eCompós

Revista da Associação Nacional dos Programas de Pós-Graduação em Comunicação

ISSN 1808-2599, v. 26, jan–dez, publicação contínua, 2023, p. 1–22.
doi.org/10.30962/ec.2709

Discurso de ódio on-line
Uma análise das políticas das plataformas digitais para moderação de conteúdo

LUIZA CAROLINA DOS SANTOS
Universidade Estadual de Ponta Grossa
Ponta Grossa, Paraná, Brasil

RENATA TOMAZ
Fundação Getulio Vargas
Rio de Janeiro, Rio de Janeiro, Brasil

DALBY DIENSTBACH
Fundação Getulio Vargas
Rio de Janeiro, Rio de Janeiro, Brasil

EURICO MATOS
Fundação Getulio Vargas
Rio de Janeiro, Rio de Janeiro, Brasil

DANIELLE SANCHES
Fundação Getulio Vargas
Rio de Janeiro, Rio de Janeiro, Brasil

ID 2709

Recebido em
12/11/2022
Aceito em
03/08/2023

/resumo

O objetivo deste artigo é compreender a moderação do discurso de ódio do Facebook, Instagram, TikTok, Twitch, Twitter e YouTube a partir da análise de seus termos de uso e suas diretrizes de comunidade. Na discussão, são analisados quatro fatores: conceituação de discurso de ódio; *categorias protegidas*; critérios de avaliação de denúncias; e sanções. Por fim, discutimos três pontos em comum às diretrizes de comunidade: os valores das plataformas; desafios contextuais e linguísticos; e interlocuções e práticas de moderação. A conclusão aponta que esses pontos suportam a governança do discurso de ódio das plataformas analisadas.

Palavras-chave: Discurso de ódio. Plataformas digitais. Moderação de conteúdo. Liberdade de expressão. Pesquisa documental.

Online Hate Speech: an Analysis of Digital Platforms Policies for Content Moderation

The aim of this article is to understand the moderation of hate speech on Facebook, Instagram, TikTok, Twitch, Twitter and YouTube by analyzing their terms of use and community guidelines. In the discussion, four factors are analyzed: conceptualization of hate speech; protected categories; criteria for evaluating complaints; and sanctions. Finally, we discuss three points in common with the community guidelines: the platforms' values; contextual and linguistic challenges; and interlocutions and moderation practices. The conclusion is that these points support the governance of hate speech on the platforms analyzed.

Keywords: Hate speech. Digital platforms. Content moderation. Freedom of expression. Documental research.

Discurso de odio en línea: un análisis de las políticas de moderación de contenidos de las plataformas digitales

El objetivo de este artículo es comprender la moderación del discurso de odio en Facebook, Instagram, TikTok, Twitch, Twitter y YouTube mediante el análisis de sus condiciones de uso y directrices comunitarias. En el debate se analizan cuatro factores: la conceptualización del discurso del odio; la definición de las categorías protegidas; los criterios para evaluar las denuncias; y las sanciones. Por último, discutimos tres puntos comunes a las directrices comunitarias: los valores de las plataformas; los retos contextuales y lingüísticos; y las interlocuciones y prácticas de moderación. La conclusión apunta que estos puntos sustentan la gobernanza del discurso de odio en las plataformas analizadas.

Palabras clave: El discurso del odio. Plataformas digitales. Moderación de contenido. La libertad de expresión. Investigación documental.

/autores

Luiza Carolina **DOS SANTOS**

Doutora em Comunicação e Informação pela Universidade Federal do Rio Grande do Sul (UFRGS) e professora substituta na Universidade Estadual de Ponta Grossa (UEPG). Coordena o Grupo de Pesquisa em Tecnologias e Culturas Digitais da Intercom.

Universidade Estadual de Ponta Grossa, Ponta Grossa, Paraná, Brasil

E-mail: luizacdsantos@gmail.com

ORCID

Renata **TOMAZ**

Doutora e mestre em Comunicação e Cultura pela Universidade Federal do Rio de Janeiro (UFRJ). Professora adjunta na Escola de Comunicação, Mídia e Informação da Fundação Getulio Vargas (FGV ECMI). Professora colaboradora do Programa de Pós-Graduação em Mídia e Cotidiano (PPGMC-UFF).

Fundação Getulio Vargas, Rio de Janeiro, Rio de Janeiro, Brasil

E-mail: renata.tomaz@fgv.br

ORCID

Dalby **DIENSTBACH**

Doutor em Estudos da Linguagem pela Universidade Federal Fluminense (UFF) e professor adjunto na Escola de Comunicação, Mídia e Informação da Fundação Getulio Vargas (FGV ECMI).

Fundação Getulio Vargas, Rio de Janeiro, Rio de Janeiro, Brasil

E-mail: dalby.hubert@fgv.br

ORCID

/autores

Eurico **MATOS**

Doutor e mestre em Comunicação e Cultura Contemporâneas pela Universidade Federal da Bahia (UFBA). Professor adjunto na Escola de Comunicação, Mídia e Informação da Fundação Getulio Vargas (FGV ECMI) e do Instituto Nacional de Ciência e Tecnologia em Democracia Digital (INCT.DD).

Fundação Getulio Vargas, Rio de Janeiro, Rio de Janeiro, Brasil

E-mail: eurico.neto@fgv.br

ORCID

Danielle **SANCHES**

Doutora em História das Ciências pela École des Hautes Études en Sciences Sociales (EHESS/Paris) em Cotutela com a Casa de Oswaldo Cruz (COC-Fiocruz) e professora adjunta na Escola de Comunicação, Mídia e Informação da Fundação Getulio Vargas (FGV ECMI).

Fundação Getulio Vargas, Rio de Janeiro, Rio de Janeiro, Brasil

E-mail: danielle.sanches@fgv.br

ORCID

Introdução

A definição de discurso de ódio e o debate sobre suas implicações sociais são muito anteriores à internet. O uso intensivo das mídias sociais, no entanto, induziu novos desdobramentos acerca do combate a formas de discriminação em ambientes digitais. Nos últimos anos, sua proeminência em espaços *mainstream* da internet 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 de violência física (SIEGEL, 2020). A estrutura das redes sociais dificulta a restrição da propagação dessas narrativas – um desafio que se apresenta tanto para as legislações nacionais quanto para a autorregulação das plataformas (RUEDIGER; GRASSI, 2021).

Neste artigo, analisaremos as políticas de combate ao discurso de ódio de seis plataformas digitais: Facebook, Instagram, Twitter, YouTube, TikTok e Twitch. Trata-se de pesquisa bibliográfica e documental – de caráter exploratório e com foco na análise de dados disponibilizados pelas plataformas nos termos de uso e nas diretrizes de comunidade – dividida em quatro partes. A primeira aponta especificidades do discurso de ódio nos ambientes digitais. Na sequência, indicaremos como as plataformas propõem a moderação desse conteúdo e suas limitações. Depois, descreveremos o percurso metodológico, seguido de uma análise das diretrizes de comunidade sobre discurso de ódio. A discussão dos dados produzidos teve como principal resultado a elaboração de três categorias de análise: os valores que sustentam as políticas, os desafios contextuais e linguísticos para classificar as violações e as práticas de moderação de conteúdo. Tais elementos, defendemos, são centrais para a compreensão da governança sobre discurso de ódio nas plataformas analisadas, conformando as políticas e as práticas de moderação, assim como seus desafios.

Discurso de ódio em ambientes digitais: definições e especificidades

O discurso de ódio é caracterizado pelo incitamento à violência por meio de ofensas e narrativas mordazes contra uma determinada pessoa ou um grupo em razão de suas características (FARIS *et al*., 2016). Ele tem um caráter discriminatório baseado na ideia de uma suposta superioridade de quem agride sobre quem sofre a agressão, o que a legitimaria (LUCCAS; GOMES; SALVADOR, 2020). Nesse sentido, o discurso de ódio pode ser tanto o ataque direto – na forma de insultos, intimidação e ameaças – quanto sua incitação. Em ambos os casos, as plataformas digitais se tornam arenas privilegiadas para a disseminação desse tipo de violência, criando um sentimento de "nós contra eles" (BRUGGER, 2007).

A complexidade das manifestações discriminatórias no ambiente digital se reflete na variedade de conceitos correlatos ao discurso de ódio. Benesch (2014), por exemplo, adota o termo *discurso perigoso* para designar enunciados que podem induzir ações violentas no mundo real. Buyse (2014), por sua vez, utiliza o conceito de *discurso do medo* para sinalizar práticas discriminatórias que fomentam o medo em grupos minoritários. Alguns autores adotam, ainda, termos como *discurso odioso* (SALEEM *et al*., 2017), *discurso nocivo* (FARIS *et al*., 2016) e *discurso extremo* (GAGLIARDONE, 2019).

Desde a publicação da Carta das Nações Unidas, que funda a Organização das Nações Unidas (ONU), existe um esforço em identificar segmentos mais atingidos por crimes de ódio e seu incitamento, reclamando "o respeito [...] sem distinção de raça, sexo, língua ou religião" (ONU, 1945, [s.p.]). A Declaração dos Direitos Humanos (ONU, 1948, [s.p.]) acrescentou a essas categorias as "de opinião política ou outra, de origem nacional ou social, de fortuna, de nascimento ou de qualquer outra situação". O Estado brasileiro, como signatário desses e outros documentos,[1] confere a eles força de lei no país e corrobora seus princípios

1 Conferir: Pacto Internacional sobre Direitos Civis e Políticos (1966), Convenção Americana sobre os Direitos Humanos (1969), Programa de Ação da III Conferência Mundial de Combate ao Racismo, Discriminação Racial, Xenofobia e Intolerância Correlata (2001).

em mecanismos legais como a Constituição Federal Brasileira (BRASIL, 1988), o Código Penal (BRASIL, 1989) e o Marco Civil da Internet (BRASIL, 2014). A ampliação do escopo de proteção sinaliza a percepção de que são dinâmicas as condições que vulnerabilizam determinados grupos de indivíduos, tornando-os alvos de discriminação (KONIKOFF, 2021).

A centralidade das plataformas digitais como mediadoras da sociabilidade na contemporaneidade particulariza a circulação de discurso de ódio, que ganha magnitude e assume uma natureza própria ao incorporar as funcionalidades que orientam a produção de conteúdo nos meios digitais. Embora o discurso de ódio *on-line* não seja intrinsecamente diferente de expressões semelhantes encontradas off-line, suas especificidades configuram um grande desafio no combate à sua propagação (GAGLIARDONE *et al*., 2015).

A *anonimidade* parcial dos usuários impõe barreiras à responsabilização por atos de ódio no contexto *on-line* e diminui a possibilidade de reação ou confronto físico entre agressor e vítima. A *invisibilidade* do agressor torna os ataques mais fáceis de serem efetivados, uma vez que os efeitos deles sobre a vítima não são visíveis para quem os realiza (BROWN, 2018).

Além disso, o caráter gregário da internet potencializa tanto a *formação de grupos de ódio*, que fomentam identidades específicas (WEAVER, 2013) e conferem projeção a seus membros (BOWMAN-GRIEVE, 2009), quanto processos de polarização. Ao priorizar nas *timelines* conteúdos consoantes à opinião do indivíduo, os algoritmos propiciam a criação de "câmaras de eco" (COLLEONI; ROZZA; ARVIDSSON, 2014). Opacos, esses regimes algorítmicos (JURNO; D'ANDRÉA, 2017) dificultam a desarticulação do discurso de ódio no ambiente *on-line*.

Por fim, o fácil *acesso a recursos de comunicação digital*, que economiza tempo e dinheiro, viabiliza a organização e a proliferação dos discursos de ódio (BROWN, 2018) e amplia sua disseminação. Ao lado das múltiplas formas de criar uma postagem contendo ofensas surgem diferentes tipos de agentes dessas publicações. *Trolls, haters* e *naysayers*[2] são agentes de intolerância cujas ações ofensivas tornam ainda mais complexa a moderação desse conteúdo.

Plataformas digitais: controle, moderação e regulação de conteúdo de ódio

Twitter, Facebook, Instagram, Twitch, YouTube e TikTok são plataformas digitais de mídias sociais baseadas na produção de conteúdo pelo usuário. Enquanto plataformas, constituem-se como um complexo arranjo de dimensões técnicas, políticas e econômicas (D'ANDRÉA, 2020) e são construídas a partir de uma infraestrutura amparada na coleta e no processamento de dados (GILLESPIE, 2018). Apesar dos diferentes formatos comportados por cada uma (fotos, vídeos, *streaming*, texto etc.), de modo geral elas hospedam, circulam e moderam o conteúdo produzido por seus usuários (GILLESPIE, 2018). Elas se autodenominam empresas de tecnologia e não de mídia, o que as libera de seguir a legislação e os mecanismos que regulam a atividade midiática em um dado país.

Embora não sejam autoras das publicações, as plataformas selecionam e hierarquizam os conteúdos disponíveis para os usuários (GILLESPIE, 2018) por meio de sua lógica algorítmica, assumindo o risco de seus efeitos. O próprio modelo de negócios das plataformas conforma práticas de discurso de ódio, pois seus padrões algorítmicos concedem maior visibilidade a postagens com mais interações (BEN-DAVID; MATAMOROS-FERNANDEZ, 2016). Soma-se a isso o fato de os ambientes digitais serem programados para promover conteúdos que geram "reações fortes", como os de ódio (LAVI, 2020). No caso de publicações ofen-

2 Para Lopes e Botelho-Francisco (2020), o *troll* é um interagente que encoraja discussões e conflitos a partir de comentários, comumente ofensivos, que deturpam a natureza das interações de um grupo. Os *haters* criam publicações independentes com discursos violentos e agressivos, como se fossem entretenimento. Já os *naysayers* consistem em manipuladores e opositores que usam o debate *on-line* para promover suas próprias agendas, desqualificando agressores e vítimas.

sivas, há uma disputa de narrativas entre os diferentes clusters cujo engajamento fomenta a monetização desses conteúdos, fazendo-os comercialmente oportunos.

Como esse conteúdo, incluindo o de ódio, é de um terceiro, não é viável fazer uma checagem prévia. Logo, a moderação só acontece após a publicação, diferentemente de uma mídia tradicional, que assume – e assina – a criação do que veicula. Contudo, mesmo não sendo autoras das publicações que suportam, as plataformas se apropriam delas para criar um ecossistema narrativo potencialmente comercializável de onde extraem vultosos lucros. Assim, a moderação de conteúdo integra não só a governança das plataformas, mas seu modelo de negócios.

De modo geral, as plataformas adotam dois tipos de moderação: a Moderação Comercial de Conteúdo (MCC) – realizada por humanos – e a Moderação Automática de Conteúdo (MAC) – realizada por máquinas. A moderação comercial pode ser definida como "uma prática organizada de escaneamento do conteúdo gerado por usuário postado em sites na internet, mídias sociais e outros espaços online"[3] (ROBERTS, 2019, p. 33, tradução nossa) e é realizada de forma manual. O principal mecanismo para acionar a MCC é a denúncia (flagging) dos próprios usuários sobre algum conteúdo[4] que viole as diretrizes da comunidade. Todavia, essa sinalização pode ter diferentes significados e apropriações, como a utilização desse recurso de maneira tática por ativistas *on-line* no combate a um determinado tipo de conteúdo e a denúncia de perfis e publicações de adversários em contextos de disputas político-partidárias (CRAWFORD; GILLESPIE, 2016).

A MCC pode ser feita por uma equipe própria em espaços da própria corporação, por equipes de empresas terceirizadas ou por plataformas de microtrabalho como a Amazon Mechanical Turk. A precariedade das condições de trabalho, nesse modelo, é uma variável relevante para a moderação de conteúdo de ódio, uma vez que os trabalhadores são expostos de forma sequencial a conteúdos violentos, são pouco especializados, frequentemente invisibilizados e muitas vezes vivem em contextos de vulnerabilidade (ROBERTS, 2019). Além disso, a concentração em determinados centros tira dos contextos social e cultural as postagens avaliadas e os próprios moderadores.

Twitter, Facebook e YouTube assinaram, em 2013, um acordo de combate ao discurso de ódio liderado pela Liga Antidifamação, uma organização sem fins lucrativos dos Estados Unidos.[5] Com isso, as três plataformas pactuaram analisar, de forma comprometida, as denúncias e os relatos de discurso de ódio em tempo hábil; explicar, de forma clara, como realizam a moderação de conteúdo para seus usuários e aplicar 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 a conteúdos de ódio. Termos de uso e diretrizes de comunidades mais claras, 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. Outros fatores também contribuíram para essas melhorias, tais como novas leis sobre conteúdo de ódio *on-line* e a interferência de governos (SILVA *et al*., 2019).

As pressões nesse sentido também levaram a um investimento cada vez maior na moderação automatizada. Esse tipo de moderação promete sanar o problema de escala, ou seja, uma alternativa teoricamente capaz de lidar com a grande quantidade de conteúdos a serem moderados com o menor esforço

3 No original: "Is the organized practice of screening user-generated content posted to internet sites, social media, and other online outlets".

4 Diferentes plataformas possuem *affordances* distintas para denúncias de conteúdo. As possibilidades de denúncia, as categorias disponibilizadas e os mecanismos de retorno aos denunciadores, por exemplo, variam consideravelmente entre as plataformas analisadas. Entretanto, esse detalhamento e essa discussão não serão feitos no presente artigo.

5 Disponível em: <https://www.adl.org/best-practices-for-responding-to-cyberhate>. Acesso em: 23 nov. 2021.

possível a partir da proceduralização de um processo "de forma a se replicar em diferentes contextos e parecer a mesma coisa"[6] (GILLESPIE, 2020, p. 2, tradução nossa). Tal estratégia pode inviabilizar o caráter social do problema, sobrevalorizando uma resposta tecnológica.

As soluções de inteligência artificial propõem a transformação de um grande conjunto de dados de treinamento em um conjunto pequeno de cálculos que possa agir sobre uma grande quantidade de conteúdo (GILLESPIE, 2020). A utilização de técnicas de machine learning para moderação de discurso de ódio incorre, entretanto, em alguns problemas: a) construção de um modelo a partir de dados anotados por anotadores humanos no passado, fruto de políticas e entendimentos localizados; b) dificuldade de compreensão de contexto e significados subculturais desses modelos; c) erros estatísticos cometidos por esses sistemas tendem a recair sobre grupos minoritários, também sub-representados em bancos de dados para treinamento (GILLESPIE, 2020).

Percurso metodológico

As plataformas digitais podem ser estudadas a partir de diferentes dimensões, como datificação, infraestrutura, modelo de negócio, governança e usos e *affordances* (D'ANDRÉA, 2020). Neste trabalho, a análise enfatiza os aspectos relacionados à governança, ou seja, as formas como as plataformas propõem a sua autorregulação, especificamente sobre discurso de ódio. Os principais documentos de governança das plataformas são os termos de uso e as diretrizes de comunidade, que irão compor nosso *corpus*.

Este trabalho apresenta uma análise documental (SÁ-SILVA; ALMEIDA; GUINDANI, 2009) com base nos termos de uso, nas diretrizes da comunidade e em alguns documentos auxiliares das plataformas Facebook, Instagram, TikTok, Twitch, Twitter e YouTube, a fim de compreender como elas moderam conteúdo de discurso de ódio. Os dados analisados correspondem às versões desses documentos disponibilizadas por essas plataformas em dezembro de 2021. Embora os termos de uso (ou serviço) analisados sequer mencionem "discurso de ódio", constituem uma documentação jurídica que estabelece as obrigações da plataforma e dos usuários, que inevitavelmente devem aceitar tais termos para fazer uso dos serviços por ela disponibilizados. É nesse documento também que a submissão às diretrizes de comunidade (regras ou políticas) é apresentada como condição de uso, sob pena de diferentes tipos de sanções, sendo a maior delas o cancelamento da conta.

As diretrizes de comunidade se configuram como o material base para nossa pesquisa, considerando que é no seu escopo que encontramos, nas seis plataformas, a definição para práticas relacionadas ao discurso de ódio. Também apresentam a forma como as plataformas negociam as expectativas de liberdade de expressão dos usuários e a sua segurança, especialmente daqueles presentes nas chamadas *categorias protegidas*. Além disso, são documentos retóricos que constroem uma imagem das próprias plataformas e que dialogam tanto com os usuários quanto com *stakeholders* e governos (GILLESPIE, 2018). Compreendemos tais documentos como uma fonte primária, registros em suporte digital cada vez mais acionados como *corpus* de pesquisas sobre a governança de plataformas (JIANG *et al.*, 2020; OBAR; OELDORF-HIRSCH, 2020).

Os termos de uso, as diretrizes de comunidade e outros textos prescritivos (sobre discurso de ódio) analisados estão associados às plataformas digitais em si, e não aos países específicos dos usuários em questão. Nesse sentido, são documentos que abrangem diversos países e regiões sem grandes alterações, com exceção do registro, nesses documentos, sobre a obrigação dos usuários de estarem em conformidade com a legislação local. Como falantes do português, priorizamos documentos idealmente acessíveis ao nosso idioma, o que implicou uma amostra predominantemente em língua portuguesa, conforme lista a seguir.

6 No original: "Such that it can be replicated in different contexts, and appear the same".

Plataforma	Documentos analisados
Facebook	• Termos de Serviço • Padrões da Comunidade • Organizações e Indivíduos Perigosos Discurso de Ódio • Como a Meta aplica as políticas • Who should decide what is hate speech in a Global Community
Instagram	• Termos de Uso • Diretrizes de Comunidade
Twitter	• Termos de Serviço do Twitter • As Regras do TwitterPolítica contra propagação do ódio • Política contra ameaças violentas • Política contra glorificação da violência • Sobre conteúdo ofensivo • Nossas opções de medidas corretivas • Abordagem de elaboração de políticas e filosofia de medidas corretivas do Twitter • Sobre exceções devido ao interesse público no Twitter
YouTube	• Termos de Serviço • Diretrizes da Comunidade • Política de discurso de ódio • A importância do contexto • Recursos limitados para determinados vídeos • Encerramento de canais e contas
TikTok	• Termos de Serviço • Diretrizes da Comunidade • Parceiros de segurança • Violações de conteúdo e banimentos
Twitch	• Termos de Serviço • Diretrizes da Comunidade • Conduta de ódio e assédio • Perguntas frequentes sobre as diretrizes de comunidade

Figura 1: Lista de documentos analisados[7]
Fonte: Elaboração própria.

Essa análise documental consistiu em dois momentos de categorização: uma *a priori* e outra *a posteriori*. Na primeira etapa, analisamos os documentos a partir de quatro categorias: definição de discurso de ódio; delimitação das *categorias protegidas*; clareza sobre métodos de avaliação das denúncias; e sanções aplicadas. Posteriormente, a articulação desses pontos nos permitiu formular categorias emergentes: posicionamento e valores das plataformas; desafios contextuais e linguísticos; e práticas de moderação. Ambas as etapas serão discutidas nos dois próximos tópicos.

7 A coleta de dados ocorreu em dezembro de 2021, mas a data do último acesso das páginas de documentação é 24 de junho de 2022. Considerando que os termos e as diretrizes são documentos editáveis, é possível que tenham passado por modificações desde o último acesso. Para garantir a possibilidade de recuperação dos documentos, buscamos no *Wayback Machine* pelo link de acesso aos documentos mais próximos do período de análise. A lista completa com *link*s para os documentos analisados está disponível em: <https://bit.ly/45xuG0X>.

O que dizem as plataformas sobre discurso de ódio

As diretrizes das seis plataformas apresentam restrição ou repúdio contra o discurso de ódio, categorizando-o como *comportamento* proibido. Em alguns casos, também tratam de discursos violentos, extremistas e perigosos ou formas organizadas de discurso de ódio. Variam, entretanto, no grau de detalhamento, na conceituação dos termos e na exemplificação das diretrizes que fornecem.

As diretrizes de Twitch e TikTok sobre *conduta* de ódio são detalhadas, com exemplos de diferentes *comportamentos* de ódio e relativamente concentradas em um único link, facilitando a leitura. Em suas diretrizes, a Twitch informa que aplica a mesma política em outros espaços da comunidade que não a própria plataforma, como eventos presenciais ou outras redes sociais. No YouTube, a página das diretrizes dá acesso à "Política de discurso de ódio". Além de definição, *categorias protegidas*, princípios de moderação, exemplos e possíveis sanções sobre discurso de ódio, esse documento oferece *link*s para outros tópicos que amplificam sua compreensão. Twitter, Facebook e Instagram impõem um pouco mais de dificuldade para os usuários encontrarem as informações. Para se chegar em "Política contra *conduta* de propagação do ódio", no Twitter, é necessário acessar, em Central de Ajuda, o documento "Regras do Twitter", que disponibiliza o link a partir da seção "Segurança". Embora mais confuso de navegar, o Twitter apresenta um conjunto de documentos mais detalhados sobre produção, circulação e moderação do discurso de ódio. É o único, por exemplo, a apresentar a filosofia que orienta sua elaboração de políticas de moderação e trata em um documento inteiro a importância do contexto de uma mensagem.

O Facebook, apesar de apresentar uma conceituação precisa dos termos utilizados, possui difícil navegação, dividindo as diretrizes sobre discurso de ódio em dois tópicos distintos e levando o usuário a navegar por outros *links* caso queira compreender melhor as penalidades ou formas de moderação da plataforma. O texto-base sobre discurso de ódio se divide em três níveis diferentes, possuindo diversos tópicos em cada um deles, mas sem exemplos para os usuários. Ainda, a plataforma informa que a versão mais atualizada das diretrizes se encontra no documento de referência em inglês, o que impõe uma barreira de acesso a falantes de outras línguas. A maior parte dos documentos sobre o tema está no *Transparency Center*, que é compartilhado com o Instagram. No *Transparency Center*, é possível acessar não apenas as diretrizes atuais sobre discurso de ódio, mas também as versões anteriores. As diretrizes são acompanhadas de relatórios e informações sobre denúncias, remoção de conteúdo e o *feedback* da plataforma.

O Instagram é o que menos detalha esses aspectos, com diretrizes sucintas e genéricas. Possui apenas termos de uso e diretrizes de comunidade, nos quais estabelece as dinâmicas relacionadas à circulação de discurso de ódio, vinculando os usuários às políticas da Meta, empresa proprietária do Instagram e do Facebook. A falta de especificidade da plataforma fortalece a dificuldade na moderação de conteúdo de ódio, uma vez que suas diretrizes não levam em conta as especificidades de formato, comunidade, gramáticas e práticas do Instagram.

As Figuras 2 e 3 apresentam a definição de discurso de ódio de cada plataforma, as *categorias protegidas* elencadas, os critérios de avaliação de denúncias e as sanções previstas.

	Twitch	TikTok	Twitter
Definição de discurso de ódio	Define conduta de ódio como "conteúdos ou atividades que promovam ou incentivem discriminação, difamação, assédio ou violência com base nas seguintes características pro-tegidas: raça, etnia, cor, casta, nacionalidade, astutas migratório, religião, sexo, gênero, i-dentidade de gênero, orientação sexual, de-ficiência, condição médica grave e status de veterano".	"Definimos discurso ou comportamento de ódio como conteúdo que ataque, ameace, incite violência contra ou, de alguma forma, desumanize um indivíduo ou grupo, com base nos seguintes atributos protegidos.". Dentro do tópico de "Extremismo violento", as dire-trizes nomeiam o ódio organizado: "refere-se a indivíduos e organizações que ataquem pessoas com base em características prote-gidas (...) Consideramos que ataques incluem ações que incitem à violência ou ao ódio, de-sumanizem indivíduos ou grupos ou acolham ideologias de ódio."	Define a propagação do ódio como o ato de "promover violência, atacar diretamente ou ameaçar outras pessoas com base" em ca-tegorias protegidas e "incitar lesões a outros com base nessas categorias."
Categorias protegidas	Raça, etnia, cor, casta, nacionalidade, astutas migratório, religião, sexo, gênero, identidade de gênero, orientação sexual, deficiência, con-dição médica grave, status de veterano e cer-tas proteções em relação à idade.	Raça, Etnia, origem nacional, religião, casta, Orientação sexual, Sexo, Gênero, Identidade de gênero, Doença grave, Deficiência, Con-dição migratória.	Raça, etnia, origem nacional, orientação sexual, sexo, identidade de gênero, religião, idade, deficiência ou doença grave; além de grupos alvo de assédio: mulheres, negros, lés-bicas, gays, bissexuais, transexuais, homosse-xuais, intersexuais, indivíduos assexuados, comunidades marginalizadas e historicamente sub-representadas.
Critérios de avaliação de denúncias	Violação das diretrizes de comunidade, con-texto, interesse público e maior dureza na análise de condutas de ódio em relação a outros conteúdos.	Violação de regras de comunidade; rein-cidência da conta; gravidade e frequência da violação; interesse público; atividades do titular da conta em outras redes sociais ou mesmo fora da internet.	Violação das regras de uso; gravidade e his-tórico da conta responsável pelo conteúdo; interesse público; contexto.
Sanções por violação	Remoção de conteúdo, aviso de infração e/ ou suspensão da conta; punição é mais severa quando o comportamento é "direcionado, pessoal, explícito ou repetido/ prolongado, in-cita mais abusos ou envolve ameaças de violência ou coerção"; Violações mais graves podem resultar em suspensão indefinida na primeira transgressão.	Advertência na primeira violação. Em caso de reincidência, a conta poderá sofrer banimento temporário com limitação gradativa de fun-cionalidades (suspensão da publicação de vídeos, comentários ou edição do perfil). Em caso de insistência, a remoção do conteúdo poderá ser seguida de banimento permanente da conta. Para a violação das políticas de to-lerância zero, o banimento da conta será automático, sem aviso, e poderá haver blo-queio do dispositivo para evitar a criação de novas contas.	Solicitação de remoção do conteúdo de violação e suspensão das funcionalidades de interação (modo leitura), cujo período poderá se alongar, dependendo da reincidência, ou mesmo resultar em suspensão permanente da conta. Ameaças violentas levam ao banimento imediato e permanente, sem aviso. Suspensão dos mecanismos de engajamento e de re-comendação.

Figura 2: Diretrizes de comunidade sobre discurso de ódio de Twitch, TikTok e Twitter
Fonte: Elaboração própria.

	Facebook	Instagram	Youtube
Definição de discurso de ódio	"Definimos discurso de ódio como um ataque direto a pessoas, e não a conceitos e insti-tuições, baseado no que chamamos de carac-terísticas protegidas."	Em sua própria página, não define discurso de ódio, mas informa que "não é aceitável incentivar a violência ou atacar alguém" com base em categorias protegidas. Definição em si apenas nos documentos compartilhados com Facebook.	"(...) conteúdo que promove a violência ou o ódio contra pessoas ou grupos com base em qualquer um dos seguintes atributos" rela-cionados a categorias protegidas.
Categorias protegidas	Raça, etnia, nacionalidade, religião, orientação sexual, casta, sexo, gênero, identidade de gênero, doença grave ou deficiência, status migratório. Idade e ocupação, se associadas a uma categoria protegida.	Raça, etnia, nacionalidade, sexo, gênero, identidade de gênero, orientação sexual, re-ligião, deficiências ou doenças.	Idade; classe social; deficiência; etnia; iden-tidade e expressão de género; nacionalidade; raça; estatuto de imigrante; religião; sexo/gê-nero; orientação sexual; vítimas de um grande evento violento e respetivos familiares; esta-tuto de veterano.
Critérios de avaliação de denúncias	Violação dos Padrões de Comunidade, contexto da mensagem, questões culturais, intenção do emissor, interesse público, existência de ameaça de dano no conteúdo.	Violação das Diretrizes de Comunidade, contexto da mensagem, questões culturais, intenção do emissor, interesse público, exis-tência de ameaça de dano no conteúdo.	Violação de regras de comunidade; rein-cidência da violação; gravidade e frequência da violação; grau de proximidade de uma violação; interesse público; contexto.
Sanções por violação	Remoção de conteúdo, restrição de acesso ao conteúdo, desativação permanente de usuá-rio.	Exclusão de conteúdo, desativação perma-nente de contas ou outras restrições.	Aviso sem penalidade na primeira violação. Nas seguintes, sequência de até três adver-tências com suspensão de funcionalidades (publicar vídeos, fazer stream, criar listas de reprodução, criar novo canal) e dos meca-nismos de engajamento e recomendação por até duas semanas. Após três advertências em 90 dias, o canal é encerrado. Um único caso de abuso grave poderá resultar no encer-ramento do canal ou da conta, sem aviso.

Figura 3: Diretrizes de comunidade sobre discurso de ódio de Facebook, Instagram e YouTube
Fonte: Elaboração própria.

As diversas formas de definir discurso de ódio em ambientes digitais repercutem em como as plataformas caracterizam o fenômeno. Por exemplo, a menção a "extremismo violento" no TikTok remete às práticas de organizações extremistas que têm como objetivo atacar indivíduos ou grupos com base em atributos protegidos. No Facebook, discurso de ódio inclui tipos distintos de ataques, como "discursos violentos ou desumanizantes, estereótipos prejudiciais, declarações de inferioridade, expressões de desprezo, repugnância ou rejeição, xingamentos e apelos à exclusão ou segregação"[8]. Em maior ou menor nível, as definições das plataformas tendem a abarcar diferentes práticas discriminatórias em ambientes *on-line* – do *comportamento* rude à incitação à violência física contra grupos ou indivíduos. Além das formas enunciativas, algumas diretrizes incluem em suas definições de discurso de ódio atitudes dos usuários. A Twitch foca suas diretrizes em *conduta* de ódio, o TikTok trata de *comportamento* de ódio, e o Twitter inclui entre as práticas de ódio o retuíte de mensagens violentas contra *categorias protegidas*.

A menção às características ou *categorias protegidas* elencadas nos documentos é um dos principais pontos em comum das plataformas analisadas. Facebook, Instagram, TikTok, Twitch, Twitter e YouTube fornecem uma lista das categorias ou grupos por elas considerados protegidos, incluindo em todas elas raça, etnia, nacionalidade, religião, sexo, gênero, identidade de gênero, orientação sexual e deficiência. Algumas, entretanto, aparecem como exceção. É o caso de classe social, presente somente no YouTube; idade, no Twitter e no YouTube; status de veterano, na Twitch e no YouTube; e status migratório, na Twitch, no TikTok, no Facebook e no YouTube. Já a condição médica grave não é protegida apenas no YouTube. O Twitter fala de comunidades marginalizadas e historicamente sub-representadas, além de ser a única plataforma que informa levar em conta a interseccionalidade das categorias. Observa-se que três das características protegidas apresentadas pelas plataformas como consenso se relacionam com as dos 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 algumas das plataformas: traços físicos e classe social (SILVA *et al*., 2016).

Nenhuma plataforma expõe, em detalhes, os procedimentos em torno das sanções aplicadas para os usuários que publicam conteúdos de discurso de ódio. De modo geral, apresentam o que consideram critérios na avaliação dos conteúdos denunciados ou suspeitos, sendo o primeiro deles a violação das diretrizes. Levam em conta também o histórico do usuário, se houve reincidência de conteúdo de ódio, se é uma violação grave ou se se trata de uma expressão contextualizada. O interesse público é o único critério de exceção que aparece para permanência do conteúdo de ódio em todas as plataformas analisadas.

Quanto às formas de sanção, as plataformas trabalham com um mesmo conjunto de ações: notificação do usuário, restrição de uso ou de acesso ao conteúdo, alteração no algoritmo de visibilidade e engajamento, remoção do conteúdo e banimento do usuário. A maioria das plataformas indica em seus termos de uso que, uma vez banido, o usuário está proibido de voltar a utilizar seus serviços. Entretanto, quase não existem diretrizes sobre como e em quais casos cada sanção será aplicada, o que fica para avaliação caso a caso. O Twitter e a Twitch fornecem algumas indicações sobre a aplicação das penalidades, como o histórico do usuário na plataforma, no caso do primeiro, e o *comportamento* do tipo repetido, prolongado e direcionado, no caso da segunda. O YouTube é a plataforma que descreve com maior clareza os procedimentos relacionados às sanções, considerando gravidade e frequência das infrações. Além disso, altera o algoritmo de recomendação e visibilidade da publicação se ela for caracterizada como conteúdos "próximos da linha de remoção"[9], isto é, próximo de violar as diretrizes de comunidade.

8 Trecho retirado do documento "Discurso de Ódio" do Facebook.

9 Trecho retirado do documento "Recursos limitados para determinados vídeos" do YouTube.

Categorias emergentes para análise de discurso de ódio e moderação de conteúdo

Com base na análise dos documentos, identificamos três aspectos que sustentam as referidas políticas de moderação. Eles configuram novas categorias de análise, por meio das quais é possível adensar a compreensão sobre como as plataformas lidam com a produção e a circulação de conteúdo de discurso de ódio. São elas: os valores que sustentam as políticas; os desafios contextuais e linguísticos para classificar as violações; e as práticas de moderação de conteúdo.

Liberdade de expressão, segurança e interesse público

O modo como as plataformas digitais se posicionam em relação ao combate de práticas discriminatórias em ambientes *on-line* se manifesta a partir da articulação entre três valores fundamentais: a liberdade de expressão, a promoção de um "ambiente" seguro e o interesse público. As seis plataformas defendem, em suas diretrizes de comunidade, a diversidade de vozes como meio para promoção de um ambiente deliberativo aberto à autoexpressão dos indivíduos. A interpretação de que a liberdade de expressão é um direito essencial relativo e não absoluto é preponderante e permeia os documentos analisados de todas as plataformas, tendo a busca por um ambiente comunicacional plural como condicionante.

A liberdade de expressão condicionada ao respeito aos demais membros da comunidade opera como um instrumento de garantia da livre manifestação de pensamento e de identidade de grupos minorizados. Na Twitch admite-se, por exemplo, a expressão de "opiniões impopulares", desde que não constituam abuso ou assédio. De acordo com as diretrizes do Facebook, "as pessoas se comunicam e se conectam mais livremente quando não se sentem atacadas pelo que são"[10]. O mesmo posicionamento aparece nos documentos do Twitter, ao reconhecerem a importância de limitar determinados tipos de discurso, pois "a capacidade de expressão de pessoas que sofrem assédio no Twitter pode ser colocada em risco". Esse pensamento constrói, narrativamente, a posição das plataformas pela autorregulação e a não interferência externa (GILLESPIE, 2018), além de se relacionar ao argumento sobre a urgência da proteção dos indivíduos.

O tensionamento gerado pela defesa do direito à liberdade de expressão condicionada aos limites estabelecidos pelos marcos legais da garantia da dignidade humana por vezes é revertido pelo discurso de promoção de um "ambiente seguro" para o exercício da pluralidade. Nas diretrizes gerais do Facebook, por exemplo, a segurança é um dos valores que justificam a limitação da liberdade de expressão, acompanhada de autenticidade, privacidade e dignidade.

De modo geral, entre as plataformas a própria organização dos documentos explicita o discurso sobre ambiente seguro e sua relação com outros valores. É comum, por exemplo, haver seções sobre segurança que apresentam relatórios de remoção de conteúdo, descrevem o funcionamento dos sistemas de denúncia de violação de regras e oferecem um *feedback* aos usuários sobre os tipos de conteúdo excluídos da plataforma. Assim, observa-se que o exercício de construção do almejado ambiente seguro para debate nas plataformas digitais é associado expressamente ao esforço de promoção de transparência sobre as políticas adotadas pela plataforma.

Além de mecanismos para garantir pluralidade, diversidade e segurança dos usuários, as plataformas digitais definem condições específicas para a admissão da circulação de discursos discriminatórios. Em todas, encontramos ao menos um critério de exceção para conteúdo de discurso de ódio que viole as diretrizes da comunidade e que, por isso, deve ser removido. Essa exceção está amparada no valor do interesse público, comumente ligado à ideia de conscientização do problema, razão pela qual deverá ser

10 Trecho retirado do documento "Discurso de Ódio" do Facebook.

mantido, podendo aparecer tanto na forma de conteúdo satírico e humorístico – como preveem Facebook, Instagram e Twitch – quanto como um conteúdo educacional, científico, artístico ou documental – como defende o YouTube. O TikTok acrescenta a essa lista conteúdos em contextos fictícios, contra-falas e conteúdo "com valor jornalístico ou que, de outra forma, permita a expressão individual sobre temas de importância social".[11] Já o Twitter vincula interesse público a figuras de grande visibilidade, como chefes de Estado,[12] e pode optar por manter o conteúdo sob a justificativa de a publicação funcionar como um elemento de responsabilização de seus autores.

Tanto o interesse público quanto a liberdade de expressão e a noção de ambiente seguro são, desse modo, valores que sustentam o modelo de autorregulação das plataformas analisadas. Compreender em que base são acionados e que dilemas suscitam pode oferecer discussões qualificadas sobre a relação entre discursos de ódio, plataformização e ética.

Desafios contextuais e linguísticos

Os documentos das seis plataformas informam que as análises contextual e linguística podem participar do processo que torna a política de moderação em ação de remoção (ou de permanência). Segundo as suas diretrizes, Facebook, Instagram, Twitter, YouTube, TikTok e Twitch tendem a manter orientações globais com critérios que seriam aplicados em contextos locais. Acompanhar a mudança nos usos e apropriações dos termos – ora usados para xingar um grupo, ora utilizados pelo mesmo grupo em práticas de sociabilidade – é um trabalho constante para os revisores que analisam denúncias ou pedidos de contestação de remoção.

Na execução desse trabalho, merecem atenção alguns aspectos pertinentes aos níveis tanto linguístico quanto discursivo do uso da linguagem que concorrem para a construção dos sentidos de um enunciado e intervêm nos esforços para se classificar uma postagem como sendo discurso de ódio. Um primeiro ponto diz respeito a componentes estruturais da língua – isto é, relativos aos seus elementos mais formais –, tal como o caráter polissêmico inerente, em princípio, ao léxico de qualquer língua natural (CANÇADO, 2012). Os vários significados que virtualmente todas as palavras de uma língua podem abrigar (principalmente quando não se vinculam a uma área de especialidade) podem orientar a codificação (ou a decodificação) dos enunciados para uma ou para outra direção. A lista de acepções que completam entradas como "burro", "asno" e "jumento" no dicionário, por exemplo, poderiam embaralhar algumas decisões quanto a uma determinada postagem ser ou não discurso de ódio.

É necessário, ainda, levar-se em consideração, nos esforços para a identificação de discurso de ódio em mídias sociais, o conjunto de condições de produção e recepção que integram o evento de realização das postagens e interferem no sentido dos textos (ORLANDI, 2005). Essas condições incluem desde aspectos situacionais da enunciação – como local, período e interlocutores da postagem – até componentes das conjunturas social, cultural e histórica dessa enunciação. Fazem parte dos primeiros, entre outras coisas, o autor da postagem, as suas intenções, os seus interlocutores e o tipo de atividade social em que a mensagem é colocada em funcionamento (BHATIA, 2004). Um exemplo de como esses componentes operam na enunciação seria uma postagem que dirigisse um elogio ou um comentário favorável a um dado indivíduo. Dependendo da relação entre os interlocutores – por exemplo, se desafetos ou rivais políticos –, ela pode manifestar, em vez disso, um enunciado irônico ou ofensivo.

11 Trecho retirado do documento "Diretrizes de Comunidade" do Tiktok.

12 Apesar da menção explícita a chefes de Estado como figuras cujos discursos possuem interesse público, o próprio Twitter já optou por remover conteúdo produzido por Jair Bolsonaro no exercício de seu mandato e realizou a suspensão de Donald Trump no início de 2021 (GONTIJO, 2021).

Nas suas diretrizes, as plataformas parecem atentar ao funcionamento dos dois níveis apresentados acima. No que se refere ao estrato lexical das mensagens que circulam nas redes, por exemplo, Twitch, Facebook e Twitter insinuam ponderações sobre traços semânticos potencialmente pejorativos ou ofensivos de palavras e expressões ao mirarem estereótipos dos grupos protegidos. Embora reconheçam que dados termos supõem, de modo hegemônico, significados ofensivos, também afirmam considerar aspectos situacionais relativos ao evento da enunciação nas suas decisões de moderação. Conforme diretrizes, as definições e decisões sobre discurso de ódio ficam a cargo da intenção e/ou da interpretação dos seus interlocutores sobre o texto que leem – ainda que a forma como se identifica a intenção ou a interpretação não esteja clara. Especificamente o Twitter e a Twitch acrescentam que comunidades podem adotar entre si palavras e expressões que, em princípio, soariam pejorativas; entretanto, atribuem-lhes outros encaminhamentos semântico-pragmáticos, como forma quer seja de empoderamento, de reafirmação de identidades ou de reconhecimento de pares.

Quando se amplia ainda mais o escopo de fatores que podem intervir na construção dos significados de uma mensagem e eventualmente promover algum discurso de ódio, pode-se listar, entre outras coisas, as formações político-ideológicas evidenciadas no evento da enunciação; os *comportamento*s e crenças hegemônicos, bem como os marginalizados, nessa conjuntura; e os papéis sociais e as relações de poder entre os indivíduos e grupos de indivíduos envolvidos na emissão da postagem. É o que tem ficado evidente, atualmente, com discussões – presentes, inclusive, nas mídias sociais – sobre o uso de pronomes e flexões de gênero de substantivos e adjetivos (masculino e feminino) quando nos dirigimos a travestis e pessoas transgêneras (VIVERITO; BERTUCCI, 2020). Uma dada escolha de pronomes para se referir a certas identidades de gênero poderia – a depender das condições em que se insere – infringir camadas de polidez, soar ofensiva ou, em última análise, incorrer em discurso de ódio. Vale salientar que todas as plataformas firmam ater-se a esses fatores como critérios elementares de moderação de conteúdo. Elas alegam cotejar "normas culturais", "o contexto e a história" ou "culturas e subculturas" das localidades em que operam para estabelecer limites aos conteúdos. Contudo, as regras que disponibilizam são pouco claras quanto à configuração e ao funcionamento desse universo; não explicitam pontualmente quais componentes compõem o que chamam de culturas ou de contextos, como acessam esses componentes ou de que modo eles participam das decisões de moderação.

Das políticas às práticas de moderação

Os documentos analisados apontam a participação de especialistas, ativistas, acadêmicos e integrantes de ONGs como interlocutores das plataformas para criar políticas e as atualizar. Não encontramos, contudo, referências específicas a educadores, pais, responsáveis, representantes governamentais ou juristas, por exemplo. Não identificamos, ainda, qualquer tipo de indicação de um trabalho cooperativo entre plataformas para um intercâmbio de informações, experiências ou estratégias de enfrentamento da proliferação de discurso de ódio. As diretrizes também não fazem menção à regulação externa. Autoridades são mencionadas apenas para dizer que, havendo a detecção de perigo contra alguém ou algum grupo, elas serão acionadas pelas plataformas.

A moderação do conteúdo de discurso de ódio tem uma relação direta com as denúncias dos usuários, que aparecem como a principal ferramenta para seu combate. O que é denunciado gera não apenas estatísticas, mas um entendimento sobre como determinados indivíduos ou grupos se sentem em relação a certos conteúdos, além de alimentar os processos de aprendizagem de máquina. A expectativa de que o usuário não só produza conteúdo, mas também o fiscalize aponta uma terceirização da prática de moderação. O sistema de denúncias se configura, nesse sentido, como um trabalho não remunerado apropriado pelas plataformas, agregando valor de um ambiente seguro ao usuário e sendo, consequentemente, atrativo para anunciantes.

Na Twitch, além do sistema de denúncia e da moderação de conteúdo institucional, a plataforma acrescenta aos produtores de conteúdo a responsabilidade de monitorar e moderar a *conduta* de ódio em seus canais, sobretudo no formato ao vivo (característico da Twitch). Para esse fim, a plataforma disponibiliza a ferramenta *AutoMod* – capaz de fazer filtragem e moderação de conteúdo nos *chats* ao vivo de forma automática, uma vez programada – e indica a participação de moderadores humanos da própria comunidade do produtor durante as transmissões. Aos produtores de conteúdo, é permitido implementar, em suas comunidades, regras mais rigorosas do que aquelas previamente estabelecidas nas políticas da plataforma.

Conforme apresentado no terceiro tópico deste trabalho, há basicamente dois tipos de moderação: a humana e a automatizada. Quanto à primeira, os documentos por nós analisados, em sua maioria, não explicitam esses processos. O Facebook e Instagram, da empresa Meta, por exemplo, são as únicas plataformas que abordam esse tema em *links* disponíveis no *Transparency Center*. Já o Twitter é a plataforma que fornece mais detalhes a esse respeito, buscando enfatizar sua moderação baseada em equipes com diferentes funções e expertises. Apesar desses esforços específicos, o silenciamento, em maior ou menor grau, sobre treinamento, *background*, processo e apoio aos moderadores de conteúdo profissionais indica uma falta de transparência das plataformas digitais acerca de suas tecnologias, arquiteturas e práticas, como aponta a literatura (GILLESPIE, 2018). Não fica claro, portanto, como as plataformas passam das suas políticas de moderação para as práticas de moderação.

Sobre a moderação automatizada, das seis plataformas analisadas, apenas o YouTube informa, nas diretrizes de comunidade, que, além das denúncias de conteúdo inadequado pelos usuários, também realiza um monitoramento próprio através de aprendizado de máquina para detectar conteúdo problemático em escala, que é posteriormente revisado por humanos. Facebook, Instagram e TikTok, em outros documentos, indicam o uso de monitoramento automatizado. No caso do Facebook e do Instagram, o monitoramento automatizado é feito de duas formas: via automação para detecção de conteúdo inapropriado e remoção automática ou via automação para avaliação de conteúdo denunciado por usuários. Sabe-se, porém, que o Twitter também utiliza mecanismos automatizados na detecção de conteúdos que violem suas diretrizes (SILVA *et al*., 2019). Entretanto, não existe, por parte das plataformas analisadas, transparência sobre os modelos utilizados para a moderação automatizada.

A automação da moderação de conteúdo pode oferecer maiores rapidez e escalabilidade aos processos, mas não sem limitações. Mesmo com o avanço tecnológico, a implementação de modelos de machine learning sempre será realizada com base em dados de decisões tomadas no passado, o que faz com que o discurso de ódio seja, na prática, uma categoria estanque – e não uma categoria em renegociação a partir das práticas e acordos da própria comunidade da plataforma. Ainda, sistemas automatizados possuem dificuldade de lidar com as questões contextuais e linguísticas discutidas no tópico anterior.

Conclusões

Neste artigo, objetivamos entender como Facebook, Instagram, TikTok, Twitch, Twitter e YouTube compreendem o discurso de ódio com base em uma análise documental de seus termos de uso e suas diretrizes de comunidade. A discussão foi realizada em duas etapas, sendo a primeira baseada em quatro categorias de análise definidas a priori. Por meio delas, identificamos que as plataformas apresentam diferentes níveis de detalhamento e de dificuldade na localização de informações sobre discurso de ódio. Apesar de apresentarem um certo consenso sobre as *categorias protegidas* de raça, etnia, nacionalidade, religião, sexo, gênero, identidade de gênero, orientação sexual e deficiência, elas se diferem em outros aspectos, como traços físicos e classe social. Nenhuma delas é suficientemente clara sobre os procedimentos de sanção, mas apresentam uma gradação comum que vai da notificação do usuário até seu banimento.

Em um segundo momento, discutimos três pontos recorrentes nos documentos analisados, os quais suportaram a elaboração de três categorias de análise *a posteriori*: valores que sustentam o posicionamento das plataformas, desafios contextuais e linguísticos e transformação das políticas em práticas de moderação. Essas recorrências se traduzem enquanto elementos que conformam as políticas de moderação das diferentes plataformas analisadas, explicitando os elementos centrais da proposta (e dos desafios) de governança em torno do discurso de ódio existente nesses espaços.

Os valores de liberdade de expressão, da manutenção de um ambiente seguro e do interesse público são pilares que conformam as políticas de moderação das plataformas. Apesar dos esforços legislativos e das políticas corporativas de moderação, prevalece ante as plataformas o seguinte desafio: como garantir a segurança de *categorias protegidas*, interferindo o mínimo possível na liberdade de expressão dos usuários? Essa questão também aponta para a contradição na forma que as plataformas tratam do assunto: por um lado, temem os efeitos da não moderação no conteúdo de ódio; por outro, são ambivalentes quanto à intervenção (GILLESPIE, 2018).

Além disso, as dificuldades impostas pela contextualidade inerente ao discurso de ódio também são uma questão levantada pelas plataformas 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. Ainda que as plataformas reconheçam a importância desses aspectos, apresentam em seus documentos poucos recursos de moderação para lidar com essas questões.

Por fim, ficam evidentes as dificuldades estruturais das plataformas para que as políticas de moderação do discurso de ódio se tornem práticas: dependência do trabalho – gratuito – de usuários para fazer denúncias; ausência de estratégias de colaboração mútua para enfrentar a discriminação; terceirização, muitas vezes precarizada, da mão de obra que realiza a moderação humana. Desse modo, políticas e práticas de moderação nas plataformas digitais pensadas globalmente carecem de mecanismos efetivos para lidarem com as particularidades do discurso de ódio nos níveis nacional e regional, especialmente no que diz respeito às subculturas.

Referências

BEN-DAVID, A.; MATAMOROS-FERNÁNDEZ, A. Hate Speech and Covert Discrimination on Social Media: Monitoring the Facebook Pages of Extreme-Right Political Parties in Spain. **International Journal of Communication**, v. 10, p. 1167-1163, 2016.

BENESCH, S. Defining and Diminishing Hate Speech. In: MINORITY RIGHTS GROUP International. **State of the World's Minorities and Indigenous Peoples**. Londres, 2014. p. 18-25.

BHATIA, V. **Worlds of Written Discourse:** a Genre-Based View. Londres: Continuum, 2004.

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. 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.

_____. Lei n. 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.

_____. Lei n. 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 em: 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 de ódio? Algumas observações sobre o Direito Alemão e o Americano. **Revista Direito Público**, v. 15, p. 117-63, 2007.

BUYSE, A. Words of Violence: Fear Speech, or How Violent Conflict Escalation Relates to the Freedom of Expression. **Human Rights Quarterly**, v. 36, p. 779-797, 2014.

CANÇADO, M. **Manual de semântica:** noções básicas e exercícios. 2. ed. São Paulo: Contexto, 2012.

COLLEONI, E.; ROZZA, A.; ARVIDSSON, A. Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data. **Journal of Communication**, n. 64, v. 2, p. 317-332, 2014.

CRAWFORD, K.; GILLESPIE, T. What is a Flag For? Social Media Reporting Tools and the Vocabulary of Complaint. **New Media & Society**, v. 18, n. 3, p. 410-428, 2016.

D'ANDRÉA, C. **Pesquisando plataformas online:** conceitos e métodos. Salvador: EDUFBA, 2020.

FARIS, R. *et al*. Understanding Harmful Speech Online. **Berkman Klein Center Research Publication**, n. 2016-21, p. 1-19, 2016. Disponível em: <https://papers.ssrn.com/sol3/papers.cfm?abstract%5Fid=2882824>. Acesso em: 1 mar. 2021.

GAGLIARDONE, I. *et al*. **Countering Online Hate Speech**. Paris: Unesco Publishing, 2015.

GAGLIARDONE, I. Extreme Speech | Defining Online Hate and Its "Public Lives": What is the Place for "Extreme Speech"?. **International Journal of Communication**, v. 13, p. 3068-3087, 2019.

GILLESPIE, T. **Custodians of the Internet:** Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. New Heaven: Yale University Press, 2018.

_____. Content Moderation, AI, and the Question of Scale. **Big Data & Society**, p. 1-5, jul.-dez. 2020.

GONTIJO, A. **Governança e moderação de conteúdo:** uma análise da plataforma Twitter entre 2018 e 2021. Dissertação (Mestrado em Comunicação Social) – Faculdade de Filosofia e Ciências Humanas, Universidade Federal de Minas Gerais, Belo Horizonte, 2021.

JIANG, J. A. *et al*. Characterizing Community Guidelines on Social Media Platforms. In: CONFERENCE COMPANION PUBLICATION of the 2020 on Computer Supported Cooperative Work and Social Computing. **CSCW'20 Companion**, Estados Unidos, p. 287-291, out. 2020. Disponível em: <https://dl.acm.org/doi/pdf/10.1145/3406865.3418312>. Acesso em: 4 fev. 2022.

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.

KONIKOFF, D. Gatekeepers of Toxicity: Reconceptualizing Twitter's Abuse and Hate Speech Policies. **Policy Studies Organization**, v. 13, p. 502-521, 2021.

LAVI, M. Do Platforms Kill?. **Harvard Journal of Law & Public Policy**, v. 43, p. 477-573, 2020. Disponível em: <https://heinonline.org/HOL/P?h=hein.journals/hjlpp43&i=492>. Acesso em: 18 maio 2022.

LUCCAS, V. N.; GOMES, F. V.; SALVADOR, J. P. F. **Guia de análise de discurso de ódio**. Rio de Janeiro: Fundação Getulio 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.

OBAR, J.; OELDORF-HIRSCH, A. The Biggest Lie on the Internet: Ignoring the Privacy Policies and Terms of Service Policies of Social Networking Services. **Information, Communication & Society**, v. 23, n. 1, p. 128-147, 2020.

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

ORLANDI, E. **Análise de discurso:** princípios e procedimentos. 5. ed. Campinas: Pontes, 2005.

ROBERTS, S. **Behind the Screen:** Content Moderation in the Shadows of Social Media. New Haven: Yale University Press, 2019.

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. Rio de Janeiro: FGV DAPP, 2021.

SÁ-SILVA, J.; ALMEIDA, C.; GUINDANI, J. Pesquisa documental: pistas teóricas e metodológicas. **Revista Brasileira de História & Ciências Sociais**, v. 1, n. 1, p. 1-15, jul. 2009.

SALEEM, H. M. *et al*. **A Web of Hate:** Tackling Hateful Speech in Online Social Spaces. arXiv, preprint arXiv:1709.10159, 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. *et al*. Analyzing the Targets of Hate in Online Social Media. In: PROCEEDINGS OF THE TENTH International AAAI Conference on Web and Social Media, 2016. Disponível em: <https://arxiv.org/abs/1603.07709v1>. Acesso em: 8 out. 2023.

_____; BOTELHO-FRANCISCO, R.; OLIVEIRA, A.; PONTES, V. 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. Disponível em: <https://periodicos.unb.br/index.php/RICI/article/view/22025/21351>. Acesso em: 2 nov 2023.

_____; BOTELHO-FRANCISCO, R. E. Gestão de conteúdo de ódio no Facebook: um estudo sobre *haters*, *trolls* e *naysayers*. **P2P e Inovação**, [S. l.], v. 6, n. 2, p. 38-56, 2020. Disponível em: <https://revista.ibict.br/p2p/article/view/5114>. Acesso em: 18 maio 2022.

VIVERITO, C. V.; BERTUCCI, P. **Inclusão de pessoas transgêneras e não binárias no local de trabalho brasileiro**. Califórnia: Out & Equal Workplace Advocates, 2020.

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.

Informações para textos em coautoria

Concepção e desenho do estudo
Luiza Santos e Renata Tomaz

Aquisição, análise ou interpretação dos dados
Luiza Santos, Renata Tomaz, Dalby Dienstbach, Eurico Matos e Danielle Sanches

Redação do manuscrito
Luiza Santos, Renata Tomaz, Dalby Dienstbach, Eurico Matos e Danielle Sanches

Revisão crítica do conteúdo intelectual
Luiza Santos e Renata Tomaz

Informações sobre o artigo

Resultado de projeto de pesquisa, de dissertação, tese
Este artigo é um desdobramento do projeto Digitalização e Democracia no Brasil, com parte de seus resultados desenvolvidos dentro do escopo do projeto. Entretanto, esta discussão apresenta uma ampliação substancial realizada de forma independente do projeto em questão.

Fontes de financiamento
Resultados parciais deste artigo foram obtidos a partir do financiamento do Ministério das Relações Exteriores da Alemanha.

Considerações éticas
Não se aplica.

Declaração de conflito de interesses
Não se aplica.

Apresentação anterior
Uma versão preliminar deste artigo foi apresentada no 44º Congresso Brasileiro de Ciências da Comunicação, em 2021, no formato *on-line*.

Agradecimentos/Contribuições adicionais:
Agradecemos a Luís Antônio de Medeiros e Gomes e a Daniel Almada Cunha pelo design das figuras presentes neste artigo.