Das políticas às práticas: análise das diretrizes de comunidade do Facebook, Instagram, YouTube e Twitter para a moderação de discurso de ódio[1]

Luiza SANTOS[2]
Renata TOMAZ[3]
Dalby HUBERT[4]
Danielle SANCHES[5]
Eurico MATOS NETO[6]
Diretoria de Análise de Políticas Públicas da Fundação Getulio Vargas (FGV DAPP), Rio de Janeiro, RJ

RESUMO

Este trabalho analisa, de forma comparativa, as diretrizes de comunidade e políticas de combate ao discurso de ódio de quatro plataformas digitais: Facebook, Instagram, Twitter e YouTube. Trata-se de pesquisa bibliográfica e documental, de caráter exploratório, cujo objetivo é oferecer possíveis tópicos de discussão capazes de aprofundar, em diferentes perspectivas, o debate sobre discurso de ódio *on-line*. Os resultados parciais do estudo apontam para a emergência de ao menos três pontos a serem problematizados a partir das dinâmicas de moderação de conteúdo em relação à discriminação nos ambientes digitais. São eles: os valores das plataformas; desafios contextuais e linguísticos; e as interlocuções e práticas de moderação.

PALAVRAS-CHAVE: Discurso de Ódio; Plataformas Digitais; Moderação de Conteúdo; Liberdade de Expressão.

1. 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 de formas de discriminação a minorias em ambientes digitais. Até pouco tempo, essas práticas eram consideradas uma atividade de nicho; entretanto, nos últimos anos, sua proeminência em espaços *mainstream* da

[1] Trabalho apresentado no GP Comunicação e Cultura Digital, XXI Encontro dos Grupos de Pesquisas em Comunicação, evento componente do 44º Congresso Brasileiro de Ciências da Comunicação.
[2] Doutora em Comunicação e Informação (UFRGS). Pesquisadora da FGV DAPP. E-mail: luizacdsantos@gmail.com
[3] Doutora em Comunicação e Cultura (UFRJ), bolsista PNPD/Capes no PPGMC/UFF e colaboradora do projeto Digitalização e democracia digital (FGV DAPP). E-mail: renatactomaz@gmail.com.
[4] Doutor em Estudos de Linguagem (UFF). Pesquisador da FGV DAPP. E-mail: dalby.hubert@fgv.br.
[5] Doutora em História das Ciências (EHESS/COC FioCruz). Pesquisadora da FGV DAPP. E-mail: danielle.sanches@fgv.br
[6] Doutor em Comunicação e Cultura Contemporâneas (UFBA). Pesquisador da FGV DAPP. E-mail: eurico.neto@fgv.br

internet[7] 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 as autorregulações das plataformas (RUEDIGER; GRASSI, 2021).

Neste artigo, analisaremos, de forma comparativa, as diretrizes de comunidade e políticas de combate ao discurso de ódio de quatro plataformas: Facebook, Instagram, Twitter e YouTube[8]. Trata-se de pesquisa bibliográfica e documental, de caráter exploratório, que apresenta resultados parciais e aprofundados de estudo mais amplo, realizado no âmbito do projeto "Digitalização e Democracia no Brasil", uma parceria entre a Diretoria de Análise de Políticas Públicas da Fundação Getulio Vargas (FGV DAPP) e a Embaixada da Alemanha em Brasília[9]. O trabalho se baseia nos dados disponibilizados pelas plataformas em seus termos de uso e diretrizes de comunidade, em fevereiro de 2021, e se dividirá em quatro partes.

Primeiro, apontaremos algumas especificidades da prática e da circulação do discurso de ódio nos ambientes digitais. Na sequência, indicaremos como as plataformas propõem a moderação desse conteúdo em seus espaços. Depois, apresentaremos e discutiremos comparativamente as diretrizes de comunidade sobre discurso de ódio das quatro plataformas. Por fim, levantaremos três pontos que, a partir da análise, se mostram problemáticos na criação de diretrizes e moderação de conteúdo desse tipo. Desse modo, objetivamos oferecer possíveis tópicos de discussão capazes de aprofundar, em diferentes perspectivas, o debate sobre discurso de ódio *on-line*.

2. 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 das suas características (COHEN-ALMAGOR, 2011; FARIS et al. 2016; PAREKH,

[7] Chamamos de *mainstream* as plataformas mais utilizadas pelos brasileiros, não caracterizadas como nichos de interesse ou como ambientes mais utilizados pelos grupos de ódio como forma de organização (como os *chans*).

[8] Segundo dados de pesquisa realizada pela *We are socia*l 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. Ver mais em: https://wearesocial.com/digital-2020. Acesso em: 12 ago. 2021.

[9] Mais informações em: https://democraciadigital.dapp.fgv.br/sobre/ .Acesso em 5 ago. 2021.

2012). Ele tem um caráter discriminatório, motivado por preconceitos baseados na ideia de uma suposta superioridade de quem agride sobre quem sofre a agressão. Segundo a definição do *Guia para análise de discurso de ódio* (LUCCAS et al., 2020, p. 4), que demarca nosso entendimento no escopo deste trabalho, trata-se de

> 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, consequentemente, legitimarem a prática de discriminação ou violência.

Essas manifestações, quando em ambientes *on-line*, possuem especificidades. Revisões sistemáticas da literatura apontam que algumas características das mídias sociais são componentes importantes de sua proliferação. A anonimidade dos usuários, ainda que parcial, 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. A invisibilidade decorrente da 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 sua 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 afinidades, 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 de ampliação de colaboradores, cultivando comunidades e valores que possibilitam o reforço de identidades no contexto digital (WEAVER, 2013) e o potencial aumento da projeção dos grupos de ódio a partir do engajamento dos seus membros (BOWMAN-GRIEVE, 2009).

Outra característica das plataformas digitais que pode contribuir para a circulação do discurso 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*][10]. 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 (COLLEONI et al., 2014). Os modos de funcionamento dos algoritmos que constroem esses regimes personalizados de visibilidade de conteúdos não são plenamente esclarecidos pelas plataformas digitais, constituindo-se como sistemas opacos para a maior parte dos usuários (JURNO; D'ANDRÉA, 2017).

Diferentemente do que ocorre nos meios de comunicação de massa, em que a checagem de conteúdo acontece antes da veiculação, nas plataformas digitais, essa moderação se dá após a publicização e, principalmente, mediante denúncia de usuários. A dificuldade da checagem prévia é outra especificidade do contexto digital. Além disso, a noção de liberdade de expressão permanece como valor principal das plataformas digitais, como veremos na seção de análise. Assim, o discurso de ódio em plataformas *on-line mainstream* é combatido apenas depois de sua circulação. Já, no contexto *off-line*, ele tende a ficar fora do circuito da grande mídia, circulando apenas de forma marginal.

Outra diferença da propagação *on-line* é que as plataformas não são as autoras das publicações. Nesse sentido, as responsabilidades não são análogas às de uma empresa de mídia tradicional. Todavia, mesmo as plataformas não sendo as produtoras desse conteúdo, elas criam o ambiente de sua difusão e obtêm lucros com a interação e a atenção dos usuários em função deles. Esse *status*, que não é nem de produtor de conteúdo no sentido tradicional das mídias e nem de espaço totalmente público de discussão, também torna mais complexo o combate ao discurso de ódio.

3. Moderação de conteúdo em plataformas digitais

Além das particularidades decorrentes do próprio meio de comunicação, as interações em cada plataforma também são moldadas a partir de seus termos de uso, documentos que dispõem sobre suas formas de funcionamento e suas regras de utilização às quais os utilizadores estão sujeitos – mesmo que, na maior parte das vezes,

[10] Uma perspectiva contrária a esta é composta por autores que defendem as ideias da exposição seletiva e da exposição inadvertida, ou seja, quando os indivíduos nas mídias sociais são expostos a conteúdos contrários aos seus pontos de vista.

eles não leiam esse contrato de serviço antes de clicarem em "Eu aceito". Em geral, esses termos apontam que os usuários devem seguir as diretrizes de comunidade e que, caso não as sigam, algumas penalidades podem ser aplicadas.

Nesses documentos, encontramos o que cada rede social entende como discurso de ódio, a sua tolerância em relação a ele 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, as plataformas se baseiem em pesquisas científicas para proporem 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 (GILLESPIE, 2020), 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 contextualizada por moderadores. As práticas de moderação das plataformas e dos moderadores humanos sobre discurso de ódio, entretanto, colocam algumas questões.

A primeira delas diz respeito à 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 (como fazem os moderadores de conteúdo), baseando-se na definição conceitual (quer 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. O Facebook pontua que a intenção do usuário ao postar também é considerada na hora da moderação do discurso de ódio, ponto que aparece também, de forma mais sútil nas diretrizes do Twitter. Isso adiciona uma camada a mais de complexidade: afinal, como determinar a intenção de outro sujeito?

É essa dinâmica situacional que impõe dificuldades práticas para a detecção e o combate ao discurso de ódio nas plataformas, muito mais do que a conceituação adotada. Somam-se a essa questão as práticas materiais pelas quais as plataformas digitais de porte global operacionalizam a moderação de conteúdo em seus espaços. As plataformas digitais não fornecem um acesso transparente às suas tecnologias, arquiteturas e práticas, o que se evidencia na moderação de conteúdo (ROBERTS, 2019).

Atualmente, a moderação de conteúdo realizada por grandes empresas de tecnologia opera no formato de moderação comercial de conteúdo (MCC), que 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" e que é realizada de forma manual (ROBERTS, 2019, p. 33). O principal mecanismo para que a MCC seja acionada é o "*flagging*" (ou denúncia) dos próprios usuários sobre um conteúdo. A MCC pode ser feita por equipes na própria empresa, através da contratação de empresas terceirizadas – muitas delas em outros países, como é o caso das Filipinas, um polo de MCC (THE CLEANERS, 2018) –, ou por plataformas de micro trabalho, como a Amazon Mechanical Turk. Assim, a precariedade das condições de trabalho dos moderadores nesse modelo também 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 cria falta de contextos social e cultural entre as postagens avaliadas e os próprios 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 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 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* e interferência de governos (SILVA et al., 2019).

As pressões nesse sentido também levaram a um investimento cada vez maior em soluções automatizadas. Esse tipo de moderação traz a promessa de sanar o problema de escala, ou seja, uma alternativa teoricamente capaz de lidar com a grande quantidade de conteúdo a ser moderado com o menor esforço possível, a partir da proceduralização de um processo "de forma a se replicar em diferentes contextos e parecer a mesma coisa" (GILLESPIE, 2020, p. 2). Tal estratégia pode invisibilizar 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) os erros estatísticos cometidos por esses sistemas tendem a recair sobre grupos minoritários, também sub-representados em bancos de dados para treinamento (GILLESPIE, 2020).

4. Diretrizes sobre discurso de ódio: análise do Facebook, Twitter, Instagram e YouTube

Na sequência, apresentamos um resumo das diretrizes de comunidade do Facebook, Instagram, YouTube e Twitter[11] sobre discurso de ódio *on-line*. As diretrizes das quatro plataformas apresentam restrição ou repúdio contra o discurso de ódio. Em

[11] Considerando a prática de atualização das referidas plataformas, é importante dizer que os dados apresentados são relativos às diretrizes e termos analisados em fevereiro de 2021.

alguns casos, também tratam de discursos violentos, extremistas ou perigosos. As plataformas variam, entretanto, no grau de detalhamento, na conceituação dos termos e na exemplificação das diretrizes que fornecem. Nesse sentido, o Instagram se apresenta como a plataforma que menos detalha esses aspectos, indicando diretrizes mais sucintas e genéricas. Embora informe aos usuários que eles também estão sujeitos às políticas do Facebook, empresa proprietária do Instagram, sua falta de especificidade alerta para a dificuldade ainda maior na moderação de conteúdo de ódio.

No entanto, há um esforço das plataformas em deixar claro, nas suas diretrizes, o que elas compreendem como discurso de ódio. O Facebook o define como ataques realizados a pessoas com características protegidas. O Instagram não deixa explícito o que compreende como discurso de ódio, porém, segue as normativas descritas nos "Produtos Facebook", além de conter, nos seus termos de uso, compromissos que o usuário deve assumir ao utilizar a plataforma. O YouTube tem uma política contra a disseminação de discurso de ódio para a produção de conteúdo e para a publicidade; foi a única plataforma em que constatamos essa especificidade. Para a produção de conteúdo, as diretrizes indicam restrições de vídeos que incentivam violência contra minorias, enquanto que, nas regras para a publicidade, é interditada linguagem de vídeo que incita o "ódio, promove discriminação, menospreza ou humilha um indivíduo ou grupo de pessoas", mesmo que tenha viés cômico. No que se refere ao Twitter, a plataforma dispõe de informações sobre comportamentos proibidos na seção "Regras e Políticas", indicando diversas categorias, entre elas a "política contra propagação de ódio", em que dispõe que "não é permitido promover violência, atacar diretamente ou ameaçar outras pessoas". Dentre as quatro redes sociais, essa é a que oferece um documento mais detalhado em torno do que é e o que não é permitido e dos próprios entendimentos sobre o tema.

A menção às características ou categorias protegidas, que podem ser elencadas ou não nos documentos, é um dos principais pontos em comum das plataformas analisadas. No caso do Instagram, não são delimitadas. 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. O Twitter é 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 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).

Das quatro plataformas analisadas, apenas o YouTube informa, nas 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. As demais informam apenas que analisam os conteúdos sinalizados por usuários de forma contextual, sem indicar outras iniciativas de controle no texto das diretrizes de comunidade. Sabe-se, porém, que tanto Facebook quanto Twitter utilizam mecanismos automatizados na detecção de conteúdos que violem suas diretrizes (SILVA et al., 2019; GILLESPIE, 2020).

Nenhuma das quatro plataformas expõe, em detalhes, os procedimentos em torno das sanções aplicadas para os usuários que publicam conteúdos de discurso de ódio e afins. O Facebook afirma que analisa a gravidade e o histórico do usuário na plataforma para tomar uma decisão, que pode variar desde notificação e restrições de uso 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 Instagram, a que menos detalha sanções e procedimentos, 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. 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 YouTube delimita que a primeira violação de diretrizes da comunidade gera apenas uma sinalização, sem penalidade. As violações seguintes geram exclusão do conteúdo (vídeo ou comentário), notificação e limitação das 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 tende a oferecer mais elementos a esse respeito. Essa transparência, todavia, é parcial, uma vez que somente a plataforma possui acesso irrestrito aos dados que informam tanto suas decisões quanto seus relatórios.

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

Quadro Comparativo - Diretrizes de Comunidade				
	Facebook	**Twitter**	**Instagram**	**YouTube**
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,	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	Idade, classe social, deficiência, etnia, identidade e expressão de gênero, nacionalidade, raça, situação de imigração, religião, sexo/gênero, orientação sexual, vítimas de um

		homossexuais, intersexuais, indivíduos assexuados, comunidades marginalizadas e historicamente sub-representadas.	ou doenças".	evento violento em grande escala e os familiares dessas pessoas, veteranos de guerra.
Critérios 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, notificação e limitação do uso temporária. Após três notificações em 90 dias, o canal é encerrado e o conteúdo deletado.

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

5. Questões emergentes no debate sobre discurso de ódio e moderação de conteúdo

Com base na análise comparativa das diretrizes de comunidade das plataformas, identificamos três tópicos de discussão sobre discurso de ódio e políticas de moderação de conteúdo. Desse modo, listamos abaixo possíveis caminhos para pesquisas futuras explorarem e aprofundarem o tema no âmbito da cultura digital.

5.1. Posicionamentos e valores: liberdade de expressão e interesse público

Ao menos dois valores norteiam as políticas de moderação nos documentos analisados: a liberdade de expressão e o interesse público. As quatro plataformas defendem a necessidade de se garantir um ambiente que congregue diferentes vozes, realize uma conversação pública e permita a livre expressão de si. Reconhecem, no entanto, que o exercício desse direito fundamental, preconizado pela Declaração Universal dos Direitos Humanos e garantido em legislações nacionais, como a Constituição Brasileira, não poderá ferir os demais – por exemplo, o da dignidade

humana. Nesse sentido, a liberdade de expressão é concebida pelas plataformas como um direito essencial relativo e não absoluto. A principal preocupação é a de que os ataques e ofensas, se não limitados, calem os grupos protegidos e descaracterizem a proposta dialógica das mídias sociais.

Um segundo ponto que ampara as políticas de moderação, definindo a remoção ou não de conteúdo, é a compreensão de interesse público. As quatro plataformas admitem que, em circunstâncias educativas, 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, com abordagens educativas, científicas ou artísticas, e à discussão pública sobre o tema, desde que a intenção do usuário, nesse sentido, esteja clara. Tanto o interesse público quanto a liberdade de expressão são, desse modo, valores que sustentam os critérios de autorregulação das plataformas. Compreender como são acionados e que dilemas suscitam pode oferecer discussões sobre a relação entre discursos de ódio, plataformização e ética.

5.2. Desafios contextuais e linguísticos

Os documentos das quatro plataformas informam que as análises contextual e linguística se impõem ao processo que torna a política de moderação em ação de remoção (ou de permanência). Facebook, Instagram, Twitter e YouTube mantêm orientações globais com critérios que serão 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.

Com base nisso, merecem atenção alguns aspectos pertinentes aos níveis tanto linguístico quanto discursivo do uso da linguagem que, inevitavelmente, intervêm em eventuais esforços de se classificar uma determinada postagem como contendo discurso de ódio. Um primeiro ponto que se observa, nesse sentido, diz respeito a componentes estruturais da linguagem, tais como o caráter polissêmico inerente, em princípio, ao léxico de qualquer língua natural (CANÇADO, 2012). De fato, os diversos significados que virtualmente todas as palavras de uma língua podem abrigar – quando não estão vinculadas a uma área de especialidade – podem orientar a codificação (ou, ainda, a

decodificação) dos enunciados que integram para uma ou outra direção. As diferentes acepções que completam entradas como "burro", "asno" e "jumento" no dicionário, por exemplo, poderiam embaralhar a decisão sobre se uma dada postagem deve ser classificada ou não como atualizando discurso de ódio.

Um segundo fator que atravessa a interpretação (e classificação) de uma postagem como conteúdo de ódio se refere, do ponto de vista do discurso, ao conjunto de condições de produção e recepção dos seus enunciados (ORLANDI, 2005). Essas condições incluem desde aspectos situacionais da enunciação até componentes das conjunturas social, cultural e histórica dessa enunciação. Fazem parte desses últimos, por exemplo, as formações político-ideológicas presentes no momento da enunciação; os comportamentos hegemônicos e os marginalizados na ocasião; as relações de poder estabelecidas entre indivíduos ou grupos de indivíduos diferentes etc. São essas condições que permitem, por exemplo, a realização efetiva de enunciados irônicos, sem polidez ou com ataques à face do interlocutor – e que atualizariam, portanto, casos de discurso de ódio.

5.3. Das políticas às práticas de moderação

As estratégias de moderação têm uma relação direta com as denúncias dos usuários, que aparecem como a principal ferramenta para o combate ao discurso de ódio *on-line*. O que é denunciado gera não apenas estatísticas, mas entendimento sobre como determinados indivíduos ou grupos se sentem em relação a certos conteúdos. Desse modo, identificamos um investimento na participação de especialistas, ativistas, acadêmicos e integrantes de ONGs como interlocutores das plataformas para criarem políticas e atualizarem-nas. Não encontramos, no entanto, nos trechos que tratam desses *stakeholders*, referências específicas a educadores, pais, responsáveis, representantes governamentais ou juristas, por exemplo. Não há, ainda, qualquer tipo de indicação de que exista um trabalho cooperativo entre plataformas para um intercâmbio de informações, experiências ou estratégias de enfrentamento da proliferação do 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. O YouTube, em um de seus relatórios periódicos, informa as solicitações de remoção por governos nacionais,

indicando se foram atendidas ou não e por quê. A presença (ou ausência) desses atores também compõe as estratégias de combate à discriminação *on-line*.

De acordo com a literatura sobre o tema (ROBERTS, 2019), há basicamente dois tipos de moderação: a humana e a automatizada. A primeira é feita de forma comercial, com frequente terceirização das atividades para mão-de-obra em outros países, com treinamento deficitário e condições de trabalho precárias, além de uma rotina que implica lidar com conteúdo pesado de forma intensa, muitas vezes em outro idioma e contexto cultural. Sobre a segunda, mesmo com o avanço tecnológico, a implementação de modelos de *machine learning* para moderação de conteúdo significa que ela 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 que está sendo sempre renegociada a partir das práticas e acordos sociais). Desse modo, os processos moderadores que tornam as diretrizes sobre discurso de ódio em ações concretas suscitam a necessidade de estudos que explicitem os elementos que os constituem, os atores que deles participam e as relações de poder que os atravessam, entre outros aspectos.

6. Considerações finais

Neste trabalho, apresentamos e analisamos as diretrizes de comunidade e termos de uso sobre discurso de ódio no Facebook, Twitter, Instagram e YouTube. A partir da análise comparativa e do suporte teórico sobre discurso de ódio em ambientes *on-line* e moderação de conteúdo em plataformas digitais, detectamos três pontos recorrentes a serem aprofundados em trabalhos futuros: o posicionamento e valores das plataformas, os desafios contextuais e linguísticos e a transformação das políticas em práticas de moderação.

Apesar dos esforços legislativos, nacionais e 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. O equilíbrio dessa relação é uma das principais questões apresentadas pelas plataformas ao abordarem o assunto: como garantir a segurança de categorias protegidas, interferindo o mínimo possível na liberdade de expressão dos usuários? O interesse público na veiculação das mensagens também compõe, conjuntamente com a

liberdade de expressão, os valores das plataformas que estão em jogo em suas diretrizes sobre discurso de ódio e que, portanto, balizam suas ações.

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. Ou seja, a cultura, a linguagem e o contexto são elementos importantes na detecção correta de discursos desse tipo, entretanto, as plataformas, apesar de reconhecerem a questão como central, apresentam poucos recursos de moderação para lidar com as questões culturais.

As práticas de moderação de conteúdo nas plataformas digitais investigadas funcionam, em grande medida, a partir de denúncias dos usuários: ou seja, a circulação ou não de discurso de ódio depende de um trabalho de vigilância do conteúdo realizado pelo usuário – portanto, gratuito. Além disso, não apresentam meios de colaboração mútua para enfrentarem a discriminação *on-line*. Ainda, na passagem das políticas para as práticas de combate efetivadas, a escala comercial da moderação impõe ao desafio contextual e linguístico a terceirização, em grau internacional, dos processos moderadores. 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

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.

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

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

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

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.

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

GILLESPIE, T. Content moderation, AI, and the question of scale. *Big Data & Society*, July-December, p. 1-5 , 2020.

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.

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 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.

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

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.

ROBERTS, Sarah. *Behind the Screen:* Content Moderation in the Shadows of Social Media. Yale University Press: New Haven, 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. Policy paper. Rio de Janeiro: FGV DAPP, 2021

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.

SILVA, L.; MONDAL, M.; CORREA, D.; BENEVENUTO, F.; WEBER, I. 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

THE CLEANERS. Direção de Hans Block e Moritz Riesewieck. Independent Lens, 2018 (95 min.).

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.