet. However, even with the efforts of social media platforms themselves to remove offensive content from their usage practices, hate speech grows in Brazil in election years, with an increase of 650% in the first half of 2022 compared to the same period of 2021[11].

Therefore, the scope that hate speech can reach when perpetrated in an electoral environment and context is possible to obscure the free conviction of the voter who is having access to this large amount of messages on social networks. This is because, as Kaleo D. Guaraty (2020) advises, "it presents itself with an appearance of truth and high value [...] But the value is prag-

11. "Complaints increase in election years; in the first six months of 2022 there were 23,947 complaints, 67.5% more than the same period in 2021". Available on: https://new.safernet.org.br/content/crimes-de-odio-tem-crescimento-de-ate-650-no-primeiro-semestre-de-2022.

matically low, since there is an identifiable consensus that hate speech is never scientifically supported, and it does not present any development to the topic involved."

Conclusion

Hate speech is extremely harmful to societies, as it transfers to public debate and makes common opinion characteristics that greatly offend the dignity of vulnerable groups, referring them to their non-existence or even denying the existence of this other. This problem can be even more deleterious when directed in the electoral context and here there are two important characteristics related to hate speech, namely: (i) the first of them is the very choice of representatives who endorse such practices, which would harm any state advance in legislative and administrative terms in the fight against hate speech and even the relativization of the problem and also (ii) political intolerance that tries to delegitimize the political-ideological choice of vulnerable groups and minorities, and may even culminate in physical violence.

The present work was not intended to delimit the parameters and limits of freedom of expression to analyze hate speech. However, for its very existence in the world, this value (or right) is essential and it is clear the understanding that only the identification of hate speech and its limits and restrictions are not adequate to create a doctrine regarding hate speech.

However, the data presented demonstrate that there are salutary characteristics when performing quantitative analysis. The first is that hate speech is not an exclusive phenomenon of social networks and, as a characteristic and reflection of life in society, it seems to increase, because there is an increase in the individual's contact with more people. Another important point and consequence of the first is that it is not possible to manipulate hate speech in order to remove it from the social context, whether physical or virtual. Hate is part of the circle of affections and its characteristic can only be manipulated, whether for greater or lesser intensity. Therefore, the

effort to modify this system goes beyond the behavior already performed by digital platforms in changing their algorithms and removing this content.

In order for the elections and, above all, voters to not be affected in their convictions, restrictions and the debate on hate speech must go through parliament. We suggest three methodological areas: (i) the first is deontological, which analyzes the impact of hate speech in conflict with fundamental values and rights; (ii) regarding the purposes and goals of hate speech regulation and (iii) empirical standards for measuring concrete damages.

The present study demonstrated that electoral hate speech in Brazil did not only stem from the 2018 general elections, as it appears. Hate speech on social media has started to show strength in electoral contexts since the 2014 elections and was the driving force of the 2016 US elections, with the use of social media as a marketing structure and legal strategy.

References

Altman, A. (1993). Liberalism and campus hate speech: A philosophical examination. *Ethics*, *103*(2), 302-317. https://doi.org/10.1086/293497

Anatel. (2021). *Panorama de dados de telefonia móvel.* https://informacoes.anatel.gov.br/paineis/infraestrutura/panorama

Brasil. (2021). *Lei nº 13.488, de 06 de outubro de 2017.* http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2017/lei/L13488.htm

Butler, J. (2021). *Discurso de ódio: Uma política do performativo*. Editora Unesp.

Carlson, C. R. (2021). *Hate speech*. The MIT Press.

Comunica que Muda. (2016). *Dossiê da Intolerância.* https://abcpublica.org.br/wp-content/uploads/2016/08/dossie_intolerancia.pdf

Chauí, M. (2000). *Mito fundador e sociedade autoritária.* Perseu Abramo.

Fadel, A. L. M. (2018). *O discurso de ódio é um limite legítimo ao exercício da liberdade de expressão?: Uma análise das teorias de Ronald Dworkin e Jeremy Waldron a partir da herança do liberalismo de John Stuart Mill*. Lumen Juris.

Freud, S. (2011). *Psicologia das Massas e Análise do Eu e outros Textos: 1920-1923*. 15. Companhia das Letras.

Guaraty, K. D. (2020). *Discurso de ódio: Conceito e hermenêutica no direito eleitoral*. (Dissertação de Mestrado em Direito e Desenvolvimento). Faculdade de Direito de Ribeirão Preto - Universidade de São Paulo, Ribeirão Preto. https://www.teses.usp.br/teses/disponiveis/107/107131/tde-02082022-103426/pt-br.php.

Lacan, J. (1998). Função e campo da fala e da linguagem em Psicanálise. In *Escritos*. Jorge Zahar.

Lawrence III, C. R. (1990). If he hollers let him go: Regulating racist speech on campus. *Duke Law Journal*, *431*(3). https://scholarship.law.duke.edu/cgi/viewcontent.cgi?article=3115&context=dlj

Matsuda, M. (1989). Public response to racist speech: Considering the victim's story. *Michigan Law Review*, *87*(8). https://repository.law.umich.edu/cgi/viewcontent.cgi?article=3417&context=mlr.

Nicolau, J. M. (2020). *O Brasil dobrou à direita: Uma radiografia da eleição de Bolsonaro em 2018*. Zahar.

Perrone, C. M. & Pfitscher, M. (2016). Discurso de ódio na internet: Algumas questões. *Redisco*, *10*(2), 146-154. https://periodicos2.uesb.br/index.php/redisco/article/view/2527/2088

Rais, D. & Tsuzuki, C. (2021). *A conexão entre o discurso eleitoral e o ódio.* Democracia e Direitos Fundamentais. https://direitosfundamentais.org.br/a-conexao-entre-o-discurso-eleitoral-e-o-odio/>.

SaferNet. (2022). *Crimes de ódio têm crescimento de até 650% no primeiro semestre de 2022.* https://new.safernet.org.br/content/crimes-de-odio-tem-crescimento-de-ate-650-no-primeiro-semestre-de-2022

Santos, M. A. & Silva, M. T. M. (2013). *Discurso do ódio na sociedade da informação preconceito, discriminação e racismo em redes sociais*. in Anais do Congresso Nacional do Conpedi/Uninove (pp. 82-99). São Paulo. Florianópolis.

Statista. (2021). *Facebook users in Brazil 2017-2025.* https://www.statista.com/statistics/244936/number-of-facebook-users-in-brazil/

Statista. (2021) *Leading countries based on Instagram audience size as of July 2021.* https://www.statista.com/statistics/578364/countries-with-most-instagram-users/

Statista. (2022). *Leading internet activities in Brazil in 2022.* https://www.statista.com/statistics/1052520/brazil-internet-activities/

Statista. (2021). *TikTok users in Brazil 2017-2025.* https://www.statista.com/forecasts/1142740/tiktok-users-in-brazil

Statista. (2021). *Twitter users in Brazil 2017-2025.* https://www.statista.com/forecasts/1146589/twitter-users-in-brazil

Statista. (2021). *Whatsapp users in Brazil 2017-2025.* https://www.statista.com/forecasts/1145210/whatsapp-users-in-brazil

Schwarcz, L. M. (2019). *Sobre o autoritarismo brasileiro.* Companhia das Letras.

Waldron, J. (2012). *The harm in hate speech.* Harvard Press.

POLITICAL ENGAGEMENT AND AGGRESSIVE USE OF SOCIAL NETWORKS. PRESIDENTIAL CAMPAIGNS IN A HIGHLY POLARIZED ELECTORAL SCENARIO

Adolfo A. Abadía
/ Universidad Icesi, Colombia

Luciana C. Manfredi
/ Universidad Icesi, Colombia

Juana L. Rodriguez
/ Universidad Icesi, Colombia

Introduction

Political science has studied electoral competition, analyzing the conflict between political parties, the role of its leaders, abstentionism, and the role of social networks as new forms of political communication (Restrepo, 2023), among others. Another part of the research has focused on the disconnection between the potential voter and politics. In fact, there is concern about political apathy and low political engagement[1] among voters (Delli Carpini, 2000).

Several authors who have analyzed the electoral behavior of young people characterize them with high levels of cynicism and apathy towards politics (Buckingham, 1997; Delli Carpini, 2000). However, the results are not conclusive since part of the research indicates that cynicism is not consistently associated with apathy

1. Political engagement includes a wide range of activities through which citizens develop and express their opinions, seek to participate in and shape the decisions that affect their lives.

and political participation (Austin & Pinkleton, 1995; Pinkleton & Austin, 2004). Colombian voters show high levels of apathy with political processes. Abstentionism has been studied as a structural phenomenon that ranges between 40% and 60% depending on the election (Barrero et al., 2013).

On another path, research on social networks shows that they function as a mechanism of communication, public opinion creation, and possibly engagement with political processes, but at the same time, as a space to generate fake news and hate speeches. Moreover, research shows that negative campaigns, which usually incite hate and polarization, achieve more engagement (Geer, 2012; Geer, 2006; Lau & Pomper, 2002; Lau, Sigelman & Brown, 2007).

The goal of this chapter is to analyze the presidential elections in Colombia 2018 and the impact of the social network Twitter® (now X)[2] during this contest, to try to explain the dynamics of electoral competition based on the communication in this social network. Additionally, an attempt is made to analyze the political interest that we will call political engagement with those candidates who are active in the use of Twitter® and who have aggressive and negative speeches. This text aims to understand and explain the new ways of doing politics in social networks. Thus, it contributes to the debate on this topic by analyzing the electoral contest and the use of social networks to generate engagement. It is important to note that aggressive and negative speeches increase as an interaction between presidential candidates.

Social networks play a fundamental role in political campaigns as they are political spaces aiming to capture the interest of potential voters for the sake of gaining power. A commonly employed tactic is negative campaigning, which seeks to highlight the opponent's weaknesses instead of highlighting one's qualities (D'Adamo & Garcia Beaudoux, 2015). Its benefits include increasing citizens' interest and attention in elections, stimulating public debate, and simplifying the decision for some voters. These campaigns show an immediate and short-term effect, generating a

2. Although the social network formerly known as Twitter® has undergone a name change and other terminology updates (Blanco, 2023), this chapter will continue to refer to it as was before, with messages being referred to as Tweets and shares as Retweets.

high level of citizen participation and attracting attention to politics, either directly or indirectly (Lau, Sigelman & Brown, 2007).

In Colombia, telecommunications infrastructure and network access allow high levels of connectivity. By 2017, the country reached a figure of 28.4 million broadband internet connections. In this context, the social network Twitter®, introduced in 2006 as a short message system intended primarily to serve as a mobile application, has now become a vast news and information network that is used worldwide by millions of people across multiple platforms (Morris, 2009).

Users of this social network, apart from being able to post tweets – messages of around 280 characters – and redirect them to other users – retweets –, but the can also be included in lists of popular topics and can choose to participate in discussions using different types of media: text, links, hashtags, trending topics, videos, and images. In this way, users, especially young people, participate in the production and reproduction of information, news, knowledge, and content, thus enriching the electoral political arena.

It is important to note that social networks, such as Twitter®, enhance political movements and citizen debates, new mode of participation (de Casas-Moreno et al., 2023), and so can be used to transmit hate speech and intolerance towards opponents, which can provoke controversy (Geer, 2012). Negative campaigns taken to the extreme can lead to adverse consequences, such as the boomerang effect, the victim syndrome, and the double harm effect (D'Adamo & García Beaudoux, 2015).

However, negative campaigns have been found to increase political participation or at least make citizens more resilient to them, rather than demobilizing electoral participation (Brooks, 2006). Hopp and Vargo's (2017) study found that during the 2012 US presidential election, increased levels of negative campaign ads correlated with increased citizen activity on Twitter®, indicating increased political participation. This highlights the importance of social networks in political participation, particularly in the context of negative political campaigns.

This study explores the impact of Twitter® on the 2018 Colombian presidential elections, analyzing the dynamics of electoral competition through communication on this platform. It also examines the level of political engagement among candidates who use negative campaign elements on this microblogging social network.

Literature review

About social networks and political behavior

Studies show that the use of Information and Communication Technologies (ICTs) in proselytizing activities improves the relationship between candidates and voters (Stanyer, 2005). For this reason, in recent times, candidates and their political parties around the world are leveraged social networks to reach the masses of citizens (Hong & Nadler, 2012). Moreover, Stanyer (2005) asserts that without ICTs, it would be very difficult for candidates and political parties to mobilize their followers and convince undecided voters.

Considering the above, it can be said that the value of social networks in the success of an election lies in the interconnection with more traditional media, such as television, radio, and newspaper, to provide a platform that allows for greater democratic participation, inclusion, and expression, especially among young voters (Essoungou, 2010).

Furthermore, social networks also have an impact, on the one hand, in reinforcing pre-existing political values, which are inherited and transmitted through interpersonal interactions. On the other hand, they also facilitate the formation of new connections, which are exclusively achieved through online communication (Di Fátima & Carvalheiro, 2023). This is particularly relevant in the context of young people, who are the most connected demographic group because they are the ones who use social networks the most for various reasons, including the search for political information.

These networks helps to inform themselves, to share information, often spreading fake news, and to interact more horizontally and democratically

with candidates. In the study conducted by Essoungou (2010), a correlation was found between the participation of young people and the dissemination of messages on social networks. The results show that social networks are an important platform for the exchange of political information among young people.

The study conducted by Manfredi and González (2019) explains how the media play a fundamental role in establishing public agenda items. As Holgado González (2003) explains, the media are crucial in the functioning of democracy since they serve as information channels for citizens who will exercise their right to vote.

For this reason, technological advances and the incorporation of new technologies into campaigns have led to part of the electoral competition taking place in non-traditional scenarios. In this sense, social networks are positioned as a new space for political competition, where more direct and personal interactions between candidates and their potential voters are evident. This type of behavior changes the electoral dynamics, influencing opinions and voting intentions.

In this way, communication becomes central to the campaign strategy. Manfredi, González, and Biojó (2019), study the dynamics of electoral competition in the communication of candidates on Twitter®. One of their contributions is related to the attacks in the media and social networks and how to respond to them when the number of resources to invest in a campaign is limited and does not imply a substantive change in their strategic campaign plan. Thus, attacks in social networks are presented as a tactic to maintain balance and establish agenda items or agenda-setting (Cohen, 2015; McCombs & Shaw, 1972).

Social networks have become an effective medium for political communication due to their easy accessibility, ability to reach large audiences, low barriers to entry, and real-time feedback potential (Lee & Xu, 2018). As a result, they are now a tool that political candidates must consider due to their versatility and immediacy in conveying messages to potential voters.

Therefore, social network platforms like Twitter® have become essential in political campaigns (Lee & Xu, 2018) as they shift the focus from political parties to individual candidates.

Spierings and Jakobs (2014) have researched the influence of social networks on electoral behavior. As stated by Lee and Xu (2018), this platform enables users to share their profiles and tweets with a wider audience. In addition, Twitter® provides a simple format that allows interaction between candidates and voters, improving their relationship over time. This can be particularly important in highly personalistic political systems, such as the one in Colombia (Carey & Shugart, 1995). According to Pérez-Curiel and García-Gordillo (2018), a political candidate's personalization can lead to a high rate of user response, including likes, retweets, and comments, which may exceed the activity of the party on the social network.

Social networks have served to gain visibility for political parties and their candidates. This visibility can have a positive correlation with the effectiveness of the results (Yamamoto, 2010). Hence, it can be said that social networks have an impact on political participation and that this relationship can be enhanced when citizens believe that they can effect change through their participation (Abramson & Aldrich, 1982). Using social networks, voters can access information more easily, as well as create a bond of trust with candidates, therefore, it is expected that political participation and political engagement will increase thanks to the use of new technologies.

It is crucial to examine the reasons for the rise of hate speech in political discourse worldwide. Restrepo (2023) argues that the rise of social networks has contributed significantly to this phenomenon, as individuals now have vast platforms to disseminate information; thus because of the decentralized, anonymous and interactive structure of such networks (Msughter, 2023).

The expansion of social networks has created new spaces for debate, facilitating access to diverse perspectives and opinions, and increasing political participation in a digitalized world. However, the increased circulation of

content promoting hate speech and hate crimes has also been observed in direct proportion to this accessibility (Amores et al., 2021).

The availability of various speeches, opinions, and ideologies, including extremist ones, provides suitable platforms for the dissemination of social, cultural, and political information that may contribute to the development of extremist practices and rhetoric. The dissemination of unverified information through social networks is a growing concern due to legal loopholes that make it difficult to distinguish between freedom of expression and hate speech (Bustos Martínez et al., 2019). This contributes to the creation of digital spaces that promote intolerance, limit exposure to different points of view, and ultimately lead to political polarization.

About the Colombia case related to this matter

The reality of the effect of hate/polarizing speeches can be evidenced in many cases, day by day, in the world political arena; however, it is significant to bring it down to the Colombian case with a highly mediatic event that marked a before and after of social networks in Colombian political campaigns: the 2016 Peace Plebiscite.

After almost 60 years of internal armed conflict in Colombia, in 2016 President Juan Manuel Santos called on Colombians to vote in a plebiscite to approve (Yes) or reject (No) the signing of the Peace Agreement with the FARC[3] after 4 years of negotiations. Both the negotiation process and the plebiscite campaign were highly publicized events in the media and the social networks of actors for and against the agreement. In the end, and to the surprise of the whole world, the 'No' option won at the polls in Colombia, although not by much (Basset, 2017).

Juárez Rodríguez & Restrepo Echavarría (2022) called Colombia a "pioneer in the rise of disinformation and manipulation structures by populist

3. Spanish acronym for *Fuerzas Armadas Revolucionarias de Colombia-Ejército del Pueblo* (FARC-EP), a far-left Colombian insurgent guerrilla organization founded in 1964. Henceforth, it will be referenced in this chapter as FARC.

movements and far-right formations" in recent decades because of the intense opposition campaign led mainly by the Democratic Center party headed by former President Álvaro Uribe, but also reinforce as a unique message with all congressmen this political party (Cifuentes & Pino, 2018).

The strategy used by the opponents that gave them the victory was guided by emotions. It was a campaign based on feelings of hate, fear, and, above all, a characterization of society between the "good Colombians" – who voted no as a symbol of protest to impunity and considered themselves as the traditional good patriots – and the "bad Colombians" – who handed Colombia over to the FARC –. Another successful strategy planned by the opponents was the creation of a discourse of suspicion and complete rejection of Castrochavismo – referring to a sort of communist movement that brought Venezuela (the neighboring country) to ruins – that would arrive in the country in case a Peace Agreement was signed.

Opposition sectors used these strategies to polarize, misinform, manipulate, and spread discourses of hate and fear to a population exhausted and hurt by an extensive conflict, and saturated with information on the negotiations for the end of the conflict to reject what would be a historic Peace Agreement in the region (Rodríguez et al., 2022).

The reason for the great success of the negative campaign used in the previous case is shown by Manfredi et al. (2019) when they argue that negative messages attract attention because they generate conflict and controversy, which allows them not to be easily ignored or forgotten. It is not a question of whether a positive campaign is preferable to a negative campaign, but of the impact it is intended to leave on potential voters.

Any candidate using a negative campaign seeks to generate a motivational impression that makes voters think that they are "strong, tough" to, in the same way, make their opponent look weak and lacking in character. By having more visibility and being seen as strong, candidates have the possibility

of discussing issues of their interest or convenience; this is, ultimately, one of the great purposes of political campaigns.

Notwithstanding the above, it is not prudent to argue that the use of negative campaigns full of hate speech is always preferable in all cases. Nor is it a matter of using negative campaigning as the only political strategy (D'Adamo & Garcia Beaudoux, 2015). What it is seen in the political arena is a mix of moments when a positive campaign is necessary and others when a negative campaign, which is more impactful, generates greater results. The 2018 presidential election in Colombia is, to some extent, an example of this.

These elections are noteworthy because they happened post plebiscite for the peace accords where the 'No' won for reasons explained above. The political environment at the time was marked by disagreements and quarrels inherited from the highly mediatized campaign for and, above all, against the 'Yes' option. Consequently, the debates of the moment revolved mainly around peace and conciliation issues, although each candidate had their own flagship issue.

Gustavo Petro and Ivan Duque were the candidates who contested most aggressively. Petro continued with the peace themes he supported in the 2016 plebiscite by being part of the pro-candidates (those who were called communists, castrochavistas) and Duque was a relatively new political actor whose theme was entrepreneurship/economy and who had the support of Álvaro Uribe, the biggest opponent of the peace agreements and one of the most active politicians in social networks (Twitter®) to express his disagreement (Suárez Álvarez, 2024).

Methods and data

This chapter presents an analysis of the data from Manfredi et al. (2019), which covers the last 90 days before the first round of the presidential election on May 27, 2018. This period corresponds to the three months allowed

for an electoral campaign in Colombia (Ley 130 de 1994). The study of the 2018 presidential electoral campaign is motivated by the highly polarized political context in which this election took place, mainly in the social network sphere (Chenou & Restrepo, 2023).

This was the first presidential election after two consecutive terms of the Santos administration, which focused on the signing and implementation of the peace agreement with the FARC. In the last two decades, the peace process has been a highly divisive and tense issue among the national political class, elite, and Colombian society in general. Twitter has been the primary social network for expressing both support and opposition to this state policy (Vallejo Mejía et al., 2022).

The selection of Twitter® messages was carried out using *Twitonomy*, a tool that allows the analysis and monitoring of trends of messages on this social network. The five presidential campaigns with the highest voting intention were followed: Humberto de la Calle (@DeLaCalleHum), Sergio Fajardo (@sergio_fajardo), Germán Vargas Lleras (@German_Vargas), Iván Duque (@IvanDuque), and Gustavo Petro (@petrogustavo), and a total of 10,603 tweets were collected for this study period.

Table 1 shows the descriptive statistics of the study, which were obtained from the twitonomy application and refined between the months of June and July 2018. In general, the two candidates who register the highest use of the social network are Gustavo Petro and Humberto de la Calle, with 3,077 and 2,123 messages, and an average of 34 and 24 tweets per day, respectively. The accounts of Petro and Sergio Fajardo are the most followed on Twitter®.

The accounts that have generated the most engagement in terms of likes and retweets according to their messages are those of Petro and Iván Duque. These two candidates were the ones who finally went to the second round on June 17 in same year, and it was Iván Duque who obtained the highest number of votes making him the 41st president of the Republic of Colombia.

Table 1. Descriptive data

Candidates	Tweets	Followers	Likes	Retweets	Activity*	Visibility**	Engagement***
de la Calle	2.123	158.295	1.339.023	391.778	24	336.060.285	247.103.040
Fajardo	1.875	1.285.416	1.253.043	403.692	21	2.410.155.000	269.783.165
Vargas Lleras	1.763	784.946	362.593	186.133	20	1.383.859.798	38.281.635
Duque	1.765	360.900	1.280.220	693.997	20	636.988.500	503.381.779
Petro	3.077	3.166.804	4.207.365	1.555.742	34	9.744.255.908	2.127.258.511
Total	*10.603*	*5.756.361*	*8.442.244*	*3.231.342*	*118*	*61.034.695.683*	*2.572.835.765*

Note: * Average number of tweets per day (Tweets/90 days). | ** Maximum possible reach (Tweets x Followers). *** Average interaction between user reactions to published tweets (Likes x Retweets / Tweets). **Source**: Own elaboration based on data from Manfredi et al. (2019).

To characterize the interactions between the candidates in terms of aggressiveness, the number of cross-mentions was identified, i.e., when a candidate included the user or directly mentioned another candidate. This type of message is part of a communication strategy, as a kind of aggressiveness at a tactical level, established through the competitive dialogue between the candidates. Thus, if a candidate sends messages that are more assertive or confrontational to their opponents, we label their behavior as aggressive (*Aggressor*). Conversely, if a candidate receives a significantly higher number of such messages, we label the recipient as aggressive (*Target*).

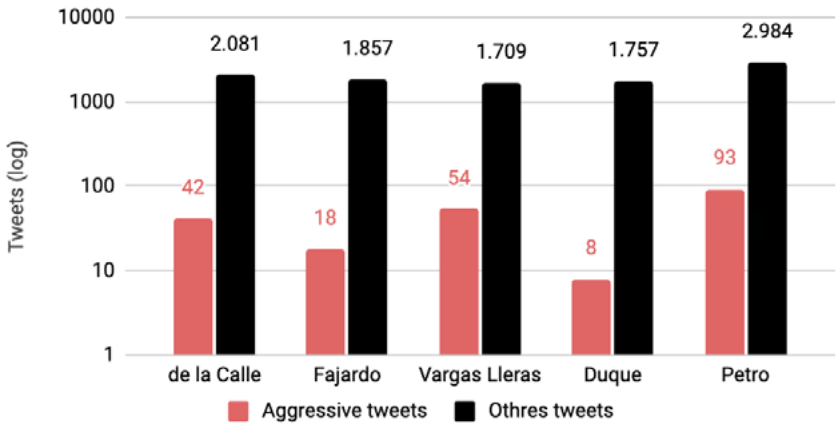
As a final methodological consideration of this study, it is proposed to study the favorability of the candidates based on the voting intentions reported in *Cifras & Conceptos (C&C)* during the campaign period. For this, the results of two findings are included, one at the beginning of the period in March 2018, whose report collection date was February 23 to 26, and one at the end of May 2018, whose report collection date was May 14 to 17 (Sonneland, 2018).

Results

The following results are based on an analysis of the 2018 Colombian presidential elections and the impact of Twitter. The study examines the level of commitment received by the main five candidates in terms of voting

intention through their aggressive messages on this social network. It considers the variation and influence of these messages on electoral support surveys during the study period.

Figure 1. Candidate non-aggressive tweets vs. aggressive tweets



Source: own elaboration based on data from Manfredi et al. (2019).

To begin with, Figure 1 shows the ratio of messages identified as aggressive to the total number of tweets published by each candidate. As can be seen, during the 3 months of monitoring, Petro was the one who published the most messages related to his electoral rivals. The candidate Duque reports the opposite case with a significantly lower value.

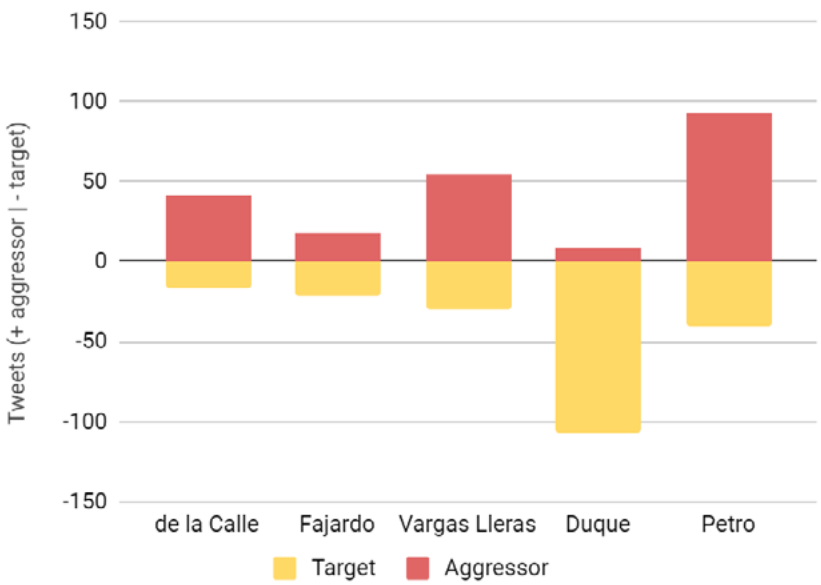
Due to the 280-character limit per message, it can be assumed that Twitter® is used as a means of communication in a strategic and calculated manner, consistent with a political message (Abadía et al., 2023). Trills, in this sense, constitute a type of textual narrative that reflects emotions resulting from the externalization of cognitive representations of the experience (Kleres, 2011).

Compared to the 2022 presidential scenario, this study reveals an aggressive tone in messaging characterized by friend-enemy relationships between candidates. This may be influenced by the peace process and

compliance with its agreement, which has led to high political polarization (Manfredi et al., 2021).

Gustavo Petro was the most aggressive candidate (see Figure 2), with the largest number of followers on Twitter®, he was the candidate who most used this social network as a means of connecting with his voters, and the least tolerant (Manfredi et al., 2019). His campaign was aggressive against his strongest opponent and having the largest number of followers, his tweets were highly reproduced and debated; his campaign was polarizing and appealed considerably to feelings and emotions that sought to belittle Duque, making him appear to public opinion as an unacceptable and morally inferior candidate, criminalizing his opinions and radicalizing the political message he communicated to his followers (Prada Espinel & Romero Rodríguez, 2019).

Figure 2. Candidates sending aggressive tweets vs. being the target of aggressive tweets.



Source: own elaboration based on data from Manfredi et al. (2019).

Ivan Duque, on the other hand, was a more passive candidate, he had a considerably lower number of followers on Twitter® than Gustavo Petro and his campaign tried to cover issues other than peace and its agreements; his speech focused on entrepreneurship and the orange economy. They were completely different campaigns, without departing from the characteristic personalistic contests based on mainly ideological attacks that increased polarization (Geer, 2006), which led to the election of Ivan Duque as the new president of the nation.

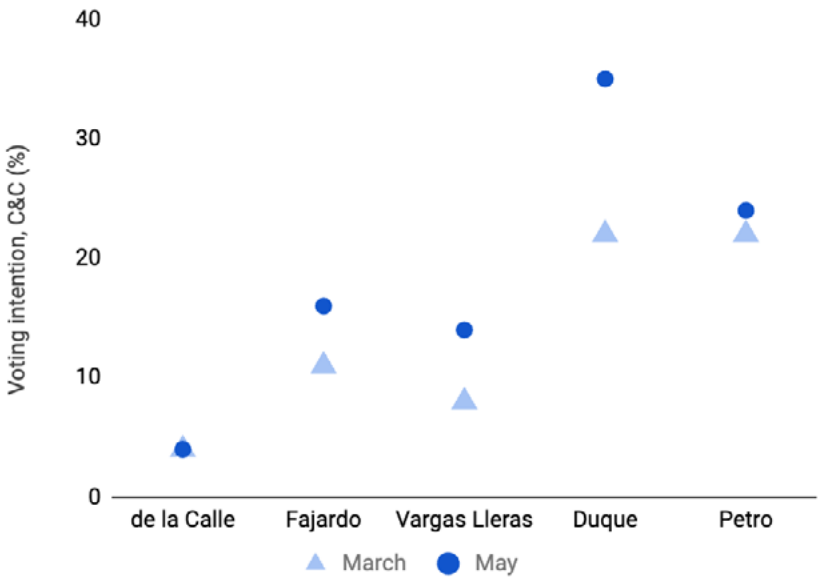
When comparing the previous comments to the results of the *Cifras & Conceptos* voting intention survey, it is evident that Duque experienced the greatest increase in support between March and May, despite starting at the same level as candidate Petro (see Figure 3).[4] Humberto de la Calle's voting intention remains at 4%, indicating that his visibility as a leader at the peace negotiation table in Havana did not necessarily translate into electoral support, as he may have anticipated before entering the 2018 presidential race.

Furthermore, although Petro and de la Calle were the candidates who published the most tweets during this period, their relative positioning in the electoral sphere was minimally impacted. In contrast to Germán Vargas Lleras, who can also be considered an aggressive candidate on Twitter®, Petro's criticism of his opponents on social networks did not significantly increase his voting intention.

This variation dialogues with the assertion that indicates that to the extent that political cleavages are deconstructed as well as links with traditional actors and organizations, the voting decision tends to become clearer as election day approaches (Luengo & Peláez-Berbell, 2017). It is important to note that the survey included in this study was conducted ten days before the May election.

4. Values correspond to the answer to the Polimetric survey question: Which of the following possible presidential candidates would you vote for? (C&C, 2024).

Figure 3. Variation in voting intention for presidential candidates, March-May 2018 (Cifras & Conceptos)

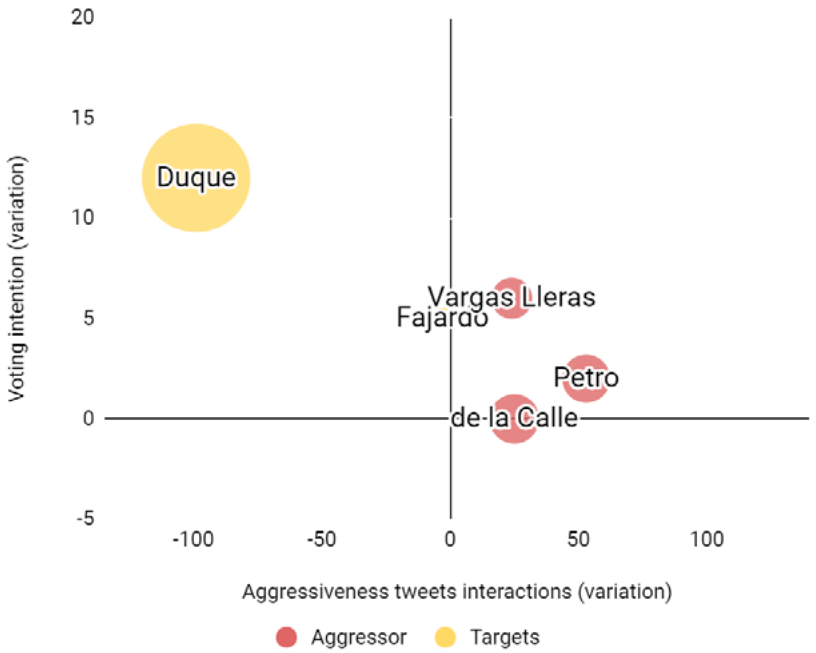


Source: own elaboration based on data from Sonneland (2018).

During this three-month period, candidate Duque experienced the greatest growth (see his circle sizes in Figure 4). Although he was the main target of other candidates' messages, this worked in his favor by increasing his visibility. This was in addition to the strategies employed by his campaign team, political party, and other strategic allies.

As an example, this presidential election has two elements that energized the issues of the electoral campaign; on the one hand, there is the 2016 Peace Plebiscite as a precedent, and on the other the figure of former president Álvaro Uribe behind the candidate Duque of the Democratic Center as an emotional effect of approval on a significant part of the voters (Milanese & Serrano Corredor, 2021).

Figure 4. Variations in candidates' voting intentions and aggressiveness in their Twitter® communications (March-May 2018)



Note: Circle sizes indicate the degree to which the *Aggressor* and *Target* profiles differ among the candidates. *Source: own elaboration based on data from Sonneland (2018) and Manfredi et al. (2019).*

From another perspective, as Murthy (2015) highlights, having an aggressive social network profile does not necessarily lead to better polling results, as demonstrated by the experience of candidate Petro in 2018. Contrary to expectations, being the target of negative appeals may increase voters' desire to seek out new information, leading them to learn more about the candidate and potentially influencing their support for them (Gelman et al., 2021).

Figure 4 shows that during the three-month study period, the patterns of social network use were more similar among the candidates Fajardo, de la Calle, Vargas Lleras, and Petro. That is, the candidates are more inclined toward an aggressive type of profile, except for Fajardo who is slightly

inclined toward a target-type profile. Additionally, the survey shows that some candidates registered a maximum growth of 6 points in the vote intention survey, while others remained stagnant between the two moments of observation, as previously stated with candidate Humberto de la Calle. This happens at some distance from candidate Duque.

Conclusions

This is a preliminary exploratory study that aims to show the engagement produced by hate speech and negative discourse in elections in polarized contexts. Overall, this study is trying to contribute to the understanding of Twitter® as an extensive tool used by political campaigns and candidates to promote their proposed political agenda, engage with their existing followers, and attract new supporters by spreading short but powerful messages. Some of these messages are related to content that tends to discursively attack the opponent by discrediting his program proposals, public appearances speeches, and campaign actions.

The results of the first round of the 2018 presidential race indicate that candidates who engage in aggressive behavior on social networks, particularly Twitter®, and receive a high volume of aggressive messages tend to perform better in the elections. It is important to note, however, that this correlation does not necessarily imply causation. We hope in the future to expand our study and contribute to the understanding of how hate speech in electoral campaigns, although it can contribute to engagement, generates a detriment to democracy.

Future studies should explore hate speech in other highly polarized political climates from a comparative perspective. Additionally, studies could examine not only the candidate's profile but also the users who are part of their support group, considering these actors as an extension of their scope of coverage.

Another area of research could involve examining hate speech beyond electoral contexts, particularly during times of government crisis when political figures are exercising their elected positions. It is important to understand

the terms under which this interaction with the audience occurs, the relationships between political actors from the same party or political line, and those in direct opposition.

As is common, this study faces some limitations when it comes to explaining the phenomenon studied. Firstly, there is no direct relationship between the link that is woven between aggressive messages on social networks, voting intention, and electoral results. However, this type of communication action aims to allocate new voters, mainly due to a strategic campaign orientation.

Another limitation is related to the change in Twitter®'s information use policy, which required us to revisit a database created in 2019. However, this highlights the importance of protecting research data, which can capture a snapshot of a specific historical moment, viewed through different lenses at different times.

Finally, it is important to note that understanding the electoral campaign solely through one source, such as Twitter® in this case, may lead to the assumption that other social networks and traditional media have equal levels of message dissemination, interaction, and engagement. While it is recognized that this is not always the case, this study focuses on one of the main channels of interaction between candidates and the electorate in recent decades, not only at the Colombian level but also worldwide.

References

Abadía, A. A., Manfredi, L. C., & Sayago, J. T. (2023). Comunicación de crisis durante la pandemia del Covid-19 y su impacto en los sentimientos de la ciudadanía. *Opinião Pública*, *29*(1), 199-225. https://doi.org/10.1590/1807-01912023291199

Abramson, P. R. & Aldrich, J. H. (1982). The decline of electoral participation in America. *American Political Science Review*, *76*(3), 502-521.

Amores, J. J., Blanco-Herrero, D., Sánchez-Holgado, P., & Frías-Vázquez, M. (2021). Detecting ideological hate on Twitter. Development and evaluation of a detector of hate speech by political ideology in Spanish-language tweets. *Cuadernos.info*, (49), 98-124. http://dx.doi.org/10.7764/cdi.49.27817

Austin, E. W. & Pinkleton, B. E. (1995). Positive and negative effects of political disaffection on the less experienced voter. *Journal of Broadcasting & Electronic Media*, *39*, 215-235.

Barrero, F., Liendo, N., Mejía, L., Orjuela, G., & Caicedo, J. (2013). *Electoral abstentionism in Colombia*. Bogotá: Registraduría Nacional del Estado Civil, Centro de Estudios en Democracia y Asuntos Electorales, Universidad Sergio Arboleda, GAP.

Basset, Y. (2017). Claves del rechazo del plebiscito para la paz en Colombia. *Estudios Políticos (Medellín)*, *52*. https://doi.org/10.17533/udea.espo.n52a12

Blanco, U. (2023). ¿Qué significa el cambio de Twitter a X y cómo afecta a usuarios? *CNN*. https://cnnespanol.cnn.com/2023/07/25/que-significa-cambio-twitter-x-como-afecta-usuarios-orix/

Brooks, D. J. (2006). The resilient voter: Moving toward closure in the debate over negative campaigning and turnout. *The Journal of Politics*, *68*(3), 684-696. https://doi.org/10.1111/j.1468-2508.2006.00454.x

Buckingham, D. (1997). News media, political socialization and popular citizenship: Towards a new agenda. *Critical Studies in Mass Communication*, *14*, 344-366.

Bustos Martínez, L., De Santiago Ortega, P. P., Martínez Miró, M. Á., & Rengifo Hidalgo, M. S. (2019). Discursos de odio: Una epidemia que se propaga en la red. Estado de la cuestión sobre el racismo y la xenofobia en las redes sociales. *Mediaciones Sociales*, *18*, 25–42. https://dialnet.unirioja.es/servlet/articulo?codigo=6963028

Carey, J. M. & Shugart, M. S. (1995). Incentives to cultivate a personal vote: A rank ordering of electoral formulas. *Electoral Studies*, *14*(4), 417-439.

Chenou, J. M. & Restrepo, E. M. (2023). Una nación dividida: Análisis del discurso político en redes sociales antes del plebiscito del acuerdo de paz con las FARC. *Análisis Político*, *36*(106), 60-84. https://doi.org/10.15446/anpol.v36n106.111038

Cifras & Conceptos – C&C. (2024). *Polimétrica – Descargables*. Productos y servicios. https://www.cifrasyconceptos.com/?page_id=3809

Cifuentes, C. F. & Pino, J. F. (2018). Conmigo o contra mí: Análisis de la concordancia y las estrategias temáticas del Centro Democrático en Twitter. *Palabra Clave*, *21*(3), 885-916. https://palabraclave.unisabana.edu.co/index.php/palabraclave/article/view/8166

Cohen, B.C. (2015). *Press and foreign policy, 2321*. Princeton University Press.

D'Adamo, O. & García Beaudoux, V. (2015, April 17). Campaigning from the opposition: Basic decalogue for choosing a good communication strategy. *El Consultor*.

de Casas-Moreno, P., Parejo-Cuéllar, M., & Vizcaíno-Verdú, A. (2023). Hate speech on Twitter: The LGBTIQ+ community in Spain. In B. Di Fátima (Ed.), *Hate speech on social media: A global approach* (pp. 143-158). LabCom Books & EdiPUCE.

Delli Carpini, M. X. (2000). Gen.com: Youth, civic engagement, and the new information environment. *Political Communication, 17*, 341-349.

Di Fátima, B. & Carvalheiro, J. R. (2023). One's heaven can be another's hell: A mixed analysis of Portuguese nationalist fanpages. *Social Sciences*, *13*(1), 29. https://doi.org/10.3390/socsci13010029

Essoungou, A. M. (2010). *Africa's social media revolution*. Bizcommunity.com.

Geer, J. G. (2006). *In defense of negativity: Attack ads in presidential*. University of Chicago.

Geer, J. G. (2012). The news media and the rise of negativity in presidential campaigns. *Political Science and Politics*, *45*(3), 422-427.

Gelman, J., Wilson, S. L., & Sanhueza Petrarca, C. (2021). Mixing messages: How candidates vary in their use of Twitter. *Journal of Information Technology & Politics*, *18*(1), 101-115. https://doi.org/10.1080/19331681.2020.1814929

Holgado González, M. (2003). The role of the media in the electoral campaign. Ámbitos, Revista Internacional de Comunicación, *10*(online).

Hong, S. & Nadler, D. (2012). Which candidates do the public discuss online in an election campaign?: The use of social media by 2012 presidential candidates and its impact on candidate salience. *Government Information Quarterly*, *29*(4), 455-461. https://doi.org/10.1016/j.giq.2012.06.004

Hopp, T. & Vargo, C. J. (2017). Does negative campaign advertising stimulate uncivil communication on social media? Measuring audience response using big data. *Computers in human behavior*, *68*, 368-377.

Juárez Rodríguez, J. & Restrepo Echavarría, N. J. (2022). Política, mentiras y discursos de odio: Colombia y España como paradigmas de las campañas de manipulación y noticias falsas en Europa y América Latina. In A. Barrientos-Báez, F. J. Herranz Fernández, & D. Caldevilla Domínguez (Eds.), *Estrategias de comunicación: Género, persuasión y Redes sociales* (1a ed., pp. 151-164). Gedisa.

Kleres, J. (2011). Emotions and narrative analysis: A methodological approach. *Journal for the Theory of Social Behaviour*, *41*(2), 182-202. https://doi.org/10.1111/j.1468-5914.2010.00451.x

Lau, R. R. & Pomper, G. (2002). Effectiveness of negative campaigning in U.S. Senate. *American Journal of Political Science*, *46*(1), 47-66.

Lau, R. R., Sigelman, L., & Brown, I. (2007). The effects of negative political campaigns: A meta-analytic reassessment. *The Journal of Politics*, *69*(4), 1176-1209.

Lee, J. & Xu, W. (2018). The more attacks, the more retweets: Trump's and Clinton's agenda setting on Twitter. *Public Relations Review*, *44*, 201-213.

Ley 130 de 1994, Dario Oficial No. 41280 (1994), por la cual se dicta el Estatuto Básico de los partidos y movimientos políticos, se dictan normas sobre la financiación y la de las campañas electorales y se dictan otras disposiciones. https://pdba.georgetown.edu/Parties/Colombia/Leyes/Ley130.pdf

Luengo, O. G. & Peláez-Berbell, J. (2017). Exploring the accuracy of electoral polls during campaigns in 2016: Only bad press? *Contemporary Social Science*, *14*(1), 43-53. https://doi.org/10.1080/21582041.2017.1393553

Manfredi, L. C. & González, J. M. (2019). Comunicación y competencia en Twitter. Un análisis en las elecciones presidenciales Colombia 2018. *Revista Estudios Institucionales, 6*(11), 133-150.

Manfredi, L. C., Abadía, A. A., & Sayago, J. T. (2021). Twitter, sentimientos y precandidatos presidenciales. Comunicación en tiempos de paro nacional. *Elecciones*, *20*(22), 309-336. https://doi.org/10.53557/Elecciones.2021.v20n22.09

Manfredi, L. C., González, J. M., & Biojó Fajardo, D. (2019). ¡Twitteo, luego existo! Un análisis de las dinámicas de competencia electoral de los candidatos a la Presidencia de Colombia 2018 en Twitter. In *Elecciones Presidenciales y de Congreso 2018. Nuevos acuerdos ante diferentes desafíos* (pp. 117-146). Konrad Adenauer Stiftung.

McCombs, M. E. & Shaw, D. L. (1972). The agenda-setting function of mass media. *Public Opinion Quarterly, 36*(2), 176-187.

Milanese, J. P. & Serrano Corredor, C. E. (2021). Realineamiento electoral. Análisis de la transferencia de votos en escenarios transicionales en Colombia. *Revista de Sociología e Política*, *29*(78), e008. https://doi.org/10.1590/1678-987321297908

Morris, T. (2009). *All a Twitter: A personal and professional guide to social networking with Twitter*. Que Publishing.

Msughter, A. E. (2023). Social media narratives and reflections on hate speech in Nigeria. In B. Di Fátima (Ed.), *Hate speech on social media: A global approach* (pp. 255-275). LabCom Books & EdiPUCE.

Murthy, D. (2015). Twitter and elections: Are tweets, predictive, reactive, or a form of buzz? *Information, Communication & Society*, *18*(7), 816-831. https://doi.org/10.1080/1369118X.2015.1006659

Pérez-Curiel, C. & García-Gordillo, M. (2018). Influence politics and fake trend on Twitter. Post-electoral effects (21D) in the framework of the Procés in Catalonia. *El profesional de la información*, *27*(5), 1030-1040.

Pinkleton, B. E. & Austin, E. W. (2004). Media perceptions and public affairs apathy in the politically inexperienced. *Mass Communication and Society*, *7*, 319-337.

Prada Espinel, O. A. & Romero Rodríguez, L. M. (2018). Polarización y demonización en la campaña presidencial de Colombia de 2018: Análisis del comportamiento comunicacional en Twitter de Gustavo Petro e Iván Duque. *Revista Humanidades*, *9*(1). https://doi.org/10.15517/h.v9i1.35343

Restrepo, C. (2023). Redes sociales y participación política en las elecciones presidenciales de 2022 en Colombia. *Análisis Político*, *36*(106), 133-164. https://doi.org/10.15446/anpol.v36n106.111058

Sonneland, H. K. (2018, June 11). *Poll Tracker: Colombia's 2018 Presidential Election*. Americas Society/Council of the Americas (AS/COA). https://www.as-coa.org/articles/poll-tracker-colombias-2018-presidential-election

Spierings, N. & Jacobs, K. (2014). Getting personal? The impact of social media on preferential voting. *Political Behavior*, *36*(1), 215-234.

Stanyer, J. (2005). Political parties, the Internet and the 2005 General Election: From web presence to e-campaigning? *Journal of Marketing Management*, *21*, 1049-1065.

Suárez Álvarez, A. V. (2024). #PazEsVotarNO. Centro Democrático y Acuerdo de Paz en Colombia en redes sociales. *Estudios Políticos (Medellín)*, *69*. https://doi.org/10.17533/udea.espo.n69a12

Vallejo Mejía, M., Gómez Céspedes, L., Lombana Bermúdez, A., & Pino Uribe, J. F. (2022). Encuadres en pugna por la paz en Twitter: El caso de las elecciones subnacionales del 2019. In F. Botero, B. Ortega, J. F. Pino Uribe, & L. Wills Otero (Eds.), *En configuración permanente: Partidos y elecciones nacionales y subnacionales en Colombia, 2018-2019* (1a ed., pp. 255-282). Universidad de los Andes: Pontificia Universidad Javeriana.

THE EUROPEAN LEGAL APPROACH TO FIGHT HATE SPEECH ON SOCIAL MEDIA

Ana Gascón Marcén

/ University of Zaragoza, Spain[1]

Introduction

The Internet is a great tool than can help to exercise human rights and it is growing in importance every day, even more as many aspects of life are being rapidly digitalised such as work or education. It is an instrument that can help access information since any point of the world and reach millions as an audience without the previous filters that traditional media had. However, it also poses serious risks for our democracies and can give a powerful loudspeaker to people that want to extend hate speech at scale.

Europe prides itself for having a robust system of human rights protection including freedom of expression but at the same time to fight strongly hate-speech. New challenges arise in the Internet due to its very nature. The Council of Europe was a pioneer in this sector and the European Union (EU) has been pressured in the last years to deal with this problem as part of the digital single market.

In this chapter, regarding the Council of Europe (Section 2), the analysis covers its soft law, the Protocol to the Budapest Convention and the case-law of the

1. Member of the research team of the project "Towards a person-centred digital transition in the European Union" (TRADIPER). This publication is part of the TED2021-129307A-I00 project, funded by MCIN/EIP/10.13039/501100011033 and the European Union's "NextGenerationEU"/PRTR.

European Court of Human Rights. Regarding the EU (Section 3), it assesses the E-Commerce Directive, the Audiovisual Media Service Directive, the Code of Conduct on countering illegal hate speech online and the Digital Services Act.

Measures to fight hate speech by the Council of Europe

The Council of Europe is the main intergovernmental organization working in Europe to promote and protect democracy, human rights and the rule of law.[2] It is considered a leader in fighting hate speech as it was the first to adopt an official definition of it in 1997.[3]

Soft-Law

In 2022, the Committee of Ministers passed a new Recommendation on hate speech updating the previous standard to the current situation and the expansion of the Internet. It called on governments to develop comprehensive strategies to prevent and fight hate speech, including the adoption of an effective legal framework and implementing adequately calibrated and proportionate measures.[4]

Following the Recommendation, States should: ensure that their legislation addressing hate speech covers online hate speech and contains clear and foreseeable provisions for the swift and effective removal of it; define and delineate the duties and responsibilities of State and non-State actors in addressing online hate speech; require internet intermediaries operating within their jurisdiction to respect human rights, including the legislation on hate speech, to apply the principles of human rights due diligence

2. The Council of Europe is formed by 46 Member States, this includes the 27 Member States of the EU.
3. In its Recommendation No. R (97)20 of the Committee of Ministers on "Hate Speech" of 30 October 1997, it defined "hate speech" as "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."
4. Recommendation CM/Rec(2022)16 of the Committee of Ministers to member States on combating hate speech of 20 May 2022.

throughout their operations and policies, and to take measures in line with existing frameworks and procedures to combat hate speech; ensure that mechanisms are in place for the reporting of cases of online hate speech to public authorities and private actors, including internet intermediaries, and clear rules for the processing of such reports. Removal procedures and conditions as well as related responsibilities and liability rules imposed on internet intermediaries should be transparent, clear and predictable and those procedures should be subject to due process. They should guarantee users the right to an effective remedy delivered through transparent oversight and timely, accessible and fair appeal mechanisms, which are subject to independent judicial review.

In addition, States should: consider the substantial differences in the size, nature, function and organisational structure of internet intermediaries when devising, interpreting and applying the legislative framework governing the liability of internet intermediaries to prevent a possible disproportionate impact on smaller internet intermediaries; establish by law that internet intermediaries must take effective measures to fulfil their duties and responsibilities not to make accessible or disseminate hate speech; have a system in place for the disclosure of subscriber information in cases where competent authorities have assessed that online hate speech is in breach of the law and authors and disseminators are unknown to the competent authorities; ensure that any disclosure of available information on their identity is in line with European and international human rights law; regularly publish reports containing comprehensive information and statistics on online hate speech, including content restrictions, and on State authorities' requests to platforms to take down content on the grounds that it is hate speech, subject to the protection of personal data in accordance with European and international standards; establish by law that relevant internet intermediaries are under an obligation to regularly produce and publish transparency reports showing disaggregated and comprehensive data on hate speech cases and content restrictions; and ensure that independent authorities, in co-operation with internet intermediaries, civil

society organisations and other stakeholders, regularly assess and improve the content moderation systems in place to improve the detection, reporting and processing of online hate speech, while eliminating the causes of unjustified content restriction and over-compliance.

Whilst this Recommendation is addressed to States, it also contains guidance for internet intermediaries. They should: identify expressions of hate speech that are disseminated through their systems; ensure that human rights law and standards guide their content moderation policies and practices about hate speech, explicitly state that in their terms of service and ensure the greatest possible transparency regarding those policies, including the mechanisms and criteria for content moderation; carefully calibrate their responses to content identified as hate speech based on its severity and elaborate and apply alternatives to the removal of content in less severe cases of hate speech; make all necessary efforts to ensure that the use of automation or artificial intelligence tools is overseen by human moderation and that content moderation considers the specificities of relevant legal, local, cultural, socio-political and historical contexts. In their efforts to take specificities into account, they should consider decentralising content moderation; appoint enough content moderators and ensure that they are impartial, have adequate expertise, are regularly trained and receive appropriate psychological support.

Internet intermediaries should furthermore ensure that trusted flaggers and fact-checkers are trained in human rights standards that apply to hate speech; establish effective co-operation with civil society organisations that work on hate speech, including on the collection and analysis of data, and support their efforts to improve policies, practices and campaigns to address hate speech; review their online advertising systems and the use of micro-targeting, content amplification and recommendation systems and the underlying data-collection strategies to ensure that they do not, directly or indirectly, promote or incentivise the dissemination of hate speech; and develop internal processes that enable them to detect and prevent risks to human rights regarding the assessment and treatment of hate speech and

should subject themselves to regular independent, comprehensive and effective human rights impact assessments and audits.

In addition, the European Commission against Racism and Intolerance (ECRI), in 2015, issued its General Policy Recommendation no. 15 on combating hate speech. Its section 7 recommended to use regulatory powers with respect to Internet providers, intermediaries and social media to promote action to combat the use of hate speech and to challenge its acceptability, while ensuring that such action does not violate the right to freedom of expression and opinion. Accordingly, they should: ensure effective use is made of any existing powers suitable for this purpose, while not disregarding self-regulatory mechanisms; encourage the adoption and use of appropriate codes of conduct and/or conditions of use with respect to hate speech, as well as of effective reporting channels; encourage the monitoring and condemnation of the use and dissemination of hate speech; encourage the use, if necessary, of content restrictions, word filtering bots and other such techniques; encourage appropriate training for moderators as to the nature of hate speech; and promote and assist the establishment of complaints mechanisms.

Section 8 recommended that States determine the particular responsibilities of authors of hate speech, internet service providers, web fora and hosts, online intermediaries, social media platforms, online intermediaries, moderators of blogs and others performing similar roles; ensure the availability of a power, subject to judicial authorisation or approval, to: require the deletion of hate speech from web-accessible material and to block sites using hate speech; require media publishers (including internet providers, online intermediaries and social media platforms) to publish an acknowledgement that something they published constituted hate speech; and enjoin the dissemination of hate speech and to compel the disclosure of the identity of those using it

Section 10 recommended to take appropriate and effective action against the use, in a public context, of hate speech which is intended or can reasonably

be expected to incite acts of violence, intimidation, hostility or discrimination against those targeted by it using the criminal law provided that no other, less restrictive, measure would be effective and the right to freedom of expression and opinion is respected. Accordingly, they must ensure that the scope of these offences is defined in a manner that permits their application to keep pace with technological developments; and cooperate with other States in tackling the transfrontier dissemination of hate speech, including in electronic format.

The Protocol to the Cybercrime Convention

The Council of Europe Convention on Cybercrime, known as the Budapest Convention, was adopted in 2001, becoming the first international treaty to fight crimes committed via the Internet. Due to resistance by the United States because of the First Amendment, its final text did not tackle hate-speech online. Conversely, in 2003, an Additional Protocol to the Convention was adopted concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems. The Protocol entails an extension of the Cybercrime Convention's scope, including its substantive, procedural and international cooperation provisions. Thus, apart from harmonising the substantive law elements, the Protocol allows the Parties to use the international cooperation tools of the Convention in this area. The Protocol has not been as successful as the Convention as it has only been ratified by 35 States, while the Convention received 68 ratifications (more than double).[5]

McGonagle (2013) underlined that the Protocol concerned primarily criminal-law measures against online hate speech and this express focus left little room for exploring civil-law and other (non-legal) remedies and responses. Banks (2010) argued that, whilst the Protocol is a laudatory endeavour, it was limited in its ability to bring together real differences in the ways in which States envisage hate speech and construct a legal framework

5. As of 28/07/2023.

through which hate based conduct may be counteracted. Gstrein (2019: 84) also commented that the practical influence of the Protocol on the discussion on hate speech or disinformation in the digital sphere seemed limited.

European Court of Human Rights case-law

In addition to the standard-setting efforts of the Council of Europe, attention should be paid to the case-law of the European Convention on Human Rights (ECHR) that oversees the application of the ECHR by its Parties which are the 46 Member States of the Council of Europe. Article 10 of the ECHR establishes that everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers.

The ECtHR has clearly stated that "freedom of expression constitutes one of the essential foundations of [a democratic] society" and "it is applicable not only to 'information' or 'ideas' that are favourably received or regarded as inoffensive or as a matter of indifference, but also to those that offend, shock or disturb the State or any sector of the population." (ECtHR Judgment of 7 December 1976, *Handyside v. the United Kingdom*, App. no. 5493/7, § 49). On the other hand, the ECtHR has also argued that "tolerance and respect for the equal dignity of all human beings constitute the foundations of a democratic, pluralistic society. That being so, as a matter of principle it may be considered necessary in certain democratic societies to sanction or even prevent all forms of expression which spread, incite, promote or justify hatred based on intolerance" (ECtHR Judgment of 6 July 2006, *Erbakan v. Turkey*, App. No. 59405/00, § 56).

The question is how the ECtHR can reconcile the protection of freedom of expression with the necessary fight against hate speech. The ECtHR has followed, depending on the case, two different approaches: to apply the exclusion from the protection of the ECHR, provided for by Article 17 (prohibition of abuse of rights), considering that hate speech negates the

fundamental values of the ECHR and therefore is not protected by it; or to apply the restrictions on protection, that the very Article 10 provides in its second paragraph, that the limits of freedom of expression should always be necessary in a democratic society and follow a legitimate objective such as in the case of hate speech, the prevention of crime or the protection of the rights of others.

Regarding online hate speech, the ECtHR has had multiple occasions to decide in recent years in cases dealing with what can be considered hate speech and liability of Internet intermediaries. As to what may constitute hate speech, the ECtHR decides on a case-by-case basis, but it has underlined some factors that national courts have to consider. Three examples are explained to show how the Court approaches this issue: *Smajić*[6], *Kilin*[7] and *Savva Terentyev*[8]. In *Smajić,* the ECtHR declared the applicant's complaint inadmissible as being manifestly ill-founded; in *Kilin,* it found no violation of Article 10, while in *Savva Terentyev* it did.

In *Smajić,* the case concerned the applicant's conviction for incitement to racial and religious hatred, discord or intolerance following several posts on an Internet forum describing military action which could be undertaken against Serb villages in the Brčko District in the event of another war. The applicant alleged that he had been convicted for expressing his opinion on a matter of public concern. However, the ECtHR found that the domestic courts had examined the applicant's case with care, giving sufficient justification for his conviction, namely that he had used highly insulting expressions towards Serbs, thus touching upon the extremely sensitive matter of ethnic relations in post-conflict Bosnian society. Furthermore, the penalties imposed (a suspended sentence and the seizure of a computer and a laptop) had not been excessive. Therefore, the interference with the

6. ECtHR Decision on inadmissibility of 8 February 2018, *Smajić v. Bosnia and Herzegovina*, App. No. 48657/16.
7. ECtHR Judgment of 11 May 2021, *Kilin v. Russia*, App. No. 10271/12.
8. ECtHR Judgment of 28 of August 2018, *Savva Terentyev v. Russia*, App. No. 10692/09.

applicant's right to freedom of expression, which had pursued the legitimate aim of protecting the reputation and rights of others, did not violate Article 10.

Kilin concerned the applicant's trial and conviction for disseminating extremist materials. The applicant had been accused of posting allegedly racist video and audio files involving neonazis, racial epithets and calls to extremism on a popular online social network. The ECtHR found that the domestic courts had convincingly demonstrated that the impugned material had incited ethnic discord and, foremost, the applicant's clear intention of bringing about the commission of related acts of hatred or intolerance. Moreover, while there was no indication that the material had been published against a sensitive social or political background, or that at the time the general security situation in Russia had been tense, those elements were not decisive in the case. Lastly, the nature and severity of the penalties imposed (a suspended eighteen-month term of imprisonment with a similar period of probation) had been proportionate. The difference between the previous cases described and *Savva Terentyev* is apparent. This case concerned the applicant's conviction for inciting hatred after making insulting remarks about police officers in a comment under a blog post. The Court held that there had been a violation of Article 10, because, while the applicant's language had been offensive and shocking, that alone was not enough to justify interfering with his right to freedom of expression. The domestic courts should have looked at the overall context of his comments, which had been a provocative attempt to express his anger at what he perceived to be police interference, rather than an actual call to physical violence against the police.

The first case in which the Court was called upon to examine a complaint about liability of an intermediary for user-generated comments on the Internet was the *Delfi* case[9]. This is a news portal run on a commercial basis, that had been held liable by the national courts for the offensive comments

9. ECtHR Judgment of 16 June 2015, *Delfi AS v. Estonia*, App. No. 64569/09.

posted by its readers below one of its online news articles about a ferry company. At the request of the lawyers of the ferry company, Delfi removed the offensive comments. Delfi acted immediately when it was notified but this was six weeks after the publication. The Estonian Courts concluded the ferry company personality rights had been violated and awarded 320 € in compensation for non-pecuniary damage. Delfi considered that its freedom of expression had been violated and appealed to the ECtHR. The ECtHR held that there had been no violation of Article 10. It noted the conflicting realities between the benefits of Internet, notably the unprecedented platform it provided for freedom of expression, and its dangers, namely the possibility of hate speech and speech inciting violence being disseminated worldwide in a matter of seconds and sometimes remaining persistently available online.

The ECtHR observed that the unlawful nature of the comments in question was clear since most of them were tantamount to an incitement to hatred or to violence against the owner of the ferry company. Consequently, the case concerned the duties and responsibilities of Internet news portals, under Article 10.2 ECHR, which provided on a commercial basis a platform for user-generated comments on previously published content and some users – whether identified or anonymous – engaged in clearly unlawful speech, which infringed the personality rights of others and amounted to hate speech and incitement to violence against them. Where third-party user comments are in the form of hate speech and direct threats to the physical integrity of individuals, the ECtHR considered that the rights and interests of others and of society as a whole may entitle States to impose liability on Internet news portals, without contravening Article 10, if they fail to take measures to remove clearly unlawful comments without delay, even without notice from the alleged victim or from third parties.

Based on the concrete assessment of these aspects and taking into account, in particular, the extreme nature of the comments in question, the fact that they had been posted in reaction to an article published by the applicant company on its professionally managed news portal run on a commercial

basis, the insufficiency of the measures taken by the applicant company to remove without delay after publication comments amounting to hate speech and speech inciting violence and to ensure a realistic prospect of the authors of such comments being held liable, and the moderate sanction imposed on the applicant company, the ECtHR found that the Estonian courts' finding of liability against the applicant company had been a justified and proportionate restriction on the portal's freedom of expression.

However, the ECtHR did not weight some factors in its assessment such as: the measures put in place by Delfi, namely the system of notice-and-take-down (never used by the owner of the ferry company) or the automatic filtering system; the liability exemptions of the E-Commerce Directive (explained below); or the effect not on the freedom of expression of Delfi but its readers and commenters.

The ECtHR emphasized that the case related to a large professionally managed Internet news portal run on a commercial basis which published news articles of its own and invited its readers to comment on them. Accordingly, the case did not concern other fora on the Internet where third-party comments can be disseminated, for example, social media where the platform provider does not offer any content and where the content provider may be a private person running the website or blog as a hobby.

Voorhoof (2015) argued that there were severe doubts if this limitation of the impact of the judgment holding an online forum liable for user generated comments is a pertinent one, reserving the (traditional) high level of freedom of expression and information only for social media, personal blogs and "hobby". According to the Grand Chamber its judgment is not to be understood as imposing a form of "private censorship". However, the judgment considers interferences and removal taken on initiative of the providers of online platforms as the necessary way to protecting the rights of others, while there are other ways that can achieve the same goal, but with less overbroad (pre-)monitoring of all user generated content or with less collateral damage for freedom of expression and information, such as

taking action against the content providers, an effectively install obligations for providers to help to identify the (anonymous) content providers in case of manifest hate speech or other illegal content. Keller (2015) considered that obliging online platforms to monitor users' comments to prevent any liability for illegal content created a new paradigm for participatory online media.

Maroni (2020) underlined that Delfi may lead to the arbitral removal of content by intermediaries to avoid liability. She argued that delegating what counts as freedom of expression to private companies risked reducing freedom of expression to a "technicality", thus neglecting the normative complexity of freedom of expression with consequences for the Internet as a pluralistic environment. Brunner (2016) stated that if the ECtHR aimed to develop rules to reduce the spread of illegal content online, the Delfi judgment was not convincing as a first step.

The ECtHR had the possibility to clarify the impact of this case in it subsequent case-law in *MTE and Index.hu*[10] (MTE), a case that concerned the liability of a self-regulatory body of Internet content providers and an Internet news portal for vulgar and offensive online comments posted on their websites following the publication of an opinion criticising the misleading business practices of two real estate websites. The applicants complained about the Hungarian courts' rulings against them, which had effectively obliged them to moderate the contents of comments made by readers on their websites, arguing that that had gone against the essence of free expression on the Internet.

The ECtHR held that there had been a violation of Article 10. It reiterated that Internet news portals had to assume duties and responsibilities, even if they were not publishers of the comments in the traditional sense. However, the ECtHR considered that the Hungarian courts, when deciding on the notion of liability in the applicants' case, had not carried out a proper balancing exercise between the competing rights involved, namely between

10. ECtHR Judgment of 2 February 2016, *Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt* v. Hungary, App. No. 22947/13.

the applicants' right to freedom of expression and the real estate websites' right to respect for its commercial reputation. Notably, the Hungarian authorities accepted at face value that the comments had been unlawful as being injurious to the reputation of the real estate websites. The ECtHR underlined that, even if some similarities could be drawn with the *Delfi* case, the comments in the present case, although offensive and vulgar, had not constituted clearly unlawful speech. Furthermore, while Index was the owner of a large media outlet which must be regarded as having economic interests, MTE was a non-profit self-regulatory association of Internet service providers, with no known such interests.

The ECtHR stated that by establishing objective liability on the side of the Internet websites, merely for allowing unfiltered comments that might be in breach of the law, would require 'excessive and impracticable forethought capable of undermining freedom of the right to impart information on the Internet'.

More than in Delfi, the ECtHR in MTE considered the negative consequences of holding Internet portals liable for third-party comments (Voorhoof, 2020), clarifying that 'such liability may have foreseeable negative consequences on the comment environment of an Internet portal, for example by impelling it to close the commenting space altogether. For the Court, these consequences may have, directly or indirectly, a chilling effect on the freedom of expression on the Internet'. Keller (2016) considered the MTE ruling as a huge step forward on a policy level. The ECtHR explicitly recognized that regulating expression and information platforms meant regulating their users' expression and information access. The ruling's core insight was that "intermediary liability" laws directly affected the rights of ordinary Internet users and could make or break their ability to speak and find information online.

However, for Voorhoof and Lievens (2016), although the ECtHR tried to reduce the problematic consequences of the approach chosen in Delfi, the judgment in MTE nevertheless reiterates the endorsement of the system

of notice-and-take-down by private online platforms deciding on the lawfulness of content. This approach risked putting the ECtHR in an isolated position, as in some jurisdictions intermediaries can only be found liable for "unlawful" content when they have failed to act following notice from a judge, a court or another independent body as to the illegality of the relevant content. Intermediary service providers are less well-placed than courts to consider the lawfulness of comments on their website domains. Especially qualifying speech as hate speech is a difficult and delicate exercise, not only for domestic courts, but also for the ECtHR. Moreover, decisions by online platforms currently lack transparency and their decision-making contains few or no procedural guarantees for those whose right to freedom of expression is interfered with.

The problem of *MTE* is that the intermediaries had to guess if the comments will be hate speech or not even if it is completely out of their control and up to the authors of the comments, because the consequences will be completely different.

Finally, the *Sanchez* case[11] concerns the criminal conviction of a local councillor who was standing for election, for incitement to hatred or violence against a group of people or an individual on the grounds of their membership of a specific religion. He was not convicted for its own publication on Facebook but for his failure to take prompt action in deleting comments posted by two other identified persons (also convicted) on his Facebook wall, as French courts considered him as a "producer". The ECtHR found no violation of Article 10.

The court underlined that the Internet is one of the principal means by which individuals exercise their right to freedom of expression, and interferences with the exercise of that right had to be examined particularly carefully, since they are likely to have a chilling effect, which carries a risk of self-censorship. Nevertheless, the identification of such a risk must not

11. ECtHR Judgment of 15 May 2023, *Sanchez* v. *France,* App. No. 45581/15.

obscure the existence of other dangers for the exercise and enjoyment of fundamental rights and freedoms. For this reason, the possibility for individuals complaining of defamatory or other types of unlawful speech to bring an action to establish liability must, in principle, be maintained, constituting an effective remedy for alleged violations.

In opinion of the ECtHR, while professional entities which created social networks and make them available to other users necessarily had certain obligations, there should be a sharing of liability between all the actors involved, allowing if necessary for the degree of liability and the manner of its attribution to be graduated according to the objective situation of each one. French law was consistent with such a view, providing in the case of the "producer" for a shared liability, subject to safeguards on implementation, while in the case of hosts liability remained limited. Moreover, the domestic courts had referred to the applicant's status as a politician and inferred from this that a special obligation was incumbent upon him; he could be expected to be even more vigilant.

Kerkhof (2023) has criticised the decision because it dramatically expands the range of people and entities that need to worry about being held liable as an internet intermediary, to the point where this could extend to everyone with a social media presence and the responsibilities falling on those potential internet intermediaries require significant legal prowess and resources, which are unattainable for most private individuals. Cotino (2023) considers that the ECtHR reinforces in a very worrying direction the Delfi doctrine of holding intermediaries responsible for content integrated by third parties. He is also concerned because this is applied to the criminal sphere and to politicians in electoral campaigns, that is, precisely where freedom of expression should be most intense. In his opinion, the ECtHR admits a very unclear regulation and jurisprudence and exaggeratedly delegates the solution of these internal issues to States and their judges.

Measures to fight hate speech by the European Union

In the EU, the main tool to fight hate speech was the Council Framework Decision 2008/913/JHA on combating certain forms and expressions of racism and xenophobia by means of criminal law. Its objective was to ensure that certain serious manifestations of racism and xenophobia were punishable by effective, proportionate and dissuasive criminal penalties throughout the EU and to improve and encourage judicial cooperation in this field. The Decision was not focused on on-line hate speech, nevertheless, it clarifies that each Member State shall take the necessary measures to ensure that its jurisdiction extends to cases where the offender commits the conduct when physically present in its territory, whether or not the conduct involves material hosted on an information system in the EU; or the conduct involves material hosted on an information system in its territory, whether or not the offender commits the conduct when physically present in the EU.

The E-Commerce Directive

The E-Commerce Directive was adopted in 2000 and its objective was to contribute to the proper functioning of the internal market by guaranteeing the free circulation of information society services between the Member States. The Directive established that the States would guarantee that the mere intermediaries could not be considered liable for the stored or transmitted data whose content had been created and shared by third parties, provided that a series of requirements were met, such as that they did not have actual knowledge of illegal activity or information and, and if they obtained such knowledge or awareness, they acted at once to remove or to disable access to the information. The Directive also prohibited States from imposing on service providers a general obligation to monitor the data they transmit or store, or a general obligation to actively search for facts or circumstances that indicate illegal activities.

As the Directive was adopted more than twenty years ago, it became outdated regarding the categories of intermediaries it included. In this time, the

content created by users and shared on the Internet has multiplied and new services of significant importance have emerged, such as cloud storage providers, social networks or marketplaces, who in many cases are not mere channels of information created by third parties, but rather order it, make recommendations, etc. However, the basic principles of the Directive are still relevant today and should be maintained, because they are a guarantee of the EU's capacity for innovation and freedom of expression. Although its operation in practice needed to be clarified or even modulated in certain cases, in addition to adding new obligations to intermediaries in aspects such as due diligence or transparency.

The Audiovisual Media Services Directive

The reform of the Audivisual Media Services Directive was adopted in 2018 to respond to the convergence between television and Internet services, so its scope would be extended to also create obligations with respect to the latter. As a result, States shall ensure that video sharing platform services under their jurisdiction take appropriate measures to protect the general public from programmes, user-generated videos and audiovisual commercial communications containing incitement to violence or hatred directed against a group of persons or a member of a group based on any of the grounds referred to in Article 21 of the Charter of Fundamental Rights of the EU and offences concerning racism and xenophobia as set out in the Framework Decision 2008/913/JHA.

Kuklis (2018) considered that there was clearly no other way to police the content moderation activities, considering the sheer scale of the operations in question, than to co-regulate the environment with platforms themselves and the Directive could create an environment where users are not only protected from the harmful content of other users, but also from overbearing or arbitrary intrusions by the platform itself. Nevertheless, according to Barata (2018), the Directive introduced a dramatic change in the way audiovisual content was regulated and monitored. Private online intermediaries could develop, interpret and enforce content rules affecting the core

elements of the right to freedom of expression within the society of each European member State. Platforms would play a fundamental role in determining the boundaries of legitimate political speech or the right to adopt and express unconventional social and cultural points of view. This role would be played under the threat of sanctions if platforms under-regulated or failed to act against dubiously legal content. For Barata, these provisions created all the incentives for a solidly State supported, privately executed, overregulation of speech. This leads to unacceptable consequences for the exercise of the right to freedom of expression in our plural and democratic European political systems.

The Code of Conduct for combating illegal hate speech on the Internet

The European Commission faced a dilemma in these matters, because there is no Article in the Treaties (TEU and TFEU) that could explicitly serve as a basis for regulating freedom of expression, although it was clear that the States and public opinion demanded that it took the initiative to fight hate speech online. The answer was to opt for non-legislative measures that were the result of dialogue and co-regulation of intermediaries. These were pushed to take measures to avoid legislative developments that would make them responsible for this type of content. Angelopoulos (2016) argued that initiatives of this nature seek to put the weight of the fight against illegal content on intermediaries, which can be counterproductive, because they do not have the same obligations to establish guarantees for the protection of fundamental rights as States.

The Commission agreed with Facebook, Microsoft, Twitter and YouTube in 2016 on a Code of Conduct for combating illegal hate speech on the Internet. In 2018, Instagram, Google+, Snapchat and Dailymotion also signed up, Jeuxvideo.com did so in 2019, TikTok in 2020, Linkedin in 2021, and Rakuten, Viber and Twitch in 2022. Althought Twitter abandoned it in 2023.

The purpose of the Code was for intermediaries to act expeditiously against illegal hate speech online, upon receipt of valid notification and within an appropriate time-frame. The companies assumed, among others, the following

commitments: to have clear and effective procedures to examine notifications related to illegal incitement to hatred that occur within the framework of the services they provide, so that they can withdraw or disable access to said content; and review most valid hate speech takedown notices within 24 hours, and remove or disable, if necessary, access to such content.

The IT Companies agreed, among others, in the following commitments: to have in place clear and effective processes to review notifications regarding illegal hate speech on their services so they can remove or disable access to such content and Rules or Community Guidelines clarifying that they prohibit the promotion of incitement to violence and hateful conduct; upon receipt of a valid removal notification, to review such requests against their rules and community guidelines and where necessary national laws transposing the Framework Decision 2008/913/JHA, with dedicated teams reviewing requests; to review the majority of valid notifications for removal of illegal hate speech in less than 24 hours and remove or disable access to such content, if necessary; to educate and raise awareness with their users about the types of content not permitted under their rules and community guidelines; to provide information on the procedures for submitting notices, with a view to improving the speed and effectiveness of communication between the Member State authorities and the IT Companies; and to encourage the provision of notices and flagging of content that promotes incitement to violence and hateful conduct at scale by experts (trusted flaggers).[12]

Alkiviadou (2019) argued that the Code served as a light at the end of the Internet hate tunnel where issues of multiple jurisdictions as well as technological realities have resulted in the task of online regulation being more than a daunting one. Podstawa (2019) shed light on the emerging role of the Internet Service Providers (ISPs) who control the virtual environment

12. Other relevant commitments were: to provide regular training to their staff on current societal developments and to exchange views on the potential for further improvement; to intensify cooperation between themselves and other platforms and social media companies to enhance best practice sharing; to work in identifying and promoting independent counter-narratives, new ideas and initiatives and supporting educational programs that encourage critical thinking; to intensify their work with civil society organizations to deliver best practice training on countering hateful rhetoric and prejudice and increase the scale of their proactive outreach to civil society organizations to help them deliver effective counter speech campaigns.

where illegal behaviours may occur. Whilst not responsible for what, *prima facie*, is published, it is argued that the ISPs are an essential element in the enforcement of hate speech criminal rules, as confirmed by the Code. This governance instrument exemplified in his opinion the essentiality of the ISPs collaboration with the traditional enforcement agents in ensuring the blocking and removal of content online, as well as in subsequent criminal proceedings. At the same time, by involving the representatives of the broader community in monitoring the implementation of the Code, the mixed hybrid governance and enforcement model offers a possible (even if imperfect) solution to the current deadlock in the regulation of Internet governance.

Bukovská (2019:10), considering the problematic legal basis and unclear process of implementation of the Code, defined it as a misguided policy on the part of the European Commission. For companies, it was likely to amount to no more than a public relations exercise, but, despite its nonbinding character, the Code could lead to more censorship by private companies (and thus undermine the rule of law) and create a chilling effect on freedom of expression on the platforms they run. For Bukovská, there were several legal questions and implications for freedom of expression under the Code, such as the delegation of responsibility for determining what is "unlawful hate speech," vague and overbroad criteria, lack of due process, and redress mechanisms for violations of the right to freedom of expression. Portaru (2017) underlined that the Code was likely to have a significant impact on free speech in the digital area by putting companies into the role of free speech regulators. According to her, any limitation on speech rights needs to be firmly grounded in the law; it cannot be based on voluntary terms of service or codes of conduct with arbitrary implementation and lacking legal redress, and the regulation of speech should not be outsourced to private and unaccountable actors.

Kuczerawy (2016) argued that the Code represented a hybrid situation, as any interference with freedom of expression resulting from its implementation cannot be attributed directly to the European Commission (as the

restrictions will be administered by the IT companies). Nevertheless, the Commission's role was more than that of a facilitator. In her opinion, by inviting private companies to restrict speech of individuals the Commission became an initiator of the interference with a fundamental right by private individuals – a type of "state interference by proxy". It was disputable whether an EU initiative which stimulates private companies to restrict freedom of expression of individuals without providing any safeguards for that right would stand scrutiny under the Charter of Fundamental Rights of the EU. If the EU wants to enlist private entities, for the purpose of efficiency, to do (at least part of) the job, they should ensure that the "arrangement" is equipped with appropriate safeguards for freedom of expression.

Another question is if the commitments of the Code are followed and if it is giving results. For it to be effective, the European Commission decided to follow it with periodical monitoring rounds.[13] There was also a progress report for the period 2016-2019[14] that assessed the progress achieved and underlined that the Code contributed to achieve quick progress, including in particular on the swift review and removal of hate speech content (28% of content removed in 2016 versus 72% in 2019; 40% of notices reviewed within 24 hours in 2016 versus 89% in 2019).[15]

The Digital Services Act

The Code of conduct was considered insufficient by several EU countries that decided to legislate the matter. The most prominent examples are the German Network Enforcement Act (N*etzwerkdurchsetzungsgesetz* or

13. See European Commission, Factsheet of the 6th evaluation of the Code of Conduct of 7 October 2021. Available at: https://ec.europa.eu/info/sites/default/files/factsheet-6th-monitoring-round-of-the-code-of-conduct_october2021_en_1.pdf
14. European Commission, Assessment of the Code of Conduct on Hate Speech on line. State of Play. 27 September 2019. Doc. 12522/19. Available at: https://ec.europa.eu/info/sites/default/files/aid_development_cooperation_fundamental_rights/assessment_of_the_code_of_conduct_on_hate_speech_on_line_-_state_of_play__0.pdf
15. The removal rate is now stable at more than 70% on average. It is important to understand that here the aim is not a 100%, because as the report clarifies the current average removal rate can be considered as satisfactory in an area such as hate speech, given that the line against speech that is protected by the right to freedom of expression is not always easy to draw and is highly dependent on the context in which the content was placed and users can wrongly flag content as hate speech that is perfectly legal.

NetzDG) which came into force in 2017 although the reporting obligation started to apply in 2018[16] and the French Act to Combat Hateful Content on the Internet (*Loi visant à lutter contre les contenus haineux sur internet* or *Loi Avia*) of 2020. The French Act gave the platforms a period of 24 hours to remove reported content that was manifestly illegal, in relation to a series of crimes (hate speech, terrorism, etc.). The European Commission was highly critical of the draft,[17] considering that the obligation of the platforms to eliminate any illegal content notified within a period of 24 hours, combined with the heavy sanctions provided, the wide variety of crimes subject to said obligation (which may require a more or less in-depth contextual assessment) and reduced notification requirements, could have dire consequences. The Commission argued that this could create a disproportionate burden on platforms and a risk of excessive content removal, undermining freedom of expression.

The Act was partially annulled by the French Constitutional Council.[18] In particular, the Council annulled subsection I of Article 1 of the Act, which provided that the administrative authority could request the platform or editors of an online communication service to eliminate certain content regarding child pornography or terrorism and if they did not do so, they could be subject to a penalty of one year in prison and a fine of 250,000 €. The limited period of time and the fact that the order came from an administrative authority that was not a judicial authority led the Council to consider that the Act violated freedom of expression and communication because it

16. The Law was mainly aimed at combating hate speech, but also applied to a whole series of categories of content considered illegal by German law. The priority target of the Law was the large social media platforms with more than 2 million users located in Germany. The Act required these platforms to provide a mechanism for users to submit complaints about illegal content. The procedure had to be easily recognizable, directly accessible and permanently available. Once they received a complaint, the platforms had to investigate if the content was illegal, if it was "manifestly illegal" they had to remove it within 24 hours, the rest had to be removed within 7 days, in a similar line to the Code of conduct. It included some safeguards such as that platforms must immediately notify the complainant and the user of any decision related to the removal or blocking of content, including the reasons for the decision. The Act also imposed transparency requirements, if a platform received more than 100 complaints per year, it had to publish semi-annual reports detailing its content moderation practices, although these have proven not to be very informative.
17. European Commission, *Notification 2019/412/F Loi visant à lutter contre les contenus haineux sur internet - Emission d'observations prévues à l'article 5, paragraphe 2, de la directive (UE) 2015/1535 du 9 septembre 2015*, C (2019) 8585 final, pp. 7 y 8.
18. Conseil constitutionnel, *Décision n° 2020-801 DC du 18 juin 2020.*

was not adequate, necessary and proportionate to the objective pursued; and, therefore, that provision was unconstitutional. It also declared unconstitutional the obligation of the platforms to remove manifestly illegal content within 24 hours after any person's complaint, because it was up to the intermediary to examine the reported content with respect to a long list of crimes, even though the constituent elements of some of them could be legally complex and require an assessment of the context of the content, for which 24 hours was excessively short. Buri (2020) argued that this should be a lesson for the following discussion of standards at EU level.

Regarding initiatives such as the *NetzDG* or the *Loi Avia*, de Streel and Husovec (2020:31) considered that the adoption of these laws increased the risks of Internal Market fragmentation. This was one of the reasons why a general updated European standard for intermediary obligations was required.

The European Commission also considered that new and innovative information society services have emerged, changing the daily lives of EU citizens and shaping and transforming how they communicate, connect, consume and do business; and the use of those services had also become the source of new risks and challenges, both for society as a whole and individuals using such services,[19] such as the spread of hate-speech at scale. The Commission first used sector-specific instruments to fill the regulatory gaps such as the Code of conduct mentioned, but also to protect copyright, fight terrorist content or disinformation. However, these sector-specific instruments were not enough, and the Commission was under constant pressure to do more. Therefore, it proposed the Regulation on a Single Market for Digital Services, also known as the Digital Services Act or DSA, adopted in 2022. The DSA imposes the obligation to act against illegal content. However, the illegal nature of such content, products or services is

19. European Commission, Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC, COM/2020/825 final.

not defined in it but results from EU or national law, as is the case with hate speech. The European Commission sought to regulate intermediaries, not content.

The DSA keeps the main elements of the E-Commerce Directive and regulates the conditions under which intermediaries are exempt from liability for the information of third parties that they transmit and store (principle of "safe harbor"), but includes the clarification that exemptions should not be disapplied when providers of intermediary services carry out voluntary own-initiative investigations or comply with the law ("good Samaritan" exception). Establishing an obligation of general supervision or active investigation for intermediaries continues to be prohibited. There are new obligations such as, in respect of orders from national judicial or administrative authorities, to act against illegal content and to provide information.

The DSA also sets asymmetric due diligence obligations on distinct types of digital service providers depending on the nature of their services and their size, to avoid the creation of disproportionate burdens. It starts with the basic obligations applicable to all intermediaries,[20] and then it creates new layers of obligations for hosting providers,[21] online platforms,[22] and very large online platforms and search engines.[23]

20. Section 1 lays down those obligations: to establish a single point of contact; to designate a legal representative in the EU; to set out in their terms and conditions any restrictions that they may impose on the use of their services and to act responsibly in applying and enforcing those restrictions; and transparency reporting in relation to the removal and the disabling of information considered to be illegal content or contrary to the providers' terms and conditions.
21. Section 2 lays down obligations for them to put in place mechanisms to allow third parties to notify the presence of alleged illegal content. Furthermore, if they decide to remove or disable access to specific information provided by a recipient of the service, it imposes the obligation to provide that recipient with a statement of reasons.
22. Section 3 lays down obligations applicable to all online platforms (but not to micro or small enterprises): to provide an internal complaint-handling system in respect of decisions taken in relation to alleged illegal content or information incompatible with their terms and conditions; to engage with certified out-of-court dispute settlement bodies to resolve any dispute with users of their services; to ensure that notices submitted by entities granted the status of trusted flaggers are treated with priority and sets out the measures online platforms are to adopt against misuse; to inform competent enforcement authorities in the event they become aware of any information giving rise to a suspicion of serious criminal offences involving a threat to the life or safety of persons; and to publish reports on their activities relating to the removal and the disabling of information considered to be illegal content or contrary to their terms and conditions.
23. Section 5 lays down obligations to manage systemic risks. Three categories of systemic risks should be assessed in-depth and the first category concerns the risks associated with the misuse of their service through the dissemination of illegal content, namely hate speech. Very large online platforms are obliged: to conduct risk assessments on the systemic risks brought about by or relating

The creation of different obligations depending on the size and functions of the intermediaries makes full sense as they are widely different. However, even considering this approach, Keller (2016) argued that it is almost certain that smaller platforms will have difficulty shouldering the DSA's burdens and, in pursuit of legitimate content moderation goals, the DSA may inadvertently sacrifice competition goals, and foreclose future diversity in platform practices and speech rules.

The DSA underlines that, while the Code of conduct sets a benchmark for the participating companies with respect to the time needed to process valid notifications for removal of illegal hate speech (24 hours), other types of illegal content may take different timelines for processing, depending on the specific facts and circumstances and types of illegal content at stake. Regarding very large online platforms and online search engines, they shall put in place reasonable, proportionate and effective mitigation measures, with consideration to the impacts of such measures on fundamental rights. Such measures may involve adapting content moderation processes, including the speed and quality of processing notices related to specific types of illegal content and, where appropriate, the expeditious removal of, or the disabling of access to, the content notified, in particular in respect of illegal hate speech or cyber violence, as well as adapting any relevant decision-making processes and dedicated resources for content moderation.

Even if the DSA does not specifically apply to hate speech given its general scope, it is clear that is going to have a big impact in how social media tackle hate speech on line as it has the potential of becoming a kind of GDPR for content moderation. There may be a Brussels effect (Bradford, 2020) and many companies may generally apply these rules even outside the EU.

to the functioning and use of their services; to take reasonable and effective measures aimed at mitigating those risks and to submit themselves to external and independent audits.

Conclusions

The Council of Europe and the EU have developed a great deal of standards to fight hate speech, including soft and hard-law. Their objective is to help their Member States to fight online hate speech in a way respectful with human rights which sometimes is a challenge, including how to tackle the hate speech disseminated through social media.

States should be mindful of the recommendations of the Committee of Ministers of the Council of Europe and ECRI that present a useful roadmap to put in place laws and measures that help them create a suitable legal framework. National laws such as the *Loi Avia* should be avoided, because they are a quick undercut that conflicts with constitutional rights. Sometimes it is not so clear what may constitute a fringe case of hate-speech and 24 hours deadlines could not be enough, as not all intermediaries are Google, Facebook or Twitter. However, platforms that have a high revenue and a business model fed by third-party generated comments should channel the proper resources to moderate comments and act expeditiously when notified about manifestly illegal content.

Social media platforms must fulfil their obligations, act transparently, respect human rights and apply the necessary due diligence. Nevertheless, the principles of the E-Commerce Directive cannot be quickly discarded. Legislation that creates overblocking incentives should be avoided and that may be the result of enhancing the liability of intermediaries. Laws should respect the fact that offensive or shocking comments are protected by freedom of expression.

It is important to monitor the results of both the Code of Conduct of the EU and the DSA to see how they affect hate speech online and if they are the right measures. The DSA brings a sizeable number of procedural improvements that could be mimicked in other parts of the world with the necessary adjustments.

References

Alkiviadou, N. (2019). Hate speech on social media networks: Towards a regulatory framework?. *Information & Communications Technology Law,* 28, pp. 19-35.

Angelopoulos, C. *et al.* (2016). *Study of fundamental rights limitations for online enforcement through selfregulation.* Institute for Information Law (IViR).

Banks, J. (2010). Regulating hate speech online. International Review of Law. *Computers and Technology, 24*(3), 233-239.

Barata J. (2018). The new Audiovisual Media Services Directive: Turning video hosting platforms into private media regulatory bodies. *Center for Internet and Society Blog*, 24/10/2018. https://cyberlaw.stanford.edu/blog/2018/10/new-audiovisual-media-servicesdirective-turning-video-hosting-platforms-private-media.

Brunner, L. (2016). The liability of an online intermediary for third party content: The watchdog becomes the monitor: intermediary liability after Delfi v Estonia. *Human Rights Law Review*, *16*(1), 163-174.

Bukovská, B. (2019). The European Commission's Code of Conduct for Countering Illegal Hate Speech Online: An analysis of freedom of expression implications. *Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression Series. https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/06/EC_Code_of_Conduct_TWG_Bukovska_May_2019.pdf*

Buri, I. (2020). The lesson of the French Constitutional Council on the fight against hate speech and the latest on the upcoming Digital Services Act. *CiTiP Blog*. https://www.law.kuleuven.be/citip/blog/the-lesson-of-the-french-constitutional-council-on-the-fight-against-hate-speech-and-the-latest-on-the-upcoming-digital-services-act/

Cotino Hueso, L. (2023). Menos libertad de expresión en internet: el peligroso endurecimiento del TEDH sobre la responsabilidad de moderación de contenidos y discurso del odio, Derecho Digital e Innovación. Digital Law and Innovation Review 16.

De Streel, A. & Husovec, M. (2020). *The e-commerce Directive as the cornerstone of the Internal Market: Assessment and options for reform*. European Parliament, Luxembourg.

Gstrein, O. J. (2019). The Council of Europe as an Actor in the Digital Age: Past achievements, future perspectives. *Festschrift der Mitarbeiter* Innen und Doktorand* Innen zum*, *60*, 77-90.

Keller, D. (2016). New intermediary liability from the Court of Human Rights: What will they mean in the real world?. *Inforrmr's Blog*. 19/0/2016 https://inforrm.org/2016/04/19/new-intermediary-liability-from-the-court-of-human-rights-what-will-they-mean-in-the-real-world-daphne-keller/#more-33797

Keller, D. (2022). The DSA's industrial model for content moderation. *Verfassungsblog: On Matters Constitutional*. https://verfassungsblog.de/dsa-industrial-model/

Kerkhof, J. (2023) Sanchez v France: The expansion of intermediary liability in the context of online hate speech. *Strasbourg Observers,* 17/07/2023 https://strasbourgobservers.com/2023/07/17/sanchez-v-france-the-expansion-of-intermediary-liability-in-the-context-of-online-hate-speech/

Kuczerawy, A. (2016). The Code of Conduct on Online Hate Speech: An example of state interference by proxy? KU Leuven. *CiTiP Blog,* 20/07/2016 https://www.law.kuleuven.be/citip/blog/the-code-of-conduct-on-online-hate-speech-an-example-of-state-interference-by-proxy/

Kuklis, L. (2018). European regulation of video-sharing platforms: What's new, and will it work?. *Media Policy Project Blog LSE,* 29/11/2018. https://blogs.lse.ac.uk/medialse/2018/11/29/european-regulation-of-video-sharing-platforms-whats-new-and-will-it-work/

Maroni, M. (2020). Chapter 11: The liability of internet intermediaries and the European Court of Human Rights. *Fundamental Rights Protection Online* (pp. 255–278). Elgar.

McGonagle, T. (2013). *The Council of Europe against online hate speech: Conundrums and challenges*. Expert Paper, MCM(2013)005. https://rm.coe.int/168059bfce

Podstawa, K. (2019). Hybrid governance or... nothing? The EU Code of Conduct on combatting illegal hate speech online. In *Use and Misuse of New Technologies* (pp. 167-184). Springer.

Portaru, A. (2017). Freedom of expression online: The code of conduct on countering illegal hate speech online. *RRDE*, 77.

Voorhoof, D. (2015). Delfi AS v. Estonia: Grand Chamber confirms liability of online news portal for offensive comments posted by its readers. *Strasbourg Observers*, 18/06/2015. https://strasbourgobservers.com/2015/06/18/delfi-as-v-estonia-grand-chamber-confirms-liability-of-online-news-portal-for-offensive-comments-posted-by-its-readers/#more-2891

Voorhoof, D. (2020). Blog Symposium "Strasbourg Observers turns ten" (2): The Court's subtle approach of online media platforms' liability for user-generated content since the 'Delfi Oracle'. *Strasbourg Observers*, 10/04/2020. https://strasbourgobservers.com/2020/04/10/the-courts-subtle-approach-of-online-media-platforms-liability-for-user-generated-content-since-the-delfi-oracle/

Voorhoof, D. & Lievens, E. (2016). Offensive online comments - New ECtHR Judgment. *ECHR Blog*, 15/02/2016. https://www.echrblog.com/2016/02/offensive-online-comments-new-ecthr.html

HATE POSTINGS ON SOCIAL MEDIA AND PEACE IMPERATIVES IN NIGERIA

Nosa Owens-Ibie
/ Caleb University, Nigeria

Eric Msughter Aondover
/ Caleb University, Nigeria

Introduction

Social media platforms have changed how quickly and easily people communicate across social and geographical boundaries. Millions of people worldwide can now quickly access any digital content, compared to the past when gatekeepers controlled and negotiated access to the mass media platforms (Appel et al., 2020). Demuyakor and Doe (2021) point out that this particular development has increased the impact and harm associated with misinformation, disinformation, and malinformation related to hate speech, in addition to improving opportunities for citizens' freedom of expression and diversity. Aondover et al. (2023) state that globally, regulators are exploring practical options to these new socio-legal challenges social media present.

In the post-truth era, social media are perceived as the fourth state (Chiluwa & Samoilenko, 2019). Social media technologies are deemed to be not just the essence but also the backbone of 21st-century democracy. Chiluwa *et al.*, (2020) observed that social media have been reflecting the happenings around the globe, starting from individuals to the entire community. Demuyakor and Doe (2021) note that if effectively monitored, social

media platforms could be helpful in not only seeking or pursuing the truth, but also reporting stories as they are, rather than changing facts or coming up with a story to suit the interest of one particular group. In this regard, social media ought to report the truth, desist from forms of hate speech; offer a voice or help to the voiceless; reflect diversity as well as improve critical judgment. At individual levels, social media is argued to gratify users' self-fulfilment as they reflect what users stand for and believe in.

Digital platforms are quickly replacing traditional communication channels on a global scale. It is incredibly difficult to establish a set of cyber norms that are globally recognized due to the variety of cultural, political, and social norms that users around the world adhere to. The prevalence of internet anonymity and the rise in predatory and harmful behaviour add to the complexity of this situation. Users have the potential to unintentionally hurt other users through hate speech, fraudulent reviews, offensive messages, and other methods (Chakraborty & Masud, 2022). With hate speech trending in dialogue and disagreements in the public domain, terrestrial or virtual, this aspect of the media ecology has unintended repercussions. Such an emphasis advances the bigger conversation on media material that represents reality (Owens-Ibie, 2019).

However, the trending nature of hate speech on social media in Nigeria and elsewhere in the world is alarming. The kind of hateful information people post on social media is undermining the collective peaceful co-existence of individuals, as people or groups. This also underscores the position of Owens-Ibie (2019) that the role of the media to either mirror or shape society, has an underlying logic. This position is anchored on the circular relationship that captures both elements in any instance and process. The social media as mirror argument, acknowledges that content is based on reality and subsisting cultural values and social trends, rather than products of the imagination. Similarly, social media content is hardly a perfect or neutral reflection of such mirrored reality. It is evident that there is a selectivity process informed by content creators, news sources, and other

media producers and gatekeepers in the information chain who often present content their own way.

This suggests that information and communication spaces are dynamic and are structured using elements that are always changing. Audiences and individuals who have had, and continue to have, an impact on the political and other settings are at the intersection of this dynamic. This setting explains why concerns and trends relating to hate speech have received more attention on social media. This chapter is a condensed collection of opinions on the problems, recent developments, and discontent surrounding hate speech on social media. It discusses aspects of hate speech on social media, from the more general to the more specific.

Context

Free expression, which is protected under Article 19 of the Universal Declaration of Human Rights, has not only been adopted as a global template, but has come to define Nigeria's democratic process. Demuyakor and Doe (2021), citing the United Nations, state that every individual has the right to freedom of expression, which includes the freedom to hold opinions without interference and to seek, acquire, and disseminate knowledge through any media, regardless of boundaries. Social media and the right to free speech are both thought to be crucial for advancing the democratic ethos. Aondover *et al.,* (2022) noted that throughout the 20th century, the most popular method of safeguarding individual freedom or right of expression, was utilizing judicial formation and protection of legal as well as constitutional rights.

For example, many democratic societies have added a clause against the use of hate speech in guarantees on freedom of speech. For instance, Article 10(2) of the European Convention on Human Rights (ECHR) provides that "the exercise of freedom of expression may be subject to such formalities, conditions, restrictions or penalties as prescribed by law, the interest of national security for the protection of the reputation or right of others." Most

doctrines that established freedom of speech and expression in Nigeria added a clause to guard against hate speech, and promote human dignity, societal cohesion and peace. Section 39(1) of the 1999 Constitution as amended in 2011, provides that "every person shall be entitled to freedom of expression." Similarly, section 45 provides that nothing in section 39 shall invalidate any law that is reasonably justifiable in a democratic society in the interest of public order, public morality and to protect the rights and freedom of other persons.

In exploring the diminishing role of facts and analysis in American public life, Kavanagh and Rich (cited in Owens-Ibie (2019) described the questions of reliability and rejection of factual information in contemporary society as indicators of "truth decay." Three of the four trends they identified in this "'decay" are manifest in technology-moderated engagements, and include: a blurring of the line between opinion and facts; increasing relative volume, and resulting influence, of opinion and personal experience over fact; and declining trust in formerly respected sources of factual information. These trends have been influenced majorly by the interface between human (with sensibilities) and technology (without sensibilities). Today, misinformation, disinformation and malinformation have spread largely through the empowerment of technological platforms like social networks and through social messaging, which in their manifestations raise questions on the extent of regulation and self-regulation of companies providing these services.

Recent changes in the media eco-system have given rise to new challenges in the media landscape that journalists, academicians, technology companies and experts are confronted with and still haven't quite resolved. One of the most prominent of such challenges is hate speech (Aondover, 2022). Recent experiences continue to show both the negative and positive usage of the social media in Nigeria and globally (Kurfi *et al.,* 2021). An anti-social media bill introduced by the Nigerian Senate on November 5 2019, sought to criminalize the use of social media in peddling false or malicious information, but was discontinued due to public criticisms. The bill prescribing death by hanging for any person found guilty of any form of hate speech

sponsored by the Senate spokesperson, Sabi Abdullahi, sought the establishment of an Independent National Commission for Hate Speeches, but was resisted by civil society groups in Nigeria.

Attempts at regulation of social media in the last few decades have made it an even more attractive option in social interactions, surpassing contributions of other innovations in the history of mass media. This flows from the foundational logic that information is necessary for effective modern social, economic, and political development. Social media have made people, being the sources, the processors as well as the end users of all information. Its powerful networks; and its speed of transmission have impacted people (Mojaye & Aondover, 2022). Information is power, information is the engine room for meaningful and sustainable development and information is also the catalyst for effective social interaction (Owens-Ibie, 2016). Therefore, the biggest benefit accruable from the use of social media is their facilitation of information flow, communication and the inherent freedom of expression.

Social media have in the process democratized and personalized accusation, aided by the powerful networks of internet communication. However, the growing menace constituted by hate speech has informed the attempts at legislative remediation. Stakeholders at a national summit on hate speech observed that social media have contributed to the spread of hate speech and writings, through the spread of gory pictures of either false or imagined wars (Ogbuoshi, *et al*., 2019). Therefore, hate speech is a serious issue that has characterized the social media era and if not checked, is capable of promoting conflict and division rather than the required cohesion necessary for the country's collective development.

Between hate speech legislation, free speech and press freedom

Freedom of expression is a natural right individuals enjoy, which is enshrined in the Constitution, local legislation and international human rights law. The antecedence of free speech dates back to the ancient Greece when the debate was about whether persons other than male landowners should

be allowed to speak in public (Oriola, 2019). The advent of mass communication through the invention of printing also attracted repression of expression through licensing laws. However, the rise of democracy has promoted international debate and facilitated legislations and conventions on free speech and press freedom.

Some of the output from the global response to the logics of human rights, are: the Universal Declaration of Human Rights (UDRH), Article 19; International Covenant on Civil and Political Rights (ICCPR); and Article 9, African Charter on People's and Human Rights (ACPHR). These documents are unequivocal on the right of every person to seek, receive and share any kind of information in any form without any hindrance. Mihajlova *et al.,* in Oriola (2019) identify the classes of information as political, artistic and commercial and the forms of communication as oral, written, artistic and any other media including new technologies. Section 39(1) of the 1999 Constitution of the Federal Republic of Nigeria as amended, provides that "every person shall be entitled to freedom to hold opinions and receive and impart ideas and information without interference."

Press freedom and free speech can therefore, be seen as two aspects of the same concept because they both derive from the basic freedom to gather, process, and transmit information without hindrance. The freedom of the media to source and publish, information as well as protect their sources, is guaranteed by international human rights legislation, the Nigerian constitution, and other legal frameworks.

The Nigerian 1999 Constitution in section 39(2) states: "without prejudice to the generality of subsection (1) of this section, every person shall be entitled to own, establish and operate any medium for the dissemination of information, ideas and opinions." It also stipulates conditions for ownership of television or wireless broadcasting stations.

Since there is no absolute freedom anywhere and the fact that a society is defined by a web of social interactions, it is evident that in the course of the

exercise of a right by an individual or a group, an infringement on the rights of others may occur. This consideration and that of national security are adduced for limitation to human rights, including the right to freedom of expression and of the press. McQuail (2010) observed that such regulations and control lead to censorship restraints or limits on publication. One of such measures with local and international advocates is the restriction on hate speech (Oriola, 2019).

Within this framework, the need to promote equality and discourage discrimination is provided in Articles 1, 2, and 7 of the UDHR and Articles 2(1), Article 20(2) and 26 of the ICCPR. For instance, Article 20(2) of the ICCPR places the obligation on States to legally prohibit "any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence." This is an express international human rights requirement for the prohibition of hate speech. The Nigerian Constitution does not expressly provide restrictions on hate speech but in section 39(3), states that nothing in this section shall invalidate any law that is reasonably justifiable in a democratic society. This means there is a limit to freedom of expression and hate speech is not freedom of expression.

Effects of hate speech

Public dialogue on hate speech assumed some prominence during the administration of former President, Muhammadu Buhari. For example, Adelakun (2017) notes that the 2015 Nigerian election campaign season saw a rise in hate speech as a result of the extensive use of divisive political advertising by political actors. Since the 2015 election campaign, hate speech as a concept has been argued to have established roots in Nigeria (Ezeibe, 2015). The All Progressive Congress (APC) presidential candidate Muhammadu Buhari was the target of a documentary that was shown on Africa Independent Television (AIT), over which the APC petitioned the National Broadcasting Commission as a "hate broadcast," requesting that AIT should be sanctioned.

According to Adelakun (2017), there is need however, for a legal and administrative conceptualization of hate speech for the purpose of setting boundaries so that simple social media comments, as well as mainstream media messages that constitute insults, slander, libel, comedy, propaganda, criticism, and legitimate protests against government policies, are not mistakenly construed as hate speech. Nigeria is multi-ethnic, multilingual, and culturally sensitive, and languages have contextual meanings and interpretations, which may have inspired the commercial. This implies the need to contextualise social and political messages as an expression that is acceptable in one culture may be viewed as insulting in another. But this is one perspective.

In line with US Legal (2016), violence is a potential consequence of hate speech. Hate speech aids in recruiting new members for an organization. According to Brown (2017: 420), the consequences of hate speech include "harm, dignity, security, healthy cultural dialogue, democracy, and legitimacy." Both individual and collective victims suffer injury. Adelakun (2017) agrees with Brown on the detrimental effects of hate speech, stating that it argues for, supports, promotes, or incites hatred against a certain group of people known for something. It makes its victims emotionally and psychologically uncomfortable, reduces social and economic mobility by fostering inequality, causes drug and alcohol abuse, and may even result in hate crimes, which can have serious repercussions for societal peace, order, and security. He also asserts that hate speech undermines the democratic concept of the free exchange of ideas since it encourages social and political exclusion of specific people and groups that Waltman and Ashely (2017) label as "out-groups."

Hate speech includes any statements, actions, gestures, writings, or displays that have the potential to instigate violence or carry out discriminatory deeds. In essence, these speeches deprive other people of their respectability and sense of order (Mrabure, 2015).

Analysis of hate postings

Okunna's (2018) position that hate speech appears to be largely associated with the ruling class and trends on social media, while not exclusively pointing in the direction of the political leadership cadre, tend to provide some evidence to corroborate this. Some examples of controversial posts on the social media handle of leaders or attributed to them are as follows:

> "The North would make the country ungovernable if President Goodluck Jonathan wins the 2011 polls. Anything short of a Northern President is tantamount to stealing our presidency" (A comment credited to a former governor of Kaduna state (Nasir El-Rufia) in 2010 on social media) (Hate speech comment on Facebook, 2022).
>
> "The Igbos are collectively unlettered, uncouth, uncultured, unrestrained and crude in all their ways. Money and the acquisition of wealth is their sole objective and purpose in life" (Femi Fani-Kayode, a former Aviation Minister, 2013). (Hate speech comment on Facebook, 2013).
>
> "They are senseless and idiots. But I will not blame them because most of them (Igbos) don't have parents. They are being produced by baby factory."(Hate speech comment on Facebook, 2022).
>
> "If Arewa youths and all the Hausa-Fulani cows don't keep to their October 1st threat, they will forever remain fools and cowards." (Hate speech comment on WhatsApp, 2022).
>
> "Bunch of illiterate almajiris wanting to lead the literates." (Hate speech comment on Facebook, 2023).
>
> "Idiot president, Fulani herdsmen that go about fully armed killing and destroying villages are not senseless to you, but armless freedom fighters are senseless...Foolish man." (Hate speech comment on Facebook, 2023).

"Buhari would likely die in office if elected, recall that Murtala Muhammed, Sani Abacha and Umaru Yar'Adua, *all former heads of state from the Northwest, had died in office"* – The Governor of Ekiti State, Peter Ayodele Fayose, (*ThisDay*, January 19, 2015, and other social media platform like Facebook).

"Wetin him dey find again? Him dey drag with him pikin mate, old man wey no get brain, him brain don die patapata – what *else is he (Buhari) after, contesting with people young enough to be his children. The old man who lacks gumption; he is completely brain dead*." – Former First Lady, Patience Jonathan, at a PDP political party rally in Kogi State (*The Express News,* 4th March, 2014 as well as social media platforms like Facebook and WhatsApp).

"God willing, by 2015 something will happen. They either conduct a free and fair election or they go a very disgraceful way. If what happened in 2011 should again happen in 2015, by the grace of God, the dog and the baboon would all be soaked in blood." – Presidential Candidate of Congress for Progressive Change, General Muhammadu Buhari (Lika Binniyat in *Vanguard Newspaper* on May, 15th 2012, which also appears on social media like Facebook).

The highlight in italics is for emphasis, and indicative of the depth of negative passion and resentment for individuals, and extended to their group, and trivialises death. It is important to note how these comments went viral. Since communication on social media trigger multiplier effects and volatilities, these postings which are attributed to political leaders and anonymous social media activists, had influences which may have been consequential. Their significance derives not only from the difficulty in measuring their impacts, but that they featured during and around the period of elections, politics and national controversies. They raise issues of emotional intelligence and levels of understanding of the consequences of communication and the use especially of easily accessible and largely unregulated social media. As Okunna (2018) has noted, without the social media, such hate

speech could fail to come alive. That means that the media are instrumental to the spread of hate speech.

The Hate Speech Bill, which was proposed by the Nigerian Senate prescribed death by hanging for any person, found guilty of any form of hate speech that results in the death of another person. The 'Hate Speech Bill' was sponsored by Senator Aliyu Sabi Abdullahi. The Bill seeks to "eliminate" hate speech and discourage harassment on the grounds of ethnicity, religion, or race among others. It prescribes stiff penalties for offences such as "ethnic hatred." "Any person, who uses, publishes, presents, produces, plays, provides, distributes, or directs the performance of any material, written or visual, which is threatening, abusive or insulting or involves the use of threatening, abusive or insulting words, commits an offence"(Okunna, 2018).

The bill prescribed a penalty for those who are found guilty of any form of hate speech that results in the death of another person after judicial processes in a Federal High Court. According to critics of the bill, "critical sections of the society like the mass media, civil society, pressure groups, the academia, writers, and creative or performing artists who expectedly will bear the main brunt of the obnoxious law have been curiously and dangerously indifferent, as only a few voices have raised the alarm." The critics of this penalty thought that death penalty should have been provided as the punishment when someone or a group is responsible for the deaths of other people. In an editorial on March 19, 2018, *The Punch* criticized the bill in the following words:

> Although promoting or inciting hatred is wrong from all angles, this bill is undesirable because it is being presented by those who are unaccustomed to and uncomfortable with the procedures and intricacies of democracy and fundamental rights. President Muhammadu Buhari never suggested such severe jail terms and penalties against the use of free speech and media freedom during his first term as a military head of state. The notorious anti-media Decree 4, which made headlines dur-

> ing the military coup he oversaw in 1984–1985, fell far short of levying fines amounting to millions of naira or establishing the death penalty. Furthermore, the British colonial rulers who created and implemented a series of anti-press and sedition laws did not consider using the death sentence to stifle free speech (p. 45).

The Newspaper pointed out that the Bill conflicts with the 1999 Constitution Chapter IV's provisions on fundamental rights, particularly Sections 38 and 39, which protect freedom of expression and the press, respectively, and freedom of thought, conscience, and religion. The fundamental law and supporting legislation sufficiently forbid the abuse of these rights and discrimination against individuals or groups on the basis of race, ethnicity, or religion. According to Shehu Sani, a former Nigerian Senator, the law outlawing hate speech would be used against free speech, and Nigerians should resist and reject it, as the bill would be used as a form of intimidation (*Sahara Reporters*, 2015).

Implications for peaceful coexistence

The social media boom is one of the most amazing inventions of the twenty-first century. According to Dauda et al. (2017), it has altered how people communicate, comprehend, and respond to social events in general and conflicts in particular. It is necessary to comprehend how hate speech affects the peaceful coexistence of different ethnic groups, and identify other fault lines in the country, given Nigeria's ethnic and religious identities and plurality. The political elites, ethnic groupings, and religious organizations have exploited these identities, and engineered movements for power and resource control by social groups who were previously repressed (Danaan, 2017), although social media can also be utilized to further development goals in a multicultural community (Kurfi *et al.*, 2021).

The trend has however been towards violent conflicts, whether electoral, communal or ethno-religious, as end product of the spread of hate speech through social media (Dauda et al., 2017). According to a report by the

Centre for Information Technology and Development (CITAD) in Aondover (2022), there has been a rise in hate speeches among Nigerians on various social media platforms. The report indicated that 60.3 per cent of hate speeches recorded came from Facebook, 5.9 per cent from newsletters and 4 per cent from blogs surveyed; 63 per cent of perpetrators of hate speeches are prominent people while 39 per cent of them are ordinary citizens (non-prominent); 35.2 per cent of the hate speeches insult people for their religion, abuse people for their ethnic or linguistic affiliation, or express contempt against people because of their place of origin (Aondover *et al.*, 2023). Findings of a study by Ende and Dzukogi (2012) indicate that verbal terror attacks directed at individuals, ethnic groups, religious institutions and regions, as stereotypes were used to describe those involved. Comments deemed as offensive employed hate speech, threats, abusive language and assassination of character.

Due to lax rules, the problem of hate speech has grown significantly in Nigeria, and apparently throughout Africa. As the use of hate speech remains largely unchecked, animosity amongst the ethnic groups that make up Nigeria has grown (Ezeibe, 2015). There are still attempts in the public space to distinguish between constructive criticisms and hate speech. The necessity to control the spread of divisive and inflammatory comments through social media is obviously unavoidable given the country's ethnic and religious fault lines. The Nigerian government has tried to respond through regulatory agencies like the NBC, initiatives through the legislature, and media statements and other pronouncements.

In August 2017, Nigeria's former Vice-President Yemi Osinbajo declared that "hate speech will no longer be tolerated, as the country's leaders' silence on this issue would be a grave disservice to the nation, its peace, and its future." We have established a clear line against hate speech; it will not be permitted, will be viewed as terrorism, and will be met with all appropriate sanctions (Aondover, 2022). The National Orientation Agency (NOA) started a campaign on social media with the hashtag "say no to hate speech", stating that, "In the last few months, our country and its people have witnessed

a disturbing trend in social and political conversations that sometimes call into question our traditional friendship, love for one another, and respect for authority," the Director General of NOA noted" (*Vanguard*, 2017).

Efforts by the federal government to address issues of hate speech have reignited debates among political and civil society organizations on what exactly qualifies as hate speech. While some political and public affairs experts suspect the motives of government as likely to infringe on people' rights (especially the right to free speech and expression), others concede the need for legislation to control hate speech. "Any law capable of hindering the freedom of expression granted under Section 39 of the 1999 constitution and the African Charter, would be illegal and unconstitutional," As Ebun-Olu Adegboruwa, a human rights attorney in Lagos, elaborated in a statement to *Premium Times,* "this is only an attempt by the APC-led administration to intimidate citizens" (Ezeamalu, 2017).

A former governor of Ekiti State, Ayodele Fayose, viewed the decision to categorize hate speech as an act of terrorism, as a scheme by the APC administration to intimidate the PDP: "This appears to be another plot to silence the opposition,", "and I make bold to say that, saying the truth concerning the country and its rulers cannot be termed as hate speech" (Aondover et al., 2022). The former governor of Rivers State, Nyesom Wike, referred to the action as a threat that was only intended to terrorize PDP opponents "...I don't know what they refer to as hate speech." I'm not sure if we should all remain silent when something is wrong" (Gogo, 2017).

Conclusion

This chapter has discussed the concerns and patterns of hate speech on social media in Nigeria. While social media provide platforms for variable expressions, concerns about its unmanageable or largely unregulated mechanisms have led to debates on the need to interrogate their uses and abuses, especially given its rising profile as a major belt for the transmission of hate speech. Such direct and inadvertent promotion of hate speech

has contributed to instabilities and violence deemed as inimical to the development goals of Nigeria. It is necessary to stop the trend of utilizing social media to instigate crises. The need to promote peaceful coexistence is imperative. While the spread of hate speech on social media poses severe challenges to peaceful coexistence, its positive potential to advance peaceful coexistence will go unrealized unless excesses associated with users who appear unable to strike a balance between their right to and freedom of expression, and the demands of emotional intelligence and conflict sensitivity, are appropriately addressed. There is need for stakeholders' consensus to enable the alignment of law and policy with reality in a way which does not attempt to prey on access or validate opportunism.

References

Adelakun, A. (2017). But, what exactly is hate speech? *Punchnicom*. https://punchng.com/but-what-exactly-is-hate-speech/#google_vignette

Aondover, E. M., Oyeleye, S. A., & Aliyu, M. A. (2023). New world information and communication order and its changing role in Nigerian Television Authority (NTA) Kano. *Unisia, 41*(1), 17-38.

Aondover, E. M. (2022). Interpretative phenomenological analysis of hate speech among editors of *Daily Trust*, *The Nation* and *The Guardian* newspapers in Nigeria. *Konfrontasi Journal: Culture, Economy and Social Changes*, *9*(2) 216-226.

Aondover, E. M. (2023). Social media narratives and reflections on hate speech in Nigeria. In: B. Di Fátima (Ed.), *Hate speech on social media: A global approach* (pp. 255-275). LabCom Books & EdiPUCE.

Aondover, P. O., Aondover, E. M., & Babele, A. M. (2022). Two nations, same technology, different outcomes: Analysis of technology application in Africa and America. *Journal of Educational Research and Review, 1*(1), 001-008. http://dx.doi.org/10.5281/zenodo.7488568

Appel, G., Grewal, L., Hadi, R., & Stephen, A. T. (2020). The future of social media in marketing. *Journal of the Academy of Marketing science*, *48*(1), 79-95.

Brown, A. (2017). What is hate speech? Part 1: The myth of hate. *Law and Philosophy*, *3*(6), 419-468. https://doi.org/10.1007/s10982-017-9297-1

Chakraborty, T. & Masud, S. (2022). Nipping in the bud: Detection, diffusion and mitigation of hate speech on social media. *ACM SIGWEB Newsletter*, *2022*(Winter), 1-9.

Chiluwa, I. E. & Samoilenko, S. A. (Eds.). (2019). *Handbook of research on deception, fake news, and misinformation online*. IGI Global.

Chiluwa, I., Taiwo, R., & Ajiboye, E. (2020). Hate speech and political media discourse in Nigeria: The case of the indigenous people of Biafra. *International Journal of Media & Cultural Politics*, *16*(2), 191–212. https://doi.org/10.1386/macp_00024_1

Danaan, G. N. (2017). Reporting diversity: Towards understanding Nigeria's ethnic and religious conflicts through the mediatisation theory. In: U. A. Pate & L. Oso (Ed), *Multiculturalism, diversity and reporting conflict in Nigeria* (pp. 75-95). Evans Brothers (Nigeria Publishers) Limited.

Dauda, S., Abubakar, A. A., & Lawan, A. K. (2017). Discursive devices, social media and conflict discourse in Nigeria. In: U. A. Pate & L. Oso (Ed.), *Multiculturalism, diversity and reporting conflict in Nigeria* (pp. 250-271). Evans Brothers (Nigeria Publishers) Limited.

Demuyakor, J. & Doe, V. A. (2022). Social media, democracy, and freedom of expression: Some evidence from Ghana. *International Journal of Political Science and Governance, 4*(1), 87-94.

Ende, S. T. & Dzukogi, A. A. (2012). Verbal terror and Nigerian online news readers comment. *The Nigerian Journal of Communication*, *10*(1), 61-76.

Ezeamalu, B. (2017, August 19). There's no 'hate speech' under Nigerian law-Lawyer. *Premium Times.* https://www.premiumtimesng.com/news/more-news/240822-theres-no-hatespeech-under-Nigerian-law

Ezeibe, C. (2015). *Hate speech and electoral violence in Nigeria*. Conference paper submitted to the Department of Political Science University of Nigeria Nsukka. http://www.inecnigeria.org/wpcontent/uploads/2015/07/Conference-Paper-by-ChristianEzeibe.pdf

Gogo, J. (2017, September 21). Hate speech: I will not be intimidated by fed govt-Wike. *Today.NG*. https://www.today.ng/news/nigeria/15745/hatespeech- intimidated-fed-g.

Kurfi, M. Y., Aondover, E. M., & Mohammed., I. (2021). Digital images on social media and proliferation of fake news on Covid-19 in Kano, Nigeria. *Galactica Media: Journal of Media Studies*, *1*(1), 103-124. https://doi.org/10.46539/gmd.v3i1.111

McQuail, D. (2010). The future of communication studies: A contribution to the debate. *Media and Communication Studies Interventions and Intersections*, *2*(7). 1-15.

Mojaye, E. M. & Aondover, E. M. (2022). Theoretical perspectives in world information systems: A propositional appraisal of new media-communication imperatives. *Journal of Communication and Media Research*, *14*(1), 100-106.

Mrabure, K. O. (2015). *Counteracting hate speech and the right to freedom of expression in selected jurisdictions*. https://www.ajol.info/index.php/naujilj/issue/view/13988

Ogbuoshi, L. I., Oyeleke, A. S., & Folorunsho, O. M. (Eds). (2019). *Opinion leaders' perspectives on hate speech and fake news reporting and Nigeria's political stability. Fake news and hate speech: Narratives of political instability*. (6th Ed.). Canada University Press.

Okunna, S. (2018). Assessment of the use of different forms of tobacco products among Nigerian adults: Implications for tobacco control policy. *Tobacco Prevention & Cessation*. https://doi.org/10.18332/tpc/87126

Oriola, O. M. (Eds). (2019). *Mainstream media reporting of hate speech and press freedom in Nigeria politics. Fake news and hate speech: Narratives of political instability*. (6th Ed.). Canada University Press.

Owens-Ibie, N. (2003). The importance of cultural affairs in communication & writing in Africa. *Four Decades in the Study of Nigerian Languages and Linguistics: A Festschrift for Kay Williamson*, (1), 397.

Owens-Ibie, N. (2016). Conflicting communication in the communication of conflict: Chibok and narratives on media representation in taking stock: Nigerian media and national challenges. *ACSPN Book Series*, (1), 69-88.

Owens-Ibie, N. (Eds). (2019). Privacy, confidentiality and the uses and misuses of social media. In: N. Owens-Ibie, M. Oji, & J. Ogwezi. (2019), *Fake news and hate speech: narratives of political instability* (pp. 5-25). Canada University Press.

Sahara Reporters. (2015, January 19). Governor fayose places death-wish advert on buhari in national newspapers, p. 17.

The US Legal. (2016). Hate speech law and legal definition. *The US Legal*. https://definitions.uslegal.com/h/hate-speech/on13/8/18

Vanguard News Online. (2017, August 12). NOA on say no to hate speech on social media platforms. *Vanguard News Online*. https://www.vanguardngr.com/2017/08/noa-sayno-hate-speech-campaign.

Waltman, M. S. & Asely, A. M. (2017). Understanding hate speech. *Oxford Research Encyclopedia of Communication*, September, 1-28. https://doi.org/10.1093/acrefore/9780190228613.013.422

THE POLITICAL USE OF HATE SPEECH THROUGH SOCIAL MEDIA IN BRAZIL

Joelma Galvão de Lemos
/ Federal University of Sergipe, Brazil

Daniel Menezes Coelho
/ Federal University of Sergipe, Brazil

Introduction

Since it was first created, the Internet has brought with it the hope for increased access to information. Many of its developers believed that people could, through it, "free themselves from both the government and the big corporations" (Castells, 2003: 26). The Internet has been, since then, provoking many changes in areas such as work organization, economic development, and access to information. Besides all of that, it has also been promoting a new dynamic in the way people interact socially, as well as new forms of social and political organization.

As examples of political mobilizations that came to be, or were operated, mainly via the internet, we can highlight a number of different ones around the world: "the online protests against the events of Tiananmen Square in China, in 1989, via computer networks operated by Chinese students abroad" (Castells, 1999: 378); the Zapatist movement which, in the 90s, used the internet to disseminate their cause and share pictures of their mobilizations; movements carried out in Tunisia, Iceland, Spain, the United States, and, more specifically, in Brazil (Castells, 2013). This last one happened in 2013,

when thousands of young people took to the streets, initially to demand no increase in public transportation fares and the creation of a public transportation policy. However, those demands were soon followed by more general calls for social and economic changes in the country.

What all of those movements have in common is that they all used online social media as a tool for mobilizing and organizing political actions that would later be carried out in the streets. However, in Brazil's particular case, it is precisely within these mobilizations that we observed the expansion of two pre-existing movements in the country: the co-optation of the movement that took to the streets by the conservative right, and the proliferation of alternative media platforms, specially on the right-wing side of the political spectrum.

Analyzing data from Twitter about the 2013 movement, "only the right-wing side of things is shown", illustrating how this chapter in Brazil's history "was being disputed and was eventually won by conservative and right--wing movements" (Medina, as cited in JornalGGN, 2016, 2). According to Javier Medina (2016), this was the only social movement organized via social media that was co-opted by the conservatives, thus clashing with the other international mobilizations of the same kind that we have mentioned before.

This usage of social media as a means of political mobilization became very important in Brazil. In the 2018 elections, the elected candidate chose to carry out his campaign mostly through the internet. Let us look more closely at how that came to be.

Hate and hate speech

Before we keep on going, it is necessary to conceptually distinguish "hate", as a regular emotion (or affection), from "hate speech", as a specific compound of language and behavior.

To hate something is a common, if not universal, human experience. We all feel hate, even within loving relationships. It is normal for us to hate

our parents, our children, our companions, our friends, and we may feel it more the stronger the bond is (Freud, 1915/1996). We hate what is strange and different, but we also hate what is alike: if something is too similar to us, we might feel the need to differentiate ourselves from it – hence hating a neighboring country, for example, or a rival soccer team. Another interesting fact about hate is that when two people share hatred towards a third party it creates a bond between them. Those bonds become stronger the more increased is our ability to hate those who are not us (Freud, 1921/1996, 1930/1996).

According to the Freudian perspective, hate would be a correlative of the death drive, just as love would be a correlative of the life drive. That is to say that hate is a force within us, that constitutes us as much as the love force does. We are, after all, talking about the constitutional forces of the living beings (Freud, 1920/1996). Thus, we cannot and should not dream of a world without hate or death drive. This also means that there is no pacifism that is not a struggle, i.e., that is not moved by the very force of hatred that it fights against. Using an example taken from the Brazilian politics context, it is worth remembering the famous speech by Ulysses Guimarães during the promulgation of the 1988 Constitution, which re-democratized the country: "We hate and we are disgusted by the dictatorship. Hate and disgust".

However, despite the difficulties, we can insist on changing the final destination of human's aggressive impulses "to such a degree that they do not need to find expression in war" (Freud, 1933/1996: 205). They can find their expression elsewhere, for example, through dialogue: even if it is difficult and challenging, even if there is conflict. After all, a conflict in which all parties can manifest themselves is preferable to extremist positions such as wars and dictatorships, in which the other is not allowed to participate as anything other than an enemy.

This defines what we are here comprehending as being the regular, everyday mundane hate.

Hate speech, on the other hand, is an entirely different concept altogether. It designates a mechanism of political influence, which instrumentalizes the feeling of hatred and the desire for violence in order to gain power. According to Schäfer, Leivas & Santos (2015: 149-150).

> Hate speech is the expression of intolerant, prejudiced and discriminatory ideas against vulnerable individuals or groups, with the intention of offending their dignity and inciting hatred based on the following criteria: age; gender; sexual orientation; cultural identity; political opinion; social origin; socioeconomic status; educational level; status of migrant, refugee, repatriate, stateless person or internally displaced person; disability; genetic characteristics; physical or mental health status, including infectious; and disabling psychological conditions, or any other condition.

This form of discourse usually does not occur in the "context of a private conversation", but it is, instead, "addressed to a group, to a collective" (Carvalho, 2017: 51).

It should not be a big surprise that those who "vociferate hate" (Dias, 2012) end up finding in social media a very prolific space for the publication of this kind of speech. Social media is also a good place for them to find like-minded individuals in order to share their speeches, to perpetrate their online attacks and harassment.

Such behaviors can easily take place because the current way social media platforms are organized has contributed to the alienation of its users: by keeping them in ideological bubbles, these platforms end up blinding them from "the experiences of other groups that are being manipulated separately" (Lanier, 2018: 105). In other words, the "[...] version of the world you are seeing is invisible to other people, who will misunderstand it, and vice versa" (Lanier, 2018: 105). This mechanism creates an "online myopia", as "most people can only find time to see what is put in front of them by algorithmic feeds" (Lanier, 2018: 98). This reduces the possibility of people understanding each other, i.e., understanding different political positions or

social opinions (Lanier, 2018), as it limits how much of the other we can actually see. This online "myopia" enhances the proliferation of hate speech, aggression, and even stalking[1] on social media.

While organized in their respective bubbles, internet users may feel like there will be no consequences to their actions, since the individuals that are responsible for the sites and applications where hate speeches are posted do almost nothing to prevent them. Numerous reports have been made regarding the posting of hate speech on social media, as well as the way these platforms operate and the effects of content suggestions and navigation guidance (Lanier, 2018). However, those in charge typically take no action, as they financially benefit from the dissemination of this type of content, demonstrating that financial gain is of greater importance to these big companies than the wellbeing of the many people affected by this problem (Córdova, 2019; Dias, 2020; Zuboff, 2020).

A little bit about Brazil

It is also necessary to contextualize Brazil a bit more, so that our readers can better understand the following topics, as each country and each people has their own particularities.

Brazil is a former colonized country, and because of that it bears the marks of the barbarism, violence, prejudices and huge social and financial gap between classes that is proper to former colonies. The Brazilian people originates from the encounter between the Original peoples (i.e., the ones that were already here when the Europeans invaded the land), the European colonizers, and the multiple enslaved African people. This miscegenation is an important factor to the nation's very culture. Most of the elite, the majority of which descend from the white Europeans that conceived the country as just a land to be explored (Ribeiro, 1995), look down upon the rest of the population, as if they were still meant to be mere servants.

1. In April 2021, Brazil approved and enacted the Law 14,132/21, which criminalizes stalking, including the online type. This new law increased the penalty for internet stalking to three years of imprisonment (Brazilian Chamber of Deputies, 2022).

Due to these historical prejudices, the financial gap between upper and lower class endures to this day (Souza, 2017).

Adding to that, the upper class has always repressed popular political movements with the use of violence, just like almost all the governments in the country's history. This was done in order to keep the people away from political discussions and decisions, as well as to keep them as far away as possible from the country's wealth (its lands, for example). The so-called "communist threat", for one, which was used to justify the civil-military coup d'etat of 1964, was nothing more than a popular, tentative organization of agrarian reform and peasant leagues. However, the elite created heavy propaganda to sell this movement as an armed guerrilla, as if it were a violent national threat, in order to contain it.

Afterwards, the period of time that comprises the military dictatorship that followed the 1964 coup was filled with horrible practices of torture in basements designed for just that. It lasted until 1985, but the upper-class refusal regarding policies aimed at reducing social inequality persisted.

Between 2003 and 2016, during the government of the *Partido dos Trabalhadores* (Workers' Party, in a free literal translation), a number of efforts were made to implement several policies aiming at social equality. During this period, we could see the elite's strong refusal embodied in a big array of demonstrations of disdain: they would, for example, state that, now filled with poor people, the airports had turned into bus terminals. Another subject that they regarded with the utmost criticism was the social welfare program called *Bolsa Família* (Family Allowance).

The *Bolsa Família* program was aimed at reducing social inequality through income distribution. Each family received a set amount of money, according to social criteria, and then would have to commit to a number of responsibilities: keeping the children in school, following prenatal care in basic health units, in case of pregnancy, and providing medical attention to children up to one year of age. This program deeply bothered the upper-class, as we have already mentioned, and some of their criticisms were delusional:

that women would intentionally try and get pregnant to be entitled to *Bolsa Família*, for example, and that the program would encourage people to stop working and consequently turn Brazil into a nation of unemployed leeches.

The program's beneficiaries became special targets of hate speech in the 2014 presidential elections. During this time, tensions were running high and the political discussions were more and more heated. However, it is only following the outcome of this election, and the non-acceptance of defeat by the candidate of the Brazilian Social Democracy Party - PSDB, that we begin to see a greater presence of vociferations and insults on social media. In 2016, two years later, this hate speech became even more noticeable by the time the impeachment/removal request was being discussed. This was the episode, as it is well known, that paved the way for the 2016 coup.

The 2016 coup happened when the president at the time, Dilma Rousseff, was impeached over allegations that she had manipulated public accounts (particularly referred to, in Portuguese, as "*pedalada fiscal*"). However, in Brazil, that very same type of accounting maneuver had taken place since the government of Fernando Henrique Cardoso (1995-2003) (Amora, 2015). Yet, neither he nor his successor, Luiz Inácio Lula da Silva (2003 - 2010), had their accounts disapproved. And, even more importantly, none of them had to face impeachment requests.

In the case of Dilma, though, the request was accepted. The senators who approved the coup – openly declaring their votes in the name and in defense of Christian values, against communism and gender ideology –, are the same politicians who never questioned the exact same practice in previous governments and whom, two days after the coup, approved a law that made the use of supplementary credits more flexible without the authorization of the National Congress. That is: they legalized the very practice that deemed Dilma guilty (IG, 2016).

Ever since this turn of events, the tension between "coxinhas" (conservatives in favor of the impeachment) and "petralhas" (government defenders and left-wing people) was increasingly present on social media spaces. But

it was during the 2018 elections, two years later, that the Brazilian internet became visibly overrun by a massive use of hate speech, fake news, and automatic bots, orchestrated by the election campaign of one presidential candidate: Jair Messias Bolsonaro. Those latter devices were used as a means to artificially increase the perceived number of his supporters, even though he already had many people actually engaging with and believing in the publications that were mainly received through WhatsApp groups and kept circulating in these spaces (Soares, 2018).

The candidate kept his supporters in a made-up social imaginary, leading them to believe that they represented the good, that they were the defenders of moral values, all the while portraying other groups as evil, as people who should be defeated and eradicated. An example of how the candidate, and later elected president, thought about the leftists, is that at one of his rallies, in Rio Branco - Acre, Jair Messias Bolsonaro held a camera tripod as if it were a machine gun, and declared: "Let's shoot the petralhada[2]!" (YouTube, 2018).

Name calling and cursing among Internet users on Facebook in 2016[3]

Name calling, cursing, and quarreling are some of the ways people may behave in order to attack others around them, and thus to obtain some satisfaction out of their aggressive and hateful desires (Pereira, 2006). These tools are also used by Brazilian people when wanting to differentiate themselves from other Brazilians, with whom, thanks to that, they share a large number of traits. According to Freud (1921/1996: 112), people "closely related keep a certain distance from each other: South Germans cannot stand North Germans; the English slander the Scots in all possible ways; Spanish people despise Portuguese people", and so on and so forth. This attempt to keep a certain distance from the other that is all too similar for comfort is what Freud calls "narcissism of small differences".

2. A slang, used pejoratively, for voters of the Workers' Party or supporters of the left-wing in general.
3. This topic is part of the Dissertation "O uso político do discurso de ódio no Brasil: um estudo de caso no Facebook (2016 - 2017)", defended and approved in the Postgraduate Program in Social Psychology at the Federal University of Sergipe.

Similarly, this need humans have to separate themselves and keep a distance from the other that is seen as too similar, can be seen when Brazilians speak ill of their Argentine and Paraguayan neighbors, for example, as well as when some Brazilians from the Southeast speak ill of Northeasterners, offending them and belittling their political positions. For example, a time when this happened openly in social media was 2014, more specifically, the moment following the results for the presidential election, when it was revealed that the northeast was mostly in favor of the Workers' Party (Estadão, 2014). Many Southeasterners reacted very aggressively through social media, cursing and name calling the Northeasterners because of that.

Despite not having drastically changed the political and economic systems, the social changes that have taken place in Brazil during these recent years have brought about significant changes[4] for lower-class people: they not only have more access to consumer goods, but also to places that were priorly reserved for those with greater economic power, such as airports, shopping malls, and even universities. Because of that, lower-class people became more visible and more present, which, for the upper-class, was all the more uncomfortable. In other words, the Brazilian elite's difficulty in coexisting with its other, in the Freudian sense, became more and more evident (Dunker, 2015; Singer, 2016; Souza, 2016).

President Dilma Rousseff's impeachment in 2016 precisely illustrates the behavior of this segment of society: permeated by competitive individualism and the old discourse demanding order, security, and against corruption (Chauí, 2016; Costa, 1989; Dunker, 2015; Souza, 2016), as well as united by the feeling of hatred (Cleto, 2016) towards the Workers' Party and their defenders, they took to the streets to demand the impeachment. Supported by the traditional Brazilian media (Lopes, 2016), they were even alongside those who called for the return of the military dictatorship.

4. These social changes took place as a result of a more distributive national public policy, marked by an increase in the minimum wage (Bresser, 2012) and a greater number of people with access to consumption (Souza, 2016).

Therefore, these people are often acting aggressively, demonstrating their antipathy and aversion towards the other (Chauí, 2016; Dunker, 2015; Souza, 2016), in order to avoid the anguish that can be caused by critical thinking, by questioning, and maybe by realizing that the other is not as strange as one supposed.

In order to exemplify this, we will analyze some posts made by internet users[5], collected on Facebook, during a moment when political acts, both in favor and against the initiation of the impeachment proceedings in 2016, were taking place on the streets.

***Netizen 1**: Clowns, idiots, and imbeciles gathered in front of FIESP.*
***Netizen 2**: Look at 'family workers' defending their rights, walking happily, singing alongside FIESP[6] and the Employers' Union.[7]*
***Netizen 3**: That's better than the CUT and MST bums who burn tires and torment these same workers.* (Comments taken from the post in the video "Por volta das 8pm, manifestantes pró-impeachment se reuniram em frente à Fiesp [...]" - El País Brasil, 2016c).

The way many protesters, both in favor and against the impeachment, treated each other shows us once again their difficulties in dealing with small differences. That is because the more closed the group to which an individual belongs, the more the hatred will be directed towards external people (Freud, 1927/1996). Thus, establishing a dialogue between those different groups will be even more difficult (Dunker, 2016 as cited in Oliveira, 2016).

5. Although users usually show their names on Facebook, we have chosen not to identify them. We have only enumerated each publication instead. The posts are presented in italics, for a better display, and their contents are kept without corrections (we have included the occasional misspellings or punctuation mistakes).
6. FIESP – Federation of Industry of the State of São Paulo (Federação da Indústria do Estado de São Paulo)
7. In Portuguese, this would heavily imply that they are betraying their own class by marching alongside the ones that employ them. It is heavier than what the English translation allows us thanks to the contexts in which both terms are used.

***Netizen 4**: Francisco be sure...I'M NOT A MISERABLE THAT IS ALSO A CRIMINAL ORGANIZATION'S PAWN and I'M SURELY NOT BREAD WITH MORTADELLA[8]...Hahahahaha.*
***Netizen 5**: Is attacking the only thing you can do? Are you demonizing everything? You're hanging out with the bishop[9] a lot, huh? Kisses for you. Avoid me in this life and in all others.*
***Netizen 6**: Camila relax I'm not from the "EDUCATORY HOMELAND"[10] of your CRIMINAL LEADER DILMA ROUSSEFF which you prefer "a thousand times". You're like every criminal PETRALHA who thinks that whoever doesn't vote for the CRIMINAL ORGANIZATION CALLED PT like you must vote for PSDB... My education towards you is great for your level, poor people. Hahahaha*
***Netizen 7**: Just answer me sir, what are you made of in order to talk like that about people you've never seen??? You are nothing but a pile of disrespect! You do not respect the thoughts of others!* (Comments taken from the video post "Na avenida Paulista, em frente ao MASP [...] [...]" (El País Brasil, 2016b).

Netizen 4's comment shows his need to make it clear to all his possible interlocutors that he is not "a miserable pawn", that is, that he belongs to another group that is way better than the ones he is pejoratively referring to: "My education towards you is great for your level, poor things". At the same time, he tries to disqualify the condition of other people to make their own decisions, since he implied that they are all easily manipulated. He even tries to offend them by using the slang "bread and mortadella" – an expression used to refer to workers, and lower-class people, as well as those who defend the Lula and Dilma governments (PT).

The offensive language that is being used in these posts shows not only the difficulty in living alongside others and respecting their positions, but also

8. Bread with Mortadella, or "Pão com Mortadela" in Portuguese, is an old slang that refers to supporters of the Workers' Party (PT). Bread with mortadella is a very cheap meal that was rumored to be given as food to the supporters of said party when they went out into marches. Calling said supporters like that meant that they were so poor they were happy to give their votes in exchange for that meal. This explanation will also be given in the text, this note aims only to give more context to the non-Brazilian reader.
9. This might refer to either Silas Malafaia or Edir Macedo, both protestant leaders in Brazil, famous for being conservative and right-wing.
10. Educational homeland or "Pátria educadora" was the slogan for Brazil during Dilma's government.

a resistance in accepting that the other also has the right to speak. This strong refusal to communicate in a fair manner is reinforced by Facebook itself, since the platform allows its users to remain in their bubbles, treading only among "equals". This demonstrates that, on Facebook, the "monological discourse, instead of giving way to a dialogical discourse, splits into a series of soliloquies, with the users no longer insisting on being heard, but also refusing to listen" (Bauman, 1998: 103). This means that the social media user publishes whatever they want, and it does not matter if they will be heard, if there will be an interlocutor: what matters is talking and, in this case, attacking.

In addition to the insults and attempts to disqualify the other's speech, there are also, in some cases, incitements to acts of violence: we observe some people defending the extermination, the annihilation of those who disagree with their ideas. This demonstrates the Ego's[11] difficulty in coexisting with what is different and, consequently, in engaging in a dialogue with the other. As it is illustrated in this comment:

***Netizen 8**: Only molotov could save them. Or "pau de arara"[12] for the communes* (Comment taken from the post of the video "Manifestantes pró-impeachment comemoram o resultado da votação na Câmara", El País Brasil, 2016a).

The defense of the annihilation, torture and death of the other offers some satisfaction for our hateful desires, but it also offers the satisfaction of the recognition that is received through the likes, shares, and even responses in the comments. These mechanisms strengthen the bond between the ones reacting to a post and the ones posting them. The members of these groups remain united first and foremost because they share the same ideals and the same hatreds, at the same time that they feel protected by a "sense of anonymity and privacy", which leads many of them to "take more risks" (Kallas, 2016: 56).

11. "The Ego. [...] the seat of consciousness and also the place of unconscious manifestations" in Freud's first topography. In the second topography, "[...] the ego is the instance of the imaginary register par excellence, therefore, of identifications and narcissism" (Chemama, 1995: 95).
12. "Pau de arara" is a torture device in which the hands and knees of the victim are tied together, making the person curl into a suspended ball of flesh.

The 2018 elections: Political use of hate speech through social media

As mentioned earlier, the candidate who won Brazil's 2018 elections managed to keep his voters in an almost "alternate reality" by bombarding them with content throughout all the most important social media platforms: Facebook, Twitter, YouTube, WhatsApp (the most famous messaging application in Brazil), among others. Bolsonaro and his team used fake news, as well as both "bots and people to forge an engagement in certain content and give visibility to certain themes, simulating a polarity that he does not have [did not have]" (Mello, 2020: 24).

The way Brazilians get information, added to the importance they give to the content shared in WhatsApp groups and other digital applications, made it possible to carry out an electoral campaign almost exclusively in a world without contradictions: a world created by the social media bubbles and their manipulation by Bolsonaro's campaign. Let us check some data to back up this argument.

In Brazil, 79% of people get information through WhatsApp; 50% through television programs; 49% via YouTube; 44% via Facebook; 38%, by news sites and 22%, by radio programs (Senado Federal, 2019). Among Brazilians, 52% trust news sent by family members on social media, and 43% trust information sent by friends via WhatsApp (Mello, 2020). Therefore, the huge importance these platforms have in Brazilian's daily lives becomes clearer. That, added to the credibility most Brazilians give to the messages received online, ended up creating a very favorable context for the dissemination of not only fake news, but even documentary pieces that attacked, defamed and persecuted some candidates participating in the 2018 election campaign (D'Ávila, 2020).

And in this very context, Jair Bolsonaro was infinitely more present digitally than the other candidates. His Facebook page had 6.9 million subscribers, while that of the candidate of the Workers' Party (PT), Fernando Haddad, had 689 thousand followers – that is, ten times less than his opponent.

On Instagram, Bolsonaro had 3.8 million followers, while Haddad had 418 thousand (Soares, 2018).

Bolsonaro and his three sons became digital influencers and started documenting their lives through YouTube and other social media platforms. This made it easy for them to communicate their ideas directly to their supporters (Mello, 2020). They already knew all too well how social media works and, because of that, they started to use it to publicize their political positions as well as to be closer to voters. This all lines up with the changes that could be noticed ever since the previous election, in which personalism, i.e., a more direct contact with the political candidate, gained even more space in the detriment of political proposals themselves (Fernandez, 2005).

According to the journalist Patricia Mello,

> WhatsApp was a key part of the approach conceived by 'Zero Two'[13] [Carlos Bolsonaro – son of Jair Bolsonaro]. Over the years, groups of supporters were formed that ended up constituting a digital army. The groups worked like transmission lists, in which the administrators – i.e., those who created the group – sent messages to the 256 members, the maximum number allowed by the tool's rules [WhatsaApp]. If a person accesses a link to subscribe to a group, he or she is likely to have a confirmation bias, that is, they are predisposed to believe the content they will receive. Group members, in turn, distribute this content to family and friends (Mello, 2020: 32-33).

Fabrício Beneveluto (2018), creator of the project "Eleições Sem Fake" (Elections without Falsehood, in a free translation), has documented what was discussed under the label of politics in the main social media platforms, as well as in WhatsApp groups, during that time. This led him to the conclusion that the elected candidate had a greater number of supporters online

13. Bolsonaro refers to his sons as if they were squad members. In Brazil's military you don't necessarily have names as the only way to be referred to (e.g., private Ryan); you might also have numbers assigned to you according to your position of power within a given squad. Thus, the first in command is zero one (01), the second in command is zero two (02), and so on.

than the Workers' Party. He also found out that the messages distributed within groups of Bolsonaro's followers were often false, and some were *meme*s and mockery, used to discredit other candidates and disseminate ideas that kept supporters on his side, pushing them to further engage in more social media spaces. Beneveluto (2018) also points out that "WhatsApp is a no man's land", as it is "difficult to track [a publication] because of the encryption of messages and, for this reason, it is a fertile field for the spreading of fake news" (Beneveluto, 2018 as cited in El País, 2018: 1).

Within WhatsApp groups, Bolsonaro's supporters engaged with every content, sharing them massively to other groups that they participated in (Cesarino, 2020). The posts and messages followed a pyramid model, from the largest group to the smallest. They were so organized that there were even members who orchestrated attacks on pages of opponents of the candidate (Santos, 2018 as cited in Simões, 2018).

According to a research group on Political Communication Technology at the State University of Rio de Janeiro (Universidade do Estado do Rio de Janeiro), which monitored WhatsApp groups during the 2018 elections, if any member of the group complained about fake news or questioned a post, they were quickly excluded from the group. Thus: "while the administrators deleted comments from the discussion that they thought could disrupt the campaign, they let loose any hate speech against certain segments of society". There were also "various threats against women and LGBT people" (Aldé, 2018 as cited in Simões, 2018).

This campaign, based on exploiting the potential of hate speech and creating the illusion of persecution – the illusion that politics is a fight of good against evil – not only guaranteed Bolsonaro's victory in that presidential election, but also caused a lot of discomfort within Brazilian society. Many, even today, are trying to elaborate, in the psychoanalytical sense, what was lived in that period.

The following passages, taken from two interviews[14], illustrate how this electoral campaign left its marks.

> *"I left some WhatsApp groups because I noticed I didn't have affinities with the group anymore. Previous to that, I was part of three family groups, as my family is very large. As the elections came to be, politics was a big issue with no way to get rid of. Everyone started to publicize their political opinions and it was a very big turmoil for me. I was the sole one in the group to position myself differently from the other members. When it got closer to the election and I realized that everyone thought differently than me, I was devastated and tired. This was mainly due to the fact that I was making a stand against everyone. This shocked me. After Bolsonaro won, I left the groups because I realized that despite it being our family group it had nothing to do with me. They didn't want to listen to me and there was no possible dialogue. They believed in what they shared. They believed in their lies"* (Interview 1).

> *"I've been a part of more [WhatsApp] groups before, but today I'm more selective with them. My family group used to discuss politics. I left the group that was my father's side of the family because of that. They kept pouring out gratuitous hatred in the group and kept saying that anyone who voted for PT was a thief, a robber, etc. At first, I even argued against that, but in order to avoid further disagreements, I preferred to leave. In person, no one addressed that with me. I can also say that the way I build relationships with other people changed a lot during this time. I walked away from people that seemed to be the most extremist ones, and I didn't come back to talk to them again because it's the kind of situation that doesn't benefit me in any way. I also distanced myself from some family members, but when we meet at family lunches, they treat me the same way they did before. They are very aggressive on social media, but in person they are very tame. The arguments were directed to the WhatsApp group"* (Interview 2).

14. These two interviews are part of Joelma Galvão de Lemos' doctoral research. The material was collected in order to study the political use of social media during the 2018 Brazilian election and its effects in social relationships (Lemos, 2023).

If we watched as Facebook was taken over by hate speech between internet users "coxinhas" and "petralhas" in 2016, in 2018 we watched as this speech reached, through WhatsApp, the most intimate groups: family and friends. When this happened, it was always cause for discomfort and suffering among the group members. Affectionate bonds were shaken, and oftentimes interactions were avoided or even completely interrupted.

Paradoxically, the same social media that makes so many connections possible contributed to the rupture of the same ties they deem to promote. In this case, they contributed to the choice of not co-existing, not engaging in a dialogue with this "other", which is familiar, but also different. This situation experienced by many Brazilians illustrates how challenging it is for humans to co-exist and accept that the similar is always a similar in difference (Kehl, 1996).

Some closing thoughts

When a series of political uprisings took place and brought the country to a halt in 2013, we watched as a new model of organization and social mobilization started to emerge. We also watched the increase in the usage of social media by conservative groups, not only in order to criticize the government, but also to incite a conservative, racist, misogynistic, and homophobic speech. It is as if on the internet, anything goes, and there's no longer any shame in making prejudiced, intolerant, and foolish positions clear (Dias, 2020).

Although some conservative groups have been successful in painting themselves as the driving force behind the 2013 manifestations and, since then, have not left social media spaces, it was only from 2018 onwards that we could notice the systematic use of several social media platforms simultaneously, in Brazil, for political purposes.

Of course, social media was also used by progressive movements, but according to journalist Rosana Pinheiro-Machado (2020), in Brazil, it is "the extreme right that beats the left-wing by W.O" when it comes to the political

use of social media platforms. Since 2013, the conservatives have never left their digital spaces. On the contrary, they increased their influence, reaching more and more people as Brusadin & Graziano (2020) also show in their work.

These right-wing activists, including Bolsonaro and his team, used social media's own organization in their favor. They took special advantage of the fact that the algorithms are programmed to direct users to the same type of contents. By knowing so, they explored the potential of hate speech to mobilize and bring people together, thereby creating a horde of hateful people. The individual, when participating in these groups, ends up identifying with the other members, mainly due to shared ideals between them, and/or because they follow the same leader. In this interaction, even if it is online, social demands and restrictions may be reduced, as people identify with others who think similarly to them. Thus, there is no need for censorship, because they come to love those in their group as they do themselves, and they can, together, direct their hatred towards those on the outside (Freud, 1921)[15].

Isolated in their online bubbles, overwhelmed by the same daily content – because that is how the social media's algorithm works –, alienated by a false sense of reality, and now not as alone as before, internet users feel authorized to express themselves in huge groups, in whatever way they want to. This can be very troubling in itself, and it becomes worse as they begin to share that same content in smaller groups (e.g., the family group), without taking into account what the other members of said group may think or feel.

The more isolated from the ones that think differently, the harder it is for the individual to really engage in real dialogue, and the likelihood of

15. Sigmund Freud (1921), in "Group Psychology and the Analysis of the Ego", demonstrated how the superego loosens its restrictions and demands, and how the ego, through identification with group members, engages in certain actions it likely wouldn't if alone. Another characteristic of individuals in a group is related to their emotions: the emotion of love is directed towards members of the same group, while hatred and aggression are directed towards those who do not belong in those groups. These characteristics identified by Freud (1921) in the analysis of offline groups also can be seen in the relationships of individuals in online groups, within their respective bubbles.

reproducing hate speech increases. As an extreme example of how far this hatred can go, we should highlight the murder of a voter who declared support for the Workers' Party candidate, committed by a Bolsonaro supporter during the first round of the 2018 elections (Brasil de Fato, 2018). This horrible example illustrates how the hate speech disseminated on social media can spill over into the streets.

Nowadays, the Internet is an indisputable part of our lives. Even those who do not engage with any social media per se make use of the internet, be it by accessing their bank's application, by going online shopping, or even by making appointments at a local doctor, since technology has reached more and more services like those. It is impossible to stay away from it.

In this context, we should also highlight that the current organization of Web 2.0 has contributed to the intensification of the ideological and political positions not only in Brazil, but in other countries as well. The fact that the web can be used as a political tool is not a new phenomenon around the globe (Orlowski, 2020).

Therefore, as the writers of this paper, we join other researchers who advocate for the regulation and transparency of the internet and social media platforms. Our efforts and our hopes are that this technology can be at the service of the life drive, that is, that the internet may be used to encourage dialogue and coexistence with the "other". If nowadays the internet works exclusively as a profit tool for many megacorporations, it is also possible that a joint effort can, in a way, force those same megacorporations to review the logic and organization behind their platforms and their algorithms, prioritizing coexistence with alterity over these financial gains.

We understand that this will be challenging, but it is necessary for our current generation to face this issue, as Shoshana Zuboff (2020) points out: every generation faces challenges and needs to find answers to the difficult questions of its time.

References

Amora, D. (2015, 26 april). Manobras fiscais na Caixa cresceram no governo Dilma. *Folha de S. Paulo*. https://www1.folha.uol.com.br/poder/2015/04/1621205-manobras-fiscais-na-caixa-cresceram-no-governo-dilma.shtml.

Após reeleição de Dilma, eleitores do Nordeste são atacados nas rede sociais. (2014, 26 october). *Estadão*. https://politica.estadao.com.br/noticias/eleicoes,apos-reeleicao-de-dilma-eleitores-do-nordeste-sao-atacados-nas-redes-sociais,1583393

Bauman, Z. (1998). *O mal-estar da pós-modernidade*. Zahar.

Bresser-Pereira, L. C. (2012). Brasil, sociedade nacional-dependente. *Novos Estudos CEBRAP*, 93. http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0101-33002012000200008

Brusadin, M. & Graziano, X. (2020). Marketing político e o Darwinismo digital. In J. Fratini (Org.), *Campanhas políticas nas redes sociais* (1 Ed., pp. 45- 54). Matrix.

Carvalho, L. S. N. R. de. (2017). *Discurso do ódio e islamofobia: Quando a liberdade de expressão gera opressão*. [Monograph, Federal University of Bahia]. https://repositorio.ufba.br/ri/bitstream/ri/24031/1/CARVALHO%2C%20Luciana%20Soares%20Neres%20Rosa%20de.%20Discurso%20do%20%C3%93dio%20e%20Islamofobia.pdf

Castells, M. (1999). *A era da informação: Economia, sociedade e cultura*. Vol.1. Paz e Terra.

Castells, M. (2003). *A galáxia da internet: Refelxões sobre a internet, os negócios e a sociedade*. Zahar.

Castells, M. (2013). *Redes de indignação e esperança: Movimentos sociais na era da internet*. Zahar.

Cesarino, L. (2020). Como vencer uma eleição sem sair de casa: A ascensão do populismo digital no Brasil. *Internet & Sociedade*, *1*(1), 92-120. https://revista.internetlab.org.br/wp-content/uploads/2020/02/Como-vencer-uma-eleic%CC%A7a%CC%83o-sem-sair-de-casa.pdf

Chauí, M. (2016). A nova classe trabalhadora brasileira e ascensão do conservadorismo. In: I. Jinkings, H. Doria, & M. Cleto (Orgs), *Por que gritamos golpe? Para entender o impeachment e a crise política no Brasil* (1. Ed., pp. 15-22). Boitempo.

Chemama, R. (Org.). (1995). *Dicionário de Psicanálise Larousse*. Artes Médicas.

Checagem no WhatsApp é o trabalho mais nobre para conter "*fake news*". (2018). *El País*. https://brasil.elpais.com/brasil/2018/09/27/politica/1537999429_399901.html

Cleto, M. (2016). O triunfo da antipolítica. In: I. Jinkings, K. Doria, & M. Cleto (Orgs.), *Por que gritamos golpe? Para entender o impeachment e a crise política no Brasil* (1. Ed., pp. 41 – 48). Boitempo.

Córdova, Y. (2019, 9 de janeiro). Como o YouTube se tornou um celeiro da nova direita radical. *The Intercept*. https://theintercept.com/2019/01/09/youtube-direita/

D'Ávila, M. (2020). *E se fosse você? Sobrevivendo às redes de ódio e fake news*. Instituto E Se Fosse Você.

Dias, M. M. (2012). *Os ódios: Clínica e política do psicanalista, seminário / Mauro Mendes Dias*. Iluminuras.

Dias, M. M. (2020). *O discurso da estupidez*. Iluminuras.

Dias, T. (2020, 14 de setembro). O dilema das redes: Sair da internet não vai salvar a internet. *The Intercept*. https://theintercept.com/2020/09/14/internet-netflix-redes/

Dois dias após impeachment, Senado aprova lei que permite pedaladas fiscais. (2016, 02 set.). *IG*. http://economia.ig.com.br/2016-09-02/lei-orcamento.html

Dunker, C. I. L. (2015). *Mal-estar, sofrimento e sintoma: Uma psicopatologia do Brasil entre muros*. Boitempo.

Eleitor de Bolsonaro mata mestre de capoeira por declarar votos no PT. (2018, 08 oct.). *Brasil de Fato*. https://www.brasildefato.com.br/2018/10/08/referencia-da-capoeira-e-da-cultura-afro-e-assassinado-apos-discussao-politica-na-ba

Entra em vigor lei que criminaliza perseguição, inclusive na internet. (2021). *Câmara dos Deputados*. https://www.camara.leg.br/noticias/742273-entra-em-vigor-lei-que-criminaliza-perseguicao-inclusive-na-internet/

Esquerda brasileira é muito ruim na internet, diz Javier Toret. (2016, 11 jul.). *JornalGGN*. https://jornalggn.com.br/politica/esquerda-brasileira-e-muito-ruim-na-internet-diz-javier-toret/

Freud, S. (1996). O instinto e suas vicissitudes. In: S. Freud (Ed.), *A história do movimento psicanalítico, artigos sobre a metapsicologia e outros trabalhos* (pp. 117 - 146). Imago.

Freud, S. (1996). Além do princípio do prazer. In: S. Freud (Ed.), *Além do princípio do prazer, Psicologia de Grupo e outros trabalhos* (pp. 13- 78). Imago.

Freud, S. (1996). Psicologia de Grupo e a análise do ego. In: S. Freud (Ed.), *Além do princípio do prazer, Psicologia de Grupo e outros trabalhos* (pp. 79 - 145). Imago.

Freud, S. (1996). O futuro de uma ilusão. In: S. Freud (Ed.), *O futuro de uma ilusão, o mal-estar da civilização e outros trabalhos* (pp. 13 - 65). Imago.

Freud, S. (1996). O mal-estar na civilização. In: S. Freud (Ed.). *O futuro de uma ilusão, o mal-estar da civilização e outros trabalhos* (pp. 67- 150). Imago.

Freud, S. (1996). Por que a guerra? In: S. Freud (Ed.), *Novas conferências introdutórias sobre a psicanálise e outros trabalhos* (pp. 190-210). Imago.

Kallas, M. B. L. de M. (2016). O sujeito contemporâneo, o mundo virtual e a psicanálise. *Revista Reverso*, 71, 55-64. http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S0102-73952016000100006

Kehl, M. R. (1996). Psicanálise, ética e política. In: M. I. França (Org), *Ética, psicanálise e sua transmissão* (pp. 109 -121). Vozes.

Lanier, J. (2018). *Dez argumentos para você deletar agora suas redes sociais*. Intrínseca.

Lemos, J. G. de. (2018). *O uso político do discurso de ódio no Brasil: Um estudo de caso no Facebook - 2016-2017*. (Master's Thesis, Federal University of Sergipe). http://ri.ufs.br/jspui/handle/riufs/10106

Lemos, J. G. de. (2023). *O uso político das redes sociais online nas eleições de 2018 no Brasil*. (PhD thesis, Federal University of Sergipe). https://ri.ufs.br/handle/riufs/18174

Lopes, M. (2016). As quatro famílias que decidiram derrubar um governo democrático. In: I. Jinkings, K. Doria, & M. Cleto (Orgs.), *Por que gritamos golpe? Para entender o impeachment e a crise política no Brasil* (1. Ed., pp. 117 – 125). Boitempo.

Mais de 80% dos brasileiros acreditam que redes sociais influenciam muito a opinião das pessoas. (2019, 10 dec.). *Senado Federal*. https://www12.senado.leg.br/institucional/datasenado/materias/pesquisas/mais-de-80-dos-brasileiros-acreditam-que-redes-sociais-influenciam-muito-a-opiniao-das-pessoas

Manifestantes pró-impeachment comemoram o resultado da votação na Câmara (2016a). Vídeo. *El País Brasil*. https://www.facebook.com/elpaisbrasil/videos/1063092710417345/

Mello, P. C. de. (2020). *A máquina do ódio: Notas de uma repórter sobre fake news e violência digital*. Companhia das Letras.

Na Avenida Paulista, em frene ao Masp, manifestantes contrários ao impeachment de Dilma Rouseff começam a se reunir no início da noite desta quarta. O fotógrafo Maurício Pisani mostra como estava o lugar por volta das 19h. (2016b). Vídeo. *El País Brasil*. https://www.facebook.com/elpaisbrasil/videos/1078123668914249/

No Acre, Bolsonaro fala em 'fuzilar a petralhada' e enviá-los à Venezuela. (2018). Vídeo. *You Tube*. https://www.youtube.com/watch?v=p0eMLhCcbyQ&ab_channel=Poder360

Oliveira, T. (2016). Polarização política, reflexo de uma sociedade murada. *Revista Carta Capital*. https://www.cartacapital.com.br/politica/polarizacao-politica-reflexo-de-uma-sociedade-murada

Orlowski, J. (2020). *The social dilemma*. Documentary. Netflix.

Pariser, E. & Helsper, E. (2021). *The filter bubble: What the internet is hiding from you*. LSE public lecture. https://www.lse.ac.uk/assets/richmedia/channels/publicLecturesAndEvents/slides/20110620_1830_the-FilterBubble_sl.pdf

Pereira, S. W. (2006). *As pulsões de morte e seus derivados: os avatares da teoria*. (PhD thesis, Federal University of Rio de Janeiro). http://www.dominiopublico.gov.br/pesquisa/DetalheObraForm.do?select_action=&co_obra=133505

Pinheiro-Machado, R. (2020, 21 de julho). Na batalha das redes, a extrema direita ganha por W.O. *The Intercept*. https://theintercept.com/2020/07/21/batalha-redes-extrema-direita-esquerda/

Por volta das 20h, manifestantes pró-impeachment se reuniram em frente à Fiesp na Avenida Paulista. O fotógrafo Maurício Pisani conta que já há bonecos infláveis e carros de som. (2016c). Vídeo. *El País Brasil*. https://www.facebook.com/elpaisbrasil/videos/1078149188911697/

Ribeiro, D. (1995). *O Povo brasileiro: A formação e o sentido do Brasil*. Companhia das Letras.

Schäfer, G., Leivas, P. G., & Santos, R. H. dos. (2015). Discurso do ódio: Da abordagem conceitual ao discurso parlamentar. *Revista Informações Legislativas*, *52*(207), 143-158. http://www2.senado.leg.br/bdsf/item/id/515193

Simões, M. (2008, 24 de outubro). Pesquisa mostra como atuam os grupos pró-Bolsonaro no WhatsApp. *Exame*. https://exame.com/brasil/pesquisa-mostra-como-atuam-os-grupos-pro-bolsonaro-no-whatsapp/

Singer, A. (2016). Por uma frente ampla, democrática e republicana. In: I. Jinkings, K. Doria, & M. Cleto (Orgs), *Por que gritamos golpe? Para entender o impeachment e a crise política no Brasil* (1. Ed., pp. 151 - 156). Boitempo.

Soares, J. (2018, 7 de outubro). Time digital de Bolsonaro distribui conteúdo para 1.500 grupos de WhatsApp. *O Globo*. https://oglobo.globo.com/politica/time-digital-de-bolsonaro-distribui-conteudo-para-1500-grupos-de-whatsapp-23134588

Souza, J. (2016). *A radiografia do golpe: Entenda como e por que você foi enganado*. LeYa.

Souza, J. (2017). *A elite do atraso: Da escravidão à Lava Jato*. LeYa.

Zuboff, S. (2020). *A era do capitalismo de vigilância: A luta por um futuro humano na nova fronteira do poder*. Intrínseca.

FREEDOM OF THE PRESS OR HATE SPEECH? REGULATING MEDIA OUTLETS IN THE POST-TRUTH ERA

Branco Di Fátima
/ University of Beira Interior, Portugal

Marco López-Paredes
/ Pontifical Catholic University of Ecuador, Ecuador

Introduction

Freedom of the press is a crucial element for the existence of modern democracy. A state can only be considered democratic if its citizens have unrestricted access, with exceptions as provided by law, to information of public interest (Briggs & Burke, 2006). Thus, media outlets perform the dual function of keeping citizens informed, providing the basis for politicizing public opinion, and monitoring the actions of governments while revealing injustices obscured by other social forces (Correia, 2011; Sousa, 2010).

Well-informed citizens also play a crucial role as active members of the community. With a high level of media literacy, individuals are theoretically better equipped to participate in the public sphere (Livingstone, 2003). In this context, the importance of media outlets lies in their ability to exercise critical oversight of the three branches of the nation-state (the executive, the legislative, and the judicial) and to provide citizens with accurate information about society (Bobbio, Matteucci, & Pasquino, 1983). Therefore, the argument that media outlets constitute the fourth estate is justified due to their direct

influence on public decisions (Castells, 2009). In other words, journalists perform the watchdog function over those in power (Benkler, 2006).

Despite the origins and recent usage of the term hate speech, concepts such as media pluralism and freedom of the press can no longer be considered untouchable. On the contrary, rules regarding hate speech and efforts to maintain the objectivity of information are usually enacted at the national level, always with consideration for human rights, especially in an ultra-mediated society (López-Paredes & Carrillo-Andrade, 2024).

Many concepts, particularly those related to media and technology, possess specific cultural, social, or legal meanings but lack universally recognized definitions. For instance, the term post-truth primarily pertains to journalism, media, and politics, yet it lacks a legal or internationally agreed-upon definition. According to the Oxford Dictionaries (2016), post-truth is defined as an adjective "relating to or denoting circumstances in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief". In this context, social media platforms play a crucial role in shaping an alternative reality influenced by algorithms and intense emotions, thereby blurring the line between truth and simulacra (Snyder, 2021; Fischer, 2021).

This chapter delves into the complex dynamics of regulating media outlets in the post-truth era, focusing on the challenges posed by the proliferation of online hate speech. These issues are particularly pronounced in highly polarized societies, where media outlets often blend opinionated narratives with factual news, sometimes neglecting the ethical principles of journalism. The chapter also uses Brazil as a case study, highlighting a country without a regulatory agency for media outlets and significantly impacted by political polarization over the last decade (Capoano, Sousa, & Prates, 2023).

The emphasis on freedom of the press has sometimes compromised other rights (Varjão, 2016), facilitating the spread of hate speech (Roozen & Shulman, 2014; Pinto, 2013). Without effective regulation of media outlets, the principle of freedom of the press can become a double-edged sword.

While freedom of the press is essential for democracy, it can also lead to violations of other rights. The unchecked power of media outlets can result in disinformation, misinformation, defamation, invasion of privacy, and the propagation of harmful content. The Brazilian example illustrates how the absence of media regulation can compromise the safety of marginalized social groups, who are particularly vulnerable to online hate attacks.

Legal matter and political concern

In the 20th century, ensuring social rights became a primary concern for legislators, policymakers, and the public (Bobbio, Matteucci, & Pasquino, 1983). Consequently, modern media outlets emerged as watchdogs, highlighting governmental deviations and failures to uphold rights such as education, health, food security, labor rights, and those stipulated in liberal constitutions (Benkler, 2006). While not novel, this phenomenon has become increasingly complex due to the widespread adoption of various digital technologies such as the internet, smartphones, and social media platforms (Castells, 2009). However, the crux of the matter remains the freedom of the press.

Marx (2006) already advocated for freedom of the press, emphasizing the importance of media regulation to ensure the quality and accuracy of published information. During his tenure as editor-in-chief of the *Rheinische Zeitung* from 1842 to 1843, opinion often dominated factual news content (Gorender, 1996; Beltrão, 1980). Therefore, Marx regarded press law as essential for legally recognizing freedom of the press (Eidt, 1998).

In the same spirit, the press law could ensure the effectiveness of other principles protecting constitutional rights (Silva, 2006), such as the presumption of innocence, privacy, and protection of minorities against hate speech. Essentially, the debate on regulating media outlets appears to arise from both legal considerations – focused on approving mechanisms for control, education, and enforcement – and political considerations – centered on the interests of diverse social actors in establishing regulatory agencies for media outlets.

In the absence of specific legislation, society must rely on self-regulation, supported by the likely adherence to journalism's code of ethics. According to Fidalgo (2000: 320), self-regulation enables journalists to establish "basic rules of conduct and strengthens their commitment to uphold them". Marx (2006: 12) already viewed this approach positively, stating that "the first requirement of freedom is self-awareness". However, he did not believe that self-regulation alone could effectively oversee the operations of media outlets, thus underscoring the need for legislation specific to the media sector.

Numerous pertinent questions emerge from the debates surrounding regulatory mechanisms: To what extent should state intervention occur in journalism? Could press legislation pose a threat to democracy? What are the epistemological boundaries that distinguish regulation from censorship? How can hate speech be effectively combated, particularly in loosely controlled virtual spaces, without compromising freedom of the press?

These inquiries stem from how various sectors position themselves in their proposed actions, their conduct of public debates, and the editorial stance of media outlets, which sometimes adopt more aggressive narratives. However, the responses from these sectors often remain ambiguous and should not be rigidly interpreted outside their original contexts. Understanding the positions of these entities requires a deeper exploration of the concepts of regulation and censorship regarding hate speech. An analysis of their convergence and divergence should begin with acknowledging that freedom of the press encompasses the right to express thoughts freely through the media (Hungria, 1995). However, this freedom must be balanced with respect for the rights of others, particularly social minorities and those most vulnerable to online hate attacks.

Even though hate speech lacks a globally accepted definition, it can generally be understood as verbal and non-verbal attacks targeting inherent characteristics of individuals, such as race, nationality, religion, culture, or sexual orientation (Di Fátima, 2023). Hate speech encompasses all forms of human language that "spread, incite, promote or justify racial hatred,

xenophobia, anti-Semitism or other forms of hatred based on intolerance" (Keen & Georgescu, 2020: 143). Rooted in the codes and values of specific cultures, hate speech functions as a violent narrative against diversity (Matamoros-Fernández & Farkas, 2021).

Social media platforms have significantly amplified the spread of hate speech, much of which thrives on disinformation (Martínez Valerio, 2022). Consequently, violent narratives from traditional media outlets, often masked as journalism, gain momentum online via platforms like Facebook, X, or YouTube (Garbe, Selvik, & Lemaire, 2023; Chekol, 2023). A primary argument put forth by haters is the defense of freedom (Amores *et al.*, 2021). This is particularly contentious because, in contexts with authoritarian governments, legislation aimed at combating hate speech has been used to penalize political dissidents and ordinary citizens who critique prevailing norms (Munoriyarwa, 2023). Here lies the paradox: Any attempt at media regulation risks being perceived as a form of censorship on freedom of the press.

The term *regulation*, derived from the Latin word *regularis*, originally refers to a ruler or measuring stick, not to the impediment or control of something, someone, or an action. Therefore, state monitoring of media outlets' activities does not inherently violate the right to information and should not be confused with censorship. Regulation serves as a standard mechanism for addressing complaints related to abuse of power, invasion of privacy, manipulation, hate speech, and various other offenses (Dalmonte, 2011).

The regulatory process is essential for ensuring pluralism and giving a voice to all parties involved in a social event, as described in many journalism manuals (Traquina, 2007). Therefore, media regulation involves implementing practical methods for monitoring news content (Fidalgo, 2006). This should not be perceived as a restriction on freedom of the press. In democratic countries, regulatory agencies review content only after it has been published or broadcast, never before. Despite concerns expressed by many digital platforms about online censorship, the evaluation of content on social media can only occur after publication.

Derived from the Latin word *censere*, the term *censorship* refers to the act of judging something, someone, or an action. Media censorship involves the complete or partial prevention of the publication of information that is of public interest. This can include mechanisms of punishment, manipulation, or persuasion used against press professionals or media companies, including social media platforms. As a result, censorship prevents certain facts from being exposed to public opinion (Morozov, 2011). Journalism manuals and codes of ethics also often emphasize the importance of preventing obstacles to the free flow of information.

The role of a censor revolves around pre-emptively managing content to minimize its potential impact on others. Censors conceal, modify, and sometimes censor messages intended for public consumption. A press can be considered unfree if it faces institutional threats and constraints that prevent the publication of information. In a democracy, journalists alone should decide whether to disclose facts to the public.

Some authors argue that the right to information is civil, political, and social simultaneously, given its paramount importance for modern society (Cepik, 2011). Access to information is primarily controlled by the state (Bourdieu, 1996), although it frequently interfaces with media outlets and internet networks (Castells, 2009). It is in this sense that Marx (2006: 60) asserts that "the free press is the omnipotent eye of the people".

An analysis of the concepts of regulation and censorship reveals that free media outlets are not exempt from responsibility for what they publish, both in terms of content and form. Freedom of the press cannot be synonymous with licentiousness, as media outlets are legally considered a public service (Briggs & Burke, 2006). According to Hungria (1955: 261), "media outlets, due to the significant interests that sometimes conflict with freedom of ideas and opinions, have been subject to specific regulations in practically all Western countries".

Ignoring this rule is tantamount to claiming that media outlets act arbitrarily, without any accountability for their mistakes (Dalmonte, 2011). The issue

is so urgent that it is being debated in various countries with diverse political contexts, such as Argentina (Califano, 2018), Portugal (Miranda, 2018) or Angola (Miguel, 2016). The lack of regulatory instruments is even more concerning when other rights are violated by the actions of the media outlets. In this sense, there are numerous instances where media outlets have propagated hate speech (Munoriyarwa, 2023; Roozen & Shulman, 2014).

Hatred can be fueled by the way media outlets report on social events. This news content is then mirrored on social media, gaining immense visibility beyond its original context. Sensationalized news coverage often exaggerates and distorts views of social minorities, creating negative representations of these communities (Saleem, Yang, & Ramasubramanian, 2016). For example, immigrants or refugees are frequently associated with increased crime and so-called deviant practices (Bruno, 2016). By spreading stereotypes about the behavior of these groups, the media incites fear and unrest in society, potentially leading to hate speech and, in more serious cases, physical aggression. There are numerous examples of these cases.

Radio Télévision Libre des Mille Collines (RTLM) in Rwanda played a pivotal role in spreading the hatred that culminated in the 1994 genocide (Roozen & Shulman, 2014). RTLM broadcast propaganda that dehumanized the Tutsi population and incited violence, resulting in one of the most brutal genocides in history. Following the September 11, 2001 attacks in the United States, media outlets like CNN and Fox News frequently presented intense and often negative coverage of Muslims and Arabs (Pervez & Saeed, 2010). This pervasive reporting disproportionately linked Muslims and Arabs with terrorism. This biased reporting contributed to a significant increase in hate crimes against these communities.

During the mid-2010s refugee crisis, several media outlets, notably UK and Australian tabloids, ran stories linking refugees to criminal activity (Parker, 2015). This negative portrayal has exacerbated xenophobia, leading to a rise in online attacks against immigrants and refugees. In Russia, for example, media outlets frequently publish homophobic and transphobic content,

using disinformation against these communities (Edenborg, 2018; Ennis, 2014). Such hate speech, propagated through media outlets, finds extensive circulation and discussion on Russian social media platforms, blogs, and Internet forums. This dissemination fosters an environment of increasing hostility and violence toward LGBTQ+ individuals.

What distinguishes news content from an opinion article, beyond narrative structure, is the use of verifiable information and the inclusion of diverse voices, respecting sources' rights to contradict each other (Traquina, 2007; Beltrão, 1980). Journalists must maintain independence and autonomy precisely to ensure that "the final product of their work (the news) is not influenced by factors outside of journalistic criteria" (Fidalgo, 2000: 326). Failure to adhere to this ethical standard compromises the right to accurate information, especially in highly polarized societies in the post-truth era.

Freedom or hate speech in Brazil

Brazil is one of the few United Nations (UN) member states without a press law. On April 30, 2009, the Federal Supreme Court (STF) abolished the existing press law with a vote of seven to four. The ministers justified their decision by stating that the law, which had been enacted during the Military Dictatorship (1964-1985), violated democratic principles. However, a new press law has not yet been voted on by the Brazilian parliament.

The STF made its second controversial decision regarding media outlets on June 17, 2009. The justices abolished the requirement for a journalist's diploma to practice the profession in the country. The justices argued that requiring a diploma violated freedom of expression, guaranteed by the Federal Constitution (1988), and hindered free access to information as stipulated in the American Convention on Human Rights (1969). Consequently, the National Federation of Journalists (Fenaj) was obliged to issue professional cards to anyone who requested one.

The STF also created a legal vacuum. Consequently, the media in Brazil exclusively self-regulate their journalistic work. Despite its undeniable public

value, the Journalists' Code of Ethics exclusively targets the operations of newsrooms (Christofoletti, 2011). Revised in 2007 by Fenaj, this document does not include sanctions for serious offenses committed by journalists, such as disseminating hate speech or prejudice against social minorities. Therefore, in most cases, these codes lack legal enforceability within the country's judicial framework (Fidalgo, 2000).

Media outlets and their professionals can be held accountable for their actions under various legal frameworks, including the Federal Constitution (1988), the Penal Code (1940), and the Consumer Defence Code (1990). However, the absence of specific legislation for the media outlets sector often turns even simple cases involving the right of reply, under Law No. 13,188, into protracted legal battles. In some instances, the stages of investigation, judgment, and publication of the sentence can take months, creating a *periculum in mora* – a risk of irreparable damage due to procedural delays. Consequently, delayed judgments often render the issue irrelevant over time, as the individuals or groups harmed by a news story rarely see their reputations restored through a retraction.

This phenomenon intensifies when social movements advocate for the establishment of a regulatory agency, facing resistance from media companies and some professionals (Filho, 2009). Supporters of creating a National Communication Council (CNC) argue that such regulation ensures freedom of the press, access to information, and protection of social minorities, including protection against hate speech. However, opponents are concerned that over time, the CNC could transform into a tool for state censorship.

The purpose of social movements is grounded in everyday empirical experiences. Research conducted by the News Agency for Children's Rights (ANDI) revealed, for instance, that in just one month, 4,500 rights violations were documented in television and radio programs across ten Brazilian states. During this period, broadcasters were accountable for over 15,760 legal infractions (Varjão, 2016).

Disrespect for human rights, particularly evident in police news coverage, has been strongly criticized. The presumption of innocence, guaranteed by the Federal Constitution (1988), is frequently overlooked (Linhares & Grotti, 2021). Furthermore, many media outlets rely solely on police sources, disregarding Article 12 of the Code of Ethics. According to Fenaj (2007: 3), journalists are obligated to hear "from the widest range of people and institutions involved in the coverage, particularly those who are the subject of accusations". Failing this, they risk perpetuating stigmas against specific social groups, often resulting in the creation of a negative image of a community.

In the Brazilian media landscape, there appears to be a significant gap between journalists' recognition of deontological principles and their actual implementation in news practice (Cepik, 2011). Despite these rules being constitutional and enforceable, instances of non-compliance are seldom met with sanctions. For this reason, social movements have advocated for the establishment of the CNC. However, media companies and a significant number of journalists oppose its creation, fearing that the council could potentially evolve into a state-serving censorship body given its legal nature (Filho, 2009). This debate has prominently featured in the political arena in recent years, especially during electoral periods.

In the 2018 elections, candidate Fernando Haddad (PT) presented a government plan that placed significant emphasis on regulating media outlets. According to Haddad (2018: 17), "all established democracies worldwide implement mechanisms for democratic regulation to support the broad exercise of the human right to communication". In contrast, former president Jair Bolsonaro's (PSL) government plan also highlighted freedom of the press but opposed the creation of specific legislation. According to Bolsonaro (2018: 7), "we are against any regulation or social control of media outlets".

In the national imagination, there exists a delicate relationship between media outlet regulation and press censorship (Haddad, 2018; Bolsonaro, 2018; Gaspari, 2002). This perception partly stems from the lingering effects of the Military Dictatorship, which governed the country for 21

years and notably suppressed media outlets and freedom of expression. Conversely, media companies harbor concerns that regulatory agencies could undermine their economic interests and political influence. These dual apprehensions – shared by both the public and corporate sectors – are evident in ongoing debates over the broader implications of press regulation (Costa, 2014; Pinto, 2013; Bôaviagem, 2011), including the development of legal frameworks to address online hate speech.

Efforts to establish press control mechanisms have encountered opposition from media companies, criticism from journalists, and attacks from political parties (Filho, 2009). One of the most notable confrontations occurred in December 2009, during the 1st National Communication Conference (Confecom). The preparations for Confecom involved regional and local events where ideas were presented for the national gathering. Delegates elected at these events endorsed approximately 630 proposals, including the creation of the CNC and press regulations.

Confecom was expected to recommend guidelines, but they may never materialize (Penna Pieranti, 2019). Consequently, television programs such as *Brasil Urgente* (Band TV) or *Cidade Alerta* (Record TV) frequently propagate hate speech against marginalized groups and social minorities, often without presuming the innocence of the accused. In the post-truth era, their presenters blend opinionated narratives with information, sometimes disregarding the ethical principles of journalism.

Consequently, hate speech directed at social minorities has shifted from traditional media to the online sphere, potentially reaching a broader audience beyond its original media outlets. Programs like *Brasil Urgente* and *Cidade Alerta* have garnered significant traction on social media in recent years. On YouTube alone, they have amassed an audience of nearly ten million followers, and their content has been viewed more than 4.7 billion times on the platform by July 2024.

Conclusions

Self-regulation is currently the only mechanism ensuring that media outlets in Brazil act responsibly. This responsibility falls to individual professionals, companies, or unions. The Code of Ethics lacks legal force and does not impose sanctions for violations of basic journalistic principles or the propagation of hate speech against social minorities.

Media companies, unions, and journalists have expressed concerns about the existence of a regulatory agency in Brazil, and their concerns seem justified. In the popular imagination, the authoritarian incursion by the Military Dictatorship remains relatively recent (Gaspari, 2002). On the other hand, social movements understand that the rights established and guaranteed in the constitution cannot be violated in the name of so-called freedom of the press.

The justification for analyzing the lack of regulatory mechanisms is the need to find a common denominator for drafting a press law, specifically one that addresses hate speech. While it is the responsibility of the legislative branch to create laws and the executive branch to enforce them, it is civil society that establishes their legitimacy.

In most cases, so-called press offenses are quickly forgotten. However, the consequences of these violations have a lasting impact on the lives of those portrayed and harmed by the news content. Sometimes, the conduct of the media outlets is judged solely by public opinion. When public opinion alone is responsible for assessing the veracity, objectivity, and impartiality of the content, it often struggles with information that is difficult to counter. "When the news is published, most readers believe it" (Silva, 2006: 58). At this point, fact-checking agencies make a valuable contribution to combating hate speech that originates in traditional media outlets and migrates to the online sphere.

Haters often argue in favor of unrestricted freedom. As a result, narratives of violence that originate from traditional media outlets, sometimes disguised as journalism, gain popularity on platforms such as Facebook, X,

and YouTube (Chekol, 2023). This debate is contentious because in some cases, laws designed to address hate speech have been used to punish political dissidents and ordinary individuals who challenge prevailing norms (Munoriyarwa, 2023). Moreover, a press law could be utilized either to combat hate speech in media outlets or to censor legitimate narratives that challenge established power.

Although freedom of the press is guaranteed in Brazil, the right to information and adherence to basic journalistic standards are often neglected by professionals and media companies. The crux of the matter lies in Fidalgo's analysis (2006: 437): the notion that the press "will be held accountable for its actions if and when they contradict the responsibilities and expectations associated with its role".

The debate on drafting a press law needs to go beyond viewing regulation as a mechanism for censorship. Many studies suggest that the involvement of all sectors – journalists, unions, citizens, media companies, social movements, etc. – is essential. This council should be independent of government influence and should not be punitive in nature. However, it could recommend criminal investigations based on the interpretation of existing legislation.

What makes the approval of legislation for the press, or the creation of a regulatory body, urgent is the ineffectiveness of self-regulation. In the name of free information, other rights are historically violated without regulation. This is not only a legal problem but also a political one. Addressing it depends on understanding different social forces. Otherwise, the modern state, which was founded on the support of a legislative body, is put at risk.

References

Amores, J. J., Blanco-Herrero, D., Sánchez-Holgado, P., & Frías-Vázquez, M. (2021). Detectando el odio ideológico en Twitter: Desarrollo y evaluación de un detector de discurso de odio por ideología política en tuits en español. *Cuadernos.info*, *49*(2021), 98-124.

Beltrão, L. (1980). *Jornalismo opinativo*. Sulina.

Benkler, Y. (2006). *The wealth of networks: How social production transforms markets and freedom*. Yale University Press.

Bôaviagem, A. A. (2011). Responsabilidade civil da imprensa. *Revista Duc In Altum*, *3*(3), 133-146.

Bobbio, N., Matteucci, N., & Pasquino, G. (1983). *Dicionário de política* (Vol. 1). Editora UnB.

Bolsonaro, J. (2018). *O caminho da prosperidade: Proposta de plano de governo*. Partido Social Liberal (PSL).

Bourdieu, P. (1996). *Razões práticas: Sobre a teoria da ação*. Papirus.

Briggs, A. & Burke, P. (2006). *Uma história social da mídia: De Gutenberg à internet*. Zahar.

Bruno, M. (2016). Media representations of immigrants in Italy: framing real and symbolic borders. *REMHU – Revista Interdisciplinar Da Mobilidade Humana*, *24*(46).

Califano, B. (2018). La regulación de la comunicación durante el primer año de gobierno de Mauricio Macri en la Argentina. *Intersecciones en Comunicación*, *1*(12), 1-16.

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.

Castells, M. (2009). *Communication power*. Oxford University Press.

Cepik, M. (2000). Direito à informação: Situação legal e desafios. *Informática Pública*, *2*(2), 43-56.

Chekol, M. A. (2023). Ethiopian socio-political contexts for hate speech. In B. Di Fátima (Ed.), *Hate speech on social media: A global approach* (pp. 227-254). LabCom Books & EdiPUCE.

Christofoletti, R. (2011). O caso do Brasil: Valores, códigos de ética e novos regramentos para o jornalismo nas redes sociais. *Cuadernos de Información*, *29*(julio-diciembre), 25-34.

Correia, J. C. (2011). *O admirável mundo das notícias: Teorias e métodos*. LabCom Books.

Costa, T. M. (2014). Conteúdo e alcance da decisão do STF sobre a lei de imprensa na ADPF 130. *Revista Direito GV*, *10*(1), 119-154.

Dalmonte, E. F. (2011). É preciso ordenar a comunicação? Questionamentos acerca da necessidade de instâncias mediadoras entre mídia e público. *Estudos em Jornalismo e Mídia*, *8*(1), 21-39.

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

Edenborg, E. (2018). Homophobia as geopolitics: "Traditional values" and the negotiation of Russia's place in the world. *Gendering Nationalism*, 67-87.

Eidt, C. (1998). *O Estado racional: Lineamentos do pensamento político de Karl Marx nos artigos da Gazeta Renana (1842-1843)*. Master's dissertation, Federal University of Minas Gerais.

Ennis, S. (2014, January 17). Homophobia spreads in Russian media. *BBC Monitoring*.

Fenaj – Federação Nacional dos Jornalistas. (2007). *Código de Ética dos Jornalistas Brasileiros*. Fenaj.

Fidalgo, J. (2000). A questão das fontes nos códigos deontológicos dos jornalistas. *Comunicação e Sociedade*, *14*(1-2), 319-337.

Fidalgo, J. (2006). *O lugar da ética e da autorregulação na identidade profissional dos jornalistas*. PhD thesis, University of Minho.

Filho, J. P. C. (2009). *Por uma Lei de imprensa*. Observatório da Imprensa.

Fischer, F. (2021). *Truth and post-truth in public policy: Interpreting the arguments*. Cambridge University Press.

Garbe, L., Selvik, L. M., & Lemaire, P. (2023). How African countries respond to fake news and hate speech. *Information, Communication & Society*, *26*(1), 86-103.

Gaspari, E. (2002). *A ditadura escancarada*. Companhia das Letras.

Gorender, J. (1996). Apresentação. In K. Marx (1996). *O capital* (Vol. 1). Nova Cultura.

Haddad, F. (2018). *Plano de governo 2019-2022: Coligação o povo feliz de novo*. Partido dos Trabalhadores (PT).

Hungria, N. (1955). *Comentários ao Código Penal* (Vol. 6). Forense.

Keen, E. & Georgescu, M. (2020). *Bookmarks – A manual for combating hate speech online through human rights education*. Council of Europe.

Linhares, É. & Grotti, V. (2021). Liberdade de imprensa e presunção de inocência: a condenação social e midiática antecipada. *Novas Teses Jurídicas I*, *8*(51), 306-320.

Livingstone, S. (2003). The changing nature and uses of media literacy. *Media@lse*, (4), 2-31.

López-Paredes, M. & Carrillo-Andrade, A. (2024). Cartografía de consumo de medios en Ecuador: De las mediaciones e hipermediaciones a una sociedad ultramediada. *Palabra Clave*, *27*(1), e2712.

Martínez Valerio, L. (2022). Mensajes de odio hacia la comunidad LGTBIQ+: Análisis de los perfiles de Instagram de la prensa española durante la "Semana del Orgullo". *Revista Latina de Comunicación Social*, (80), 364-388.

Marx, K. (2006). *Liberdade de imprensa*. L&PM.

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

Miguel, N. K. G. (2016). *A regulação da imprensa em Angola*. Master's dissertation, New University of Lisbon.

Miranda, J. (2018). Regulação participada e regulação em parceria como resposta aos desafios da profissão. *Media & Jornalismo*, *18*(32), 31-42.

Morozov, E. (2011). *The net delusion: The dark side of internet freedom*. PublicAffairs.

Munoriyarwa, A. (2023). Mapping social media hate speech regulations in Southern Africa: A regional comparative analysis. In B. Di Fátima (Ed.), *Hate speech on social media: A global approach* (pp. 203-226). LabCom Books & EdiPUCE.

Oxford Dictionaries (2016). *Word of the Year 2016*. OxfordLanguages.

Parker, S. (2015). 'Unwanted invaders': The representation of refugees and asylum seekers in the UK and Australian print media. *eSharp*, *23*(Myth and Nation), 1-21.

Penna Pieranti, O. (2019). Confecom, 10 anos depois: Um debate necessário sobre a implementação das propostas aprovadas. *Chasqui*, (141), 275-288.

Pervez, S. & Saeed, S. (2010). Portrayal of Muslims and Islam in the talk shows of CNN and Fox News (2007-2009). *Journal of Media Studies*, *25*(2), 122-140.

Pinto, I. L. F. (2013). Liberdade de expressão, lei de Imprensa e discurso do ódio – Da restrição como violação à limitação como proteção. *Revista de Direito Administrativo & Constitucional (A&C)*, *13*(53), 195-229.

Roozen, B. & Shulman, H. C. (2014). Tuning in to the RTLM: Tracking the Evolution of Language Alongside the Rwandan Genocide Using Social Identity Theory. *Journal of Language and Social Psychology*, *33*(2), 165-182.

Saleem, M., Yang, G. S., & Ramasubramanian, S. (2016). Reliance on Direct and Mediated Contact and Public Policies Supporting Outgroup Harm. *Journal of Communication*, *66*(4), 604-624.

Silva, W. C. (2006). Da lei à ética: Mecanismos de limitação à liberdade de imprensa. *Diálogos Possíveis*, *5*(2), 49-67.

Snyder, T. (2021). *The American abyss. A historian of fascism and political atrocity on Trump, the mob and what comes next*. The New York Times.

Sousa, J. P. (2010). Relembrando o contexto histórico: 1644-1974. In J. P. Sousa (Ed.), *O pensamento jornalístico português: Das origens a Abril de 1974* (pp. 4-56). LabCom Books.

Traquina, N. (2007). *O que é jornalismo*. Quimera.

Varjão, S. (2016). *Violações de direitos na mídia brasileira: Pesquisa detecta quantidade significativa de violações de direitos e infrações a leis no campo da comunicação de massa* (Vol. 3). ANDI – Comunicação e Direitos.

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

Adolfo A. Abadía is professor in the Department of Political Studies at Universidad Icesi (Cali, Colombia), Junior researcher member of the Nexus group (A1-Minciencias). His research interests are electoral studies, subnational politics, and, lately, scholarly communications. He holds a Master's degree and a BA in Political Science from the same University.

Ana Gascón Marcén holds a PhD and is a senior lecturer at the University of Zaragoza where she teaches Public International Law and European Union Law. She was also a civil servant of the Information Society Department of the Council of Europe. Her main research topics are Human Rights and the regulation of Information and Communication Technologies, specially freedom of expression online, personal data protection and intermediary liability.

Caitlin Ring Carlson, Ph.D., is a Professor and Chair of the Communication and Media Department at Seattle University. Her primary research interests are in media law and policy as they pertain to new media, freedom of expression and social justice. Her current work focuses on hate speech. She is also interested in women's freedom of expression, including women's press freedom and women's media ownership. She is the author of the book "Hate Speech," which was published by MIT Press in 2021. Dr. Carlson is also a member of the author team for textbook, The Law of Journalism and Mass Communication (8th ed.), published by Sage in 2023. Her work has been published in leading academic journals such as Communication Law & Policy, the Journal of Media Law and Ethics, and First Amendment Studies.

Daniel Menezes Coelho is a professor, psychologist and psychoanalyst. He teaches both at the Department of Psychology and at the Postgraduate Studies in Psychology at Federal University of Sergipe (UFS). There, he conducts research projects that aim to make use of psychoanalysis, both in its theory and in its approach to clinical processes, to address cultural, social, and political issues. At the same institution, he has a strong presence in the Psychology Department's clinic school, where he frequently serves as coordinator and permanently acts as a clinical supervisor for undergraduate and graduate students. He is a co-organizer of two recently published books: A vida para além das mortes - elaborações psicanalíticas da pandemia (Life beyond so many deaths - psychoanalytic elaborations on the pandemic) and Psicanálise, Gênero, Fronteiras (Psychoanalysis, Gender, Borders), both published by Editora Devires in 2024.

Eric Msughter Aondover, PhD, is a lecturer in the Department of Mass Communication at Caleb University, Imota, Lagos. Aondover is a communication scholar with specialization in Media and Communication Research. He has published papers in several national and international scholarly journals and attended and participated in several conferences and workshops on communication, media, and journalism.

Joelma Galvão de Lemos is a psychologist and psychoanalyst. She has a master's degree in Social Psychology and a PhD in Psychology, both granted by the Postgraduate Studies in Psychology at Federal University of Sergipe (UFS). Her research has evolved from psychoanalysis and its dialogue with political, social, cultural, and economic issues in Brazil, mainly focusing on the analysis of hate speeches published on social media during the impeachment of President Dilma Rousseff and the 2018 Brazilian presidential elections. She also investigates the subjective changes that the use of social media has been causing in our modern society. She teaches at Projeto de Constituição do Campo Psicanalítico de Moçambique, linked to both the Laboratório de Teoria Social, Filosofia e Psicanálise of University of Sao Paulo (LATESFIP-USP) and the Comunidade Psicanalítica de Moçambique (COPSIMO). She also works in private practice, providing psychological counseling, supervision, and clinical case discussions.

Juana L. Rodriguez is BA student in Political Science with an emphasis in international relations at Universidad Icesi (Cali, Colombia). Her interests are electoral studies, political marketing, and international relations studies.

Kaleo Dornaika Guaraty holds a Bachelor's and Master's degree in Law and Development from FDRP – University of São Paulo. He is a member of the Brazilian Academy of Electoral and Political Law and a practicing lawyer.

Luciana C. Manfredi is Associate professor in the Department of Marketing and International Business at Universidad Icesi (Cali-Colombia). Her research interests are political and social marketing and political communication. She holds a Ph.D. in Management from Tulane University, a Master's degree in Management from Tulane University, a Master's degree in Strategic Communication Management Rome Business School, an MBA from Icesi University, and a BA in Political Science from Universidad de Buenos Aires, Argentine.

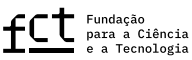
Marco López-Paredes holds a PhD in Communication Sciences with postdoctoral studies. He is an expert in corporate communications and marketing, social networks, and branding. He is a speaker at international events and the Director of the Communication Observatory (OdeCom). He is also a highly productive researcher at the Pontifical Catholic University of Ecuador (PUCE), specializing in the direction and management of scientific journals globally. He is accredited and ranked by the National Secretariat of Higher Education, Science, Technology, and Innovation of Ecuador and the Secretariat of Science of Portugal. He is the author of more than 50 scientific articles and twenty book chapters.

Nosa Owens-Ibie, PhD, is a Professor of Communication, Media and Development in Caleb University, Imota, Lagos, Nigeria. Owens-Ibie has published in scholarly journals both local and international. He coordinated the establishment and is General Secretary of the Association of Communication Scholars & Professionals of Nigeria (ACSPN). A consultant biographer and former newspaper columnist, he has consulted for WHO, UNICEF, UNFPA, IOM, UNESCO, ActionAid Nigeria and other private, government and public entities.

Rubens Beçak holds a Masters and PhD in Constitutional Law and is a Full Professor at Faculty of Law of Ribeirão Preto at University of São Paulo (FDRP-USP). He obtained his undergraduate and postgraduate degrees from the Faculty of Law of São Paulo at the University of São Paulo (FD/USP). Additionally, he serves as an Associate Professor at USP and coordinates the Graduate Program in Constitutional and Electoral Law at USP (lato sensu). He is also a visiting professor at the University of Salamanca (USAL), affiliated with the Master's program in Brazilian Studies, and co-editor of the international publication Revista de Estudos Brasileiros.

Tiago Augustini de Lima graduated in Law from the Faculty of Law of Ribeirão Preto at the University of São Paulo (FDRP/USP). He is currently a Master's candidate in Law at FDRP/USP, a member of the Tutorial Study Program (PET/USP), a member of the Electoral Law study group at FDRP/USP, and a practicing lawyer.

DOI FCT - LABCOM
https://doi.org/10.54499/UIDB/00661/2020



This is the second book of the **Online Hate Speech Trilogy**. The work focuses on the legal challenges of combating toxic language and retaliating against those who spread hate on the Internet. Although the need for fighting violent narratives appears evident, given the role of hate in eroding trust and fragmenting the social fabric, there are many sensitive layers to the matter. The authors analyse the weaknesses of platform self-regulation, the European Union's legal approach to combating online hate, the use of toxic language as a political weapon in Latin America, and the risks it poses to peace in Africa.

LabCom Books
www.labcom.ubi.pt

Editorial Universidad Icesi
www.icesi.edu.co/editorial