COMPOLÍTICA 9ª EDIÇÃO
Democracia e opinião pública em tempos de fake news
24 a 28.05
2021

POLÍTICAS DE REGULAÇÃO DE CONTEÚDOS EM MÍDIAS SOCIAIS: Remoção de discursos antidemocráticos e o caso de Donald Trump[1]

CONTENT POLICY REGULATIONS IN SOCIAL MEDIA: Antidemocratic speech removal and the case of Donald Trump

Daniel Jorge Teixeira Cesar[2]

***Resumo**: Este artigo analisa as políticas e práticas de moderação de conteúdos em mídias sociais sobre discursos de ódio e notícias falsas, e compara a atuação das plataformas, especificamente o Facebook e o Twitter, para compreender como ambas lidam com conteúdos violentos e desinformação e como reagiram às publicações de Donald Trump no contexto do ataque ao Capitólio em Janeiro de 2021. O estudo foi realizado a partir da metodologia de rastreamento de processos para organizar a análise documental dos regulamentos e comunicados oficiais das duas redes sociais. O resultado da investigação aponta que há diferenças no tratamento de conteúdos e nas políticas das plataformas, e que apesar de definir os conteúdos proibidos, abrem brechas que permitem que discursos de ódio e notícias falsas possam permanecer no ar. O artigo conclui com um argumento crítico sobre a falta de transparência e a baixa efetividade das medidas tomadas pelas plataformas.*

***Palavras-Chave:** Moderação de conteúdos. Mídias sociais. Governança de plataformas.*

***Abstract**: This paper analyses content moderation policies and practices in social media concerning hate speech and fake news, and compares measures applied by platforms, specifically Facebook and Twitter, in order to understand how both deal with violent content and disinformation and how they reacted to Donald trump's publications in the context of the attack on Capitol in January 2021. The study was realized using process tracing methodology to organize a document analysis of regulations and official communications from publicized by those social networks. The result of the investigation points that there are differences on how content is treated and between platform policies, and that despite defining forbidden content, there are loopholes that allow hate speech and fake news to remain online. The*

1 Trabalho apresentado ao Grupo de Trabalho Políticas e Governança da Comunicação do VIII Congresso da Associação Brasileira de Pesquisadores em Comunicação e Política (VIII COMPOLÍTICA), realizado online, de 24 a 28 de maio de 2021.

2 Doutorando do Programa de Pós-Graduação em Comunicação da Universidade de Brasília. Contato: danieljtc@gmail.com

paper concludes with a critical argument over lack of transparency and the low effectiveness of actions taken by platforms.

Keywords: *Content moderation. Social media. Platform governance.*

1. Introdução

Sites de mídias sociais se consolidaram na última década como fontes de informação para milhões de pessoas, compondo espaços privados na Internet onde os usuários se intercomunicam para divulgar e amplificar suas mensagens. Este é o serviço oferecido por empresas como Facebook e Twitter: fazer circular em suas redes o conteúdo criado pelos usuários em troca da exploração comercial de seus dados. As plataformas sustentadas por estas empresas possuem normas quanto ao que é permitido em suas redes para regular os conteúdos produzidos pelos usuários de modo a evitar a divulgação de desinformação e mensagens de ódio.

Considerando as diferenças de público e formato das redes, bem como as diferenças sobre a formação e aplicação das políticas de remoção de conteúdos, é possível afirmar que apesar do estabelecimento de regras e dos recursos das plataformas para moderar, é cada vez mais comum encontrar notícias falsas e ameaças violentas nas plataformas. A proliferação de mensagens com cunho antidemocrático como estas pode influenciar os debates em uma esfera pública automatizada (Pasquale, 2017) e a formação da opinião pública sobre os processos democráticos. Em situações em que as normas e a moderação falham, conteúdos que estimulam violência, preconceito ou desinformação podem ser amplificados e causar danos à democracia. Mais recentemente discursos de ódio e conteúdos de violência por parte de governantes se tornaram parte da paisagem nas mídias sociais e representam um risco para a manutenção de Estados democráticos.

Partindo da hipótese de que as comunicações em mídias sociais podem afetar e influenciar o debate político, a proposta do artigo é pesquisar as políticas e práticas de regulação de conteúdos nas plataformas de mídias sociais mencionadas,

e investigar quais os limites impostos pelas empresas para a moderação comercial de conteúdos, considerando que há falhas que permitem a divulgação em grande alcance de mensagens antidemocráticas, discursos de ódio, desinformação e notícias falsas.

O foco deste estudo é analisar as políticas e práticas de moderação de conteúdos e comparar a atuação das mídias sociais na remoção das publicações e perfis de Donald Trump de suas redes após a invasão ao Capitólio em Janeiro de 2021. Os fatos ocorridos no período eleitoral americano e nas semanas posteriores nos servem aqui de ponto de partida para ilustrar o problema da moderação das redes sociais, espaços privados de formação e circulação da opinião pública, em remover conteúdos proibidos de suas plataformas.

A ferramenta metodológica escolhida para coletar, organizar e analisar os dados é o rastreamento de processos (Collier, 2011; Falleti, 2016), que busca estabelecer mecanismos causais que expliquem um fenômeno a partir do exame de evidências em forma de dados qualitativos como documentos e relatórios internos das plataformas, publicações oficiais e depoimentos de ex-funcionários, por exemplo.

As plataformas possuem políticas próprias de autorregulação dos conteúdos postados pelos usuários e determinam sanções como a remoção da postagem infratora e a suspensão do perfil na rede. As plataformas determinam os tipos de conteúdos que devem ser apagados, porém defendem que seus regulamentos seguem princípios de liberdade de expressão e assim abrem brechas para divulgação de mensagens ofensivas. É preciso considerar que há conteúdos que extrapolam o direito de livre expressão e casos em que é necessária a exclusão de perfis, como no episódio envolvendo Trump. Dessa maneira, é preciso investigar se as atitudes tomadas pelas empresas para remover conteúdos antidemocráticos condizem com os regulamentos e práticas de autorregulação das plataformas, visto que não há regulação governamental consistente para decidir sobre as comunicações em mídias sociais.

De modo específico, os objetivos destes artigo são: identificar as medidas práticas e formas como as plataformas lidam com conteúdos proibidos; identificar os

possíveis discursos antidemocráticos e violações às normas cometidas por Trump; comparar as políticas e práticas com a expulsão de Trump das plataformas; e estabelecer uma análise crítica sobre o funcionamento da moderação de conteúdos nas plataformas.

2. Moderação de conteúdos em mídias sociais

A Internet na última década se organizou sobre plataformas comerciais que oferecem serviços variados desde troca de mensagens até transporte pessoal. Os serviços são oferecidos em troca da exploração comercial dos dados oferecidos pelos usuários. Gorwa (2019) estabelece que corporações como Facebook e Twitter são responsáveis pela manutenção de suas plataformas enquanto espaços de interação e interconexão entre usuários e estabelecem regras de governança para regular as políticas de cada empresa quanto a exploração dos dados dos usuários e utilização dos serviços que oferecem, em consonância com outros enquadramentos regulatórios como as legislações de Estados nacionais. Um exemplo é a Seção 230 do Communications Decency Act de 1996, que retira a responsabilidade da plataforma sobre a informação publicada. Isso fornece bases legais para que as plataformas operem sob guisa de liberdade de expressão enquanto são livres para regular ou não os conteúdos que circulam por suas redes.

As empresas criam espaços privados e tomam decisões sobre os conteúdos produzidos por usuários que circulam em suas redes, operando como intermediários e curadores de conteúdo. Isso leva à necessidade de regras para manter os espaços seguros e promover o modelo de negócios em que a plataforma se baseia. Assim, as plataformas são dependentes do conteúdo gerado pelos usuários para manutenção do interesse dos mesmos sobre seus espaços e assim explorar financeiramente os dados. Nos interessa aqui a governança de plataformas quanto à moderação de conteúdos e as formas de autorregulação exercidas.

Denardis (2015) descreve um conjunto de problemas envolvendo governança de plataformas, incluindo questões de privacidade e exploração de dados, disputas entre Estados e empresas quanto a inovações tecnológicas produzidas pelas empresas e formação e acesso a uma esfera pública digital. Mantendo o foco nesta última questão, a autora considera que enquanto sempre houve espaços privados de deliberação e debate públicos, com as plataformas de mídias sociais há um maior potencial de alcance para mensagens pois não há barreiras físicas ou geográficas. Dessa forma, é possível argumentar que a divulgação de informações falsas e discursos de ódio amplificados pelas mídias sociais pode afetar os debates na esfera pública e trazer danos à democracia

No caso das plataformas de mídias sociais há restrições quanto ao conteúdo de mensagens que são permitidas circular nas redes. Cada plataforma criou seu próprio conjunto de regras sobre o que é permitido em suas redes, formando jardins murados onde as empresas determinam o que pode ser acessado pelos usuários, criando espaços privados de formação da opinião pública e influência sobre a esfera pública em sua modalidade automatizada (Pasquale, 2017). Decisões sobre moderação de conteúdos podem impactar nos discursos e formação da opinião pública pelo acesso e qualidade da informação disponibilizada e divulgada nas mídias sociais. Plataformas podem moderar conteúdos de maneiras diferentes, como pelo uso de rótulos para indicar conteúdos fraudulentos, pela remoção de postagens e suspensão de perfis infratores de maneira temporária ou permanente de acordo com a gravidade da infração.

Para regular seus espaços, cada plataforma utiliza sistemas algorítmicos para organizar e selecionar o conteúdo que é visto pelos usuários. Autores como Pariser (2012) e Silveira (2018) já descreveram os métodos de filtragem de informação e o potencial de modulação de comportamentos dos usuários em relação ao tipo e qualidade de informação acessada. Para além disso, as plataformas possuem regras estabelecidas em documentos como os Padrões da Comunidade do Facebook e as Regras do Twitter para determinar a proibição de divulgação de mensagens com conteúdos de violência, ameaças, desinformação, discursos de ódio, pornografia e abuso infantil entre outros tipos de discurso ofensivo.

Para exercer governança sobre conteúdos e limitar o potencial de mensagens com potencial danoso aos direitos humanos, sociais e civis, além de filtrar o que o usuário pode acessar e determinar o que é proibido em suas redes, as plataformas utilizam ferramentas como inteligências artificiais e moderadores humanos para filtrar e remover conteúdos que não devem estar nas redes.

Autores como Tarleton Gillespie (2014) e Sarah Roberts (2019) já descreveram etapas do processo de moderação de conteúdos pelo aspecto do trabalho humano realizado por moderadores para excluir conteúdos proibidos pelos regulamentos das plataformas. Entre as medidas de moderação podemos citar o uso de denúncias pelos usuários, comumente chamada de *flagging* (Crawford & Gillespie, 2016) e o trabalho de moderadores humanos (Roberts, 2019) em detectar e remover conteúdos proibidos. Ocorre que há falhas na remoção de conteúdos, e postagens que, pelas regras, deveriam ser apagadas, permanecem nas redes. É possível questionar que isso se deve ao volume e quantidade de postagens, à dificuldade em monitorar as mídias sociais pelo trabalho humano ou de máquinas e à falta de transparência sobre a criação e execução das normas.

Roberts define categorias para a moderação, dividindo entre voluntária – isto é, realizada pelos próprios usuários – e comercial – realizada por uma empresa terceirizada contratada para revisar os conteúdos nas redes de acordo com o documento regulador das políticas da plataforma. A autora descreve o trabalho dos moderadores de conteúdo

Além da opacidade dos documentos e regras que determinam de maneira vaga os conteúdos proibidos nas redes das plataformas, os métodos de remoção também não são de conhecimento do grande público, visto que não há comunicação sobre o uso de algoritmos de detecção eremoção ou do trabalho de moderadores humanos, mesmo em casos em que o conteúdo é denunciado pelos próprios usuários. Plataformas não são transparente quanto ao funcionamento de denúncias (flagging) e se sabe apenas que parte significativa da moderação é realizada por sistemas de inteligência artificial programados para reconhecer e deletar conteúdos antes que sejam postados. Há também trabalho de moderadores humanos que avaliam e retiram do ar conteúdos proibidos segundo as regras de cada plataforma. Em seu

trabalho, Roberts retrata os problemas relacionados à terceirização do trabalho de moderação, visto que não é feito pelas plataformas. Tanto o Twitter quanto o Facebook contratam empresas e deslocam as funções de moderação para contratos terceirizados. Segundo a autora, isso leva a problemas de precarização das condições de trabalho como danos psicológicos pela exposição a conteúdos de violência extrema e abuso infantil

As plataformas, em geral, procuram o equilíbrio entre proteger a comunidade e a liberdade de expressão, bem como entre a busca por preservar o interesse dos usuários na rede e manter o modelo de negócios de exploração de dados, mas a literatura aponta que há opacidade nos métodos e práticas de remoção de conteúdos. Desde 2012 há controvérsias em torno do Facebook quanto a essa questão, e apenas em 2018 a plataforma divulgou seus Padrões da Comunidade para esclarecer os tipos de conteúdos proibidos na rede. Além disso, como apontado por Roberts a moderação é aplicada por funcionários terceirizados que assinam acordos que os impedem de divulgar detalhes sobre a remoção de conteúdos. Sobre isso Zwart destaca que:

> As instruções internas do Facebook quanto à moderação de conteúdos e material de treinamento para moderadores são mantidas em sigilo, conhecidas apenas por vazamentos na mídia. Tais documentos confidenciais foram criticados por sua falta de consistência e aparente falta de aplicação, permitindo a permanência (ou não) de conteúdos no Facebook de acordo com a identidade do emissor do discurso – por exemplo, a afirmação 'pessoas brancas são racistas' seria deletada enquanto afirmação equivalente sobre outros grupos étnicos seria permitida. (Zwart, 2018 p. 285, tradução própria)

A moderação de conteúdos, de maneira geral, para as plataformas, possui uma série de problemas e dificuldades que abrangem a terceirização dos moderadores e o volume de postagens, visto que não é possível avaliar todos os milhões de conteúdos publicados mesmo com sistemas de inteligência artificial e moderadores humanos. Segundo as plataformas, a razão pela qual não definem com clareza o funcionamento dos mecanismos de remoção é que o conhecimento do funcionamento das regras pode levar usuários a explorá-las ao limite e postar

mensagens ofensivas dentro das normas e assim evitar que sejam removidas das redes.

3. Metodologia de pesquisa

O objeto deste artigo – as políticas e práticas de moderação de conteúdos em plataformas faz parte de estudo mais amplo de tese de doutorado em que documentos internos e relatórios publicados pelas empresas em suas páginas oficiais foram coletados e analisados segundo metodologia de rastreamento de processos. Essa metodologia consiste em interpretar dados qualitativos referentes a um processo na busca de mecanismos causais que expliquem seu resultado. Os dados analisados aqui são referentes ao caso específico das práticas e políticas envolvendo as postagens que levaram à exclusão dos perfis de Donald Trump nas redes.

O rastreamento de processos tem sua origem na área da Psicologia na década de 1970, nos estudos cognitivos sobre processos decisórios individuais, e foi adaptada para as Ciências Políticas na década de 1980 para estudar a formação de políticas públicas. O método propõe o exame de dados qualitativos que podem servir como evidência de um processo para descrever e analisar mecanismos causais que geram fenômenos complexos como a moderação de conteúdos.

Diferentes autores apresentam abordagens diversas sobre o método e indicam as possibilidades e variações de seu uso, de forma mais rigorosa quanto à definição de teorias e hipóteses iniciais ou mais abertas e flexíveis para formar hipóteses a partir da teoria e dos dados coletados. No livro *Case Studies and Theory Development in the Social Sciences* de 2005, Bennet e George (2005) dedicam um capítulo a explicar os fundamentos do rastreamento de processos, deixando claras suas possibilidades, seu uso e as limitações do método. Estabelecem como conceito que o rastreamento de processos procura explicar os mecanismos causais de fenômenos.

Posteriormente, autores como Collier (2011) e Falleti (2016) trataram de apresentar definições mais refinadas que ora se complementam, ora divergem em pontos essenciais. Todos, porém, concordam que se trata de um método qualitativo para testar ou desenvolver hipóteses a partir da análise de evidências a fim de identificar a relação entre mecanismos causais na produção de um resultado e assim explicar fenômenos específicos.

Collier sugere que o método pode ser utilizado para pesquisas comparativas entre um número pequeno de objetos, e define o rastreamento de processos como um instrumento analítico para inferir as causas de um fenômeno a partir do exame sistemático de evidências selecionadas e analisadas à luz de teorias e hipóteses propostas pelo pesquisador. O autor descreve testes de inferência causal, formulados anteriormente por Andrew Bennet, que indicam a suficiência e a necessidade de uma evidência para relacionar as variáveis observadas, bem como sua utilização para explicar causas. Esses testes fornecem rigor ao método pois determinam se as evidências coletadas podem suportar as inferências do pesquisador. São delimitadas em quatro tipos de testes:

1) O teste de palha ao vento ou *Straw in the wind* indica uma evidência que torna uma hipótese mais plausível, mas não é necessária ou suficiente para comprová-la. Dos quatro testes é o mais fraco, mas pode conter informações importantes para a construção de teorias. Sua força está na quantidade de evidências deste tipo coletadas, que podem indicar uma tendência.
2) O teste de aro ou *Hoop test* determina evidências que podem afirmar a relevância de uma hipótese, mas não são suficientes para comprová-la. Pode, por outro lado, descartar hipóteses que não podem ser confirmadas pela evidência. Não é, portanto, suficiente para comprovar uma afirmação, mas é necessária para estabelecer critérios para validar e aceitar uma explicação.
3) No teste de arma fumegante ou *Smoking gun se* confirma se uma evidência é suficiente para suportar uma hipótese e pode enfraquecer

outras hipóteses concorrentes, mas não é necessária para inferir causalidade.

4) O último teste, o duplo decisivo ou *double decisive* é o tipo mais forte e mais incomum pois indica que a evidência é ao mesmo tempo necessária e suficiente para comprova a hipótese.

Falleti reforça a importância da teoria para guiar a investigação e destaca o rastreamento de processos como um método histórico-comparativo nas Ciências Políticas. A autora define também o rastreamento de processos guiado por teoria, ou *theory guided process tracing*, como uma ferramenta que permite demonstrar, testar e gerar teorias a partir das evidências coletadas, indicando uma maior flexibilidade sobre como o método se relaciona com o uso de teorias e hipóteses. Estabelece também as formas extensiva e intensiva da metodologia, que diferem pela inclusão ou exclusão das causas e consequências na cadeia causal analisada. No caso do intensivo apenas as variáveis e mecanismos causais são estudados, enquanto o extensivo considera também as causas e consequências na observação. Para Falleti o rastreamento de processos é um conceito guarda-chuva que abrange as explicações sobre fenômenos a partir da elaboração do histórico ou teste de hipóteses.

Trata-se portanto de um método focado em análise, não apenas na descrição ou reconstrução da sequência de fatos. O rastreamento de processos busca abrir a caixa-preta de fenômenos complexos observados e compreender, a partir das múltiplas variáveis, quais são os resultados produzidos por ações e eventos. No caso do objeto estudado aqui, a partir da compreensão de que a moderação de conteúdos constitui um processo identificável pela criação e aplicação de normas definidas pelas plataformas, busca-se estudar comparativamente o que dizem as regras e que medidas são cabíveis quando há infração às normas.

Para estudar a reação das plataformas que excluíram as páginas de Trump após a invasão ao Capitólio, foram coletados e analisados dados qualitativos, na forma de documentos reguladores das políticas de conteúdos e comunicados oficiais das plataformas estudadas, com objetivo de buscar evidências sobre o

funcionamento da regulação de perfis e postagens em mídias sociais e a atuação da moderação neste caso específico. Para determinar a utilidade e validade das informações usou-se os testes para estabelecer que as evidências apontadas nos documentos fornecem bases para compreender o processo de remoção de conteúdos por parte das plataformas. No caso tratado aqui as evidências são aprovadas no teste do aro, visto que são necessárias mas não suficientes para comprovar a afirmação de que há problemas na moderação de conteúdos. O acumulo de evidências fortalece esta afirmação ao avaliar como se deu o processo de remoção em comparação aos documentos analisados.

4. Revisão dos documentos e dados em discussão

No dia 6 de Janeiro de 2021 houve uma tentativa de invasão ao Capitólio durante a etapa de certificação dos resultados e oficialização da eleição de Joseph Biden e a confirmação da derrota de Donald Trump. O evento é significativo por dois motivos: por marcar um raro momento em que a democracia dos Estados Unidos esteve ameaçada por uma multidão violenta comandada por um líder político; e porque foi organizado e transmitido por meio de plataformas de mídias sociais, utilizadas para divulgar informações e influenciar as pessoas ao ato. Trump, então ainda presidente dos Estados Unidos, utilizou suas redes sociais para postar vídeos e mensagens estimulando seus eleitores a invadir o Capitólio para contestar a vitória de seu opositor.

Durante seu mandato, Trump ficou conhecido por utilizar as mídias sociais para se comunicar e mobilizar seus eleitores, e em mais de uma ocasião publicou mensagens de cunho violento e discriminatório contra minorias sociais, notícias falsas, informações falsas sobre Covid e sobre os resultados das eleições. Apesar das múltiplas infrações não houve maiores reações por parte das plataformas para remoção de conteúdos até o ataque ao Capitólio. Nessa ocasião, Trump usou suas redes para convocar os eleitores a se insurgirem e, como consequência, todas as

contas oficiais de mídias sociais de Trump foram bloqueadas. Foi a primeira vez que plataformas de mídias sociais apagam conteúdos e perfis oficiais de um político líder de Estado em exercício.

As plataformas de mídias sociais estudadas aqui possuem conjuntos de regras definidas sobre conteúdos problemáticos como os postados por Trump. Revisaremos a seguir o que dizem as políticas de Padrões da comunidade do Facebook e as Regras do Twitter para analisar quais conteúdos são proibidos e quais as sanções para a postagem desses tipos de conteúdos. Analisaremos também notas oficiais e documentos de ambas as empresas que esclarecem as ações tomadas para a derrubada dos perfis de Trump.

Os documentos reguladores do Twitter e do Facebook servem como guias de referência para determinar os tipos de conteúdos proibidos em suas redes. Nas descrições das normas de ambas as empresas estão definidos tipos específicos referentes à violência direcionada, ameaça e incitação ao crime, divulgação de notícias falsas sobre Covid e resultados eleitorais entre as violações cometidas por Trump nos anos recentes. Em comparação, os Padrões da Comunidade do Facebook estabelecem apenas os tipos de conteúdos proibidos, mas não deixam claro qual é a sanção ou penalidade para perfis que infringem as normas, ou quais métodos utiliza para limitar a exposição das mensagens. Além das normas que definem conteúdos proibidos, ambas as plataformas possuem também políticas que indicam o tratamento diferenciado a conteúdos segundo interesse público quando se trata de líderes políticos mesmo em contextos de violação das normas.

O Twitter define com maior especificidade em suas regras os conteúdos proibidos e é mais transparente quanto aos métodos de remoção e punição para infratores. Dá mais detalhes nas páginas de ajuda que fazem referência às regras, indicando a possibilidade de diminuir o alcance da postagem, limitar as respostas ao tweet e banimento temporário ou permanente da conta. Seus documentos indicam o uso de inteligência artificial para detectar e remover mensagens com informações sensíveis, ameaças violentas, assédio direcionado e discurso de ódio, enquanto os documentos do Facebook não definem com clareza quais os mecanismos de remoção. Ao longo do ano de 2020, o Twitter utilizou etiquetas para indicar que

postagens do então presidente continham desinformação e glorificação da violência como forma de aviso de que se tratava de conteúdo questionável. Nestes casos a plataforma incluiu um texto com aviso de desinformação e restringiu a divulgação do tweet apenas para ser repassado com comentário do usuário.

Por se tratar de uma pessoa pública e presidente de um país, as mensagens na plataforma do Twitter receberam tratamento diferenciado e permaneceram online, segundo as regras da empresa sobre interesse público que impede que postagens de líderes de Estado sejam deletadas[3]. A norma, publicada em 2019 no blog da empresa, determina que, se a postagem é de interesse público, pode permanecer no ar com um rótulo de aviso de conteúdo mesmo que viole as regras da plataforma e determina que haverá sanção apenas para casos em que a postagem promova terrorismo ou violências direcionadas e específicas, compartilhamento de informações ou imagens privadas, automutilação e exploração de menores.

A primeira reação do Twitter às postagens de Trump foi bloquear a conta por um período de 12 horas e apagar as mensagens por violação das regras sobre integridade cívica e ameaças violentas. Pela recusa de Trump em apagar e manter as postagens na página e para evitar novas violações, a plataforma decidiu por suspender a conta permanentemente em 8 de Janeiro pela infração da política de glorificação da violência:

> Devido às tensões crescentes nos Estados Unidos, e a escalada na conversação global sobre as pessoas que atacaram violentamente o Capitólio em 6 de janeiro de 2021, esses dois tweets devem ser lidos tanto no contexto amplo de eventos do país e no modo como as declarações do presidente podem ser mobilizadas por diferentes públicos, incluindo incitação à violência, quanto no contexto de um padrão de comportamento dessa conta nas semanas recentes. Após verificar a linguagem nos tweets em relação à nossa política de glorificação da violência, nós determinamos que esses tweets violam esta política e o usuário @realDonaldTrump deve ser suspenso permanentemente do serviço
>
> (https://blog.twitter.com/en_us/topics/company/2020/suspension.html, tradução própria)

Os Padrões da Comunidade do Facebook definem as restrições sobre conteúdos com incitação ou declarações de cunho violento, reproduzidas por

3https://help.twitter.com/en/rules-and-policies/public-interest

indivíduos ou organizações e direcionadas a pessoas ou grupos sociais, e sobre discurso de ódio e violência explícita mais especificamente na área de Conteúdos Questionáveis do documento.

Apresentam as variações de discursos de ódio proibidos em três categorias: conteúdos visando indivíduos ou grupos étnicos de maneira degradante, comparando a animais e objetos; indivíduos ou grupos com base em características físicas e intolerância a transtornos ou deficiências; e discursos de exclusão e segregação social, política e econômica.

A plataforma registra as atualizações da política em uma página separada para apontar reforços ao texto das restrições por meio de definições mais específicas que possam oferecer ameaça a grupos sociais e indivíduos.

Ofensas anteriores de Trump contendo notícias falsas, como a divulgação de resultados não oficiais que indicariam sua vitória nas eleições não sofreram sanção por parte do Facebook, apenas um aviso na forma de etiqueta para avisar sobre a contagem dos votos, semelhante aos rótulos usados pelo Twitter. O Facebook descreve sem detalhes a implementação de inteligência artificial para detectar e sancionar conteúdos que violem as políticas e impedir a divulgação de postagens com conteúdos de ódio e incitação à violência, indicando falta de transparência nos documentos.

Assim como o Twitter, o Facebook define em seus regulamentos que os Padrões da comunidade não se aplicam igualmente a todos. Desde 2016 há registro de que postagens “dignas de notícia”, significativas ou de importância para o interesse público não sofrem as mesmas sanções, e em 2019 a questão foi formalizada em uma postagem da página oficial da empresa:

> Quando determinamos que algo é digno de nota, nós avaliamos o valor de interesse público do discurso em relação ao risco de dano. Quando comparamos esses interesses, consideramos um número de fatores, incluindo circunstâncias específicas de países, como se há uma eleição a caminho ou se o país está em guerra; a natureza do discurso, incluindo se há relação com governança ou políticas; e a estrutura política do país, incluindo se o país possui uma imprensa livre. Na avaliação do risco a dano nós consideramos a severidade do dano. Conteúdo que possui potencial de incitar violência, por exemplo, pode representar risco à segurança que supera o valor de interesse público . Cada uma dessas avaliações será

> holística e compreensiva em natureza e em consideração aos padrões de direitos humanos internacionais
>
> (https://about.fb.com/news/2019/09/elections-and-political-speech/, tradução própria)

O Facebook reagiu imediatamente à postagem no perfil de Trump e tomou providências para buscar e remover todos os conteúdos de incitação à invasão e de questionamento das eleições buscando diminuir os riscos de violência, segundo a plataforma. Anunciou também medidas para maior controle dos administradores de grupos sobre o que podem postar, desativaram comentários em postagens que incitam violência ou reproduzem discurso de ódio e utilizaram mecanismos de Inteligência Artificial para demover conteúdos que violam as políticas, isto é, limitar a divulgação e diminuir o alcance das mensagens.

A reação imediata do Facebook sobre as postagens de Trump foi remover o conteúdo após o ocorrido e bloquear a conta por 24 horas. No dia seguinte, suspendeu o perfil por tempo indefinido, mas manteve outras postagens apenas com avisos de sinalização sobre o conteúdo até o banimento da conta em 7 de Janeiro.

Diferentemente do Twitter, uma das medidas adotadas pelo Facebook foi repassar a decisão do banimento do perfil para ser revisada pelo recém-criado conselho de supervisão da plataforma, um comitê financiado de maneira independente, organizado em novembro de 2018 e formado por profissionais de diferentes áreas do conhecimento para analisar as decisões sobre remoção de conteúdo da empresa, que por estatuto reconhece a autoridade do comitê e se submete às decisões sobre remoção de conteúdos. A página oficial do conselho descreve o trabalho do grupo da seguinte maneira:

> “O objetivo do comitê é promover a liberdade de expressão por meio da tomada de decisões independentes e baseadas em princípios com relação ao conteúdo no Facebook e no Instagram e por meio da emissão de recomendações sobre a política de conteúdo relevante da empresa do Facebook.” (https://www.oversightboard.com)

O conselho aceitou o caso em 21 de Janeiro e determinou que estabeleceria um veredito em até 90 dias, mas sem data definida para emitir resolução.

Percebemos, portanto, as semelhanças e diferenças no tratamento das postagens de Trump. Ambas as empresas definem em suas plataformas os conteúdos proibidos de acordo com normas para manter as redes seguras e livres de violência e ameaças, considerando que, para políticos eleitos e pessoas públicas, há regras que definem interesse público em manter conteúdos mesmo que violem as normas. Há uma busca pelo equilíbrio entre liberdade de expressão e interesse público, definidos nos documentos reguladores das plataformas, e a preservação da segurança e de informações relevantes nas redes mantidas pelas empresas, mas não há transparência sobre os processos de localização e remoção de conteúdos, se por moderadores humanos ou algoritmos.

Jack Dorsey e Mark Zuckerberg declararam publicamente perspectivas opostas sobre a checagem de fatos sobre conteúdos postados por Trump. Enquanto Dorsey defende que o Twitter exerça checagem de fatos, Para Zuckerberg o Faceook não pode ser um árbitro da verdade, se eximindo de responsabilidade sobre os conteúdos que circulam em sua rede.

5. Conclusão

As plataformas de mídias sociais operam a partir de uma lógica de empresa onde os dados oferecidos pelos milhões de usuários se convertem em capital financeiro em troca dos serviços de interconexão e interatividade para divulgação dos conteúdos publicados em suas redes. As plataformas buscam manter o interesse e engajamento de usuários ao mesmo tempo em que procuram manter as redes seguras e livres de conteúdos ofensivos como discursos de ódio e notícias falsas.

O estudo elaborado aqui mostra que as mídias sociais possuem poder de amplificar o alcance de mensagens e é necessário estabelecer mecanismos de governança que impeçam que mensagens de cunho antidemocrático ganhem espaço nas redes. Esses mecanismos se realizam na forma de autorregulação das

plataformas na definição e aplicação de normas de segurança sobre conteúdos proibidos.

No caso de Donald Trump, cujas postagens, mesmo representando risco para a ordem democrática, foram mantidas no ar até o limite, as plataformas agiram de acordo com suas regras, permitindo a divulgação de mensagens de incitação à violência, preconceito e notícias falsas até o que foi considerado aceitável pelas empresas. As páginas e postagens foram suspensas e apagadas somente após a realização de intento criminoso por parte dos eleitores de Trump, em função do risco que a divulgação de mensagens de violência originadas por um ex-presidente com milhões de seguidores pode oferecer para a democracia. Tanto o Twitter quanto o Facebook possuem conjuntos de regras que foram violadas repetidas vezes por Trump antes de ter suas páginas removidas das redes em decorrência dos eventos de 6 de Janeiro.

Após utilizar as redes durante todo o mandato para divulgar notícias falsas, questionar os resultados das eleições e promover violência, as plataformas baniram os perfis em uma atitude sem precedentes em se tratando de Donald Trump. Até então as únicas sanções haviam ocorrido na forma de retirada da postagem ou de sinalização quanto ao conteúdo. No caso da invasão ao Capitólio as plataformas justificam a exclusão do perfil com base no interesse público que determina que nenhuma postagem seria apagada até oferecer risco real à segurança.

É possível afirmar que há problemas de governança nas plataformas e como lidam com conteúdos ilícitos, pela falta de transparência nos processos de remoção e pelo volume de conteúdos ofensivos que circulam das redes apesar das políticas e práticas de moderação. O modo como as plataformas de mídias sociais são projetadas permite alcance para mensagens de todos os tipos, e usuários de grande notoriedade e milhões de seguidores podem influenciar a cultura e a política.

As normativas são opacas e com definições rasas, sem detalhamento de quem são os responsáveis por criar as normas e sobre a aplicação das políticas, pois não há esclarecimentos acerca das remoções, o que abre brechas para que conteúdos ofensivos possam continuar nas redes. As motivações por trás das normas são a proteção e segurança da comunidade, mas há abertura nas regras

que permitem a divulgação de mensagens de ódio e notícias falsas que podem ir contra os interesses dos usuários e cujas postagens são protegidas.

Há pouca informação também sobre os processos de remoção, se realizados por algoritmos ou moderadores humanos. Tanto o Facebook quanto o Twitter apresentam problemas em diferentes aspectos da moderação na questão de escala e volume e na opacidade das políticas e práticas. Embora o Twitter apresente esforços nos documentos para ser mais transparente e explicativo que o Facebook, as plataformas não estabelecem em que casos os revisores de conteúdo agem, e essa opacidade em torno dos métodos pode eximi-las de responsabilidade ao mesmo tempo que protegem a marca.

É necessário maior governança por parte do Estado para garantir que as plataformas ofereçam mais regulação sobre os conteúdos postados pelos usuários, visto que o modelo de autorregulação sustentado pelas empresas é insuficiente para impedir que discursos amplificados possam aumentar a intolerância e oferecer risco à democracia pela influência na política e na esfera pública que podem ser causadas pela pouca regulação das plataformas.

Referências

BEACH, D.; PEDERSEN, R. B. **Process-Tracing Methods. Foundations and Guidelines**. Ann Arbour: University of Michigan Press, 2013.

BENNETT, A.; CHECKEL, J. T. **Process tracing: from philosophical roots to best practices**. In: Process tracing. From Metaphor to Analytic Tool. Cambridge: Cambridge University Press, 2015.

CHARBONNEAU, É. et al. **Process Tracing in Public Administration: The Implications of Practitioner Insights for Methods of Inquiry**. In International Journal of Public Administration, v. 40, 2017.

COLLIER, D. **Understanding process tracing**. PS - Political Science and Politics, v. 44, n. 4, 2011.

CRAWFORD, Kate, GILLESPIE, Tarleton. **What is a flag for? Social media reporting tools and the vocabulary of complaint**. In New Media & Society. v. 18 n. 3, 2016

DENARDIS, L., Hackl, A.M. **Internet governance by social media platforms**. Telecommunications

Policy. v. 39 n. 9, 2015

FALLETI, T. G. **Process tracing of extensive and intensive processes**. New Political Economy, v. 21, n. 5, 2016.

GILLESPIE, Tarleton. **Governance of and by platforms** in SAGE handbook of social media, 2017

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

GORWA, Robert.. **What is Platform Governance?** In Information, Communication & Society. v. 22 n. 6, 2019

GORWA, Robert. T**he platform governance triangle: conceptualising the informal regulation of online content**. In Internet Policy Review, v. 8 n. 2, 2019

GORWA, Robert. BINNS, Reuben. KATZENBACH, Christian. **Algorithmic content moderation: Technical and political challenges in the automation of platform governance**. In Big Data & Society v. 7 n. 1, 2020

PASQUALE, Frank. **A Esfera Pública Automatizada** in Revista eletrônica do Programa de Mestrado em Comunicação da Faculdade Cásper Líbero. v. 1 n. 39, 2017

ROBERTS, Sarah T. **Behind the Screen: Content Moderation in the Shadows of Social Media**. Yale University Press, 2019

SILVEIRA, Sérgio Amadeu da. SOUZA, Joyce. AVELINO, Rodolfo. **A Sociedade de Controle: Manipulação e modulação nas redes digitais**. São Paulo. Hedra, 2018

VAN DJICK, Jose. **The Culture of Connectivity: A Critical History of Social Media**. Oxford; New York. Oxford University Press, 2013

ZWART, Melissa de. **Keeping the neighbourhood safe: How does social media moderation control what we see (and think)?** in Alternative Law Journal, v. 43 n.4, 2018

https://www.facebook.com/communitystandards

https://about.fb.com/news/2016/10/input-from-community-and-partners-on-our-community-standards/

https://about.fb.com/news/2019/09/elections-and-political-speech/

https://about.fb.com/news/2021/01/responding-to-the-violence-in-washington-dc/

https://about.fb.com/news/2021/01/referring-trump-suspension-to-oversight-board/

https://help.twitter.com/pt/rules-and-policies/twitter-rules

https://blog.twitter.com/en_us/topics/company/2020/suspension.html

https://blog.twitter.com/en_us/topics/company/2019/worldleaders2019.html

https://help.twitter.com/en/rules-and-policies/public-interest