

DISCURSO DE ÓDIO EM AMBIENTES DIGITAIS

Definições, especificidades e contexto
da discriminação on-line no Brasil a partir
do Twitter e do Facebook



Rio de Janeiro, Março 2021

Apoio



Embaixada
da República Federal da Alemanha
Brasília

DISCURSO DE ÓDIO EM AMBIENTES DIGITAIS:

Definições, especificidades e contexto da discriminação
on-line no Brasil a partir do Twitter e do Facebook

Rio de Janeiro

FGV DAPP

2021

Dados Internacionais de Catalogação na Publicação (CIP)
Ficha catalográfica elaborada pela Biblioteca Mario Henrique Simonsen/FGV

Discurso de ódio em ambientes digitais : definições, especificidades e contexto da discriminação on-line no Brasil a partir do Twitter e do Facebook / Coordenação Marco Aurélio Ruediger, Amaro Grassi. – Rio de Janeiro : FGV DAPP, 2021.
1 recurso online (60 p.) : PDF

Inclui bibliografia.
ISBN: 978-65-86845-12-9

1. Mídia digital - Censura. 2. Políticas públicas. 3. Discurso de ódio na Internet. 4. Liberdade de expressão. 5. Redes sociais on-line. 6. Discriminação. I. Ruediger, Marco Aurelio, 1959- . II. Grassi, Amaro. III. Fundação Getulio Vargas. Diretoria de Análise de Políticas Públicas.

CDD – 302.23

Como citar

RUEDIGER, M. A.; GRASSI, A. (Coord.). **Discurso de ódio em ambientes digitais : definições, especificidades e contexto da discriminação on-line no Brasil a partir do Twitter e do Facebook**. Policy paper. Rio de Janeiro: FGV DAPP, 2021.

EXPEDIENTE



Fundada em 1944, a Fundação Getulio Vargas nasceu com o objetivo de promover o desenvolvimento socioeconômico do Brasil por meio da formação de administradores qualificados, nas áreas pública e privada. Ao longo do tempo, a FGV ampliou sua atuação para outras áreas do conhecimento, como Ciências Sociais, Direito, Economia, História, Matemática Aplicada e Relações Internacionais, sendo referência em qualidade e excelência, com suas dez escolas.

Edifício Luiz Simões Lopes (Sede)
Praia de Botafogo 190, Rio de Janeiro
RJ - CEP 22250-900
Caixa Postal 62.591 CEP 22257-970
Tel (21) 3799-5498 | www.fgv.br

Primeiro presidente e Fundador

Luiz Simões Lopes

Presidente

Carlos Ivan Simonsen Leal

Vice-Presidentes

Francisco Oswaldo Neves Dornelles (licenciado)
Marcos Cintra Cavalcanti de Albuquerque (licenciado)



Diretor

Marco Aurelio Ruediger

Doutor em Sociologia e Diretor de Análise de Políticas Públicas da Fundação Getulio Vargas (FGV/DAPP). Seus principais campos de interesse são a sociologia política, a comunicação e redes sociais e a inovação tecnológica com seus impactos na democracia. É membro consultivo das iniciativas D4D do National Democratic Institute e do Social Science One. Atualmente está à frente da criação da Escola de Comunicação, Mídia e Informação da Fundação Getulio Vargas e do projeto “Digitalisation and Democracy in Brazil”, realizado com o apoio do Ministério das Relações Exteriores da Alemanha e da Embaixada da Alemanha em Brasília.

E-mail: marco.ruediger@fgv.br

FGV DAPP

(21) 3799-6208

www.dapp.fgv.br | dapp@fgv.br

Coordenação de Pesquisa

Marco Aurelio Ruediger

Amaro Grassi

Pesquisadores

Dalby Hubert

Danielle Sanches

Eurico O. Matos Neto

Luiza C. Santos

Lucas Roberto da Silva

Polyana Sampaio Barbosa

Renata Tomaz

Revisão técnica

Renata Tomaz

Projeto gráfico

Luis Gomes

Daniel Cunha

SUMÁRIO

SUMÁRIO EXECUTIVO	6
SÍNTESE DOS RESULTADOS	6
APRESENTAÇÃO	7
METODOLOGIA	9
CONCEITUAÇÕES E DEFINIÇÕES TEÓRICAS	11
1. O que é discurso de ódio?	11
2. Abordagens e enquadramentos jurídicos sobre discurso de ódio no mundo e no Brasil	15
3. Discurso de ódio em ambientes digitais	19
4. Diretrizes das plataformas digitais sobre discurso de ódio	24
ANÁLISE DO DEBATE SOBRE DISCURSOS DE ÓDIO NO TWITTER E NO FACEBOOK	35
CONCLUSÕES	53
REFERÊNCIAS	56

SUMÁRIO EXECUTIVO

As práticas de discurso de ódio on-line se apresentam como um desafio para as democracias, especialmente por acionar questões em torno de um de seus pilares, a liberdade de expressão. Buscando compreender as dinâmicas de negociação entre segurança de grupos protegidos e garantia da liberdade de expressão, este estudo investiga o debate público a partir do Twitter e do Facebook em torno do discurso de ódio e da censura. Para isso, primeiro apresentamos um panorama geral das definições e legislações que informam a compreensão sobre discurso de ódio, discutindo as especificidades do contexto on-line e as formas de definição e auto-regulação propostas pelas plataformas. Em seguida, a partir de coleta de dados em redes sociais, especificamente o Twitter e o Facebook, analisamos o debate público em torno do discurso de ódio e da censura, considerando especialmente os picos de discussão sobre os temas e os acontecimentos a eles associados. Por fim, identificamos os principais atores e interações ocorridos no pico máximo de debate sobre discurso de ódio e de censura, a partir de grafos que apresentam uma análise estrutural das redes sociais. O estudo utiliza fontes bibliográficas, documentais e dados coletados no Facebook e no Twitter, posteriormente analisados qualitativa e quantitativamente.

Palavras-chave: discurso de ódio; plataformas digitais; censura; liberdade de expressão; redes sociais.

SÍNTESE DOS RESULTADOS

- A revisão teórica combinada com a análise das plataformas revelou que, embora existam diretrizes para sanções de contas que publicam conteúdos ofensivos é muito difícil e complexa a detecção desse tipo de postagem;
- A pesquisa se focou na coleta de dados sobre discurso de ódio e censura, capturando cerca de 11 milhões de postagens no Twitter e 34 mil postagens no Facebook, entre novembro de 2020 a janeiro de 2021;

- No que se refere ao debate sobre discurso de ódio no Twitter e no Facebook, as postagens se concentraram em torno de temas como misoginia e racismo. O pico de postagens sobre o tema, em ambas plataformas, ocorreu em 20 de novembro de 2020, Dia da Consciência Negra. As publicações foram intensificadas pela discussão em torno do racismo estrutural com a notícia da morte por espancamento, no dia anterior, de um consumidor negro no Supermercado Carrefour, em Porto Alegre (RS);
- O discurso sobre censura e liberdade de expressão foi intensificado pelas sanções impostas pelas plataformas em contas, como as do ex-presidente dos Estados Unidos Donald Trump, cujas postagens questionando o resultado das eleições presidenciais foram associadas à invasão do Congresso americano por setores extremistas. A suspensão de contas mobilizou as redes na abordagem sobre a censura nas plataformas em torno da expressão de opinião dos seus usuários, tanto no Twitter como no Facebook.

APRESENTAÇÃO

A manutenção dos regimes democráticos depende, dentre outros fatores, de um equilíbrio saudável entre liberdade de expressão e combate à intolerância, especialmente quando direcionada a grupos minorizados. Tão importante quanto a possibilidade de livre expressão de cada sujeito é a manutenção de um ambiente democrático no qual discursos que representam um ataque à diversidade sejam adequadamente contidos, como os discursos de ódio ou violentos. As possibilidades de expressão pública geradas pelas mídias sociais colocam novos desafios a essa já controversa equação.

Correntes jurídicas diferentes compreendem o papel do Estado na manutenção desse ambiente de formas distintas: países como os Estados Unidos, por exemplo, protegem a liberdade de expressão acima de tudo; já países como a Alemanha colocam limites mais rígidos naquilo que pode ser dito. No contexto on-line, é preciso ainda incluir as posições das próprias plataformas digitais onde muitos desses discursos circulam, que são, em geral, contrárias às regulações estatais e defensoras das auto-regulações.

Considerando a necessidade de um ecossistema onde coexistam liberdade de expressão e segurança dos diversos grupos, a questão dos discursos de ódio é primordial quando falamos de democracia digital. Sua abordagem tem sido crescente tanto em pesquisas mais amplas, como as que investigam a desinformação (BARBOSA et al., 2020), quanto em mapeamentos mais específicos (NERIS et al., 2020), que analisam seus aspectos constitutivos. Neste estudo, apresentamos um panorama geral das definições e legislações em torno do discurso de ódio, discutindo as especificidades do contexto on-line e as formas de definição e auto-regulação propostas pelas plataformas. A partir de coleta de dados em redes sociais, especificamente, no Twitter e no Facebook, analisamos o debate público em torno do discurso de ódio e da censura levando em conta especialmente os picos de menções aos temas e os acontecimentos a eles associados. Realizamos, ainda, uma análise dos principais atores e interações que foram mobilizados a partir dos momentos de maior atividade no debate sobre discurso de ódio e sobre censura, com base em grafos que apresentam uma representação estrutural das redes sociais.

Este é o terceiro *policy paper* lançado por meio do projeto **Digitalização e Democracia no Brasil**, uma parceria entre a Diretoria de Análise de Políticas Públicas da Fundação Getúlio Vargas (FGV DAPP) e a Embaixada da Alemanha no Brasil. Com duração até 2022, o projeto utiliza-se da experiência em pesquisa aplicada de redes sociais e da expertise de análise de políticas públicas da FGV DAPP, construída ao longo da última década.

O objetivo é desenvolver estratégias de enfrentamento e compreensão sobre os novos e iminentes desafios da democracia brasileira — agora representada por um cenário consolidado de extremismo digital e de suas consequências para a sociedade. Nesse sentido, concentra-se, em primeiro lugar, no fortalecimento dos valores democráticos e no desenvolvimento de mecanismos e conhecimentos necessários para promover um debate público digital consistente e profícuo.

METODOLOGIA

Este trabalho utilizou fontes bibliográficas, documentais e dados coletados em plataformas de redes sociais digitais - especificamente o Facebook e o Twitter. A pesquisa bibliográfica informa a primeira e segunda parte deste documento, apresentando o estado da arte da discussão em torno de discurso de ódio em ambientes digitais. A pesquisa documental informa a primeira parte do trabalho, em torno das diretrizes legais vigentes no Brasil, e a segunda parte do trabalho, que aborda especificamente a posição das plataformas em torno do discurso de ódio que ocorre em seu meio. Para o mapeamento dos conceitos de discurso de ódio e diretrizes de comportamento estabelecidas pelas plataformas, foram consultados os Termos de Uso e as Diretrizes de Comunidade do Twitter, Facebook, Instagram e YouTube. Ambos estão disponíveis de forma aberta para consulta em abas específicas do site de cada plataforma, sendo periodicamente atualizados. Além disso, *links* adicionais encontrados nessas páginas também foram consultados sempre que pertinente. Os registros encontrados neste trabalho dizem respeito aos Termos de Uso e Diretrizes de Comunidade em vigor durante o mês de fevereiro de 2021 nas respectivas plataformas.

No que diz respeito à análise de dados em redes sociais que completa este estudo, a abordagem empírica lançou mão de métodos automáticos de coleta e de classificação de postagens em plataformas de redes sociais. Em um primeiro momento, com vistas à coleta dos dados no Twitter, foi elaborada uma lista de termos que pudesse dar conta do escopo temático sobre o qual o estudo se debruça – isto é, discurso de ódio e, eventualmente, liberdade de expressão. Parte dessa coleta foi realizada na aplicação Trendsmap, que permite o monitoramento em tempo real de postagens com maior engajamento no Twitter, em termos tanto de compartilhamentos quanto de comentários e respostas. Além disso, essa lista de termos foi utilizada para filtrar as postagens feitas durante o período de análise – entre 16 de novembro de 2020 e 06 de fevereiro de 2021 –, as quais foram armazenadas em um banco de dados PostgreSQL.

A partir da coleta inicial de dados no Twitter, procedeu-se, então, à classificação dos dados alcançados, com base nas regras linguísticas construídas sobre o tema (RUEDIGER, 2017). Essa etapa constituiu na elaboração de sintaxes de busca orientadas pelos dois tópicos centrais deste estudo, que são "discurso de ódio" e "liberdade de expressão" (e "censura"). Sintaxes de busca equivalem a um conjunto de valores linguísticos, articulados por meio de operadores booleanos (ou lógicos), que se baseia em elementos semântico-pragmáticos com que um dado tópico é geralmente instanciado. Essas sintaxes de busca foram, então, executadas em uma busca textual no banco de dados. Para a coleta e a classificação dos dados no Facebook, por sua vez, o estudo adaptou as sintaxes de busca elaboradas para a classificação dos dados do Twitter – para os dois temas já mencionados, ou seja, "discurso de ódio" e "liberdade de expressão" –, a fim de executá-las no Crowd Tangle, uma aplicação de monitoramento de postagens no Facebook vinculada à própria plataforma.

Em que pese o fato da metodologia aplicada se amparar na construção de regras linguísticas estruturadas, o tema alvo da nossa pesquisa, o debate sobre discurso de ódio, é extremamente complexo e, por essa razão, algumas vezes observamos a presença de postagens que lançaram mão de vocabulário simplificado, mas que tinham significados diversos. Esse desafio fez com que tivéssemos que estabelecer alguns filtros após as coletas realizadas, focando nos picos de menções para analisar o engajamento do debate. A análise qualitativa desses momentos em que a discussão se intensifica se torna, nesse sentido, uma estratégia metodológica para lidar com as limitações dos resultados da extração dos dados a partir da elaboração das sintaxes de busca.

CONCEITUAÇÕES E DEFINIÇÕES TEÓRICAS

1. O que é discurso de ódio?

A definição de discurso de ódio – on-line ou off-line – é extremamente difícil de se ter. Esse tema tem sido muito debatido por acadêmicos, de diferentes áreas de atuação, juristas e legisladores, sem que se chegue a um consenso. Pode-se afirmar que o discurso de ódio é motivado por preconceitos, ofensas, discursos mordazes contra uma determinada pessoa ou um grupo em razão das suas características (COHEN-ALMAGOR, 2011; FARIS et al. 2016). No entanto, Sellars (2016) aponta que, apesar da existência de uma extensa literatura sobre as causas e os efeitos desse tipo de discurso, ainda há uma lacuna sobre a definição e a sistematização do termo.

Pode-se afirmar que uma grande variedade de tópicos pode se enquadrar na definição de discurso de ódio, por exemplo, as calúnias e insultos que são facilmente identificáveis. No entanto, a linguagem tem muitas nuances e nem sempre os conteúdos podem necessariamente ser considerados discurso de ódio pelo locutor ou destinatário alvo. Outro ponto que merece destaque diz respeito à utilização de códigos, principalmente pelas comunidades on-line, para vociferação de ódio, o que dificulta ainda mais a detecção do discurso. A literatura sobre o tema também nos revela a existência de discursos chamados de violentos, aqueles que incitam a violência no ambiente off-line contra grupos vulneráveis (SIEGEL, 2020).

Nesse sentido, as definições existentes de discurso de ódio podem ser extremamente amplas, como as que tratam de uma variedade de discursos dirigidos contra um determinado grupo ou indivíduo com base em suas características físicas ou em seus gestos fora dos padrões normativos estabelecidos (PAREKH et al., 2012). Por outro lado, estão as definições que resultam em prejuízo. As definições restritas acerca do discurso de ódio abordam que ele está diretamente associado ao incitamento à violência em massa ou dano físico contra um grupo externo (BENESCH, 2013). Essa dualidade de recorte (definição

ampla e definição restrita) reflete a dificuldade de se obter uma definição que aborde adequadamente a pluralidade de fenômenos que podem ser considerados discurso de ódio.

Apesar da complexidade inerente ao conceito e dos desafios que a questão coloca, adotamos, neste estudo, a definição proposta pelo “Guia para análise de discurso de ódio”, segundo a qual os discursos de ódio “são manifestações que avaliam negativamente um grupo vulnerável ou um indivíduo enquanto membro de um grupo vulnerável, a fim de estabelecerem que ele é menos digno de direitos, oportunidades ou recursos do que outros grupos e indivíduos membros de outros grupos, e, conseqüentemente, legitimar a prática de discriminação ou violência” (LUCCAS; GOMES; SALVADOR, 2020, p. 4). De acordo com Nandi (2018), os grupos que são alvos de discurso de ódio são comumente os integrantes de minorias e em situação de vulnerabilidade social e econômica, como negros, indígenas, homossexuais, mulheres e minorias religiosas. Ainda segundo o autor, a veiculação do discurso de ódio se relaciona com a intimidação da vítima ou das vítimas em suprir algum direito e por essa razão é direcionado a indivíduos ou grupos que estejam em destaque nos debates acadêmicos, na jurisprudência ou debates legislativos. Para Nandi, o discurso de ódio tem sempre um viés político já que o ataque cumpre a função de deslegitimar pleitos.

Um ponto que merece destaque neste debate são as especificidades do discurso de ódio on-line, pois este envolve instigadores, diferentes tipos de alvos, diversos motivos e táticas. Além disso, existe a questão de os agressores terem seus próprios seguidores e com isso aumentar seu alcance para atingir determinados indivíduos ou grupos. Esse é um ponto que preocupa os acadêmicos no debate sobre o discurso de ódio, pois a fala que incita a violência é diferente quando realizada por um único indivíduo de quando utilizada enquanto uma espécie de “campanha de ódio”, coordenada por uma multidão digital (SELLARS, 2016).

Nesse aspecto, as discussões vêm sempre acompanhadas da relação entre liberdade de expressão versus discursos de ódio. Apesar de se tratar de um debate no campo do Direito,

podemos afirmar que a liberdade de expressão prevista em tratados internacionais e legislações nacionais tem sido cada vez mais interpretada, em diferentes áreas, como um direito não absoluto, mas limitado a outras garantias, havendo responsabilidades e restrições no que se refere às liberdades individuais (LUNA; SANTOS, 2014). Porém, raramente se faz uma vinculação direta entre liberdade de expressão e violência, pois de acordo com Antoine Buyse (2014), é muito difícil concordar sobre onde estão os limites legais e sociais da liberdade de expressão.

O discurso do medo é, portanto, diferente do discurso de ódio. Mais do que dirigir diretamente palavras de ódio a indivíduos pertencentes a determinados grupos, o discurso do medo é endereçado geralmente a membros do próprio grupo social do seu emissor. O objetivo é criar internamente um sentimento de temor contra indivíduos externos àquela comunidade. Por meio deste mecanismo, o medo legitima a violência. A emergência de discursos violentos, portanto, seria ativada pelo sentimento de medo alimentado dentro dos grupos sociais. Antoine Buyse, que é professor de Direitos Humanos na Universidade de Utrecht, lembra que em contextos em que há leis rígidas contra o discurso de ódio, os indivíduos tendem a evitar manifestações diretas de apelo à violência contra determinados grupos sociais e passam a adotar formas mais indiretas de promoção de hostilidade contra determinadas comunidades.

É interessante pontuar que o discurso violento impinge medo em grupos minorizados ou que se pretende atingir, muitas vezes maior, do que a violência em si (violência armada, violência física). O discurso sobre o medo é um recurso utilizado como uma violência preventiva, uma espécie de estabelecimento de posições entre aqueles que imprimem esse discurso e aqueles que são os alvos do mesmo.

Sobre essa questão dos discursos de medo é importante chamar a atenção para o impacto que este exerce sobre um grupo ou uma população inteira. As mensagens que diariamente observamos em canais de mídias (televisivas, impressas e digitais), sobre a atuação de criminosos nas cidades brasileiras ou a existência de ataques orquestrados por grupos terroristas na Europa faz com que o discurso sobre o extermínio desses grupos - a ideia

advinda do senso comum de que “Bandido bom é bandido morto” - muitas vezes seja endossada pela população, sobrepujando-se à ideia de justiça e prisão. De acordo com Anna Simons e John Mueller (2001), isso se deve à incorporação do discurso do medo da ação de um grupo por outro, o que pode levar a reações extremas, fazendo com que violências sejam aceitas em prol do bem estar comum¹.

Outra distinção conceitual significativa é o entendimento do que seja discurso de ódio e do que configura discurso extremista. Segundo relatório publicado pela UNESCO (ALAVA; FRAU-MEIGS; HASSAN, 2017), o uso da violência extrema é mais comum de ser detectado na internet sendo esta um vetor para a proliferação da radicalização de discursos e ideologias extremistas. Todavia, apesar de permitir a propagação desses discursos, de acordo com o mesmo relatório, não há uma relação de causalidade entre discursos radicalistas on-line e atos reais de violência extrema. Esse debate sobre até que ponto postagens e mensagens públicas ou privadas no ambiente virtual são transferidas para práticas sociais cotidianas de determinados grupos nos leva a refletir se caberia às companhias de mídias sociais o combate a esses discursos, moderando o conteúdo das publicações, pressionando o campo jurídico para criação de leis que as permitam regular a atuação on-line dos seus usuários, ou se deveríamos pensar em uma estratégia conjunta de regulação desse tipo de conteúdo.

Por fim, o discurso de ódio não é um fenômeno homogêneo. Antes disso, discursos e práticas discriminatórios contra minorias estão relacionados diretamente aos contextos sociais e políticos em que se expressam. A pluralidade na forma como o discurso de ódio se manifesta em diferentes contextos é um fator que torna ainda mais árdua tarefa de formular estratégias de combate à disseminação de ódio seja nos ambientes on-line ou off-line. Por esse motivo, é mais adequado abordar o fenômeno como “discursos de ódio”, no plural, uma vez que essa pluralidade e estreita relação com o contexto são definidoras de sua prática nos contextos sociais.

¹ O exemplo utilizado aborda questões que afetam toda a sociedade, no entanto, qualquer grupo que se sinta perseguido, oprimido ou afetado por ações de outro grupo pode utilizar o discurso do medo ou incorporá-lo e aceitar o recurso de violências físicas e armadas.

2. Abordagens e enquadramentos jurídicos sobre discurso de ódio no mundo e no Brasil

A noção cada vez mais complexificada de discursos de ódio, incluindo seu uso no plural, é muito tributária dos esforços realizados no campo jurídico. Os tratados internacionais, particularmente posteriores às duas grandes guerras mundiais, têm alimentado as discussões e os debates que cobram cada vez mais dos Estados mecanismos legais para coibir e punir ações motivadas pelo desejo de desqualificar, silenciar, invisibilizar e desumanizar grupos vulneráveis. Embora os tratados internacionais não ofereçam definições do que seja um discurso de ódio, eles estabelecem parâmetros fundamentais para sua identificação e, conseqüentemente, para o combate a seus efeitos diretos e indiretos.

Essas cartas, pactos e declarações se baseiam no princípio da dignidade humana, fortemente evocada após os “sofrimentos indizíveis” da guerra, conforme abertura da Carta das Nações Unidas, documento que funda a Organização das Nações Unidas (ONU, 1945). O texto defende o “respeito aos direitos humanos e às liberdades fundamentais para todos, sem distinção de raça, sexo, língua ou religião”. E é justamente essa ideia de que a condição humana é suficiente para garantir a igualdade de direitos e o gozo das liberdades que fornece base para o entendimento jurídico do combate aos discursos de ódio. A Declaração dos Direitos Humanos (ONU, 1948), amplia o enfrentamento à discriminação nos âmbitos “de opinião política ou outra, de origem nacional ou social, de fortuna, de nascimento ou de qualquer outra situação” (Art. 2º) e avança as linhas de ação “contra qualquer incitamento a tal discriminação” (Art. 7º) ou ataques à “honra e reputação” (Art. 12º), rechaçando também, nesse sentido, os discursos.

Na segunda metade do século XX, a oposição aos crimes de preconceito e aos mecanismos de incitação se intensificaram. A Convenção Internacional sobre a Eliminação de Todas as Formas de Discriminação Racial (BRASIL, 1969), adotada pela ONU em 1965, introduziu o termo “ódio racial” no debate, que logo se estendeu aos âmbitos nacional e religioso no

Pacto Internacional sobre Direitos Civis e Políticos (BRASIL, 1992a), adotado pela ONU em 1966, e na Convenção Americana sobre os Direitos Humanos (BRASIL, 1992b), adotada pela Organização dos Estados Americanos em 1969. Os três documentos defenderam que a apologia ao ódio racial, nacional e religioso se tornasse um crime legalmente punível. Em 1992, o Brasil promulgou tanto o Pacto Internacional quanto a Convenção Americana, conferindo-lhes, assim, força de lei no país.

A Declaração e o Programa de Ação da III Conferência Mundial de Combate ao Racismo, Discriminação Racial, Xenofobia e Intolerância Correlata (2001) demandaram por parte dos países medidas menos punitivas e mais preventivas e protetivas. Reconheceu, ainda, a necessidade de uma multiplicidade de atores no combate ao “ódio” e elencou suas diversas formas de manifestação, bem como seus principais alvos. Destacou o papel das diferentes mídias, especialmente as digitais, na disseminação de ideias que incitam o ódio a grupos vulneráveis, por um lado, e na promoção de ações que o enfrentem, por outro. Além disso, cobrou dos países, no Artigo 147 do Plano de Ação, que, ao garantirem o direito de liberdade de pensamento e de expressão, também adotem e apliquem “com maior abrangência possível, legislação adequada para se ajuizar os responsáveis pelo incitamento ao ódio racial ou à violência através das novas formas de informação e tecnologias de comunicação, incluindo a Internet”.

A liberdade de expressão é reconhecida pelos tratados internacionais - bem como por legislações nacionais, a exemplo da Constituição Brasileira de 1988 - como um direito fundamental para o estabelecimento pleno de regimes democráticos. No entanto, o aumento das possibilidades de disseminação dos discursos de ódio no ambiente digital tem impulsionado, nas primeiras décadas do século XXI, as discussões sobre o papel que o Estado tem de garanti-la, mas também de limitá-la quando ferir princípios como o da dignidade humana. Os contrários a essa intervenção acreditam que intervir na livre manifestação de ideias, incluindo o discurso de ódio, viola tal liberdade (DWORKIN, 2006; 2009), fragilizando com isso os pilares democráticos. Já os defensores de um Estado que imponha limites à liberdade de expressão argumentam que tal ação é necessária no combate à intolerância (WALDRON, 2012), um entrave ao estabelecimento da dignidade

humana e, nesse sentido, ao próprio espírito democrático. A primeira posição costuma estar mais ligada aos teóricos estadunidenses, comumente embasados no princípio da neutralidade do Estado. Os países europeus, que vivenciaram a experiência de duas grandes guerras e do holocausto, tendem a um enquadramento mais rígido quanto aos limites dessa liberdade (BRUGGER, 2007; CONSANI, 2015)

Nesse sentido, não é simples oferecer normas jurídicas que dêem conta das dinâmicas sociais que envolvem os discursos de ódio e dos contextos em que eles emergem. As chamadas *soft laws*, também conhecidas como *soft norms* ou direito brando, funcionam como suporte para superar as limitações identificadas nos tratados internacionais. São encaminhamentos que visam sanar impasses, no âmbito no direito internacional, por exemplo nos casos em que a liberdade de expressão é acionada para proteger o discurso de ódio. É o caso da Relatoria Especial para Liberdade de Expressão, criada pela Organização dos Estados Americanos para dar recomendações a partir do disposto na Convenção Americana de Direitos Humanos.

A legislação brasileira se alinha mais a uma concepção europeia que privilegia a igualdade de direitos e a dignidade humana frente ao direito da liberdade de expressão. No Brasil, é crime a discriminação por raça, cor, etnia, religião e procedência nacional, nos termos da Lei 7.716 (BRASIL, 1989), que também torna puníveis atos divulgadores de nazismo. [O Projeto de Lei 7582/2014](#), que tipifica crimes de ódio e intolerância, amplia o escopo dos grupos alvos para classe e origem social, condição de migrante, refugiado ou deslocado interno, orientação sexual, identidade e expressão de gênero, idade, situação de rua e deficiência. Até que esses critérios sejam incluídos na legislação, casos de discursos de ódio que os envolvam costumam ser julgados à luz de princípios da Constituição Federal (BRASIL, 1988), como a dignidade humana (Art. 1º) e a igualdade perante a lei (Art. 5º), que inclui a igualdade de gênero e a não submissão a tratamento desumano ou degradante (Art. 5º).

É importante registrar que o Código Penal (BRASIL, 1940) tipifica o crime de injúria, agravado pela Lei 10.741 (BRASIL, 2003) quando consta de “elementos referentes a raça,

cor, etnia, religião, origem ou a condição de pessoa idosa ou portadora de deficiência”. Outro dispositivo legal importante no enfrentamento aos discursos de ódio no Brasil é a Lei 12.735 (BRASIL, 2012). Ela expande a legislação que define crimes de preconceito e discriminação, na medida em que inclui “condutas realizadas mediante uso de sistema eletrônico, digital ou similares”, indicando o esforço de tornar a legislação vigente aplicável aos contextos digitais.

Em se tratando de ambiente virtual, o mais importante dispositivo legal brasileiro é o Marco Civil da Internet, que entrou em vigor com a aprovação da Lei 12.965 (BRASIL, 2014). Embora o texto afirme no Artigo 2º que “A disciplina do uso da internet no Brasil tem como fundamento o respeito à liberdade de expressão”, o parágrafo único do Artigo 3º esclarece que “Os princípios expressos nesta Lei não excluem outros previstos no ordenamento jurídico pátrio relacionados à matéria ou nos tratados internacionais em que a República Federativa do Brasil seja parte”. O Brasil, nesse sentido, se insere em um entendimento sócio-jurídico mais amplo de que a garantia da liberdade de expressão não dá base para o livre discurso de ódio.

Mais recentemente, começou a tramitar no Congresso o [Projeto de Lei 2.630, de 3 de julho de 2020](#), que institui a Lei Brasileira de Liberdade, Responsabilidade e Transparência na Internet - a chamada “Lei das Fake News”. O projeto, de autoria do senador Alessandro Vieira (Cidadania-SE), foi aprovado pelo Senado em 2020 e segue em discussão na Câmara dos Deputados. Dentre as alterações propostas ao texto original, algumas se relacionam com o tema do discurso de ódio, como a que exige a publicação, pelas plataformas, de relatórios trimestrais de acesso público sobre as contas e conteúdos removidos e a permissão para exclusão de conteúdo sem notificação do usuário em casos de incitação à violência e dano imediato de difícil reparação². A proposta vem recebendo críticas

² Até o momento, no Brasil, o Marco Civil da Internet (BRASIL, 2014) permite que a remoção de conteúdo seja feita pelas plataformas sempre que uma publicação violar as diretrizes de comunidade, assinadas pelos usuários no momento da criação da conta. Tudo aquilo que não estiver contemplado nas diretrizes não pode ser removido, mesmo mediante denúncia, exceto quando exista uma ordem judicial. Na Alemanha, por exemplo, uma polêmica legislação específica para a administração de conteúdos ilegais no contexto on-line, a chamada Network Enforcement Act, dispõe que é dever das plataformas digitais a remoção de conteúdo de ódio em um período de 24 horas, correndo o risco de elevadas multas (SILVA et al., 2019).

especialmente em relação ao estabelecimento de mecanismos de vigilância e aos possíveis efeitos para o ecossistema da internet, como discutidos pelo [Coletivo Intervenções](#) e a [Revista Piauí](#).

As abordagens e enquadramentos jurídicos assinalados acima, portanto, acumulam entendimentos caros para o enfrentamento da produção e disseminação de discursos de ódio nas interações em rede do contexto brasileiro. Prevalece nos mecanismos legais vigentes e propostos a ideia de que a liberdade de expressão não é um direito absoluto e, desse modo, deve ser garantida em consonância com direitos fundamentais, amparados por princípios constitucionais como o da dignidade humana.

3. Discurso de ódio em ambientes digitais

A questão do discurso de ódio é muito anterior à internet, porém, a proliferação das mídias digitais e o uso intensivo de redes sociais on-line trazem questões específicas para sua discussão e combate. Até pouco tempo as práticas de discurso de ódio on-line eram consideradas uma atividade de nicho, entretanto, nos últimos anos, sua proeminência e presença em espaços *mainstream* da internet³ tornam esse tema cada vez mais visível. Intensificam esse quadro as crescentes consequências off-line de ações coordenadas de discurso de ódio on-line, que se concretizam em ataques (SIEGEL, 2020) como [o caso brasileiro de Suzano](#), na região metropolitana de São Paulo⁴. Nesta seção, debatemos como o discurso de ódio ganha novas nuances quando se manifesta em ambientes

³ Entendemos como plataformas digitais *mainstream* aquelas que são mais utilizadas pelos brasileiros, não se caracterizando como nichos muito restritos de interesse ou mesmo como plataformas mais utilizadas pelos grupos de ódio como forma de organização (como os chans). Em pesquisa da We are Social com a Hootsuite, figuram entre as dez redes sociais mais utilizadas pelos brasileiros em 2020, em ordem decrescente de usuários, Facebook, Whatsapp, YouTube, Instagram, Facebook Messenger, LinkedIn, Pinterest, Twitter, TikTok e Snapchat.

⁴ Massacre ocorrido em 13 de março de 2019 na Escola Estadual Professor Raul Brasil, na cidade de Suzano, São Paulo. Dois atiradores, Guilherme Monteiro e Luiza Henrique de Castro, ex-alunos da escola, mataram cinco alunos e dois funcionários, o tio de um deles. Na sequência, um deles matou o outro e se suicidou. Ambos eram usuários de um chan - um tipo de fórum anônimo - onde discutiram e receberam orientações sobre o ataque. O envolvimento das comunidades de ódio presentes nos chans e seus impactos em situações reais foi amplamente discutido na imprensa brasileira.

digitais. Para tanto, iniciaremos apontando algumas características dos ambientes digitais que podem contribuir para a prática e a circulação de discursos de ódio, considerando as especificidades das mídias digitais e as formas de organização humana a partir das mesmas. A estrutura das redes sociais coloca dificuldades para a restrição da propagação dessas narrativas, um desafio que se apresenta tanto para as legislações nacionais quanto para as auto-regulações das plataformas, tópico a ser abordado na seção seguinte.

Para que a questão possa ser endereçada adequadamente, precisamos primeiro compreender no que o discurso de ódio on-line se difere do off-line. Revisões sistemáticas da literatura apontam que algumas características das mídias digitais são componentes importantes de sua proliferação, nomeadamente: a anonimidade, a invisibilidade, a criação de comunidades por afinidades (inclusive de ódio) sem barreiras geográficas, baixo custo de tempo e dinheiro para veiculação desse tipo de discurso e a instantaneidade que os meios digitais possibilitam (BROWN, 2018).

A combinação desses fatores tanto potencializa a discriminação on-line, ofertando novas possibilidades para comunidades movidas pelo ódio, quanto explica parcialmente a presença desses discursos nesse contexto. A anonimidade dos usuários, ainda que parcial (quando não falamos de *deep web*), se expressa especialmente na remoção de barreiras de responsabilização pelos atos de ódio no contexto on-line e na diminuição da possibilidade de reação ou confronto físico entre agressor e vítima. Ainda, a invisibilidade expressa pela não presença visual do agressor e da vítima torna os ataques mais fáceis de serem efetivados, uma vez que os efeitos dos mesmos na vítima não são visíveis para quem o realiza. Assim, o discurso de ódio on-line, por sua invisibilidade e anonimidade parcial *pode parecer* menos real, com menos implicações do que de fato possui (BROWN, 2018).

O papel da internet na criação de comunidades por afinidade, que cultivam sentimentos de pertencimento em torno de características ou interesses comuns de sujeitos de origem geograficamente distinta, também gera implicações na proliferação do discurso de ódio. A facilidade de acesso a recursos de comunicação digital, que economiza tempo e dinheiro tanto na organização quanto na proliferação dos discursos de ódio também é um fato

considerável (BROWN, 2018). Grupos de ódio utilizam a internet como forma sistemática de recrutamento e ampliação de colaboradores, cultivando comunidades e valores que possibilitam o reforço de identidades no contexto digital (WEAVER, 2013). A partir desses espaços é possível criar e alimentar o sentimento de comunidade entre os envolvidos, promovendo o estreitamento dos laços e das identidades, com o potencial de aumento da projeção dos grupos de ódio a partir do engajamento dos seus membros (BOWMAN-GRIEVE, 2009). Esse engajamento pode ser visualizado nas redes sociais nas já mencionadas campanhas de ódio, por exemplo, a partir da coordenação de esforços de seus membros para manifestação em torno de um mesmo tópico ou ataque de um determinado ator na rede, de forma simultânea.

Por fim, a possibilidade de ação conforme nossos impulsos também cria um tipo de proliferação de discurso de ódio que ocorre por ocasião, sem tempo de pensar efetivamente naquela ação. Esse tempo, na mídia tradicional, é geralmente dado pelos custos de transação da produção e circulação de discursos deste tipo, como produzir, imprimir e circular um panfleto. Atualmente, basta pegar o celular, abrir o aplicativo do Twitter, escrever uma mensagem de ódio e apertar enviar. Assim, o modo de funcionamento da internet encoraja o tipo de discurso de ódio espontâneo (BROWN, 2018).

Outra característica das plataformas digitais que pode contribuir para a circulação de discursos de ódio é sua lógica de funcionamento baseado em algoritmos cuja gestão da visibilidade dos conteúdos está relacionada diretamente às preferências dos usuários (GILLESPIE, 2018) - é o que alguns autores chamam de “câmara de eco” (“echo-chambers”)⁵. Ao priorizar nas *timelines* conteúdos consoantes à opinião do indivíduo, os algoritmos podem também produzir uma percepção distorcida acerca dos cenários sociais, além de induzir a polarização por meio da criação de bolhas ideológicas. Os modos de funcionamento dos algoritmos que constroem esse regimes personalizados de visibilidade de conteúdos não são plenamente esclarecidos pelas plataformas digitais,

⁵ Uma perspectiva contrária a esta é composta por autores que defendem a ideia da exposição seletiva e exposição inadvertida. Ou seja, quando os indivíduos nas mídias sociais são expostos a conteúdos contrários aos seus pontos de vista. Ver mais deste debate em Barberá (2015).

constituindo-se como sistemas opacos para a maior parte dos usuários (JURNO; D'ANDRÉA, 2017).

Ao agregar pessoas com pensamentos, posições e opiniões semelhantes, a lógica algorítmica das redes sociais pode, segundo este argumento, induzir os usuários a uma leitura enviesada do clima de opinião predominante na sociedade. Se, por meio das minhas redes sociais, vejo que a maioria das pessoas aceita e, por vezes, recompensa determinados tipos de comportamentos sociais, tendo a imaginar que aquele comportamento pode ser considerado como padrão socialmente adequado. Nessa lógica, ao ser exposto a discursos de ódio em sua *timeline*, há um risco de que esse tipo de conteúdo seja interpretado como socialmente aceito em uma dada comunidade. As pessoas, portanto, tenderiam a se sentir confortáveis em emitir discursos de ódio quando percebem que esse tipo de comportamento nas redes pode gerar recompensas sociais naquele ambiente (curtidas, compartilhamentos, interação etc).

De forma oposta ao controle de conteúdo que ocorre nos meios de comunicação de massa, onde a checagem acontece antes da veiculação, nas plataformas digitais os discursos dos usuários são checados apenas após a sua publicização e, de forma geral, mediante denúncia de usuários. Isso ocorre também por especificidades do contexto digital: o volume de conteúdo massivo produzido por milhares de usuários cotidianamente torna a checagem prévia um impeditivo. Além disso, a noção de liberdade de expressão permanece como valor principal das plataformas digitais, como veremos na seção seguinte. Decorre disso que o discurso de ódio em plataformas on-line *mainstream* acontece primeiro e é combatido apenas depois de sua circulação. Já no contexto off-line, o discurso de ódio tende a ficar fora do circuito da grande mídia, circulando apenas de forma marginal.

Outra diferença importante entre as formas de propagação em plataformas digitais e em empresas de mídia tradicional é que, no primeiro caso, as empresas não são as produtoras do conteúdo que circula no ambiente, enquanto no segundo caso, sim. Por esse motivo, não é possível apenas transpor ou comparar as formas de funcionamento, de controle e de

combate ao discurso de ódio nos dois contextos. Nesse sentido, as responsabilidades também não são análogas. Entretanto, mesmo que as plataformas digitais não produzam conteúdo discursivo, elas criam o ambiente onde essas mensagens são difundidas e, principalmente, obtêm ganhos financeiros com a interação e a atenção dos usuários em função delas. Por isso, as plataformas digitais também não são como espaços públicos comuns, uma vez que são empresas e, por isso, locais privados em certa medida responsáveis pelo que ocorre naquele espaço. Esse *status* que não é nem de produtor de conteúdo no sentido tradicional das mídias e nem de espaço totalmente público de discussão, torna as soluções para o combate ao discurso de ódio (entres outros tópicos) mais complexas.

Pesquisas anteriores mostram que a proporção entre conteúdo de discurso de ódio em plataformas *mainstream* é baixa quando comparada ao total de conteúdos circulando nas mesmas plataformas. Entretanto, as estratégias de proliferação desse discurso geram visibilidade, fazendo com que grande parte dos indivíduos on-line já tenha sido exposta a esse tipo de conteúdo. Comunidades organizadas para ataques baseados em discurso de ódio utilizam como alvo principalmente pessoas públicas, tais como artistas, políticos e jornalistas, além de atuarem de forma coordenada, amplificando o alcance (SIEGEL, 2020).

Atualmente, entretanto, revisões sistemáticas de literatura apontam que a maior parte dos estudos em torno do tema do discurso de ódio on-line se debruçam sobre as comunidades organizadas em plataformas de nicho. Por isso, nosso conhecimento sobre os atores que proliferam discurso de ódio de maneira informal, especialmente em plataformas amplamente utilizadas, ainda é limitado. Visando contribuir com esta lacuna, este relatório apresenta um estudo temporal no contexto brasileiro do discurso de ódio nas plataformas Twitter e Facebook. Segundo dados de [pesquisa realizada pela We are social e Hootsuit](#) em 2020, o Facebook possui 130 milhões de usuários brasileiros e o Twitter possui 16,6 milhões. Conjuntamente com YouTube e Instagram, compõem os principais locais de exposição de ideias de forma aberta e debate público no contexto digital brasileiro.

É importante delimitar a diferença entre definição de discurso de ódio e categorização dos tipos de discurso de ódio. A definição está mais próxima de conceito, e, por isso, é abstrata, podendo abarcar diversos contextos específicos. Já a classificação é justamente o processo de determinar o que é e o que não é um discurso de ódio a partir de uma ocorrência real, baseando-se na definição conceitual (seja jurídica, acadêmica ou operacional das plataformas). Assim, uma definição que pode parecer conceitualmente clara sobre discurso de ódio, pode se tornar obscura em situações concretas, que dependem de contexto, de usos de linguagem, de formas de agir de determinadas culturas e inclusive de apropriações e significados linguísticos específicos de comunidades.

É essa dinâmica situacional que impõe dificuldades práticas para a detecção e combate ao discurso de ódio nas plataformas, muito mais do que a conceituação. Em plataformas digitais como o Twitter e Facebook, os sujeitos são alvos de discurso de ódio prioritariamente em função de características específicas, como orientação sexual, classe, gênero, etnia e características físicas (SILVA et al., 2016). Algumas dessas características serão entendidas pelas plataformas como características a serem protegidas, em consonância com os marcos legais, ainda que nem sempre essas características sejam exatamente as mesmas.

4. Diretrizes das plataformas digitais sobre discurso de ódio

Quando no contexto on-line, o discurso de ódio apresenta algumas especificidades, como aquelas discutidas na seção anterior. Além das particularidades decorrentes do próprio meio de comunicação, as interações em cada plataforma também são moldadas a partir dos Termos de Uso que as mesmas possuem. Os Termos de Uso de um serviço on-line são documentos que dispõem sobre as formas de funcionamento das plataformas e sobre as regras de utilização às quais os utilizadores estão sujeitos. Quando utilizamos um serviço, estamos sujeitos e concordamos com seus termos de uso – mesmo que na maior parte das

vezes os usuários não leiam esse contrato de serviço⁶. Plataformas de mídias sociais em geral apontam nos Termos de Uso que os usuários devem seguir as Diretrizes de Comunidade - e que, caso não sigam, algumas penalidades podem ser aplicadas. Por isso, na sequência iremos descrever, comparar e analisar as diretrizes de comunidade das plataformas digitais mais utilizadas no Brasil, onde constam as definições das próprias plataformas para o discurso de ódio.

Os Termos de Uso e as Diretrizes de Comunidade - conjunto de regras que determinam o comportamento esperado dos usuários - são documentos que se modificam de plataforma para plataforma. No que diz respeito ao discurso de ódio, aos discursos perigosos ou violentos, é justamente nesse documento que encontramos o que uma rede social entende como cada um desses conceitos, a sua tolerância em relação a eles e a forma como negociam as expectativas de liberdade de expressão dos usuários e a segurança dos mesmos, especialmente daqueles presentes nas chamadas categorias protegidas. Ainda que por vezes se baseiam em pesquisas científicas para propor medidas em torno do assunto, como indicam o Facebook e o Twitter, não existe um entendimento comum ou padronizado entre elas.

Apesar dos esforços específicos de criação de formas automatizadas para a detecção de discurso de ódio em redes sociais, o funcionamento prioritário indicado pelas plataformas analisadas é através de denúncias de outros usuários. Ou seja, ao se deparar com uma publicação que viole em algum aspecto as Diretrizes de Comunidade, o usuário pode denunciar o conteúdo, que será posteriormente analisado de forma contextual por moderadores. Twitter, Facebook e YouTube são signatários, desde 2013, do acordo de combate ao discurso de ódio liderado pela Liga Anti-difamação, uma organização sem fins lucrativos dos Estados Unidos. A partir do documento [Best Practices for Responding to Cyberhate](#), as três plataformas digitais se comprometem com algumas práticas, entre elas: analisar de forma comprometida as denúncias e relatos de discurso de ódio em tempo

⁶ O consentimento em ambientes digitais é baseado no princípio "Notice and Choice", ou seja, os serviços notificam os usuários sobre os potenciais riscos envolvidos no uso da plataforma e os usuários manifestam sua aprovação e consentimento clicando em um "Eu aceito". Mais sobre este princípio e suas implicações para o debate sobre privacidade em ambientes digitais, ver Sloan e Warner (2014).

hábil; explicar de forma clara como realizam a moderação de conteúdo para seus usuários e aplicarem as sanções previstas de forma consistente e justa; ofertar formas simplificadas de denúncia de conteúdo de ódio (SILVA et al., 2019).

Esse acordo é um dos impulsionadores das modificações e melhorias implementadas pelas plataformas digitais desde 2015 no combate ao conteúdo de ódio: termos de uso e diretrizes de comunidades mais claros, implementação de relatórios de moderação de conteúdo e desenvolvimento de técnicas automatizadas e pró-ativas de detecção de discurso de ódio são algumas das mudanças observadas ao longo do tempo no tratamento desse tema por parte das plataformas digitais. Outros fatores também contribuíram para essas melhorias, tais como: novas leis sobre conteúdo de ódio on-line (como o caso da Alemanha), vazamentos de dados que se tornaram públicos, ataques terroristas movidos por aglutinações on-line e interferência de governos (SILVA et al., 2019).

Na sequência iremos apresentar um resumo das Diretrizes de Comunidade atuais do [Facebook](#), [Twitter](#), [Instagram](#) e [YouTube](#) para compreender o nível de detalhe nas conceituações e as formas de lidar com o discurso de ódio e suas variáveis no contexto on-line. Apesar dos esforços legislativos, tanto nacionais quanto internacionais, e das Diretrizes de Comunidade e Termos de Uso das plataformas, as relações entre o direito de expressão dos sujeitos e os discursos de ódios não são sempre claras e esbarram em entendimentos diversos, conforme discutido anteriormente. O equilíbrio dessa relação é uma das principais questões apresentadas pelas plataformas ao abordar o assunto: como garantir a segurança de categorias protegidas interferindo o mínimo possível na liberdade de expressão dos usuários?

Além disso, as dificuldades impostas pela contextualidade inerente ao discurso de ódio também são uma questão levantada por Twitter, Facebook, Instagram e YouTube em suas páginas: palavras historicamente associadas a um sentido degradante podem ser ressignificadas por grupos sociais, novas formas de incitação ou agressão verbal surgem,

assim como formas decifradas. Ou seja, a cultura é um elemento importante na detecção correta de discursos deste tipo.

Visão Geral

Facebook, Twitter, Instagram e YouTube estão entre as dez principais plataformas digitais utilizadas por brasileiros para produção e consumo de conteúdo. Ao investigarmos os Termos de Uso e Diretrizes de Comunidade de cada uma delas, percebemos que as quatro mantêm em suas diretrizes algum nível de restrição ou repúdio contra o discurso de ódio. Em alguns casos, inclui-se também diretrizes contra discursos violentos, extremistas ou perigosos. As plataformas variam, entretanto, no nível de detalhamento, conceituação dos termos e exemplificação das diretrizes que fornecem.

Nesse sentido, o Instagram se apresenta como a plataforma que menos detalha esses aspectos, indicando [Diretrizes de Comunidade](#) mais sucintas e genéricas. Entretanto, informa aos usuários que também estão sujeitos às políticas do Facebook, uma vez que as plataformas fazem parte de um mesmo grupo. A falta de especificidade é um alerta para a dificuldade ainda maior de uma moderação que é delicada pelos desafios do contexto e da liberdade de expressão.

Outro ponto em comum às plataformas analisadas é a menção às características ou categorias protegidas e grupos específicos, que podem ser elencados ou não. No caso do Instagram, não são delimitados. Facebook, Twitter e YouTube fornecem uma lista das categorias ou grupos que estão inclusos como protegidos, assinalando variações de: idade, gênero, orientação sexual, etnia, raça, religião e situação de imigração. Algumas, entretanto, não são consenso e aparecem exclusivamente em uma ou duas das plataformas. O Facebook e o Twitter incluem características físicas ou doenças, o Twitter fala de comunidades marginalizadas e historicamente sub-representadas, e o YouTube considera também classe social e veteranos de guerra como categorias e grupos protegidos. Observa-se que três das características protegidas apresentadas pelas

plataformas como consenso se relacionam com aquelas associadas aos grupos mais propensos a serem vítimas de discurso de ódio segundo a literatura: orientação sexual, gênero e etnia (SILVA et al., 2016). Contudo, duas das características listadas pela literatura como mais vulneráveis compõem as diretrizes de apenas uma das quatro plataformas: traços físicos e classe social (SILVA et al., 2016).

Pontuamos ainda que das quatro plataformas analisadas, apenas o YouTube informa em suas [Diretrizes de Comunidade](#) que, além das denúncias de conteúdo inadequado por parte de usuários, também realiza monitoramento próprio através de um sistema de sinalização automática para detecção de conteúdos que violam as diretrizes da comunidade. É possível que essa maior preocupação do YouTube em verificar os conteúdos postados esteja relacionada à monetização de vídeos, já que as publicações com conteúdo ofensivo não são passíveis de serem monetizadas pelo ambiente. As demais plataformas informam apenas que analisam os conteúdos sinalizados por usuários de forma contextual, sem indicar outras iniciativas de controle e monitoramento no espaço dedicado às Diretrizes de Comunidade. No entanto, sabe-se que tanto o Facebook quanto o Twitter utilizam mecanismos automatizados na detecção de conteúdos que violem suas diretrizes de comunidade (SILVA et al., 2019).

Na sequência, detalharemos as Diretrizes de Comunidade do Facebook e do Twitter no que diz respeito ao discurso de ódio e tópicos relacionados, uma vez que essas plataformas são os ambientes nos quais realizamos as coletas de dados que informam as análises presentes nesse documento.

f	Facebook
O Facebook ressalta nas Diretrizes de Comunidade que o “comprometimento com a expressão é uma prioridade”, sinalizando um alinhamento maior com a perspectiva americana de preservação de liberdade de expressão. É justamente a partir do argumento da liberdade de expressão que justifica suas diretrizes em relação ao	

discurso de ódio, no sentido de “garantir que todas as vozes sejam valorizadas”, incluindo as chamadas categorias protegidas.

Define discurso de ódio como “ataque direto a pessoas baseado no que chamamos de características protegidas: *raça, etnia, nacionalidade, religião, orientação sexual, casta, sexo, gênero, identidade de gênero e doença grave ou deficiência*”. Na sequência, o texto dá enquadramento ao termo ataque, como “discursos violentos ou degradantes, estereótipos prejudiciais, declarações de inferioridade, expressões de desprezo, repugnância ou rejeição, xingamentos e apelos à exclusão ou segregação”. Nesse sentido, a plataforma consegue fornecer um entendimento claro de discurso de ódio, ainda que, na prática, a avaliação desse entendimento pelos moderadores dependa de fatores contextuais e eventualmente até pessoais. Ao apontar as características protegidas, aponta também os grupos mais suscetíveis a esse tipo de ataque e o foco de atenção na moderação.

As diretrizes apontam três níveis diferentes para o discurso de ódio: o nível 1 inclui discurso violento, comparações degradantes, deboche e negacionismo de situações históricas referentes às categorias protegidas; o nível 2 é constituído por afirmações de inferioridade e declarações desmerecedoras das categorias protegidas; e o nível 3 condutas de segregação ou incentivo à exclusão de categorias protegidas. Em uma [aba separada](#), o Facebook também indica não tolerar exaltação e celebração de violência e conteúdo explícito em qualquer formato.



Twitter

O Facebook ressalta nas Diretrizes de Comunidade que o “comprometimento com a expressão é uma prioridade”, sinalizando um alinhamento maior com a perspectiva americana de preservação de liberdade de expressão. É justamente a partir do argumento da liberdade de expressão que justifica suas diretrizes em relação ao

discurso de ódio, no sentido de “garantir que todas as vozes sejam valorizadas”, incluindo as chamadas categorias protegidas.

Define discurso de ódio como “ataque direto a pessoas baseado no que chamamos de características protegidas: *raça, etnia, nacionalidade, religião, orientação sexual, casta, sexo, gênero, identidade de gênero e doença grave ou deficiência*”. Na sequência, o texto dá enquadramento ao termo ataque, como “discursos violentos ou degradantes, estereótipos prejudiciais, declarações de inferioridade, expressões de desprezo, repugnância ou rejeição, xingamentos e apelos à exclusão ou segregação”. Nesse sentido, a plataforma consegue fornecer um entendimento claro de discurso de ódio, ainda que, na prática, a avaliação desse entendimento pelos moderadores dependa de fatores contextuais e eventualmente até pessoais. Ao apontar as características protegidas, aponta também os grupos mais suscetíveis a esse tipo de ataque e o foco de atenção na moderação.

As diretrizes apontam três níveis diferentes para o discurso de ódio: o nível 1 inclui discurso violento, comparações degradantes, deboche e negacionismo de situações históricas referentes às categorias protegidas; o nível 2 é constituído por afirmações de inferioridade e declarações desmerecedoras das categorias protegidas; e o nível 3 condutas de segregação ou incentivo à exclusão de categorias protegidas. Em uma [aba separada](#), o Facebook também indica não tolerar exaltação e celebração de violência e conteúdo explícito em qualquer formato.

O Twitter dispõe sobre os comportamentos proibidos no uso da plataforma na seção “Regras e Políticas”, indicando três categorias de interesse para este estudo: “política contra glorificação da violência”, “política contra propagação de ódio” e a “política contra ameaças violentas”. Dentre as quatro redes sociais, é aquela que oferece um documento mais detalhado em torno do que é ou não permitido e dos entendimentos do Twitter em torno do tema.

As [Políticas contra glorificação da violência](#) esclarecem que o usuário “não pode glorificar, comemorar, elogiar ou tolerar crimes e eventos violentos em que pessoas foram vítimas por pertencerem a um grupo protegido ou os autores de tais atos”, seguido de uma definição do que entende por glorificação e uma lista de exemplos dessas ações. Essa categoria da plataforma se associa diretamente com o que discutimos em torno dos discursos perigosos.

Já na [Política contra a propagação de ódio](#), o Twitter foca seus esforços em grupos que são mais suscetíveis sofrer assédio on-line, especialmente quando possuem características interseccionais: “mulheres, negros, lésbicas, gays, bissexuais, transexuais, homossexuais, intersetais, indivíduos assexuados, comunidades marginalizadas e historicamente sub-representadas”. A política determina que “não é permitido promover violência, atacar diretamente ou ameaçar outras pessoas” com base nas características protegidas. Também não é permitido imagens de propagação de ódio, tais como aquelas associadas historicamente a grupos de ódio, ou conteúdo que faça referência a ataques violentos de grupos protegidos.

A [Política contra ameaças violentas](#) explica que o usuário do Twitter “não pode ameaçar de violência um indivíduo ou um grupo de pessoas”, especialmente no formato de declaração de intenção, incluindo violência sexual. De acordo com os documentos, a plataforma avalia algumas dessas expressões, incluindo variantes de “eu vou te matar”, de forma contextual, para exclusão de figuras de linguagem.

Interesse público

As quatro plataformas admitem que, em circunstâncias especiais de interesse público ou função de conscientização e educação sobre o tema, podem permitir que conteúdos em desacordo com os padrões da comunidade em torno do discurso de ódio permaneçam no

ar, mesmo que tenham recebido denúncias. Os exemplos utilizados são postagens que visam à conscientização pública em torno de uma questão e à discussão pública sobre o tema, desde que a intenção do usuário, nesse sentido, esteja clara.

Sanções por violação das diretrizes

Nenhuma das quatro referidas plataformas expõe em detalhes os procedimentos em torno das sanções aplicadas para os usuários que publicam conteúdos de discurso de ódio e afins de forma clara ou transparente. Elas são vagas ao pormenorizar o que acontece com os que violam as diretrizes de uso: Instagram é a plataforma que menos detalha sanções e procedimentos, e o YouTube é o que tende a oferecer mais elementos a esse respeito. Essa transparência, claro, é parcial, uma vez que a plataforma é quem possui acesso irrestrito aos dados que informam tanto suas decisões quanto seus relatórios.

O Facebook afirma que analisa a gravidade e o histórico do usuário na plataforma para tomar uma decisão, que pode variar desde notificação, restrições de uso da plataforma até desativação completa do perfil. A intenção do usuário faz parte da avaliação, tanto para a remoção do conteúdo quanto para a sanção. Em caso de risco real de danos físicos ou de segurança pública, as autoridades são notificadas.

O Twitter diz levar em conta apenas o histórico dos usuários na própria plataforma, indicando que uma primeira infração leva à remoção do conteúdo e potencialmente ao impedimento de uso temporário da conta. Infrações persistentes levam à suspensão permanente do perfil – mas o Twitter não fornece o número de repetições para que isso aconteça.

O Instagram informa que a penalidade de infração de suas diretrizes é a remoção do conteúdo e eventual cancelamento da conta do usuário, sem especificação de qualquer critério utilizado como base para essa tomada de decisão. Já o YouTube delimita que a

primeira violação de Diretrizes da Comunidade gera apenas uma sinalização, sem nenhuma penalidade. As violações seguintes geram exclusão do conteúdo (vídeo ou comentário) além de uma notificação, que também [limita as ações](#) do usuário na plataforma pelo período de uma ou duas semanas. Após três notificações em um período de 90 dias, o canal é encerrado e todos os seus vídeos deletados.

Apesar de não explicitar o que leva em conta na hora de analisar as denúncias, a partir do detalhamento dos procedimentos fica evidente que a plataforma considera a gravidade e a frequência das infrações na hora de aplicar penalidades. O YouTube também mantém um [relatório de transparência](#) de fácil acesso sobre suas sanções em torno das violações das diretrizes de comunidade. De acordo com o relatório, entre outubro de 2020 e dezembro de 2020, foram 2.055.515 canais removidos. Dessas remoções, o segundo principal motivo apontado foi “conteúdo abusivo ou incitação ao ódio”, com 8,3% do total. O primeiro motivo, com mais de 75% do total de remoções, foi por *spam*, conteúdo enganoso ou golpes. O quadro abaixo sistematiza os principais achados da investigação comparativa entre as diretrizes de comunidade do Facebook, Twitter, Instagram e YouTube.

Tabela 1 - Quadro comparativo das Diretrizes de Comunidade do Facebook, Twitter, Instagram e YouTube

Quadro Comparativo - Diretrizes de Comunidade				
	Facebook	Twitter	Instagram	YouTube
Diretrizes sobre discurso de ódio	Sim	Sim	Sim	Sim
Definição de discurso de ódio	“[...]ataque direto a pessoas baseado no que chamamos de características protegidas.”	“[...]promover violência, atacar diretamente ou ameaçar outras pessoas com base” nas categorias protegidas.	Não apresentada.	“[...] conteúdo que promove violência ou ódio contra indivíduos ou grupos com base em qualquer um dos seguintes atributos” protegidos.

Categorias e grupos protegidos	Raça, etnia, nacionalidade, religião, orientação sexual, casta, sexo, gênero, identidade de gênero e doença grave ou deficiência.	Categorias: raça, etnia, origem nacional, orientação sexual, sexo, identidade de gênero, religião, idade, deficiência ou doença grave. Grupos: mulheres, negros, lésbicas, gays, bissexuais, transexuais, homossexuais, intersexuais, indivíduos assexuados, comunidades marginalizadas e historicamente sub-representadas.	Não define categorias ou grupos como protegidos, mas informa que “nunca é aceitável” o “ataque a alguém com base em raça, etnia, nacionalidade, sexo, gênero, identidade de gênero, orientação sexual, religião, deficiências ou doenças”.	Idade, classe social, deficiência, etnia, identidade e expressão de gênero, nacionalidade, raça, situação de imigração, religião, sexo/gênero, orientação sexual, vítimas de um evento violento em grande escala e os familiares dessas pessoas, veteranos de guerra.
Crítérios de avaliação de denúncias	Gravidade, intenção e histórico do usuário na plataforma.	Histórico dos usuários na própria plataforma.	Não especificados.	Gravidade e frequência das infrações.
Sanções de violação	Varia de notificação, restrições de uso da plataforma até desativação completa do perfil.	Remoção do conteúdo e possível impedimento de uso temporário da conta. Suspensão da conta apenas após infrações persistentes.	Remoção do conteúdo e eventual remoção do usuário da rede.	Primeira violação: sinalização sem penalidade. Nas seguintes, exclusão do conteúdo (vídeo ou comentário) além de uma notificação e limitação das possibilidades de uso da plataforma por tempo determinado. Após três notificações em 90 dias, o canal é encerrado e o conteúdo deletado.
Interesse público	Conteúdo que viole as diretrizes pode ser mantido por finalidades de conscientização, combate ou educação sobre o tema.	Conteúdo que viole as diretrizes pode ser mantido por finalidades de conscientização, combate ou educação sobre o tema.	Conteúdo que viole as diretrizes pode ser mantido por finalidades de conscientização, combate ou educação sobre o tema.	Conteúdo que viole as diretrizes pode ser mantido por finalidades de conscientização, combate ou educação sobre o tema.

Fonte: Diretrizes da Comunidade de cada plataforma | Elaboração: FGV DAPP

ANÁLISE DO DEBATE SOBRE DISCURSOS DE ÓDIO E CENSURA NO TWITTER E NO FACEBOOK

Nesta seção apresentaremos as análises sobre o debate dos discursos de ódio nas redes. Coletamos postagens do Twitter e do Facebook pelo fato dessas redes concentrarem um número expressivo de usuários no Brasil e terem pontos de semelhanças no que se refere a características permitidas para as postagens e formas de engajamento.

Os dados coletados no Twitter e no Facebook foram categorizados a partir de dois grupos principais: o **debate sobre discurso de ódio** e o **debate acerca da censura** em cada uma das plataformas, de forma separada. Tanto no Twitter quanto no Facebook, o pico do debate sobre discurso de ódio ocorreu na mesma data, 20 de novembro de 2020. No que se refere ao debate sobre discurso de ódio percebemos que os usuários das redes sociais analisadas tendem a compreender o tema como sendo a publicação de ofensas - no universo on-line e off-line - contra um grupo socialmente minoritário. Por essa razão, debates sobre racismo e misoginia são constantemente detectados. É interessante ressaltar que, com base nos dados coletados, o debate nas redes sobre esse tema é sempre impulsionado por situações ocorridas no off-line e que permitem a identificação da vítima e/ou do agressor.

Quando observamos os dados coletados, constatamos que o debate sobre discurso de ódio é continuamente acompanhado pela discussão sobre a prática de censura ou ataque à liberdade de expressão, assim como a bibliografia acerca do tema aponta. É recorrente, entre comentários e publicações analisados, a ideia de que o bloqueio de contas por postagens consideradas ofensivas significa o cerceamento da liberdade individual de livre manifestação de opinião. Neste mapeamento, observamos que esse debate está em torno das sanções impostas a contas de usuários nas redes sociais, pois o foco está instaurado nas ações tomadas pelas plataformas em “derrubar” postagens, suspender ou bloquear perfis, como no caso do ex-presidente dos Estados Unidos Donald Trump.

É a partir desse panorama de discurso de ódio e liberdade de expressão que as postagens nas redes sociais se concentraram quando realizamos a extração dos dados sobre esse tema. Com base nesses dados, escolhemos analisar o pico das menções, as datas, ao longo do período coletado, onde o volume de postagens sobre os temas foram maiores. Assim, estabelecemos uma análise quali-quantitativa acerca dos dados já que em um primeiro momento mapeamos todas as publicações sobre a temática e num segundo momento focamos no pico dessas publicações para compreender o engajamento, ou melhor, para ilustrar qual postagem, notícia ou usuário contribuíram para que o debate crescesse.

O nosso olhar sobre o engajamento se dirigiu ao alcance do *post* e, por isso, observamos todas as possibilidades que uma publicação tem em envolver os usuários do Facebook e do Twitter, para além das chamadas “métricas de vaidade”, definidas por Rogers (2019) como aquelas que auferem apenas *likes* e o total de seguidores de um determinado perfil. Assumimos, em contrapartida, uma postura de análise crítica das redes sociais que vai além da mera contagem de métricas das publicações ou perfis (SANCHES, 2020). Com base nessa perspectiva, operacionalizamos uma análise qualitativa a partir de casos identificados por meio do levantamento quantitativo de interações como curtidas, compartilhamentos e comentários de uma determinada conta para compreender o alcance da postagem e a centralidade do autor e ou da publicação nas discussões da rede. A seguir, discutiremos os resultados dessa análise nos debates ocorridos no Twitter e no Facebook, gerando, primeiro, uma investigação dos picos das postagens sobre discurso de ódio e posteriormente sobre censura nessas redes.

Debate sobre discurso de ódio e censura no Twitter

Entre os dias 16 de novembro de 2020 e 06 de fevereiro de 2021, o discurso de ódio foi assunto de aproximadamente 11,6 milhões de postagens no Twitter. O debate foi impulsionado pelas comemorações do Dia da Consciência Negra e o assassinato de um consumidor negro por um segurança do Supermercado Carrefour, em Porto Alegre (RS), e

em razão da invasão do Capitólio, nos Estados Unidos, e a consequente suspensão das contas do ex-presidente norte-americano Donald Trump, nas redes sociais. Como fica evidente, os dois temas que mais impulsionaram o debate no Twitter são discurso de ódio (debate acerca do racismo estrutural) e liberdade de expressão (debate sobre censura). Assim, como a literatura sobre o tema nos mostra, o debate sobre ofensas on-line tem sempre no seu verso questionamentos sobre o cerceamento das liberdades individuais e as censuras impostas pelas plataformas contra expressões de opinião dos usuários.

No período analisado, outras notícias também contribuíram de forma expressiva com este debate. Sobre o discurso de ódio, [o episódio de assédio sofrido pela deputada Isa Penna \(PSOL-SP\) foi permeado por uma discussão sobre a vulnerabilidade feminina e a misoginia em diferentes espaços](#)⁷. No que se refere à liberdade de expressão, podemos apontar a publicação realizada pelo Ministério da Saúde sobre o tratamento precoce contra a Covid-19, considerada pelo Twitter como mensagem enganosa.

Como apontado, a coleta de dados sobre discurso de ódio on-line ou ofensas vem sempre acompanhada com postagens sobre a liberdade do usuário em poder se manifestar livremente. Nesse aspecto, durante o período analisado foi intenso o debate sobre a divulgação da suspensão de contas ou da remoção de conteúdo em sites de redes sociais, como Facebook, Instagram ou, ainda, o próprio Twitter, por violação de diretrizes das plataformas. Algumas postagens – partindo de perfis de políticos ou blogueiros com algum alinhamento partidário-ideológico ou identitário – questionam a decisão dos sites, alegando que o conteúdo bloqueado não violaria padrões das comunidades ou que as razões do bloqueio não se sustentariam, e que, na verdade, ele seria alvo de campanhas difamatórias por parte de grupos contrários. O banimento é eventualmente comparado a outros casos, em que os usuários identificam postagens que também contenham conteúdo sensível – por exemplo, ofensas e ameaças a determinados políticos ou a grupos

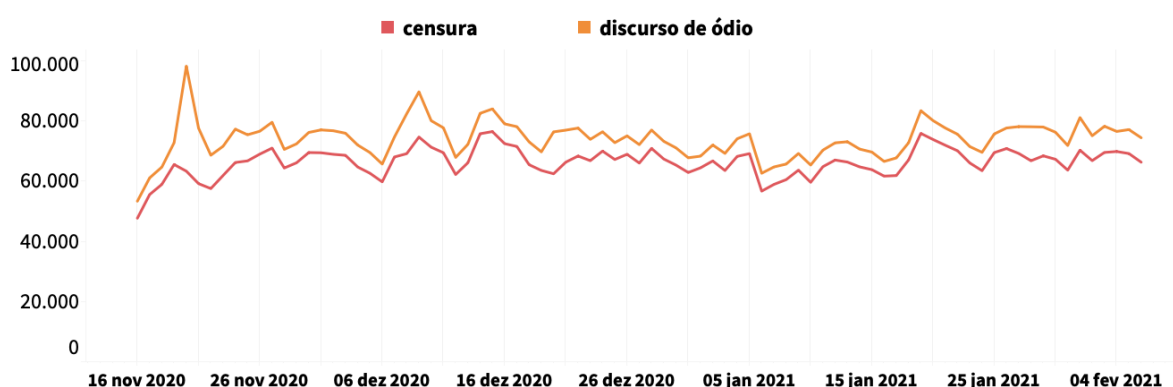
⁷ Durante uma sessão da Assembleia Legislativa de São Paulo, no dia 16 de dezembro de 2020, a deputada Isa Penna (PSOL-SP) foi alvo de assédio, tendo sofrido importunação sexual do deputado Fernando Cury (Cidadania-SP). A importunação diz respeito ao contato físico forçado em região íntima do corpo da deputada.

partidários, disseminação de notícias ou de informações falsas –, porém não foram removidas pelas plataformas nem tiveram as contas de seus autores suspensa.

O gráfico a seguir mostra a evolução do debate no Twitter e, com base nos picos apontados nele, marcadamente nos dias 20 de novembro (sobre discurso de ódio) e nos dias 15 de dezembro e 19 de janeiro de 2021 (sobre liberdade de expressão), iremos construir uma análise qualitativa baseada no engajamento acerca dos temas. Iremos levar em conta os compartilhamentos, números de curtidas e número de seguidores de cada conta a fim de detectar as publicações que impulsionaram o debate.

Gráfico 1 - Volume de interações das categorias no Twitter

Período: 16 de novembro de 2020 a 06 de fevereiro de 2021



Fonte: Twitter | Elaboração: FGV DAPP

No que se refere ao debate sobre discurso de ódio, observamos uma centralidade acerca do racismo estrutural presente na sociedade brasileira. Essa discussão foi impulsionada por duas razões que se retroalimentam: o Dia da Consciência Negra, comemorado em 20 de novembro, e [o assassinato de João Alberto Freitas por seguranças do Supermercado Carrefour, no Rio Grande do Sul](#). A discussão se a morte de João Alberto era um crime de racismo ou era fruto da truculência e agressividade do segurança deram o tom da discussão. O engajamento desse debate ocorreu devido à publicação desse fato por

diferentes canais de notícias, repercutindo ainda mais em razão da manifestação contra o assassinato, realizada na Avenida Paulista, em São Paulo, também no dia 20 de novembro de 2020. Postagens do Mídia Ninja, Quebrando o Tabu e de outras contas que abordaram o protesto avolumaram ainda mais o debate sobre racismo nas redes.

No que se refere às discussões sobre censura e liberdade de expressão no âmbito das plataformas, observamos uma quantidade expressiva de *tweets* no dia 15 de dezembro de 2020 sobre o tema, cerca de 76.467 *posts*. Ao localizar o debate constatamos duas postagens que o engajaram: a primeira do presidente Jair Bolsonaro e a segunda da deputada federal Carla Zambelli (PSL-SP). A da deputada fazia referência explícita ao [Projeto de Lei 2630/2020, de autoria do deputado federal Alessandro Vieira \(Cidadania\)](#), já abordado na seção 2 deste relatório. De acordo com a publicação, o Senado Federal estava com uma consulta pública em aberto tendendo para o voto no sim, enquanto a deputada e alguns de seus aliados desejavam que seus apoiadores votassem não. Essa lei prevê, dentre outras medidas, que as plataformas regulem e estabeleçam sanções contra postagens falsas e ofensivas. A proposta foi encarada como censuradora e contra as liberdades de expressão individuais. A partir dessas postagens, outras publicações mostraram momentos em que o Twitter não permitiu curtidas ou comentários em alguns conteúdos considerados, pela plataforma, contrários aos Termos de Uso ou às Diretrizes da Comunidade.

Exemplo claro desse debate foi a grande mobilização sobre a decisão do Twitter em bloquear a conta oficial de Donald Trump, ocorrida ao longo das primeiras semanas do mês de janeiro de 2021, após o ex-presidente americano postar um vídeo em que, a julgar pelos comentários relativos ao conteúdo, encorajava apoiadores a invadirem o Congresso dos Estados Unidos em protesto ao resultado da eleição americana de 2020, que deu a vitória ao seu adversário, Joe Biden. A decisão da plataforma se baseou na alegação de que o vídeo poderia incitar atos violentos entre os manifestantes. Mais tarde, outros sites de redes sociais – como Facebook, Instagram e Snapchat – também anunciaram a suspensão das contas de Trump nas respectivas plataformas.

A suspensão de contas do ex-presidente americano em diferentes sites incitou uma discussão sobre impasses entre discurso de ódio e liberdade de expressão nas redes sociais. Alguns perfis críticos aos protestos no Capitólio, em Washington, D.C., aprovaram a decisão do Twitter, condenando a "mensagem de ódio" propagada por Trump e alegando que a liberdade de expressão – bandeira levantada por defensores do ex-presidente – deve ter limites. É importante destacar ainda que diversos órgãos de imprensa que cobriram o acontecimento mostraram que muitos manifestantes que invadiram o Congresso americano vestiam trajes ou portavam símbolos reconhecidos como representações de supremacia branca, fato que corrobora com aqueles que defendem limites para a livre opinião.

A opinião, no entanto, que dominou o debate acerca desse episódio – aventada, principalmente, por perfis de políticos, blogueiros e influenciadores conservadores – é de que a decisão do Twitter foi arbitrária. Com o argumento de que Trump estaria sendo perseguido, esse discurso identificou na atitude da plataforma uma estratégia de grupos de esquerda para tentar tomar o poder nos Estados Unidos. Postagens apontaram incoerência na decisão do site, na medida em que mantêm ativas contas de outros governantes "ditadores" na rede social – como os perfis de Nicolás Maduro, presidente da Venezuela, e de Hassan Rouhani, presidente do Irã. Essa parte do debate também classificou como hipócritas comentários críticos à mensagem de Trump, alegando que, apesar de se declararem defensores da democracia, estariam celebrando um ato de censura contra o ex-presidente. Alguns usuários, [impulsionados pela publicação do influenciador Allan dos Santos](#), classificaram a atitude como sendo absurda, ditatorial e silenciadora de vozes, o que impulsionou o debate.

Outro ponto mobilizador do debate sobre censura e liberdade de expressão foi a catalogação da postagem realizada pelo Ministério da Saúde do Brasil, sobre o tratamento precoce contra a Covid-19, como sendo enganosa. O engajamento se deu, a partir de uma [publicação realizada pelo portal Brasil 247](#), sobre o inquérito aberto pelo procurador do Ministério Público de Goiás a respeito do caso. Embora o debate neste ponto esteja permeado por postagens sobre a Covid-19 e a seriedade da doença, ele também esteve

marcado pela interferência da plataforma em uma conta oficial, governamental, apontando para os riscos dessa atitude face à liberdade de expressão.

Na seção a seguir, iremos compreender como esses perfis interagiram e em quem e em que se localizou a centralidade do debate. No entanto, o que podemos depreender a partir do exposto é que o debate sobre censura foi presente em todo o período de análise tendo sido impulsionado por figuras políticas de relevância nacional.

Análise dos atores no Twitter

No dia 20 de novembro de 2020, o assassinato de João Alberto Silveira de Freitas em um supermercado de Porto Alegre, no dia anterior, protagonizou o debate sobre discurso de ódio no Twitter. A circulação de posicionamentos e interpretações particulares sobre o caso permitiu a identificação de grupos específicos no contexto.

Laranja – 11,21% das interações

42

Azul – 6,63% das interações

Sob a influência de políticos de direita, blogueiros e celebridades de orientação conservadora, grupo considera precipitado o entendimento do assassinato de João Alberto como crime racial. Postagens compartilham uma suposta declaração de uma delegada envolvida na investigação descartando indícios de racismo no caso. Alguns perfis também recuperam um vídeo de 2012 em que, durante uma entrevista, o ator americano Morgan Freeman questiona a existência do Mês da Consciência Negra perguntando quando seria o Mês da Consciência Branca. Fora do escopo do caso de João Alberto, destacam-se, ainda, a divulgação de processos judiciais movidos pela ex-atleta Ana Paula Henkel por conta de ofensas e ataques sofridos por ela nas redes sociais, além da divulgação do direito de resposta concedido ao influenciador Olavo de Carvalho após declarações da jornalista Thaís Oyama ao seu respeito.

Verde claro – 4,98% das interações

Na esteira do assassinato de João Alberto, grupo com canais de entretenimento divulga outros casos de racismo sofridos por profissionais e influenciadores negros no Brasil. Destaca-se, nesse contexto, a morte de um funcionário de um supermercado em Recife (PE), em agosto de 2020, cujo corpo foi escondido sob guarda-sóis enquanto o estabelecimento funcionava normalmente; o caso da youtuber @badgallore, que teria sido humilhada em uma agência bancária após ser questionada sobre a alta quantidade de dinheiro na sua conta; e o caso de um gerente de uma loja de departamentos, em Minas Gerais, quando uma cliente teria duvidado da sua posição na empresa em função de ele ser negro.

Vermelho – 3,17% das interações

A partir de reações ao assassinato de João Alberto, perfis de ativistas negros condenam a eventual relativização do racismo e a culpabilização das vítimas de violência. Contestando a suposta declaração de um delegado responsável pela investigação – que teria dito não haver indícios de racismo no caso –, postagens questionam o judiciário brasileiro,

comparando a afirmação com a alegada sentença de "estupro culposo" no julgamento de agressão sexual sofrida pela blogueira Mariana Ferrer, encerrado em setembro de 2020.

Roxo – 2,69% das interações

Mobilizado por perfis de ativistas e influenciadores negros, grupo demonstra revolta e indignação com o espancamento até a morte de João Alberto e, sobretudo, com o fato de o crime ter ocorrido nas vésperas do Dia da Consciência Negra no Brasil. Com o argumento de que não haveria o que comemorar na data em questão, muitos perfis divulgam dados sobre a realidade da população negra no país para evidenciar que o crime não configuraria um caso isolado, senão que seria um retrato da situação dos negros no Brasil.

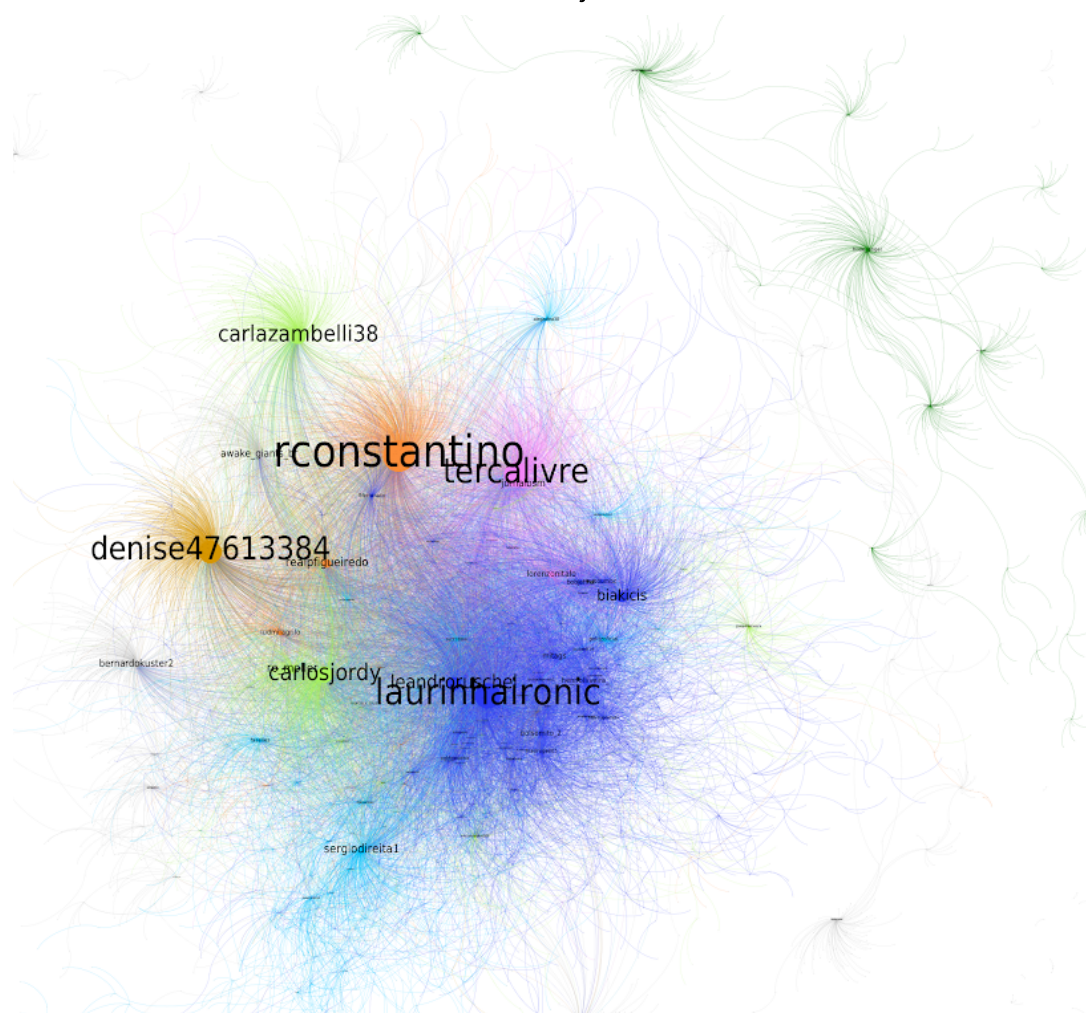
Verde escuro – 2,50% das interações

Composto por perfis de influenciadores digitais, grupo ataca a rede de supermercados em que ocorreu o assassinato de João Alberto – destacando que essa não seria a primeira vez que a empresa seria palco de atrocidades – e insiste na exposição dos seguranças responsáveis pelo crime e não da vítima.

No que diz respeito à liberdade de expressão – e, eventualmente, à censura –, o debate no Twitter se concentrou em diferentes episódios para, sobretudo, questionar as decisões tomadas por diferentes redes sociais para tentar moderar o conteúdo que circula nas suas plataformas.

Grafo 2 - Mapa de interações do debate sobre censura no Twitter

Período: 14 e 15 de janeiro de 2021



Fonte: Twitter | Elaboração: FGV DAPP

Azul escuro – 39.23% das interações

Mobilizado por políticos de direita e perfis de humor alinhados com a direita, grupo convoca seguidores a aderirem à campanha **#silenceday**, em protesto contra um suposto cerceamento da liberdade de expressão promovido pelas redes sociais, como Facebook, Twitter, Instagram e YouTube. Em resposta à moderação de uma série de conteúdos publicados, sobretudo, por perfis conservadores, a campanha instrui os usuários a não postarem, curtirem ou compartilharem nas plataformas por um dia.

Verde claro – 11,17% das interações

Diante da crise sanitária que assolou Manaus (AM) em janeiro, no contexto da pandemia de Covid-19, grupo composto por políticos de direita, blogueiros e influenciadores digitais da ala conservadora contesta uma eventual tentativa da mídia tradicional e de políticos de esquerda de responsabilizar o governo federal pelo episódio. Acusando a imprensa brasileira de espalhar notícias falsas, perfis argumentam que o colapso na saúde da capital amazonense deveria ser colocada na conta do governador do Estado.

Rosa – 9,03% das interações

Orbitando em canais de mídia alternativa alinhados com a direita, grupo critica decisões de algumas redes sociais – como Twitter e Facebook – pela moderação de canais oficiais e de conteúdo publicado pelo então presidente americano Donald Trump a respeito das eleições presidenciais dos Estados Unidos, em 2020. Com o argumento de que a decisão das plataformas configuraria censura, perfis também mobilizaram a campanha **#silenceday**.

Laranja – 8,89% das interações

Convocando aderência à campanha **#silenceday**, grupo ancorado em perfis de jornalistas conservadores ataca ações judiciais que tentam impedir a juíza Ludmila Lins Grilo de publicar nas suas redes sociais mensagens a favor da flexibilização do isolamento social e do uso de máscaras no contexto da pandemia de Covid-19. Essa parte do debate alega que as ações atentariam contra a liberdade de expressão da magistrada.

Verde escuro – 3,11% das interações

Contando com perfis de jornalistas, comentaristas políticos e influenciadores digitais, publicações ironizam a campanha **#silenceday** e contestam a afirmação de que a moderação de postagens e canais do ex-presidente americano Donald Trump nas redes sociais atentaria contra o princípio da liberdade de expressão. Criticando declarações de

políticos ligados ao governo federal minimizando a pandemia de Covid-19 ou contrariando recomendações de autoridades sanitárias, postagens exigem providências das plataformas de redes sociais, como suspensão de contas e remoção de conteúdo.

Azul claro – 12,34% das interações

Mobilizado por perfis de políticos de direita e canais de informação e influenciadores digitais conservadores, grupo se recusa a aderir à campanha **#silenceday** com o argumento de que a iniciativa não garante respeito à liberdade de expressão. Em vez disso, perfis julgam importante usuários irem às redes sociais no esforço de questionar o discurso da mídia tradicional que atribui ao governo federal a crise sanitária em Manaus. Na esteira da mobilização, postagens insistem na responsabilização do governador do Amazonas pelo colapso na saúde da região.

Amarelo – 4,16% das interações

Perfis de usuários comuns repercutem a notícia de que o governo russo teria criticado a decisão de redes sociais de suspender postagens e contas do ex-presidente americano Donald Trump, e que Vladimir Putin estaria disposto a combater tentativas das plataformas de cercear a liberdade de expressão no seu país. O grupo também compartilha a informação de que, em função da suposta censura contra Trump, sites como Twitter e Facebook estariam sofrendo prejuízos financeiros.

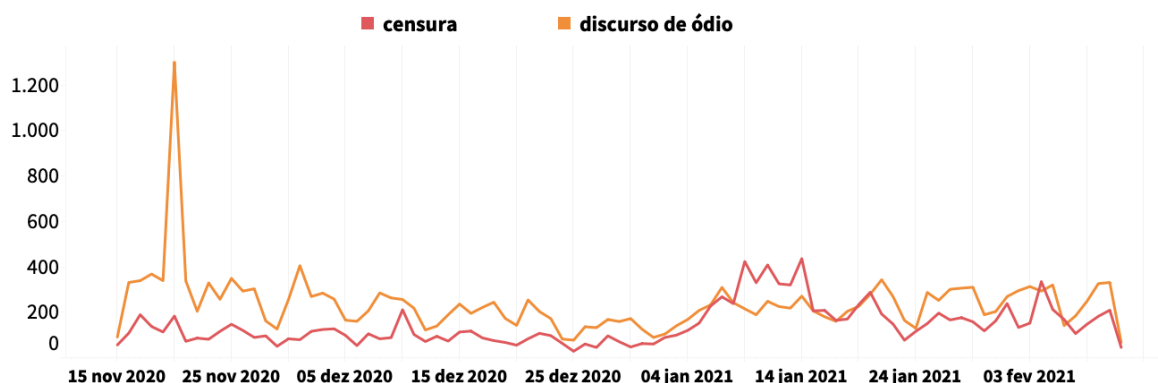
Debate sobre discurso de ódio e censura no Facebook

Esta seção apresenta os resultados da análise realizada no Facebook entre os dias 15 de novembro de 2020 e 11 de fevereiro de 2021. Ao todo, foram coletadas 21.149 postagens que tratavam sobre discurso de ódio e 12.773 sobre os temas de liberdade de expressão ou censura. A evolução de ambos debates foi observada paralelamente, considerando a relação direta entre os dois assuntos, conforme tratamos anteriormente. Pensar essa

articulação importa para este trabalho na medida em que cresce a demanda por mecanismos para coibir os discursos de ódio e, de maneira correlata, impor limites à liberdade de expressão.

Gráfico 2 - Evolução do debate de discurso de ódio e censura no Facebook

Período: 15 de novembro de 2020 a 11 de fevereiro de 2021



Fonte: Facebook | Elaborado por FGV DAPP

Começando pelas postagens relacionadas ao discurso de ódio, o *corpus* indicou alguns debates específicos que se destacaram no período analisado. O tema aparece como assunto caro em alguns perfis que apontavam a necessidade de combater discursos discriminatórios na rede, como [em compromisso assumido por uma vereadora eleita na cidade de Jarú \(RO\)](#). Ou ainda no convite de uma igreja em Osasco (SP) que se propôs a abordar a [cultura do ódio](#) em uma de suas celebrações. Os discursos de ódio também foram mencionados em páginas que consideram seu debate uma forma de priorizar relações sociais [marcadas pelas ofensas](#) em detrimento das demais. De modo específico os dados revelaram engajamento em discussões relacionadas a racismo, misoginia e homofobia.

Considerando que a amostra compreendeu os dois turnos de votação das eleições municipais brasileiras, as publicações, sobretudo nos meses de novembro e dezembro, foram afetadas por essa conjuntura. Falou-se, por exemplo, na [primeira mulher negra](#)

[eleita vereadora de Joinville \(SC\)](#), que sofreu injúrias raciais e foi ameaçada de morte nas redes. Há também postagens criticando um vídeo do humorístico *Porta dos Fundos* que faz referência à vereadora mais votada de Curitiba (PR), [dando uma conotação sexual](#) ao seu resultado bem-sucedido.

Assim, como aconteceu na coleta do Twitter, o maior pico do debate sobre o tema se encontra no mês de novembro, no dia 20, o Dia da Consciência Negra, quando foram contabilizadas 1.301 postagens sobre discurso de ódio na plataforma, mais de cinco vezes a média de 237,6 publicações ao dia, no período analisado. A data, por si só, impulsiona a produção de conteúdo, sobretudo de figuras públicas, a respeito da necessidade de combater o racismo. No entanto, a morte de João Alberto, o consumidor negro vítima de um espancamento na véspera do feriado, no estacionamento do Supermercado Carrefour, em Porto Alegre (RS), elevou os níveis de engajamento.

Ao longo do dia, as publicações se concentraram em, por um lado, discutir o caso como exemplo de racismo estrutural e, por outro, combater a ideia de que o crime tenha tido essa motivação. Embora a maioria das postagens sobre o episódio o tenham vinculado à data comemorativa como evidência de uma frágil compreensão da igualdade de direitos no país, a publicação que mais produziu engajamento foi uma realizada pelo [perfil do presidente Jair Bolsonaro](#). No texto, com 41.787 compartilhamentos, ele afirmou que o ódio estaria nos que querem dividir o povo brasileiro, por meio do ressentimento, tornando-o vulnerável a mecanismos de controle.

A abordagem foi acolhida pelos que defendem que falar de racismo no Brasil é um modo de gerar discórdia entre os brasileiros e alimentar um certo projeto esquerdista do caos. No entanto, a coleta identificou postagens de perfis não conservadores, para os quais tratar o caso de João Alberto como racismo invisibiliza [a luta de classes](#) e reduz as possibilidades de enfrentamento das desigualdades. Para uma perspectiva, o debate sobre racismo seria nocivo porque corrói as relações, fragilizando uma suposta unidade nacional. Para outra, ele é simplista porque estaria reduzindo as possibilidades de resistência a um ativismo pontual que não alteraria as estruturas.

Das dez publicações mais compartilhadas no dia, sete tratavam especificamente do assassinato de João Alberto em sua relação direta com o racismo. Entre os argumentos para essa associação apareceram dados estatísticos que revelam como os brasileiros negros são mais assassinados do que os brancos. Em quase todas as postagens surgiram indicações de como o Carrefour poderia e deveria ser punido judicialmente, mas também sobre como as pessoas poderiam mostrar sua indignação por meio de um boicote ao supermercado. As notas de repúdio foram em boa medida criticadas como paliativas e alguns lembraram outros episódios envolvendo a empresa, tanto em casos de racismo quanto de morte. De maneira ampla, foram recorrentes as convocações a que as pessoas se envolvessem no enfrentamento do racismo individual e coletivamente.

Ao todo, seis publicações se repetem, ainda que ocupando posições diferentes, quando alterados os critérios de engajamento (número de *likes*, compartilhamentos ou comentários). Dessas, quatro são de figuras públicas: além de Jair Bolsonaro, estavam entre os principais impulsionadores do debate, no dia do pico, os artistas Nando Reis (Mais Brasil) e Maria Rita e o padre Reginaldo Manzotti. Quando observados em separado, os diferentes critérios de engajamento permitem identificar outros perfis de figuras públicas. A segunda postagem com mais *likes* na amostra, por exemplo, é do jogador brasileiro Fred, do Manchester United. A publicação da blogueira de culinária Rita Lobo aparece entre os dez com mais engajamento quando acionado o critério de número de comentários. Com exceção do presidente, todos abordaram o Dia da Consciência Negra como um dia de reflexão sobre o sofrimento trazido pelo preconceito racial.

O debate sobre discurso de ódio, no período analisado, foi impulsionado, portanto, no Dia da Consciência Negra, 20 de novembro, e acabou sendo combinado com as postagens sobre o assassinato de João Alberto no Rio Grande do Sul. De modo geral, os atores entenderam o homicídio como crime motivado por preconceito racial não apenas porque a vítima era negra, mas pelo modo violento como ele foi espancado, negando-lhe o direito de defesa. Como mostra a vasta literatura do tema, a base para a definição de discurso de ódio é a ideia de uma superioridade, segundo a qual determinados grupos não merecem

ser tratados com igualdade. É a vulnerabilidade social desses grupos, materializada em características específicas, que os torna alvo dos ataques e discursos de ódio.

Sobre o debate de censura no Facebook, para além de postagens que abordaram o assunto em contextos específicos como o religioso e o artístico, chamam atenção dois tipos de publicações no intervalo de tempo analisado: as que acusam as plataformas, sobretudo Twitter e Facebook, de ferirem a liberdade de expressão na rede; e as que celebram decisões notadamente jurídicas, caracterizadas por termos como “vitória da liberdade de expressão”.

No primeiro grupo, figuram postagens que defendem haver uma ação articulada de grupos progressistas, de viés esquerdista, atuando nos meios de comunicação, nas plataformas e em instituições públicas, especialmente no judiciário, com o suposto objetivo de dismantelar os movimentos conservadores de direita. Um dos assuntos mais comentados nesse tópico foi a impossibilidade de publicar o *link* de uma matéria do *New York Post*, sob alegação de informações não checadas. O texto fazia acusações ao filho do então candidato democrata nas eleições presidenciais estadunidenses, Joe Biden. As postagens analisadas acusavam Facebook e Twitter de fazer uma “edição pró-democrata” de seu conteúdo e, com isso, censurar o que escapasse a tais critérios.

Um segundo grupo de publicações que chamam atenção fazem referência a ações judiciais motivadas pela acusação de discursos de ódio e conteúdo de desinformação, entre outros assuntos. Está nesse grupo uma série de postagens que comemoram a decisão favorável da justiça desportiva em relação ao [grito “#ForaBolsonaro” da jogadora de vôlei Carol Solberg](#), durante uma entrevista. Há também engajamento nas publicações que celebraram o fato de o STF não considerar ilícitas as postagens do deputado federal [Eduardo Bolsonaro \(PSL-SP\) contra o Movimento de Trabalhadores Sem Teto \(MTST\)](#).

As postagens sobre censura alcançaram uma média diária de 145,5, um fluxo menor que o do discurso de ódio. No entanto, nos dias 9, 11 e 14 de janeiro, o volume de postagens do debate sobre censura superou o de discurso de ódio. A mudança foi provocada pelas

notícias de que o então presidente dos Estados Unidos Donald Trump teve suas contas canceladas permanentemente pelo Twitter e, por tempo indeterminado, por Facebook e Instagram. Nesses três dias, o debate sobre censura teve uma alta de publicações, registrando um total, respectivamente, de 423, 408 e 436 postagens.

Três movimentos marcaram as discussões sobre censura. O primeiro deles diz respeito à suspensão das contas de Trump das referidas plataformas e do banimento do Parler, que mantinha perfil do ex-presidente, das lojas de aplicativo da Google, da Apple e da Amazon. As ações foram interpretadas como uma censura ao movimento conservador de maneira geral, um ataque a um legítimo representante da extrema direita. Em segundo, houve forte engajamento em publicações para promover outras plataformas: o Parler, antes de seu cancelamento; o Telegram; e o Paatria. E, no dia 14, ainda repercutindo a notícia dos dias 9 e 11, as postagens foram impulsionadas pelo #silenceDay, cujo objetivo seria passar todo o dia 15 de janeiro sem postar, curtir, marcar ou realizar qualquer outro tipo de interação no Twitter, no Facebook, no Instagram, no YouTube e no WhatsApp em protesto ao posicionamento das plataformas.

Nesse conjunto de postagens, a que obteve maior engajamento foi publicada pela [deputada federal Carla Zambeli \(PSL-SP\)](#), em que ela comemora o anúncio de que Trump criaria uma nova plataforma “com outras empresas para combater a censura contra a direita”. A publicação obteve 29.524 compartilhamentos e 13.825 comentários. Além de Zambeli, outros perfis de figuras políticas foram identificados na amostra entre os autores das publicações que mais impulsionaram o debate, particularmente o presidente Jair Bolsonaro, as deputadas federais Bia Kicis (PSL-DF) e Caroline de Toni (PSL-SC) e os deputados federais Filipe Barros (PSL-PR) e Carlos Jordy (PSL-RJ). Os perfis República de Curitiba e Jovem Pan News também fizeram postagens que impulsionaram as discussões. Diferentemente do debate sobre discurso de ódio, a pauta da censura foi marcada por maior homogeneidade em seus atores, revelando que, no período analisado, o debate foi em muito articulado por grupos conservadores de direita do campo político.

A análise dos debates sobre discurso de ódio e censura no Facebook indica o importante papel de uma agenda pública e das mídias noticiosas no impulsionamento de debates públicos. A combinação de uma data comemorativa que compreende uma série de ações e celebrações, por parte de governos e sociedade civil, com a trágica morte de João Alberto, e a repercussão internacional do cancelamento das contas de Trump alimentaram a publicação dos usuários na rede. Ambos debates ganharam lastro por meio das contas de figuras públicas da política, da música, do esporte, da religião e da internet, no caso do discurso de ódio, e da política, sobretudo, no da censura. Isso indica a relevância do envolvimento de nomes com grandes redes para que uma discussão ganhe volume e alcance diferentes instâncias nas sociedades. Também deixa evidente a responsabilidade de autoridades que ocupam cargos públicos no modo como direcionam o curso do debate.

CONCLUSÕES

Embora as ofensas, ameaças e discriminações que compreendem os discursos de ódio tenham efeitos subjetivos, considerando o grande sofrimento psíquico que produzem, seu escrutínio não poderá abrir mão dos impactos políticos que geram. O silenciamento e a invisibilidade que sofrem os grupos vulnerabilizados enfraquecem suas demandas e seus pleitos, alijando-os dessa forma dos processos democráticos de decisão. É nesse sentido que identificar e combater os discursos de ódio importa para a democracia.

Este trabalho apresentou uma revisão de literatura sobre o tema com o objetivo de elencar os principais entendimentos sobre o que são os discursos de ódio na academia e no campo jurídico, suas especificidades no ambiente digital e como são definidos pelas plataformas digitais em seus protocolos de auto-regulação. O levantamento bibliográfico e documental revelou que, se por um lado, é possível conceituar os discursos de ódio teoricamente, por outro, é bastante complexo defini-los em situações concretas. O desafio

se intensifica, nas páginas on-line, com memes, vídeos, comunidades organizadas e léxicos diversos, entre outros aspectos.

O material mostrou que, além da adoção dos tratados internacionais que combatem as diferentes discriminações, é necessário que os países formulem uma legislação que combine seus contextos sociais, históricos e, sobretudo, culturais com as especificidades do universo on-line. Nesse sentido, o desafio não se mostrou menor para as plataformas. Ao detalhar - umas mais, outras menos - em seus Termos de uso e Diretrizes de Comunidade como percebem as diversas manifestações de discursos de ódio, elas exibem uma série de lacunas sobre como abordar e coibir seus autores. Uma hipótese a esse respeito é a de que as plataformas globais precisam conhecer as particularidades socioculturais em que seus usuários estão inseridos para desenvolver encaminhamentos exequíveis e mais produtivos no enfrentamento dos discursos de ódio.

A análise empírica do estudo consistiu no mapeamento do debate dos discursos de ódio e sua relação com a liberdade de expressão, no Twitter e no Facebook, entre os meses de novembro de 2020 e fevereiro de 2021. Ainda que ambos sejam redes sociais digitais com expressivo número de usuários, inclusive no contexto brasileiro, compreendemos que apresentam limitações em termos de representatividade tanto em relação ao quantitativo de interações digitais quanto à diversidade de atores que ensejam. Assim, a análise aqui presente oferece uma visão parcial do complexo ecossistema de criação, proliferação e manutenção de discursos de ódio on-line.

No âmbito das preocupações metodológicas, um desafio que se impõe à elaboração das sintaxes de busca – para a coleta e a classificação dos dados a serem analisados – se refere às sutilezas que marcam a conceituação e o escopo semântico-pragmático do tema sobre o qual o estudo se debruça, isto é, o tópico "discurso de ódio". Visto que há grande dificuldade em se determinar, de maneira conclusiva e indiscutível, o que, de fato, configura discurso de ódio – ou, ainda, discurso ofensivo – e, principalmente, como esse discurso é instanciado no debate público nas redes sociais, identificar as estratégias

linguísticas e discursivas relativas a esse fenômeno se mostra uma tarefa de fragilidades inegáveis.

Com base na coleta de dados e na análise realizadas, foi possível perceber que o fluxo do debate de discurso de ódio e censura se intensificou durante uma data comemorativa e a partir da repercussão de notícias no âmbito nacional e internacional. Isso leva a crer que a existência de uma agenda pública e da atuação das mídias noticiosas aqueceram as discussões em curso nas plataformas investigadas. Além desses elementos, as postagens sobre o tema por parte de figuras públicas detentoras de uma grande rede produziram um alto volume de interações, o que sinaliza o papel politicamente estratégico das figuras públicas na produção e alcance do debate.

Considerando a evolução do debate on-line, no período analisado, o estudo apontou para uma tensão entre a compreensão estrutural e pessoal da discriminação, o que significa dizer que, para uma parcela dos atores, isso é central no amadurecimento democrático, e por isso deve estar na arena pública; mas, para outra, trata-se de um discurso que desestabiliza o campo político sem, contudo, produzir mudanças efetivas. Por fim, as sanções operadas pelas plataformas, interferindo no fluxo de publicações e na atuação de determinados atores, intensificou o debate sobre liberdade de expressão, levando a uma problematização não de sua legitimidade, mas de seus limites.

Ao identificar os argumentos e atores que constituem tal debate, este trabalho lança bases para novas frentes de investigação sobre a disseminação dos discursos de ódio nas redes. Entre elas, estão o surgimento de novas plataformas em resposta às sanções realizadas, a ampliação da diversidade lexical associada às variadas possibilidades de linguagem, a relação dos procedimentos de auto-regulação das plataformas com os contextos socioculturais em que são aplicados e as estratégias dos atores sociais para desarticular as discriminações on-line.

REFERÊNCIAS

ALAVA, S.; FRAU-MEIGS, D.; HASSAN, G. **Youth and violent extremism on social media**: mapping the research. Paris: Organização das Nações Unidas para a Educação, a Ciência e a Cultura, 2017.

BARBERÁ, P. How social media reduces mass political polarization: evidence from Germany, Spain, and the US. In: AMERICAN POLITICAL SCIENCE ASSOCIATION ANNUAL MEETING, 2015. **Proceedings**. American Political Science Association, São Francisco, 2015. Disponível em http://pablobarbera.com/static/barbera_polarization_APSA.pdf. Acesso em: 26 fev. 2021.

BARBOSA, B.; MARTINS, H.; VALENTE, J. **Fake news**: como as plataformas enfrentam a desinformação. São Paulo: Intervezes, 2020.

BENESCH, S. Dangerous speech: A proposal to prevent group violence. **Dangerous Speech Project proposal paper**. February 23, 2013. Disponível em: <https://dangerousspeech.org/wp-content/uploads/2018/01/Dangerous-Speech-Guidelines-2013.pdf>. Acesso em: 01 mar 2021.

BOWMAN-GRIEVE, L. Exploring Stormfront: a virtual community of the radical right. **Studies in Conflict and Terrorism**, v. 11, n. 31, p. 989-1007, 2009.

BRASIL. Decreto-lei nº 2.848, de 7 de dezembro de 1940. Código Penal. Disponível em: http://www.planalto.gov.br/ccivil_03/decreto-lei/del2848compilado.htm. Acesso em: 26 fev. 2021.

BRASIL. Decreto nº 65.810, de 8 de dezembro de 1969. Promulga a Convenção Internacional sobre a Eliminação de todas as Formas de Discriminação Racial. Disponível em: http://www.planalto.gov.br/ccivil_03/decreto/1950-1969/D65810.html. Acesso em: 26 fev. 2021.

BRASIL. Constituição da República Federativa do Brasil. Brasília, 1988. Disponível em: http://www.planalto.gov.br/ccivil_03/constituicao/constituicaocompilado.htm. Acesso em: 25 fev. 2021.

BRASIL. Lei nº 7.716, de 5 de janeiro de 1989. Define os crimes resultantes de preconceito de raça ou de cor. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/l7716.htm. Acesso em: 26 fev. 2021.

BRASIL. Decreto nº 592, de 6 de julho de 1992a. Promulga Atos Internacionais. Pacto Internacional sobre Direitos Civis e Políticos. Disponível em: http://www.planalto.gov.br/ccivil_03/decreto/1990-1994/d0592.htm. Acesso em: 26 fev. 2021.

BRASIL. Decreto nº 678, de 6 de novembro de 1992b. Promulga a Convenção Americana sobre Direitos Humanos (Pacto de São José da Costa Rica), de 22 de novembro de 1969. Disponível em: http://www.planalto.gov.br/ccivil_03/decreto/d0678.htm. Acesso em: 26 fev. 2021.

BRASIL. Declaração e programa de ação [da III Conferência Mundial de Combate ao Racismo, Discriminação Racial, Xenofobia e Intolerância Correlata]. Brasília, DF: Ministério da Cultura, 2001. Disponível em: http://www.unfpa.org.br/Arquivos/declaracao_durban.pdf. Acesso em: 26 fev. 2021.

BRASIL. Lei nº 10.741, de 1 de outubro de 2003. Dispõe sobre o Estatuto do Idoso e dá outras providências. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/2003/l10.741.htm. Acesso em: 26 fev. 2021.

BRASIL. Lei nº 12.735, de 30 de novembro de 2012. Altera o Decreto-Lei nº 2.848, de 7 de dezembro de 1940 – Código Penal, o Decreto-Lei nº 1.001, de 21 de outubro de 1969 – Código Penal Militar, e a Lei nº 7.716, de 5 de janeiro de 1989, para tipificar condutas realizadas mediante uso de sistema eletrônico, digital ou similares, que sejam praticadas contra sistemas informatizados e similares; e dá outras providências. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2012/lei/l12735.htm. Acesso em: 26 fev. 2021.

BRASIL. Lei nº 12.965, de 23 de abril de 2014. Estabelece princípios, garantias, direitos e deveres para o uso da Internet no Brasil. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l12965.htm. Acesso: 26 fev. 2021.

BROWN, A. What is so special about online (as compared to offline) hate speech? **Ethnicities**, v. 18, n. 3, p. 297-326, 2018.

BRUGGER, W. Proibição ou proteção do discurso do ódio? algumas observações sobre o direito alemão e o americano. **Revista de Direito Público**, n. 15, p. 117-136, 2007.

BUYSE, A. Words of violence: "fear speech," or how violent conflict escalation relates to the freedom of expression. **Human Rights Quarterly**, v. 36, n. 4, p. 779-797, 2014.

COHEN-ALMAGOR, R. Fighting hate and bigotry on the Internet. **Policy and Internet**, v. 3, n. 3, 2011. pp.: 1-26.

CONSANI, C. F. Democracia e os discursos de ódio religioso: o debate entre Dworkin e Waldron sobre os limites da tolerância. **Etic@**, v. 14, n. 2, p. 174-197, 2015.

DWORKIN, R. **O direito da liberdade**. São Paulo: Martins Fontes, 2006.

DWORKIN, R. Foreword to *Extreme Speech and Democracy*. In: HARE, I.; WEINSTEIN, J. (eds). **Extreme Speech and Democracy**. New York: Oxford University Press, 2009, p. v-ix;

FARIS, R.; ASHAR,, A.; GASSER, U.; JOO, D. Understanding harmful speech online. **Berkman Klein Center Research Publication**, n. 2016-21, 2016. Disponível em: <https://papers.ssrn.com/sol3/papers.cfm?abstract%5Fid=2882824>. Acesso em: 01 mar 2021.

FERNANDEZ-MATAMORO, A.; FARKAS, J. Racism, hate speech, and social media: a systematic review and critique. **Television & New Media**, v. 22, n. 2, p. 205-224, 2021.

FORTUNA, P.; SOLER-COMPANY, J.; WANNER, L. Toxic, hateful, offensive or abusive? What are we really classifying? An empirical analysis of hate speech datasets. In: CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 12, 2020. **Proceedings**. European Language Resources Association, Marseille, 2020. p. 6786-6794.

FUNDAÇÃO GETULIO VARGAS. Diretoria de Análise de Políticas Públicas. **Monitor de temas**. Rio de Janeiro: FGV DAPP, 2015. Disponível em: <http://bibliotecadigital.fgv.br/dspace/handle/10438/15262>. Acesso em: 26 fev. 2021.

GILLESPIE, T. A relevância dos algoritmos. **Parágrafo**, v. 6, n. 1, p. 95-121, 2018.

JURNO, A. C.; D'ANDRÉA, C. (In)visibilidade algorítmica no “feed de notícias” do Facebook. **Revista Contemporânea**, v. 15, n. 2, p. 463-484, 2017.

LUCCAS, V. N.; GOMES, F. V.; SALVADOR, J. P. F.. **Guia de análise de discurso de ódio**. Rio de Janeiro: Fundação Getúlio Vargas, 2020. Disponível em: <https://www.conib.org.br/wp-content/uploads/2019/11/Guia-de-An%C3%A1lise-de-Discurso-de-%C3%93dio.pdf>. Acesso em: 26 fev. 2021.

LUNA, N e SANTOS, G. Liberdade de expressão e discurso do ódio no Brasil. **Revista Direito e Liberdade**, v. 16, n. 3, 2014, p. 227-255. Disponível em: http://www2.esmarn.tjrn.jus.br/revistas/index.php/revista_direito_e_liberdade/article/view/780. Acesso em: 01 mar 2021.

NANDI, J. **O combate ao discurso de ódio nas redes sociais**. Trabalho de Conclusão de Curso (graduação). Universidade Federal de Santa Catarina (UFSC), Campus Araranguá, Graduação em Tecnologias da Informação e Comunicação, Araranguá, 2018. Disponível em: <https://repositorio.ufsc.br/handle/123456789/187510>. Acesso em: 01 mar 2021.

NERIS, N. (coord); VALENTE, M.; CRUZ, F.; OLIVA, T. **Outras vozes: gênero, raça, classe e sexualidade nas eleições de 2018**. InternetLab, São Paulo, 2019. Disponível em: https://www.internetlab.org.br/wp-content/uploads/2019/10/OutrasVozes_2018.pdf. Acesso em 10/03/2021.

ONU (ORGANIZAÇÃO DAS NAÇÕES UNIDAS). **Carta das Nações Unidas**. São Francisco: ONU, 1945. Disponível em: <https://www.un.org/en/charter-united-nations/index.html>. Acesso em: 26 fev. 2021.

ONU (ORGANIZAÇÃO DAS NAÇÕES UNIDAS). **Declaração Universal dos Direitos Humanos**. Paris: ONU, 1948. Disponível em: <https://www.un.org/en/universal-declaration-human-rights/>. Acesso em: 26 fev. 2021.

PAREKH, B. Is there a case for banning hate speech? In: HERZ, M.; MOLNAR, P. (eds.). **The Content and Context of Hate Speech: Rethinking Regulation and Responses**. Cambridge: Cambridge University Press, 2012, p. 37-56.

PEARSON, J.; CHILD, J.; DEWITT, L.; KAHL, D. Facing the fear: an analysis of speech-anxiety content in public-speaking textbooks. **Communication Research Reports**, v. 24, p. 159-168, 2007.

ROGERS, R. Engajados de outra maneira: As Mídias Sociais - Das Métricas de Vaidade à Análise Crítica. In: OMENA, J. J. (ed.). **Métodos Digitais: Teoria e Prática**. Lisboa: Editora da Universidade Nova de Lisboa, 2019. p, 73-96.

RUEDIGER, M. (coord.). **Nem tão #simples assim: o desafio de monitorar políticas públicas nas redes sociais**. Rio de Janeiro: FGV DAPP, 2017. Disponível em: <http://hdl.handle.net/10438/19436>. Acesso em 01/03/2021.

SANCHES, D. Construindo grafos de conhecimento (knowledge graphs) para análise de redes sociais: uma análise sobre discurso de ódio em Portugal. In: WORKSHOP ON MEDIA, INFORMATION AND DATA SCIENCE, 1, 2020. **Anais**. Universidade Federal de Goiás, Goiânia, 2020.

SELLARS, Andrew. Defining Hate Speech. **Berkman Klein Center Research Publication**, n. 2016-20, Boston Univ. School of Law, Public Law Research Paper, p. 16-48, 2016. Disponível em: <https://ssrn.com/abstract=2882244> or <http://dx.doi.org/10.2139/ssrn.2882244>. Acesso em: 01 mar 2021.

SHEPHERD, A. Extremism, free speech and the rule of law: evaluating the compliance of legislation restricting extremist expressions with article 19 ICCPR. **Utrecht Journal of International and European Law**, v. 33, p. 62-83, 2017.

SIEGEL, A. Online hate speech. In: PERSILY, N.; TUCKER, J. (orgs.). **Social media and democracy**. Cambridge: Cambridge University Press, 2020. p.: 56-88.

SILVA, L. R.; BOTELHO-FRANCISCO, R. E.; OLIVEIRA, A. A.; PONTES, V. R. A gestão do discurso de ódio nas plataformas de redes sociais digitais: um comparativo entre Facebook, Twitter e Youtube. **Revista Ibero-americana de Ciência da Informação**, v. 12, n. 2, p. 470-492, 2019.

SILVIA, L.; MONDAL, M.; CORREA, D.; BENEVENUTO, F.; WEBER, I. Analyzing the targets of hate in online social media. In: **Proceedings of the Tenth International AAI conference on Web and Social Media**, 2016. Disponível em: <https://arxiv.org/abs/1603.07709v1>.

SIMONS, A.; MUELLER, John. The Dynamics of Internal Conflict. **International Security**, n. 24, v. 4, p. 187-192, 2001.

SLOAN, R. H.; WARNER, R.. Beyond notice and choice: privacy, norms, and consent. **Journal of High Technology Law**, v. 14, p. 370, 2014.

WALDRON, J. **The harm in hate speech**. Cambridge: Harvard University Press, 2012.

WEAVER, S. A rhetorical discourse analysis of online anti-Muslim and anti-Semitic jokes. **Ethnic and Radical Studies**, v. 3, n. 36, p. 483-499, 2013.

WEINSTEIN, J.; HARE, I. General introduction: free speech, democracy, and the suppression of extreme speech past and present. In: HARE, I.; WEINSTEIN, J. (eds.). **Extreme speech and democracy**. Nova York: Oxford University Press, 2009, p. v-ix.