



**AS POLÍTICAS DAS  
GRANDES PLATAFORMAS  
REFERENTES A  
DISCURSO DE ÓDIO  
DURANTE A COVID-19**

Ana Laura Pérez

Publicado em 2021 pela Organização das Nações Unidas para a Educação, a Ciência e a Cultura, 7, place de Fontenoy, 75352 Paris 07 SP, França, e o Escritório Regional de Ciências da UNESCO para a América Latina e o Caribe, UNESCO Montevidéo, Luis Piera 1992, 2º andar, 11200 Montevidéo, Uruguai.

© UNESCO 2021  
MTD/CI/2021/PI/01/REV1



Esta publicação está disponível por acesso aberto sob a licença Attribution-ShareAlike 3.0 IGO (CC BY-SA 3.0 IGO) (<http://creativecommons.org/licenses/by-sa/3.0/igo/>).

Ao usar o conteúdo desta publicação, os usuários aceitam os termos de uso do repositório de acesso aberto da UNESCO ([www.unesco.org/open-access/terms-use-ccbysa-sp](http://www.unesco.org/open-access/terms-use-ccbysa-sp)).

Os termos usados nesta publicação e a apresentação dos dados que nela aparecem não implicam nenhuma posição da UNESCO quanto à situação jurídica dos países, territórios, cidades ou regiões, nem quanto a suas autoridades, fronteiras ou limites.

As ideias e opiniões expressas neste trabalho provêm dos autores e não necessariamente reflectem as opiniões da UNESCO nem comprometem a Organização.

Coordenação editorial: Sandra Sharman  
Design gráfico: Trigeon.

Esta publicação contou com o apoio da OBSERVACOM.

# CONTEÚDO



Sumário executivo

**04**



Introdução

**05**



O discurso de ódio na Internet

**07**



2020: um “tsunami de ódio e xenofobia”

**10**



As políticas das plataformas referentes

**16**

- Facebook e a remoção de conteúdo de ódio
- A moderação do discurso de ódio no Twitter
- YouTube e a moderação do discurso de ódio na pandemia

**17**

**25**

**28**



Conclusões

**33**

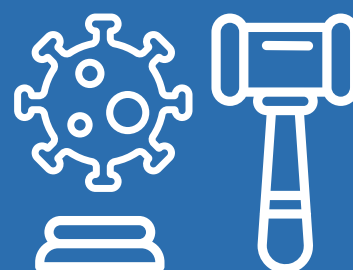


# SUMÁRIO EXECUTIVO

Este documento aborda um aumento de postagens consideradas como discurso de ódio desde a chegada da pandemia da COVID-19 no Facebook, Twitter e YouTube. Embora desigual, esse aumento pode ser atestado com base nos relatórios de transparência das diferentes plataformas e no crescimento registrado na moderação desse tipo de conteúdo desde março de 2020.

Considerando-se que, nesse mesmo período, e como consequência das medidas de isolamento adotadas pela maioria dos países do mundo, as plataformas tomaram a decisão de aumentar o uso de ferramentas de inteligência artificial em seus processos de moderação, não é possível assegurar com certeza que esse crescimento tenha sido registrado devido a um aumento na criação e publicação de mensagens ou por uma mudança nos sistemas de detecção que afetou os resultados de um ano para o outro.

# INTRODUÇÃO



A chegada da pandemia de COVID-19 em 2020 teve impactos que vão além dos serviços de saúde e das populações em todo o mundo. Impactos que, em alguns casos, ainda não conseguimos analisar totalmente e que provavelmente levaremos anos para determinar com certeza.

Alguns estudos registram o aumento de conteúdo classificável como discurso de ódio dentro das plataformas para alguns grupos específicos como resultado da pandemia de COVID-19, além de haver evidências de um crescimento do número de conteúdo removido das redes sociais com esse rótulo.

Desde 2020, as plataformas e redes sociais fizeram mudanças substanciais em seus critérios de moderação, adicionaram novas cláusulas a suas normas comunitárias e tiveram que aumentar o peso da moderação automática em seus processos normais devido ao fato de que muitos de seus funcionários tiveram que passar a trabalhar remotamente. A isso se acrescentou a entrada definitiva no debate público da preocupação com o impacto do discurso de ódio nessas plataformas e suas possíveis consequências na ocorrência de episódios de violência.

Este ano, o Twitter, o Facebook e o YouTube, em diferentes medidas, deixaram de tentar manter-se imparciais em relação ao debate público que ocorria em suas plataformas, chegando ao ponto, por exemplo, de bloquear as contas de um presidente em exercício durante os últimos dias de seu mandato.

Se forem analisados os relatórios de transparência das plataformas sociais, percebe-se que, durante 2020, o conteúdo categorizado como discurso de ódio cresceu significativamente nas redes sociais, bem como a eliminação de postagens por esse motivo. A falta de informações suficientemente desagregadas sobre o que cada uma das plataformas analisadas entende como discurso de ódio, bem como sobre processos de decisão e taxas de erro na tomada de decisões, dificulta a determinação dos motivos desse crescimento.

As plataformas, por sua vez, declararam publicamente ter experimentado problemas nos processos de moderação em decorrência do envio de milhares de funcionários dedicados a essas áreas para suas casas e a consequente decisão de aumentar o peso da moderação automatizada e dos sistemas de inteligência artificial. Também reconheceram a possibilidade de que isso tenha resultado em um aumento das taxas de erro como consequência de problemas no software de aprendizado de máquina para entender o contexto em que muitos conteúdos são criados e as diferenças entre essas capacidades e as capacidades dos moderadores humanos.

O Facebook, em particular, tanto nessa plataforma quanto em sua irmã Instagram, registrou aumentos exponenciais de conteúdos rotulados como discurso de ódio durante o período em que a COVID-19 se instalou no mundo. É notável que o aumento seja registrado de forma significativa a partir do segundo trimestre de 2020, quando, na maioria dos países do mundo, os governos começaram a implementar medidas de distanciamento físico sustentado e quarentenas ou lockdowns.

No entanto, não há dados para estabelecer se as razões para esse aumento também podem incluir uma mudança nos critérios de revisão dos conteúdos e a mudança para um modelo de moderação de conteúdos mais agressivo, nem se houve um aumento efetivo do discurso de ódio nas redes sociais a partir de 2020.

Este trabalho busca analisar o aumento do discurso de ódio na Internet desde a chegada da pandemia de COVID-19 ao mundo e as ações implementadas pelo Facebook, Twitter e YouTube, definindo seu escopo, efeitos, motivos e possíveis consequências.



# O DISCURSO DE ÓDIO NA INTERNET

Discurso de ódio é um termo complexo e díspar em sua definição, não havendo um acordo nem nas diferentes plataformas, nem nos governos e seus regulamentos e leis. O **documento Estratégia e Plano de Ação das Nações Unidas para Combater o Discurso de Ódio, assinado pelo secretário-geral das Nações Unidas, António Guterres** define discurso de ódio como “qualquer forma de comunicação verbal, por escrito ou por meio de comportamento, que seja um ataque ou use linguagem depreciativa ou discriminatória em relação a uma pessoa ou um grupo com base em quem eles são ou, em outras palavras, com relação a sua religião, etnia de origem, nacionalidade, raça, cor, ancestralidade, gênero ou outro fator de identidade”. Acrescenta-se que “em muitos casos, o discurso de ódio está enraizado na intolerância e no ódio ou gera-os, e, em certos contextos, pode ser degradante e divisionista”.

Do ponto de vista jurídico, o direito internacional não proíbe o discurso de ódio como tal, mas o “incitamento à discriminação, hostilidade ou violência”, sendo que o primeiro foi definido pelas Nações Unidas como “uma forma de expressão muito perigosa, uma vez que é explícita e deliberadamente pretende dar origem a discriminação, hostilidade e violência, que também pode provocar ou incluir atos de terrorismo ou crimes atrozes”. Por isso, como explicado no documento citado, o direito internacional não exige que

os Estados proíbam o discurso de ódio que não possa ser enquadrado como incitação. Porém, as Nações Unidas alertam que “mesmo quando o discurso de ódio não é proibido, pode ser prejudicial”.

“Em todo o mundo, estamos testemunhando uma onda preocupante de xenofobia, racismo e intolerância, com aumento do antissemitismo, ódio aos muçulmanos e perseguição aos cristãos. As redes sociais e outras formas de comunicação estão sendo exploradas como plataformas para promover a intolerância. Os movimentos neonazistas e pró-supremacia branca estão avançando, e o discurso público está tornando-se uma arma para obter ganhos políticos com uma retórica incendiária que estigmatiza e desumaniza as minorias, os migrantes, os refugiados, as mulheres e todos aqueles rotulados como ‘os outros’. E não é um fenômeno isolado nem a estridência de quatro indivíduos à margem da sociedade. O ódio está espalhando-se, tanto nas democracias liberais quanto nos sistemas autoritários, e, a cada regra que é quebrada, os pilares da nossa humanidade comum são enfraquecidos. O discurso de ódio constitui uma ameaça aos valores democráticos, à estabilidade social e à paz”, assegura a Organização das Nações Unidas no documento.

Da mesma forma, no **documento Recomendação Geral Nº 15 Relativa à Luta contra o Discurso de Ódio e Memorando Explicativo da Comissão Europeia contra o Racismo e a Intolerância (ECRI) do Conselho da Europa**, define-se o discurso de ódio como “o uso de uma ou mais formas específicas de expressão, como, por exemplo, a defesa, a promoção ou a instigação do ódio, a humilhação ou o desprezo de uma pessoa

---

ou grupo de pessoas, bem como o assédio, o descrédito, a disseminação de estereótipos negativos ou a estigmatização ou ameaça em relação à referida pessoa ou grupo de pessoas e a justificativa dessas manifestações com base em uma lista não exaustiva de características pessoais ou estados, que incluem a raça, a cor, o idioma, a religião ou as crenças, a nacionalidade ou a origem nacional ou étnica, bem como a ancestralidade, a idade, a deficiência, o sexo, o gênero, a identidade de gênero e a orientação sexual”.

De acordo com a definição, o discurso de ódio “não tem como objetivo apenas incitar o cometimento de atos de violência, intimidação, hostilidade ou discriminação, mas também atos que podem razoavelmente produzir esse efeito” e “motivos que vão além da raça, cor, idioma, religião ou crenças, nacionalidade, origem étnica ou nacional e ancestralidade”. Acrescenta-se, ainda, que o alcance do termo “expressão” se refere a “discursos verbais e publicações em qualquer uma de suas formas, inclusive o uso de meios eletrônicos e a sua divulgação e armazenamento”. Assim como o fato de que o discurso de ódio “pode assumir a forma oral ou escrita, ou qualquer outra forma, como pinturas, sinais, símbolos, desenhos, música, peças de teatro ou vídeos”, e “também engloba o uso de comportamentos específicos, como gestos para comunicar uma ideia, mensagem ou opinião”. A definição inclui “a negação, a banalização, a justificção ou o perdão público de crimes de genocídio, crimes contra a humanidade ou crimes em caso de conflito armado cujo cometimento seja provado depois que os tribunais proferem sentença ou exaltação de pessoas condenadas por tê-los cometido”.

Vários países do mundo têm legislação que proíbe o discurso de ódio e que geralmente se concentra na incitação ao ódio contra as pessoas com base em suas características de identidade.

Na América Latina, a abordagem tende a ser muito focada no legislativo e, como afirma Marianne Díaz Hernández em **seu trabalho “Discurso de ódio na América Latina: tendências de regulamentação, papel dos intermediários e riscos para a liberdade de expressão”**, na maioria dos casos, aposta-se na sanção penal direta, na sanção penal auxiliar (como agravante de um crime principal) e na proibição, que, sem criar sanções penais, estabelece medidas de reparação. Díaz Hernández acrescenta que vários países da América Latina (Costa Rica, El Salvador, Peru, Argentina, Bolívia e Uruguai, para citar alguns) “tipificaram a incitação ao ódio como crime em sua legislação penal geral”.

Por sua vez, entre os países que optaram pelo modelo sancionatório, nem todos caracterizam a incitação ao ódio com base nos mesmos parâmetros, sendo que alguns requerem a presença real ou potencial de um dano para configurar o crime. Nesse sentido, a Comissão Interamericana de Direitos Humanos tem enfatizado que “por princípio, em vez de restringi-los, os Estados devem promover mecanismos preventivos e educacionais, bem como debates mais amplos e aprofundados, como uma medida para expor e combater estereótipos negativos”.



---

No entanto, há um consenso de que o discurso de ódio pode desempenhar um papel na criação de condições para a violência contra grupos específicos da sociedade. O acadêmico Alexander Tsesis **argumenta que a principal motivação do discurso de ódio intimidatório é perpetuar e aumentar as inequidades existentes.**


“Embora a circulação de discurso de ódio intimidador nem sempre configure a existência de violência discriminatória, ela estabelece uma razão para atacar grupos particularmente desfavorecidos”, diz ele.

---

Os atos de violência contra os Rohingya em Mianmar, por exemplo, mostram o papel que as postagens no Facebook com conteúdo de discurso de ódio desempenharam no processo. Em 2018, **uma investigação da Reuters realizada em conjunto com o Centro de Direitos Humanos da UC Berkeley School of Law** detectou a existência de mais de 1.000 postagens definindo os Rohingya ou outros muçulmanos como cães, vermes e estupradores.

**Este conteúdo foi criado e distribuído no início de uma campanha de limpeza étnica e crimes contra a humanidade levada a cabo pelas Forças Armadas de Mianmar que resultou na fuga de 740.000 pessoas da etnia Rohingya para Bangladesh.**





# 2020: UM “TSUNAMI DE ÓDIO E XENOFOBIA”

**Em maio de 2020, o secretário-geral das Nações Unidas, António Guterres, advertiu que a pandemia de COVID-19 havia desencadeado um “tsunami de ódio e xenofobia” no mundo, “buscando bodes expiatórios e fomentando o medo”, e convocou a “agir agora para fortalecer a imunidade das nossas sociedades contra o vírus do ódio”.**

“Migrantes e refugiados foram vilipendiados como a fonte do vírus e, portanto, foram negados acesso a tratamento médico”, disse ele, acrescentando que os idosos foram retratados como “caricaturas desprezíveis” que “sugerem que também são as mais prescindíveis”. Os jornalistas, profissionais de saúde, trabalhadores humanitários e defensores dos direitos humanos também “estão sendo agredidos pelo simples fato de fazer seu trabalho”.

Nesse sentido, a Alta Comissária das Nações Unidas para os Direitos Humanos, Michelle Bachelet, **expôs, na 13a Sessão do Fórum sobre Questões de Minorias, em novembro de 2020**, que as redes sociais têm significado novas “oportunidades para o exercício de liberdades fundamentais, como a expressão, a associação e a participação, expandindo-

as a extensões sem precedentes”, mas “essa expansão trouxe consigo novas e significativas ameaças ao espaço cívico e aos direitos das pessoas”.

“Uma delas é o discurso de ódio, amplamente difundido em várias plataformas sociais na Internet. As minorias têm sido alvo de incitação desproporcional à discriminação, à hostilidade e à violência. Isso pode levar a tensões, distúrbios e ataques contra indivíduos e grupos. Também pode ser usado para servir a interesses políticos e contribuir para um clima de medo entre as comunidades minoritárias”, declarou Bachelet.

A alta comissária disse que “os mesmos direitos que as pessoas têm fora da Internet também devem ser protegidos dentro dela”, e acrescentou que as empresas que possuem redes sociais “têm a responsabilidade de prevenir, mitigar e remediar as violações dos direitos humanos que causam ou contribuem para que ocorram”.

“As empresas de redes sociais têm a opção de retirar ou deixar o material online. Elas também podem sinalizar um conteúdo, adicionar material compensatório, avisar a quem está divulgando e sugerir moderação. A eliminação só seria justificada nos casos mais graves. Quaisquer soluções propostas para lidar com o discurso de ódio nas redes sociais devem funcionar para preencher uma enorme lacuna na transparência e na responsabilidade democrática na tomada de decisões das plataformas. Não devemos apenas esperar que sigam as diretrizes de direitos humanos, mas também precisamos de mecanismos para monitorar e avaliar suas ações, disse Bachelet.

Na mesma linha, o Relator Especial das Nações Unidas para a liberdade de religião ou crença, Ahmed Shaheed, denunciou, em abril de 2020, a disseminação nas plataformas sociais de uma teoria da “conspiração” que afirma “que os judeus ou Israel são os responsáveis por criar e difundir o vírus da Covid-19”.

*“A luta contra os discursos de ódio na Internet não terá sucesso se a mídia em geral ou as redes sociais não levarem a sério os relatórios sobre ódio cibernético dirigido contra judeus e outras minorias. (...) A mídia deve remover qualquer publicação que incite ao ódio ou à violência, além de identificar e relatar notícias falsas”, apontou. Shaheed acrescentou que “nestes momentos extraordinariamente difíceis, é mais necessário que nunca garantir que todas as pessoas possam exercer, sem qualquer temor e na maior medida possível, seu direito à liberdade de religião ou crenças, ao mesmo tempo em que se protege a saúde pública”.*

Durante a pandemia, houve um aumento do discurso de ódio nas redes sociais. Primeiro, em fevereiro, o alvo foi sobretudo a comunidade chinesa, porque foi nesse país que nasceu a COVID-19. Depois, o ódio se voltou contra o uso de máscaras, a ponto de culpar a população LGBTIQ por ser a suposta origem de um vírus considerado castigo divino. De acordo com **um estudo realizado pela empresa Light**, o discurso de ódio contra a China ou cidadãos chineses cresceu 900% no Twitter, e o tráfego para sites que

espalham discurso de ódio ou postagens específicas contra a comunidade chinesa ou asiática aumentou 200%.

Em muitos casos, essas expressões também surgiram dos líderes políticos de diferentes partes do mundo, tanto em suas plataformas sociais (com milhões de seguidores) quanto fora delas. **O uso do termo “vírus chinês”** em suas redes sociais pelo então presidente dos Estados Unidos, Donald Trump, e o uso do termo “vírus de Wuhan” pelo então secretário de Estado, Mike Pompeo, podem ter incentivado **o uso do discurso de ódio nos EUA**.

Em fevereiro de 2020, Luca Zaia, o prefeito da região italiana de Veneto, um dos primeiros epicentros da pandemia, disse aos jornalistas que o país lidaria com o vírus melhor que a China devido à **“higiene que tem nosso povo (...), os cidadãos italianos, a formação cultural que temos, de tomar banho, lavar-nos, lavar muitas vezes as mãos (...), enquanto todos nós já vimos os vídeos com chineses comendo ratos vivos”**.

Em abril do mesmo ano, o então ministro da educação do Brasil, **Abraham Weintraub**, **sugeriu** em um tuíte que a pandemia fazia parte do “plano de dominação mundial” do governo chinês.

Essa intensificação da retórica racista nas redes sociais e na mídia coincide com o aumento dos ataques contra esses mesmos grupos registrados em várias partes do mundo. **No Reino Unido, pessoas asiáticas têm sido espancadas** e transformadas em alvo de chacota e de acusações de propagar o coronavírus. Duas mulheres atacaram

---

estudantes chinesas na Austrália, espancaram-nas, chutaram uma delas e gritaram **“voltem para China”** e **“malditas imigrantes”**. Na Espanha, dois homens agrediram fisicamente **um jovem estadunidense de origem chinesa** até deixá-lo em coma por dois dias. No estado americano do Texas, um homem com uma faca **atacou uma família birmanesa**, acusando-os de ser um fator de contágio do coronavírus.

Na África, foram relatados incidentes de discriminação e ataques contra pessoas asiáticas acusadas de serem portadoras de coronavírus, bem como contra estrangeiros em geral no **Quênia, Etiópia e África do Sul**.

Na América Latina, também foram registrados casos. No Brasil, a mídia noticiou a ocorrência de **casos de assédio e rejeição** contra pessoas de origem asiática. Em um desses episódios, uma estudante de direito denunciou ter sido vítima de racismo e xenofobia por uma passageira do metrô no Rio de Janeiro. “Essa mulher esperou que eu fosse até a porta do vagão para gritar ‘olhe para a chinesa que está indo embora, porca chinesa’, ‘nojenta’ e ‘ela fica aqui e nos contamina a todos’”, postou Marie Okabayashi no Twitter com um vídeo da agressora.

**A historiadora mexicana Yuriko Valdez**, de origem chinesa e autora do documentário “O legado de minha raça. Chineses e mestiços

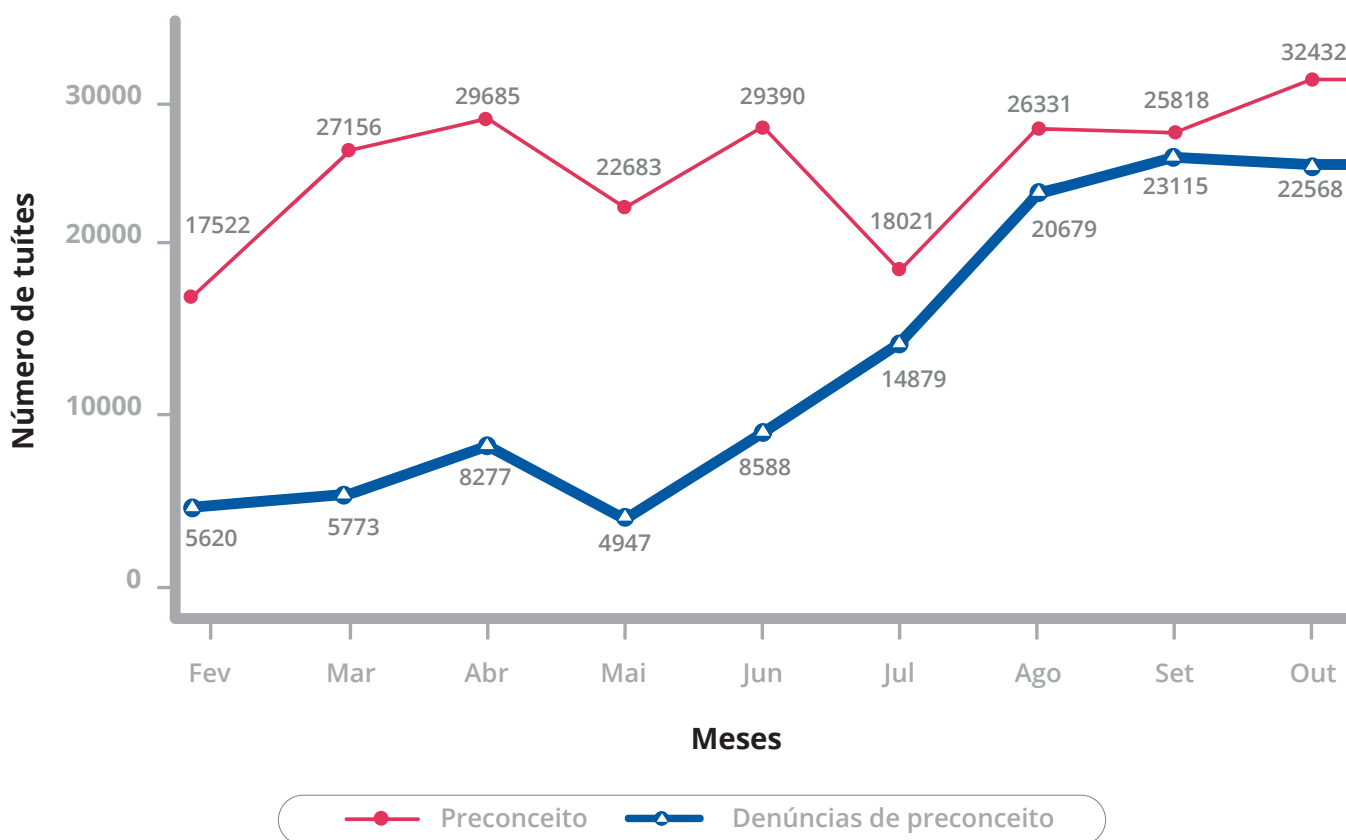
em Mexicali”, alerta sobre a proliferação de atitudes xenófobas por parte da comunidade daquela cidade, bem como os inúmeros comentários racistas nas redes sociais em publicações que falavam de comemorações como o Ano Novo Chinês em 25 de janeiro. Aos comentários usuais de “os chineses comem ratos e cães”, somaram-se “chineses porcos” ou “eles vão nos infectar porque a China é o foco da infecção do coronavírus”. Reações no mesmo sentido, por parte de “pessoas orgulhosas que alegam ser verdadeiramente de Mexicali”, apresentaram-se na promoção da abertura de uma exposição da Associação China no Zoológico Bosque da Cidade: “Os chineses não merecem uma homenagem”, “estão doentes de coronavírus”, entre outras mensagens que Valdez relata.

“As expressões de racismo e xenofobia relacionadas à COVID-19 em plataformas digitais incluem assédio, discurso de ódio, proliferação de estereótipos discriminatórios e teorias da conspiração. Não é surpreendente que os líderes que tentam atribuir a COVID-19 a certos grupos nacionais ou étnicos sejam os mesmos líderes populistas nacionalistas que fizeram da retórica racista e xenófoba o centro de suas plataformas políticas”, disse E. Tendayi Achiume, relatora especial sobre **as Formas Contemporâneas de Racismo, Discriminação Racial, Xenofobia e Formas Correlatas de Intolerância**.

A Unidade de Migração do Banco Interamericano de Desenvolvimento (BID) realizou um estudo entre fevereiro e dezembro de 2020 em que acompanhou conversas sobre **imigrantes no Twitter**. Nesta pesquisa, sete países da região considerados receptores importantes de migrantes foram monitorados: Argentina, Chile, Colômbia, Costa Rica, Equador, Panamá e Peru. Nesse monitoramento, foram coletados tuítes com termos como asilo, xenofobia, migrante, imigrante, refugiado, exilado e, uma vez feita a coleta, um algoritmo classifica-os em oito categorias mutuamente exclusivas. As primeiras sete categorias incluem tuítes que expressam preconceito

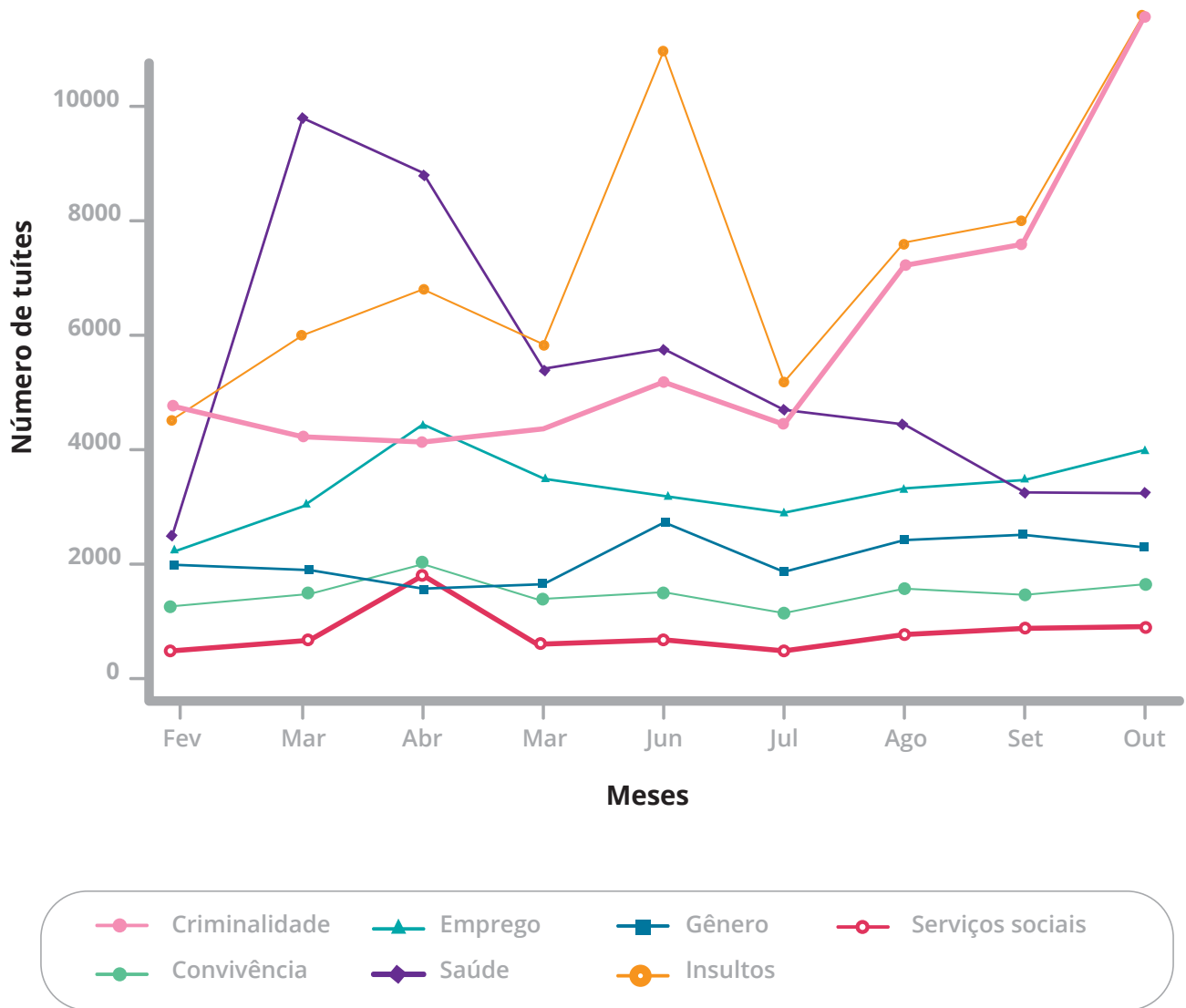
em relação aos migrantes nas áreas de Criminalidade, Emprego, Gênero, Serviços Sociais, Convivência, Saúde e Insultos em geral. A oitava categoria inclui os tuítes em que esses preconceitos são denunciados ou repudiados, conforme explicitado na investigação.

Com base nisso, e usando os tuítes de fevereiro como base pré-pandemia, o estudo encontrou um aumento de 70% das expressões de preconceito em relação aos migrantes em dois meses, passando de 17.522 tuítes mensais em fevereiro para 29.685 em abril.



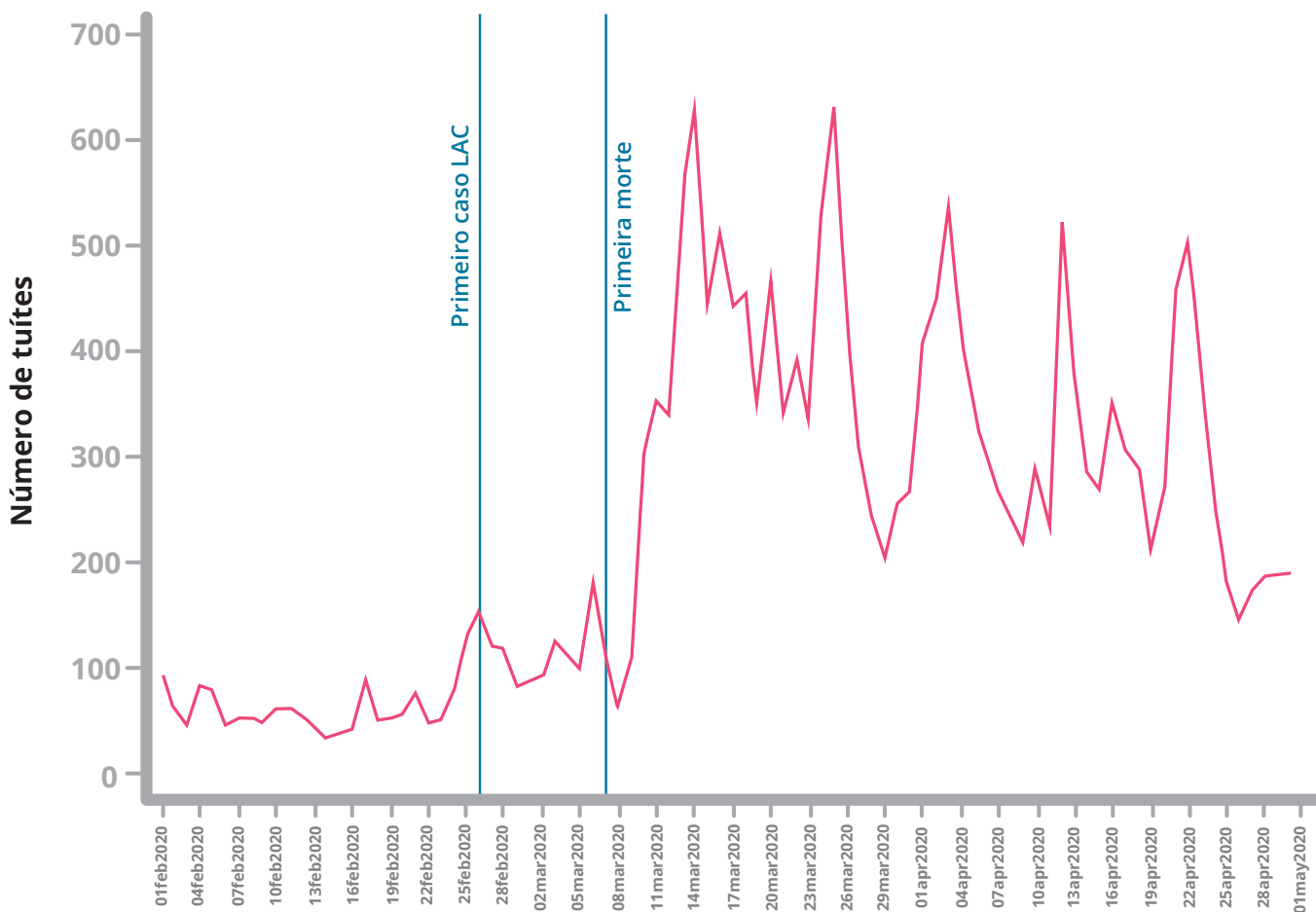
Fonte: Unidade de Migração do BID com base em dados gerados pelo Citibeats.

De acordo com o estudo, grande parte desse aumento registrado entre fevereiro e abril é explicado por preconceitos associados à área de saúde, “explicados principalmente pelo medo de que os migrantes transmitam a doença ou causem um colapso nos sistemas de saúde”.



Fonte: Unidade de Migração do BID com base em dados gerados pelo Citibeats.

O estudo do BID afirma que “esses preconceitos foram desencadeados após a primeira morte por COVID-19 ser anunciada na região”, **em março de 2020, na Argentina.**



Fonte: Unidade de Migração do BID com base em dados gerados pelo Citibeats.

Nos meses seguintes, os autores do estudo afirmam que encontraram oscilações nos níveis de xenofobia ou preconceito, mas sempre superiores aos níveis pré-pandemia de fevereiro. Em outubro, registrou-se um aumento, neste caso explicado por outros fatores (como a criminalidade) que não têm relação direta com a pandemia, e uma queda nos tuítes referentes a questões de saúde.



# AS POLÍTICAS DAS

## PLATAFORMAS REFERENTES AO DISCURSO DE ÓDIO DURANTE A PANDEMIA

**“Acredito firmemente que o Facebook não deve ser o árbitro da verdade de tudo o que as pessoas dizem na Internet”.**

**Essa frase, dita por Mark Zuckerberg** repetidamente ao longo dos anos, é um bom resumo da atitude das plataformas sobre a moderação de conteúdos até 2020. Mesmo depois das eleições de 2016 nos Estados Unidos, o Facebook, o Twitter e o YouTube enfrentaram sérias críticas por seu papel na disseminação de desinformação, ódio e teorias da conspiração, mas permaneceram muito resistentes a tomar medidas a esse respeito.

Em 2020, isso mudou. O Facebook, o Twitter e o YouTube fizeram mudanças em suas

normas comunitárias e termos de operação, o que haviam resistido a fazer por anos, desde rotular como falsas informações de contas públicas até excluir postagens de um presidente em exercício dos Estados Unidos e eliminar sua conta.

Em junho de 2020, a morte do afro-americano George Floyd como resultado de sua prisão nas mãos de quatro policiais de Minneapolis gerou uma onda de protestos em todo o mundo contra o racismo e a brutalidade policial. O então presidente dos Estados Unidos, Donald Trump, fez **uma série de postagens em suas plataformas sociais e, em uma delas em particular, escreveu: “Quando começa o saque, começa o tiroteio”**. Isso foi interpretado por grande parte da comunidade afro-americana como uma ameaça aos manifestantes. O Twitter resolveu ocultar o conteúdo. O Facebook, não.

Em meio às críticas, o diretor executivo do Facebook, Mark Zuckerberg, escreveu uma postagem explicando seus motivos para manter o post de Trump ativo. “Discordo veementemente do que o presidente disse sobre isso, mas acho que as pessoas deveriam ver por si mesmas, porque em última instância a responsabilidade dos que ocupam cargos de poder só pode ser cumprida quando seu discurso é abertamente analisado”, escreveu.

Semanas depois, um grupo de empresas, entre elas a Unilever, a Coca Cola, a Verizon e a Honda, anunciaram o início da campanha Stop Hate for Profit e a suspensão por um mês da compra de publicidade na plataforma. O vice-presidente de mídia da Unilever, Luis Di Como, disse que continuar



a anunciar “nessas plataformas no momento não agregaria valor para as pessoas e a sociedade”. **“Dada a polarização atual e as eleições nos Estados Unidos, tem que haver muito mais conformidade com as normas na área de discurso de ódio”, reclamou.**

“Respeitamos profundamente a decisão de qualquer marca e continuamos focados no importante trabalho de remover o discurso de ódio e entregar informações cruciais sobre a votação”, foi a resposta de Carolyn Everson, vice-presidente do grupo de negócios globais do Facebook, na segunda-feira. “Nossas conversas com empresas e organizações de direitos civis são sobre como podemos ser uma força para o bem, juntos”.

No entanto, em janeiro de 2021, Trump foi indefinidamente suspenso do Twitter e do Facebook, e alguns de seus vídeos foram removidos do YouTube, por disseminar mensagens denunciando uma suposta fraude eleitoral nas últimas eleições norte-americanas dirigidas a apoiadores que invadiram o Capitólio em Washington, gerando fortes episódios de violência e medo entre legisladores e funcionários, bem como **a morte de pessoas.**

“Os eventos chocantes das últimas 24 horas demonstram claramente que o presidente Donald Trump pretende usar seu tempo restante no cargo para impedir uma transição pacífica e legal de poder a seu sucessor eleito, Joe Biden”, escreveu Zuckerberg em uma mensagem no Facebook para explicar a decisão por trás do bloqueio.

---

## FACEBOOK E A REMOÇÃO DE CONTEÚDO DE ÓDIO

---

Em suas Normas Comunitárias, o Facebook define especificamente o discurso de ódio como “um ataque direto às pessoas pelo que denominamos ‘características protegidas’: raça, etnia, nacionalidade, deficiência, religião, classe, orientação sexual, sexo, identidade de gênero e doença grave”.

*“Definimos um ataque como a linguagem violenta ou desumanizante, os estereótipos prejudiciais, as declarações de inferioridade, as expressões de desprezo, a repulsa ou rejeição, os insultos ou incitamentos para excluir ou segregar. Consideramos a idade uma característica protegida quando mencionada em conjunto com outra característica protegida. Também protegemos refugiados, migrantes, imigrantes e solicitantes de asilo contra ataques graves, embora permitamos comentários e críticas relacionadas às políticas de imigração. Da mesma forma, oferecemos certas proteções para características, como profissão, quando são mencionadas em conjunto com uma característica protegida”, explica o Facebook.*

---

E acrescenta-se: “Temos consciência de que, por vezes, as pessoas partilham conteúdos que incluem linguagem que incita ao ódio emitida por outra pessoa com a intenção de o reprovar ou de conscientizar outras pessoas. Em outros casos, a linguagem que, de outra forma, violaria nossas regras pode ser usada de forma autorreferencial ou motivacional. Nossas políticas são projetadas para dar espaço a esses tipos de linguagem, mas exigimos que a intenção seja clara. Caso contrário, o conteúdo pode ser removido”.

A empresa categoriza o discurso de ódio em três níveis, de acordo com a gravidade do que é postado na rede social. O Nível 1 é todo aquele “conteúdo dirigido a uma pessoa ou grupo de pessoas com características protegidas” que inclui “linguagem que incita a violência ou que a apoia, tanto na forma escrita como visual” e “linguagem ou imagens desumanizantes em forma de comparações, generalizações ou declarações baseadas em comportamentos inadequados (na forma escrita ou visual) em relação a: insetos, animais que são culturalmente percebidos como inferiores do ponto de vista intelectual ou físico, sujeira, bactérias, doenças e fezes, predadores sexuais, infra-humanos, criminosos sexuais e violentos, outros criminosos (incluindo, por exemplo, ‘ladrões’, ‘bandidos’), afirmações que negam a existência, ridicularização do conceito de crimes de ódio, de atos desse tipo ou de suas vítimas, mesmo que não apareça uma pessoa real na imagem”.

Também se consideram discurso de ódio de Nível 1 “certas comparações, generalizações ou declarações com base em comportamentos desumanizantes (tanto escritos quanto visuais), inclusive pessoas

negras e macacos ou criaturas semelhantes a macacos, pessoas negras e máquinas agrícolas, caricaturas de pessoas negras com rostos pintados de negro, pessoas judias e ratos, pessoas judias comandando o mundo ou instituições importantes como cadeias de meios de comunicação, a economia ou o governo, negação ou distorção de informações sobre o Holocausto, pessoas muçulmanas e porcos, pessoas muçulmanas e sexo com cabras ou porcos, pessoas mexicanas e criaturas semelhantes a vermes, mulheres como objetos domésticos ou referindo-se às mulheres como propriedade ou ‘objetos’, bem como ‘referindo-se a pessoas transgênero ou de gênero não binário como se não fossem seres humanos ou dalits ou pessoas de castas ‘baixas’ ou registradas como servos”.

Os discursos de ódio de Nível 2 para o Facebook são aqueles que se referem a grupos protegidos e incluem “generalizações que denotam inferioridade (tanto escrita quanto visual)”, bem como “deficiências físicas” relacionadas à “higiene, incluindo, entre outras: ‘imundo’, ‘sujo’”, de “aparência física”, como “feio, horrível”, a “deficiências mentais”, como “bobo”, “estúpido”, “idiota”, referidas à educação como “analfabeto”, “inculto”, à saúde mental, como “doente mental”, “retardado”, “louco”, “demente”, e às “deficiências morais” relacionadas a “traços de personalidade considerados negativos culturalmente”, incluindo, mas não se limitando a: “covarde”, “mentiroso”, “arrogante”, “ignorante”, e “termos depreciativos relacionados à atividade sexual”, como “prostituta”, “cadela”, “pervertido”. Também se incluem no Nível 2 “expressões que denotam insuficiência”,

---

como “inútil”, “inservível”, “expressões de superioridade ou inferioridade em relação a outra característica protegida”, expressões relacionadas ao afastamento da norma, como “anormal”, e “expressões de desprezo”, como o “reconhecimento de intolerância a características protegidas”, como “homofóbico”, “islamofóbico”, “racista”, bem como “expressões que indiquem que uma característica protegida não deveria existir” e “expressões de ódio”, “rejeição” e “repulsa”, como “odeio” ou “não respeito”, “não gosto”, “não me importo”, “asqueroso”, “nojento”, “desagradável” etc. Também incluídos nesta categoria estão os insultos “relacionados aos órgãos genitais ou ao ânus para se referir a uma pessoa”, “frases ou termos ofensivos com a intenção de insultar”, “termos ou frases que incitam à participação em atividades sexuais ou que fazem referência a contato com os órgãos genitais ou o ânus, ou com fezes ou urina”.

Por fim, o Facebook categoriza como discurso de ódio Nível 3 o conteúdo em imagens ou texto que se refere a “segregação na forma de incitações, declarações de intenção, defesa ou apoio, ou declarações de aspirações ou condições em relação à segregação”, “exclusão na forma de incitações, declarações de intenção, defesa ou apoio, ou declarações de aspirações ou condições que incluam exclusão explícita, ou seja, atos como expulsar determinados grupos ou indicar que não têm permissão, exclusão política, ou seja, negar o direito de participação política, exclusão econômica, ou seja, negar o acesso a benefícios econômicos e limitar a participação no mercado de trabalho, exclusão social, ou seja, atos como negar

o acesso a determinados espaços (físicos e virtuais) e serviços sociais” e “conteúdos que descrevem negativamente ou destacam pessoas por meio de estigmas, que são definidos como palavras inerentemente ofensivas usadas como rótulos pejorativos”.

Em julho de 2020 e como resultado **da campanha Stop Hate for Profit**, mais de 1.200 empresas de todo o mundo se juntaram em um boicote publicitário contra as principais plataformas, exigindo um aumento da moderação do discurso de ódio, bem como a suspensão da publicidade de contas que promovam a discriminação contra grupos específicos. Uma das principais demandas das organizações e empresas envolvidas foi a eliminação de todas as contas de Trump.

Entre as demandas da coalizão, afirma-se que as plataformas devem eliminar “grupos ou páginas voltadas para a supremacia branca, milícias, antissemitismo, islamofobia e conspirações violentas”, “aumentar os recursos destinados ao monitoramento de grupos com discursos de ódio e de violência”, “mudar a política das plataformas para proibir qualquer página de evento que convoque às armas”, bem como “comprometer 5% de sua receita anual para o financiamento de um fundo independente que apoie iniciativas acadêmicas e de organizações que lutem contra o racismo, o ódio e a divisão causada pela inação do Facebook”.

Com base nessas reclamações, o Facebook publicou, em junho de 2020, **uma postagem na qual respondeu a alguns dos pedidos da Stop Hate for Profit**. Em relação ao pedido

---

da organização de “criar uma moderação separada composta por especialistas em ódio baseado em identidade para usuários que expressem terem sido atacados”, o Facebook garantiu que “relatos de discurso de ódio no Facebook já são automaticamente canalizados para um conjunto de revisores com capacitação específica em nossas políticas de ódio baseado em identidade em 50 mercados, cobrindo 30 idiomas”, e também “consultas com especialistas em ódio baseado em identidade para formular e desenvolver as políticas que esses revisores capacitados aplicam”. Também anunciaram sua “intenção de incluir a prevalência do discurso de ódio em futuros Relatórios de Conformidade com as Normas Comunitárias (CSER, na sigla em inglês), enquanto não houver mais complicações com a COVID-19”.

No mesmo mês, o vice-presidente de políticas públicas do Facebook, Richard Allan, escreveu uma coluna na qual abordou as diferenças na definição de discurso de ódio em diferentes partes do mundo e as dificuldades que a plataforma enfrentava para **detectá-lo adequadamente e tomar medidas a respeito**. “Não existe uma resposta universalmente aceita para quando alguém cruza a linha. Embora alguns países tenham leis contra o discurso de ódio, suas definições variam significativamente. Na Alemanha, por exemplo, as leis proíbem a incitação ao ódio; você pode ser alvo de uma batida policial por postar esse tipo de conteúdo na Internet. Nos Estados Unidos, por outro lado, mesmo os tipos de discurso mais vis são legalmente protegidos pela Constituição americana”, escreve Allen. “Pessoas que vivem no mesmo país – ou na casa ao lado – costumam ter diferentes níveis de tolerância ao discurso.

Para alguns, humor seco a respeito de um líder religioso pode ser considerado uma blasfêmia e um discurso de ódio contra todos os seguidores dessa religião. Para outros, uma batalha baseada em insultos de gênero pode ser uma maneira mutuamente agradável de compartilhar uma risada. É normal uma pessoa postar coisas negativas sobre pessoas de uma determinada nacionalidade enquanto compartilha dessa nacionalidade? O que aconteceria se um jovem se referisse a um determinado grupo étnico usando insultos raciais citando a letra de uma música?”, questiona-se o executivo do Facebook.

Allen também se refere no texto a erros na remoção de conteúdo que foi erroneamente classificado como discurso de ódio. “Se fracassamos na remoção do conteúdo que vocês denunciam como discurso de ódio, sentimos que não estamos cumprindo os valores de nossos Normas Comunitárias. Quando removemos algo que vocês publicam e acham que é um ponto de vista razoável, pode parecer censura. Sabemos como as pessoas se sentem mal quando cometemos esses erros e, por isso, trabalhamos constantemente para melhorar nossos processos e explicar as coisas com mais profundidade”, afirma o executivo do Facebook.

Ele acrescenta que os erros do Facebook na moderação de conteúdo “causaram grande preocupação em algumas comunidades, inclusive aqueles grupos que acreditam que agimos – ou deixamos de agir – por vieses”. “No ano passado (2019), Shaun King, um proeminente ativista afro-americano, postou um e-mail que tinha recebido contendo um discurso de ódio e incluindo insultos racistas. Excluimos a postagem de King por

---

engano por não reconhecer inicialmente que estava sendo compartilhada para condenar o ataque”, disse. Em julho, Nick Clegg, vice-presidente de Assuntos Globais e Comunicações do Facebook, escreveu **um artigo no qual afirmava que a empresa havia assumido várias medidas a partir das quais tinha registrado progressos significativo na eliminação do discurso de ódio em sua plataforma.** “Um relatório recente da Comissão Europeia descobriu que o Facebook avaliou 95,7% das denúncias de discurso de ódio em menos de 24 horas, mais rápido que o YouTube e o Twitter”, escreveu Clegg. “No mês passado, informamos que encontramos quase 90% do discurso de ódio que excluimos antes que alguém o denunciasse, em comparação com 24% há pouco mais de dois anos. Tomamos medidas contra 9,6 milhões de conteúdos no primeiro trimestre de 2020, contra 5,7 milhões no trimestre anterior. E 99% do conteúdo do ISIS e da Al Qaeda que removemos é removido antes que alguém nos informe”, declarou.

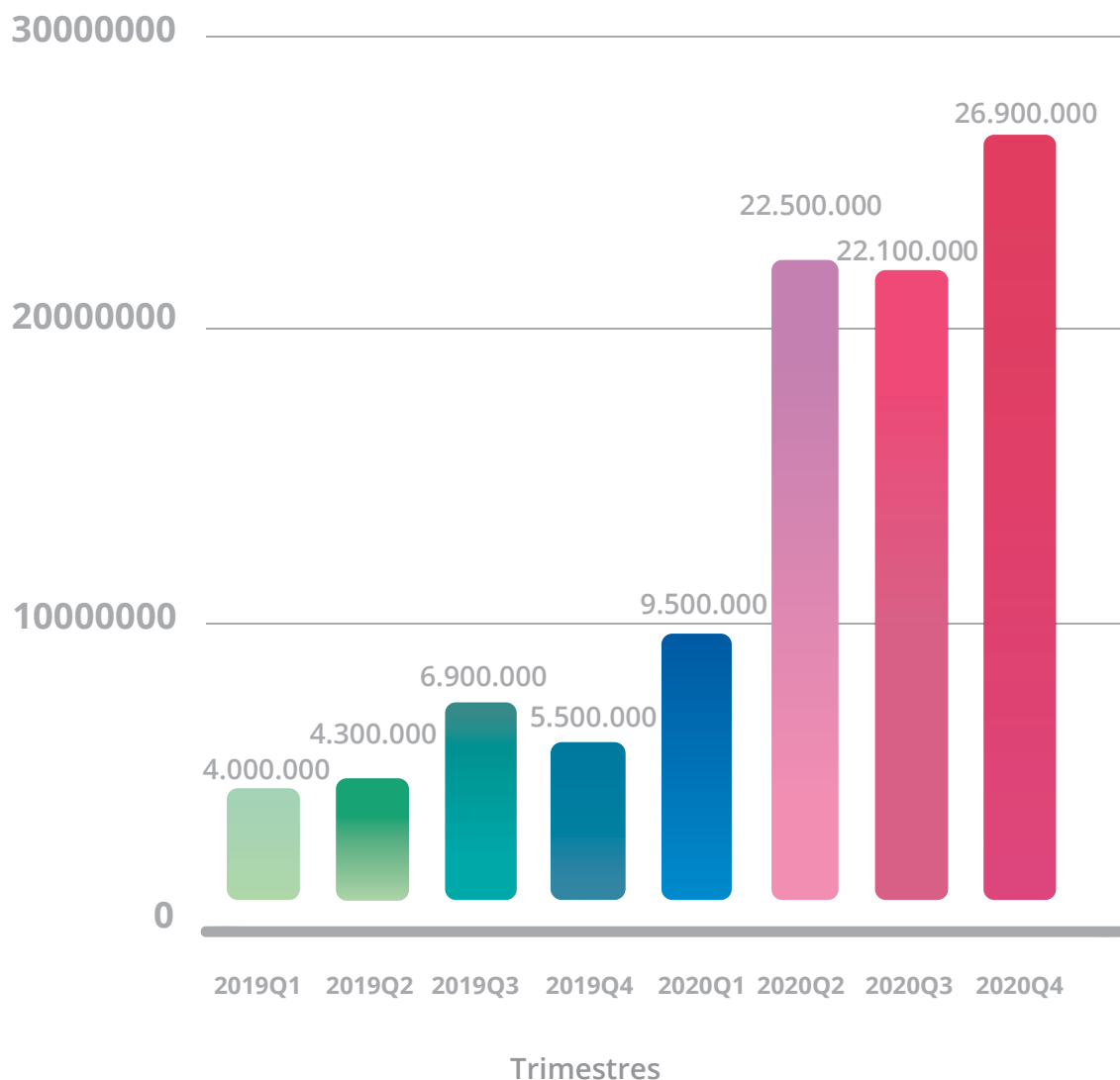
**De acordo com o Community Standards Enforcement Report (CSER), publicado em fevereiro de 2021**, o número de conteúdos em que o Facebook atuou aumentou de 20.700.000 em 2019 para 81.000.000, o que significa um aumento de quase 300% na quantidade de conteúdo categorizado como discurso de ódio entre um ano e o seguinte.

Em novembro, o Facebook começou a medir a prevalência de discurso de ódio na rede social e detectou que, entre julho e setembro, esse número estava entre 0,10% e 0,11%. Isso significa que, de cada 10.000 visualizações de publicações feitas na rede social, entre 10 e 11 seriam categorizadas como discurso de ódio de acordo com o Facebook. Entre outubro e dezembro de 2020, esse número diminuiu para 0,07% a 0,08% de prevalência. Em seu relatório, o Facebook não esclarece se essa mudança se deve a um aumento nas publicações em geral, a uma diminuição efetiva da categoria de discurso de ódio em relação ao total ou a uma mudança nos critérios ou processos de detecção.

Se olharmos para 2020 em profundidade, poderemos detectar um aumento muito significativo no número de conteúdos com discurso de ódio sobre os quais o Facebook tomou medidas a partir do segundo trimestre de 2020. Entre janeiro e março, atuou-se sobre 9.500.000, enquanto, nos meses seguintes, o número dobrou, passando para 22.500.000 entre abril e junho, 22.100.000 entre julho e setembro, e 26.900.000 entre outubro e dezembro.

De acordo com o relatório CSER, o aumento do número de conteúdos detectados, bem como da percentagem de proatividade na detecção, deve-se “principalmente à melhoria dos sistemas tecnológicos de detecção em árabe e espanhol” e “à expansão da automatização em português”.

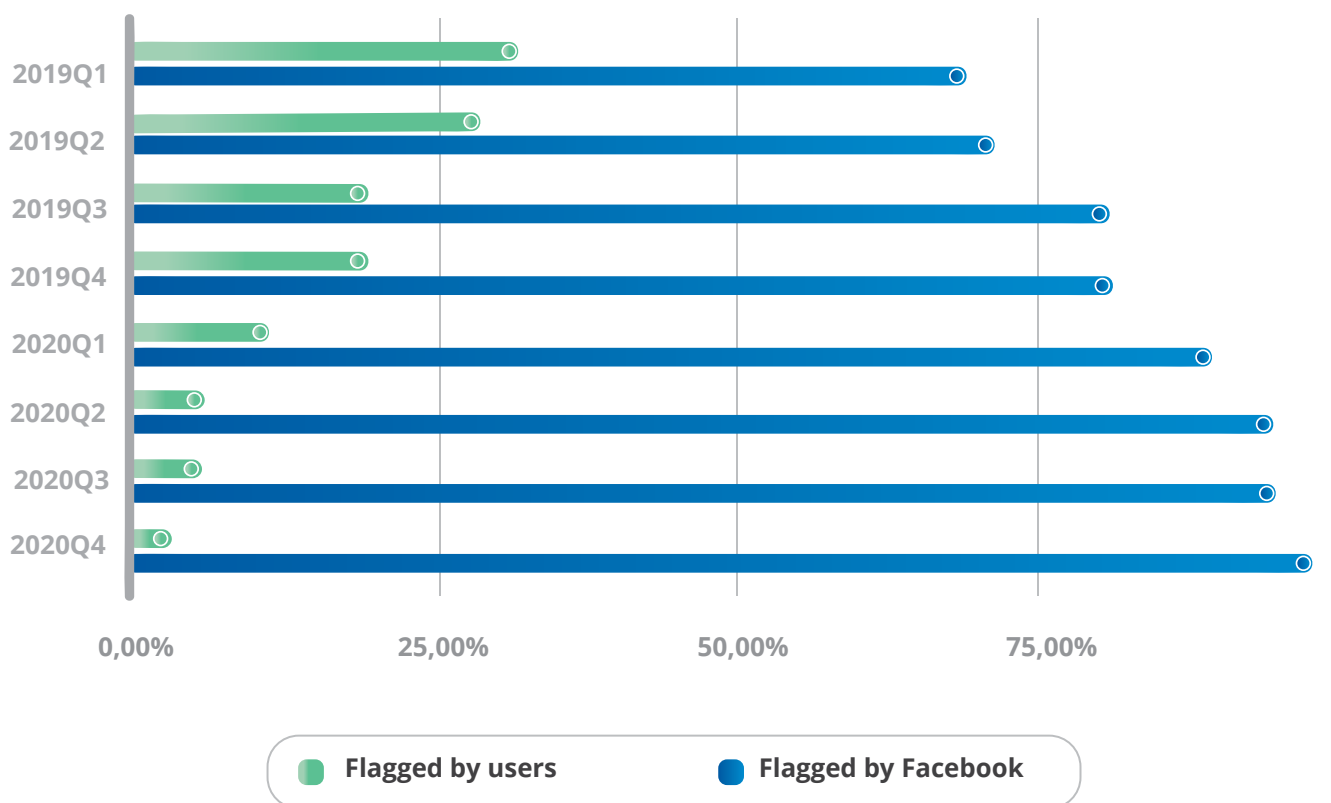
## Conteúdo sobre o qual foram tomadas medidas por incitação ao ódio



Outro aspecto que vale destacar é o aumento do percentual detectado pelo Facebook em relação ao relatado pelos usuários sobre o total de conteúdo acionado em função da categoria de discurso de ódio. Com pequenas oscilações, o peso dos sistemas internos do Facebook no total de conteúdo com discurso de ódio vem aumentando constantemente desde 2018, atingindo praticamente o total no último trimestre de 2020.

No último trimestre de 2017, o Facebook agiu em 1.700.000 por discurso de ódio, dos quais 76,4% foram detectados a partir de reclamações de usuários. Em 2020, essa relação foi revertida. Entre janeiro e março, 89,3% do conteúdo categorizado como discurso de ódio veio dos sistemas de detecção do Facebook, e algo semelhante aconteceu entre abril e junho (94,7%), julho e setembro (94,7%), e outubro e dezembro (97,1%).

## Quantos conteúdos com “discurso de ódio” foram



Algo semelhante aconteceu com o Instagram, também propriedade do Facebook, onde a ação sobre conteúdo de ódio é medida desde o último trimestre de 2019. Entre janeiro e março de 2020, o Instagram detectou e tomou medidas em 578.000 conteúdos por entender que incorreram em sua definição de discurso de ódio, indo para 3.200.000 entre abril e junho, 6.500.000 entre julho e setembro, e 6.600.000 entre outubro e dezembro de 2020.

No primeiro trimestre do ano, 57,1% do conteúdo foi detectado a partir de reclamações de usuários, enquanto, no

trimestre seguinte, a relação mudou radicalmente, e as reclamações dos usuários passaram a ser apenas 15,1% do total de ações realizadas na categoria de discurso de ódio. Essa relação se manteve nos seguintes trimestres: 5,2% entre julho e setembro, 4,9% entre outubro e dezembro.

Em meados de março de 2020, e após contínuos pedidos dessas equipes em decorrência das medidas de isolamento devido à pandemia da COVID-19, o Facebook decidiu enviar seus mais de 15.000 moderadores de conteúdo, em 20 lugares diferentes, para trabalhar em suas casas.

---

O diretor do Facebook, Mark Zuckerberg, disse naquela semana que o Facebook seria forçado durante a pandemia que atinge a maior parte do mundo a “apoiar-se mais ativamente no software de inteligência artificial para tomar as decisões de moderação de conteúdo”. A empresa também garantiu que realizaria treinamento em tempo integral para que prestassem “atenção extra” ao conteúdo “altamente sensível”. Ele alertou que os usuários “devem esperar mais quantidade de erros enquanto o Facebook melhora o processo, em parte porque apenas uma fração dos humanos continuaria participando e porque o software toma decisões mais ingênuas que os humanos, o que pode gerar “falsos positivos”, incluindo a remoção de conteúdo que não deveria ter sido removido. “Isso criará um trade-off contra alguns tipos de conteúdo que não apresentam risco físico iminente para as pessoas”, **disse Zuckerberg**.

Em novembro de 2020, o Facebook **anunciou mudanças em seus sistemas de moderação que implicam** um aumento da presença de moderação automatizada nas primeiras etapas de contato com o conteúdo. Chris Palow, engenheiro e integrante da equipe de Integridade do Facebook, admitiu durante uma coletiva de imprensa que “a inteligência artificial nunca será perfeita” e “tem seus limites” em termos de separar o discurso de ódio daquele que não é, por exemplo, por ser paródico ou humorístico. “O sistema busca combinar inteligência artificial com revisão humana para reduzir o número de erros”, explicou. O Facebook não divulga os números percentuais de conteúdo categorizado incorretamente como conteúdo que deve ser removido.

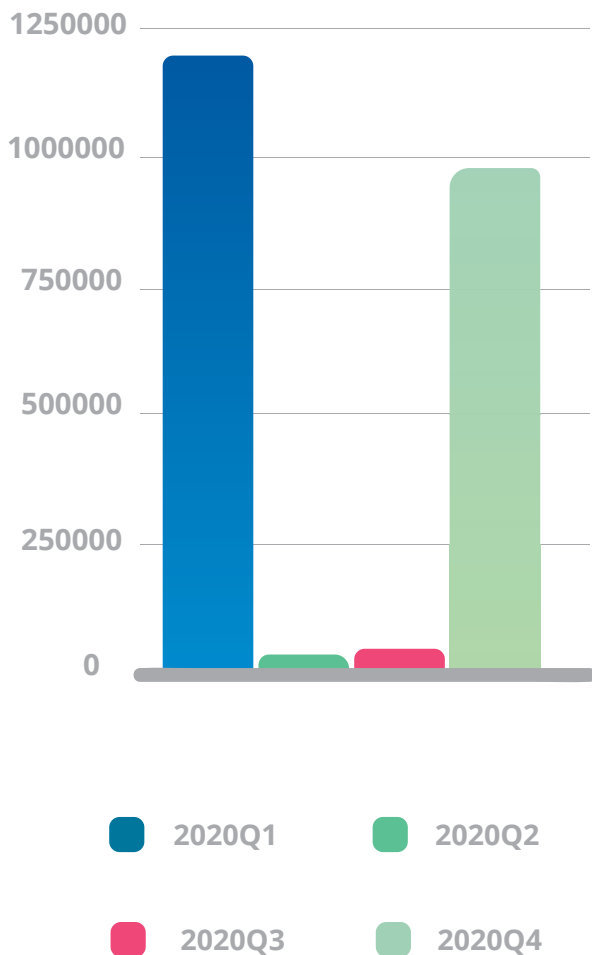
Meses depois, em fevereiro de 2021, o chefe de Política de Conteúdo Orgânico do Facebook, Varun Reddy, disse que a plataforma estava passando por problemas devido à ausência de moderadores humanos no processo de moderação de grande parte de seu conteúdo. A inteligência artificial aprende com moderadores humanos, explicou ele, e essa redução na presença humana mudou **“o quão efetiva é a inteligência artificial ao longo do tempo”**.

“Estamos trabalhando com os provedores para ter a maior capacidade de retorno virtual possível (...) Ainda não estamos no ponto de partida, mas, desde que começou o lockdown em 25 de março, esperamos que os sistemas voltem à eficiência total nos próximos meses”, disse Reddy em fevereiro deste ano.

Outro aspecto afetado pelo isolamento dos funcionários do Facebook foi o processo de recursos contra conteúdo que os usuários entendem que foi removido injustamente. “Devido a uma redução temporária em nossa capacidade de revisão como resultado da COVID-19, nem sempre podemos oferecer aos usuários a opção de entrar com um recurso. Ainda demos às pessoas a opção de nos dizer que discordam de nossa decisão, o que ajudou na revisão em muitos desses casos e na restauração do conteúdo nos casos em que fosse apropriado”, afirma o Facebook em seu **relatório CSER**. Aí se pode constatar que, entre abril e junho de 2020, quase não houve recursos, atingindo apenas 70.000 em todo o mundo durante esses 6 meses, quando, no trimestre anterior, haviam chegado a 1.200.000. No período seguinte, entre outubro e dezembro, os recursos chegaram a 984.200 mil casos.



## Recursos referentes a conteúdos acionados



Em 2020, o Facebook também atingiu números recordes de conteúdo restaurado em relação aos períodos anteriores, de 483.400 conteúdos em 2019 para 703.200 em 2020. Destes últimos, o Facebook restabeleceu 589.300 sem nenhum recurso.

## A MODERAÇÃO DO DISCURSO DE ÓDIO NO TWITTER

Em dezembro de 2020, o Twitter anunciou uma atualização de suas regras para combater a disseminação do discurso de ódio em sua plataforma e apoiou sua decisão em “investigações que associam a linguagem desumanizante à violência fora da Internet”. Em 2019, o Twitter atualizou suas regras sobre discurso de ódio para incluir religião e casta como grupos protegidos. Em março de 2020, adicionaram idade, deficiências e doenças, e, em dezembro de 2020, **anunciaram a proibição de linguagem que desumanize as pessoas por motivos de sua raça, etnia ou nacionalidade.**

Na publicação, foi incluída uma série de exemplos para ilustrar aqueles discursos que não seriam permitidas após o anúncio:

*“Todos os (nacionalidade) são baratas que vivem dos benefícios do Estado e devem ser expulsas”, “Gente que é (raça) são sanguessugas e só servem para uma coisa”, “Existem muitos (nacionalidade, raça, etnia) vermes em nosso país e eles devem ir embora”, “Todos (faixa etária) são sanguessugas e não merecem nosso apoio”, “Pessoas com (doenças) são ratos que poluem tudo ao seu redor”, “(Grupo religioso) deveria ser punido. Não estamos fazendo o suficiente para nos livrar desses animais fedorentos”.*

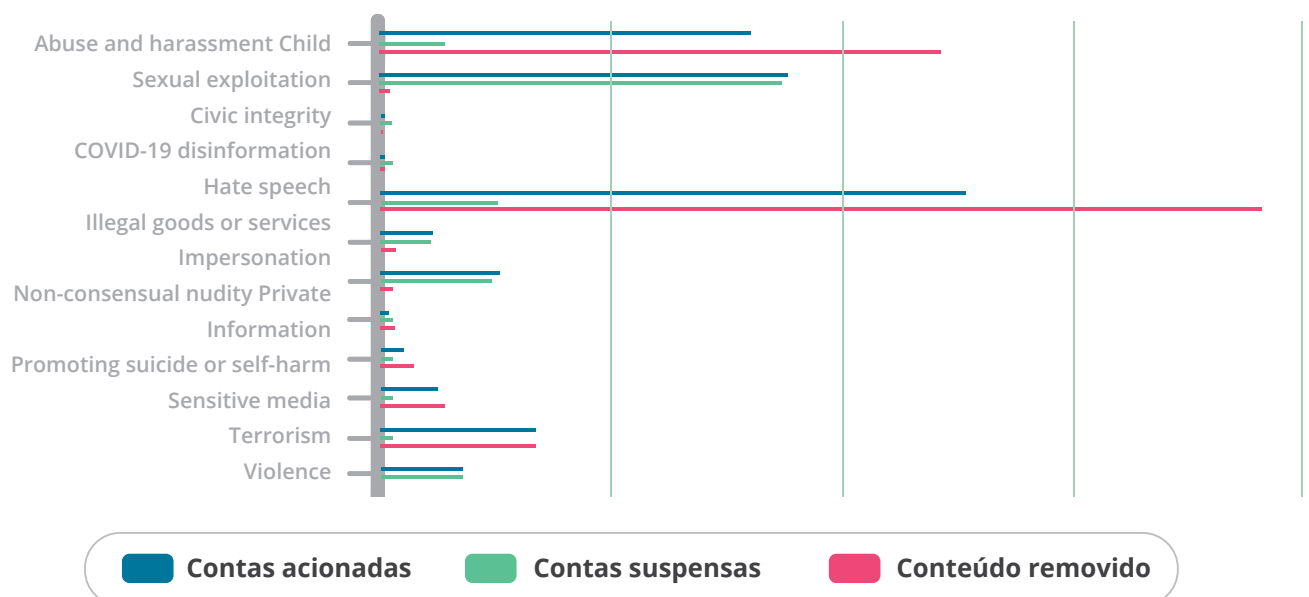
Em outubro de 2019, a atual vice-presidente norte-americana, Kamala Harris, publicou **uma carta aberta** ao diretor executivo do Twitter, Jack Dorsey, na qual exigia a moderação de algumas das publicações do então presidente Donald Trump porque, em sua opinião, violavam as normas comunitárias da rede social, inclusive as referentes ao discurso de ódio. “Nenhum usuário, independentemente de seu emprego, sua riqueza ou status, deve ser isento de seguir as regras de uso do Twitter”, argumentou Harris na carta.

No mesmo ano, **um estudo da Universidade de Nova York (NYU)** mostrou uma correlação entre o número de tuítes racistas e o número de crimes de ódio racistas ocorridos em 100 cidades dos Estados Unidos. “Acho que há um sentimento nos tuítes encontrados que está relacionado ao favorecimento de um ambiente que favoreça esses crimes”, argumenta Rumi Chunara, um dos autores do estudo. Ele acrescenta que, ao contrário, **“ter conversas produtivas melhora o meio ambiente e os resultados”**.

**“Atualmente, o sistema permite muito facilmente assediar e abusar de outras pessoas”, disse Dorsey em 2019**, acrescentando que “um dos problemas é o tamanho do peso atribuído aos seguidores e às curtidas”.

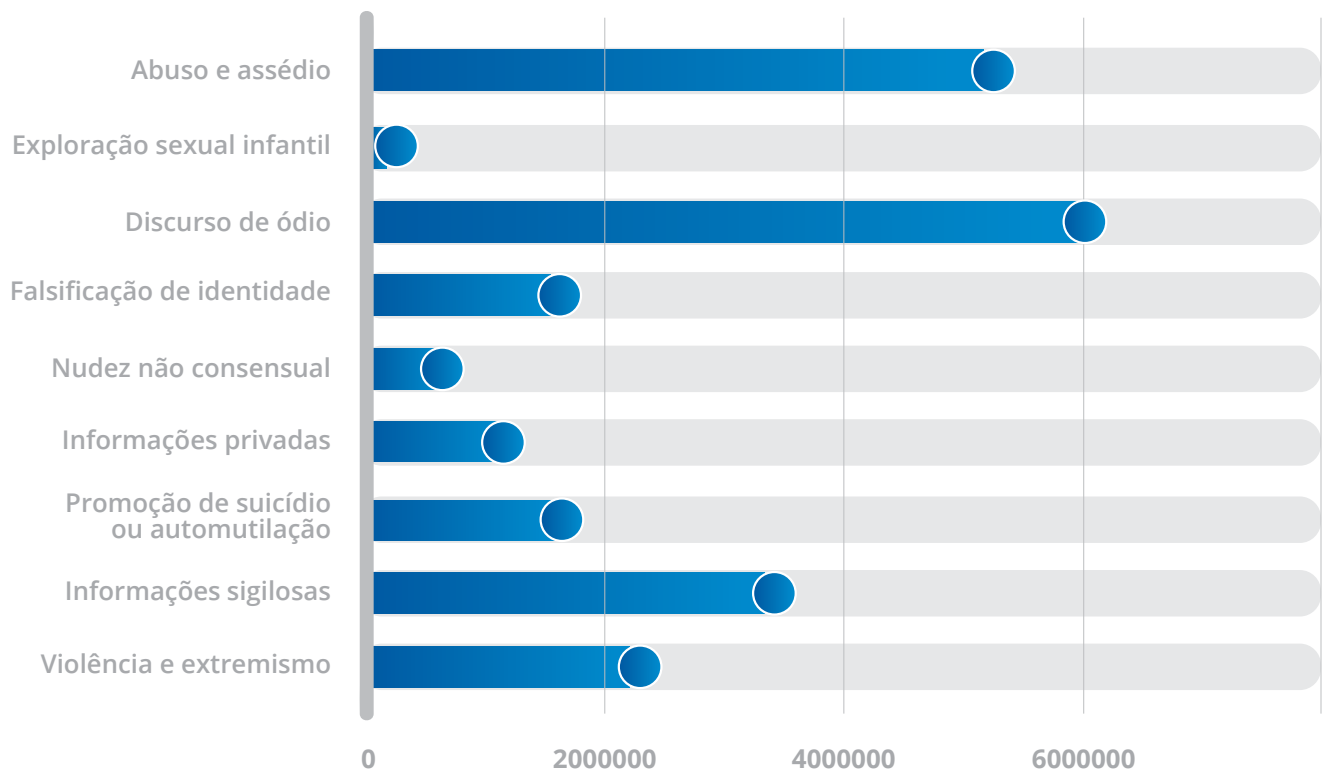
O que aconteceu em 2020 e durante a pandemia da COVID-19? De acordo **com o último relatório disponível** do Twitter Transparency Report, entre janeiro e junho daquele ano, foram realizadas ações sobre 1.940.082 contas, das quais 925.954 foram suspensas e 1.927.063 conteúdos foram removidos. Um número muito semelhante de conteúdo foi removido no mesmo período de 2019 (1.914.471), mas menor no caso de contas suspensas (687.397).

### Contas acionadas de janeiro a junho, por motivo



Do total, foram tomadas medidas em 645.416 contas com base em conteúdo rotulado como discurso de ódio, o que representa 33,2% das contas nas quais foram tomadas medidas. 12.400.000 contas foram notificadas no período de janeiro a junho, das quais cerca de metade (6.055.642) foi por motivos de discurso de ódio. De acordo com o relatório, foram denunciadas 30% mais contas que no mesmo período do ano anterior.

### Contas denunciadas janeiro - junho



O Twitter relata uma redução de 35% nas contas para as quais medidas foram tomadas por motivos de discurso de ódio em comparação com o período anterior, embora reconheça que, dadas as circunstâncias, as equipes concentraram-se em revisar o conteúdo que poderia causar danos ou que estava relacionado a informações errôneas sobre COVID-19, além de registrar "Atrasos significativos em todas as outras áreas".

**Em abril de 2020, o Twitter publicou uma postagem em seu blog informando sobre algumas mudanças como consequência da decisão de enviar a grande parte de seus funcionários para suas casas, atendendo às medidas de distanciamento social impostas por governos de todo o mundo.**

Uma parte dessas medidas foi o “aumento do uso de aprendizado de máquina e automação para realizar uma ampla gama de ações em conteúdo potencialmente abusivo e manipulador”. “Queremos ser claros: enquanto trabalhamos para tornar os sistemas consistentes, pode haver momentos em que a falta de contexto nos leve a cometer erros. Como resultado, não suspenderemos permanentemente com base apenas em sistemas de moderação automatizados. Em vez disso, continuaremos a procurar oportunidades onde as verificações de moderação humana tenham o maior impacto”, afirma o texto.

Nele, o Twitter relatou que, durante a pandemia de COVID-19, a tecnologia automatizada seria usada para “chamar a atenção sobre o conteúdo que tem maior probabilidade de causar danos e que será revisado primeiro” e para “identificar proativamente as violações das regras antes de serem denunciadas. As equipes aprendem com base em decisões anteriores e, portanto, com o tempo, a tecnologia pode ajudar a classificar o conteúdo ou revisar contas automaticamente”. Para conteúdo que “requer contexto adicional, como informações enganosas sobre a COVID-19”, o Twitter garante que suas equipes “continuarão a revisar manualmente os relatórios”.

A rede social esclarece que os tempos de resposta aos relatórios se estenderão “além dos tempos normais” e admite que “devido ao fato de que os sistemas automatizados não possuem todo o contexto nem o insight das equipes humanas, erros serão cometidos”.

---

## YOUTUBE E A MODERAÇÃO DO DISCURSO DE ÓDIO NA PANDEMIA

---

**No YouTube, a última atualização das Normas Comunitárias** sobre discurso de ódio data de 2019. Atualmente, a definição do que a empresa de propriedade do Google entende como discurso de ódio pressupõe “conteúdo que promove violência e ódio contra indivíduos ou grupos com base em qualquer um dos seguintes atributos: idade, casta, deficiência, etnia, identidade de gênero, nacionalidade, raça, status de imigração, religião, sexo ou gênero, orientação sexual, vítimas de um evento violento ou seus familiares e veteranos”.

Nas Normas Comunitárias, acrescenta-se que o YouTube não permite “que indivíduos ou grupos com essas características sejam desumanizados, afirmem que são física ou mentalmente inferiores, ou elogiem ou exaltem a violência contra eles”, nem “permitimos o uso de estereótipos que incitem ou promovam o ódio com base nessas características, nem insultos raciais, étnicos, religiosos ou outros cujo objetivo principal seja a promoção do ódio”, que “alegue a superioridade de um grupo sobre aqueles que possuem alguma das características mencionadas acima para justificar a violência, a discriminação, a segregação ou a exclusão”, ou “negue que ocorreram eventos violentos bem documentados”.

---

Em março de 2021, o YouTube participou de um acalorado debate sobre suas políticas de moderação de discurso de ódio ao remover um vídeo do comentarista Steve Crowder por considerar que violava suas políticas relacionadas à divulgação de informações incorretas sobre a COVID-19. Nesse vídeo, Crowder fez uma série de comentários sobre a decisão do governo republicano de conceder subsídio aos agricultores de minorias raciais, por considerá-los historicamente excluídos das políticas de auxílio ao setor. **Os comentários incluíam caracterizações sobre as maneiras de falar, mover-se e pensar dos afro-americanos.**

Após reclamações de diferentes organizações que defendem os direitos das minorias raciais, o YouTube emitiu um comunicado no qual assegurou que suas “políticas proíbem conteúdo que promova o ódio contra grupos com base em sua raça”, mas “embora seja ofensivo, esse vídeo de Steven Crowder não viola essas políticas”.

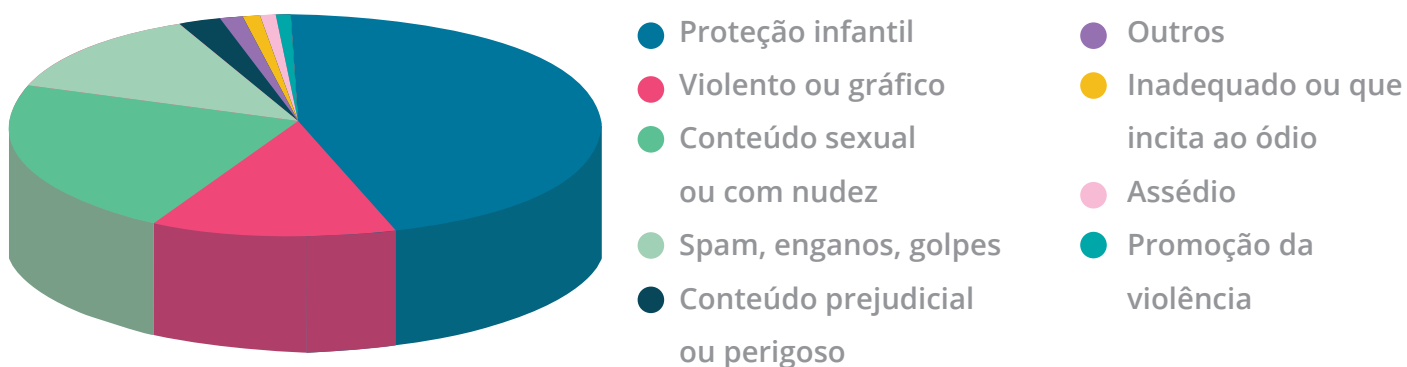
Em abril de 2021, o YouTube divulgou informações nas quais assegurou que havia aprimorado seus sistemas de detecção de discurso de ódio na plataforma. “Não queremos que o YouTube seja uma plataforma que possa causar danos atrozes ao mundo”, **disse o diretor de produto da plataforma, Neal Mohan.**

No YouTube, o fenômeno parece ser mais complexo de detectar. Na verdade, não está claro com base nos dados disponíveis que tenha havido um aumento significativo do discurso de ódio nessa plataforma, embora episódios isolados tenham sido registrados com destaque na mídia e na opinião pública.

Entre abril e junho de 2020, o YouTube removeu 11.401.696 vídeos, sem contar mais de 30.000.000 de vídeos removidos como resultado da eliminação de 1.998.635 canais no mesmo período. Desses mais de onze milhões de vídeos, apenas 552.062 vídeos foram removidos sem o uso de sistemas de detecção automática. Entre julho e setembro, foram eliminados 7.872.684 vídeos, e apenas 481.721 sem detecção automática, e, entre outubro e dezembro, 9.321.948 vídeos foram eliminados, e apenas 521.866 sem o uso de sistemas automatizados para detecção de violações das regras do YouTube.

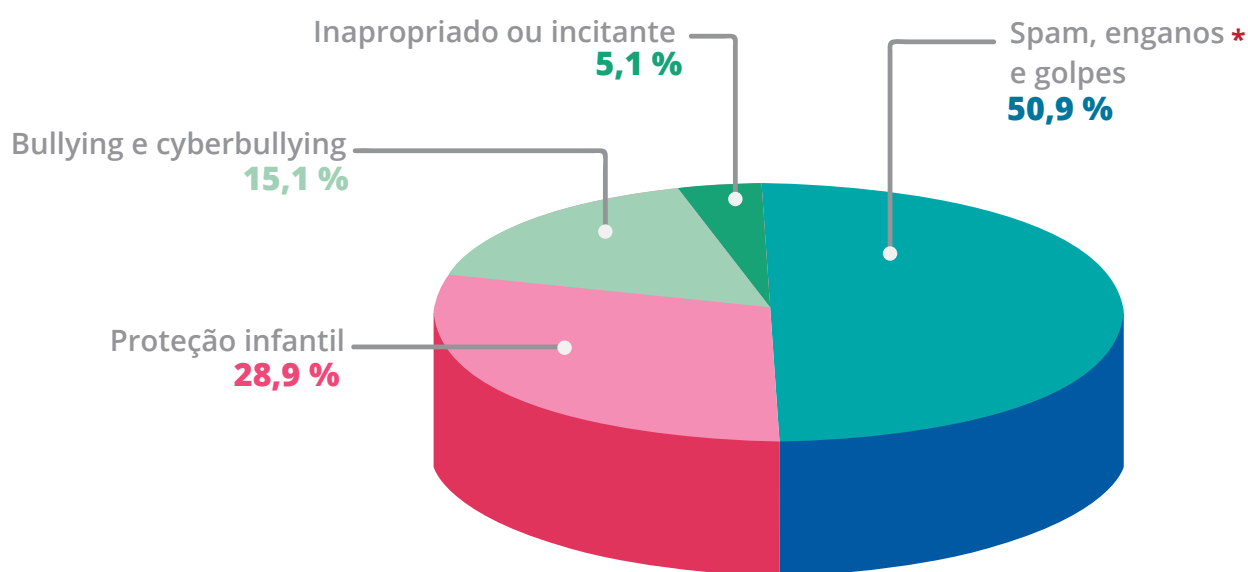
Quanto aos motivos, o discurso de ódio não ocupou um espaço significativo, atingindo o total de 97.362 vídeos eliminados no último trimestre de 2020, embora tenha registado um ligeiro aumento entre o trimestre abril-junho e os dois seguintes, passando de 0,7% para mais de 1% dos vídeos deletados.

## Vídeos removidos, por motivo, outubro - dezembro



Se for analisada a eliminação de comentários nos vídeos, sendo que 906.196.160 foram retirados no último trimestre de 2020, a motivação “incitamento ao ódio” sobe para 5% entre os motivos especificados para a eliminação desse conteúdo. Isso implica que, no último trimestre de 2020, mais de 46 milhões de comentários foram removidos do YouTube porque os sistemas de moderação automatizados entenderam que eles violavam as regras da plataforma em relação à categorização de “discurso de ódio”.

## Comentários retirados, de acordo com o motivo da retirada Out-Dez



---

Quando o YouTube enviou seus moderadores de conteúdo para casa em março devido à pandemia de COVID-19 e expandiu drasticamente o uso de seus filtros automatizados, isso levou a uma duplicação de vídeos que foram removidos no segundo trimestre de 2020. O crescimento deixou o debate aberto na rede social de propriedade da Alphabet sobre a precisão dos processos de moderação automatizada.

“Em resposta à situação da COVID-19, tomamos medidas para proteger nossa equipe externa e reduzir o pessoal presencial nos escritórios. Como resultado, e de forma temporária, estamos usando mais tecnologia para realizar algumas das tarefas que normalmente são feitas por revisores de carne e osso, por isso estamos removendo mais conteúdo que pode não violar nossas políticas. Isso influencia algumas das métricas neste relatório e provavelmente continuará a influenciar as métricas daqui para frente”, escreveu a empresa em um blog que acompanha [seu relatório de transparência do último trimestre](#). “Como a responsabilidade é nossa prioridade máxima, escolhemos a segunda opção: usar a tecnologia para ajudar em parte do trabalho que os revisores normalmente fazem”, explicou o Google.

No relatório do segundo trimestre, o YouTube admitiu que o aumento na remoção de conteúdo foi devido à empresa “aceitar um nível mais baixo de eficácia para garantir que estava removendo o máximo possível de conteúdos”.

“Uma das decisões que tomamos no início da pandemia, quando ficou claro que as máquinas não seriam tão precisas quanto os humanos, seria errar por ter certeza de que os usuários estariam protegidos, mesmo que isso poderia resultar em um número um pouco maior de vídeos baixados da plataforma”, garantiu o chefe de produto do YouTube, Neil Mohan, à [publicação especializada norte-americana Mashable](#).

Em setembro, o YouTube anunciou que moderadores humanos começariam a retornar aos escritórios e trabalhar na revisão dos sistemas de moderação para tentar voltar aos números do início de 2020.

Como pôde ser visto anteriormente, o uso de sistemas de detecção automática, segundo seus próprios executivos, fez com que o YouTube eliminasse inúmeros conteúdos que de fato não violavam suas Normas Comunitárias, chegando a dobrar o número de recursos, que passou de 166.000 no primeiro trimestre para 325.000 no segundo trimestre de 2020.

Ao contrário do Facebook, o YouTube não reduziu a atenção aos processos de entrada com recurso e continua mantendo os prazos de processamento anteriores à COVID-19. Isso significa que o número de vídeos restituídos após recursos também aumentou de 41.000 para 161.000 naquele período. [Isso significou um aumento na taxa normal de restituição de vídeos do YouTube, normalmente em 25% dos recursos, para quase a metade.](#)

---

**Em seu relatório de transparência, o YouTube detalha seu processo específico de moderação de discurso de ódio** e aborda algumas das dificuldades específicas que esse tipo de conteúdo apresenta em relação a outros tipos também proibidos pelas Normas Comunitárias.

“A política sobre a incitação ao ódio é complexa para ser aplicada em grande escala, uma vez que as decisões que devem ser tomadas requerem uma análise matizada do contexto e um entendimento completo da linguagem em questão. Para aplicar nossa política de maneira consistente, expandimos nossa equipe de revisão especializada no assunto e em temas linguísticos. Além disso, estamos implementando o aprendizado de máquina para detectar conteúdo que incite ao ódio e enviá-lo à equipe de revisão, e aplicamos as lições que aprendemos a outros tipos de conteúdo, como extremismo violento. Às vezes erramos, então temos um processo de recursos para criadores que acham que seu conteúdo foi removido indevidamente. Estamos constantemente avaliando nossas políticas e diretrizes de aplicação e continuaremos a colaborar com os especialistas e a comunidade para fazer mudanças quando necessário”, afirmam.

O YouTube acrescenta que, além de “remover conteúdo” que viola as Normas Comunitárias, eles estão trabalhando para “reduzir as recomendações de conteúdo que está no limite de violar” suas diretrizes. **“Há muito tempo, nós também temos diretrizes de conteúdo adequado para anunciantes, em que é proibida a exibição de anúncios em vídeos que incluam conteúdo que incite ao ódio”, sustentam.**





# CONCLUSÕES

Sob inúmeras pressões políticas, sociais e da mídia, o Facebook, o YouTube e o Twitter fizeram nos últimos meses mudanças em suas Normas Comunitárias relacionadas ao discurso de ódio e tomaram decisões às quais, em anos anteriores, pareciam resistir fortemente e que significam um aumento substancial em suas funções como reguladores do que pode e não pode ser dito nesses novos espaços públicos.

É difícil saber se essas mudanças tiveram sucesso e até mesmo definir o sucesso diante de medidas que as próprias plataformas admitem que não esteja claro se têm funcionado adequadamente durante a pandemia da COVID-19. Essas medidas incluíram, em alguns casos como Facebook e Instagram, ações altamente restritivas e até com poucas garantias, como o desaparecimento virtual dos processos de entrada com recurso por vários meses. Isso significou não apenas a eliminação de conteúdos de interesse público, mas também a perda do direito de exigir a revisão por parte de milhares de usuários na América Latina.

Além da falta de elementos para determinar de forma cabal cada um dos motivos que explicam essa mudança de critérios, o fato é que, em 2020, as plataformas tomaram decisões e fizeram alterações na forma e nos processos por meio dos quais moderam os conteúdos. Essas mudanças nos processos e nas Normas Comunitárias que os regulamentam significaram uma virada dramática no que diz respeito à forma como o Facebook, o Twitter e o YouTube trataram o conteúdo criado pelos usuários até agora.

Dois fenômenos parecem ter ocorrido este ano, um aumento muito significativo de publicações com conteúdo geralmente considerado “discurso de ódio” da pandemia de COVID-19 nas redes sociais. O Facebook é a rede social em que, pelo menos com base nos dados disponibilizados pelas próprias plataformas, registrou-se o maior crescimento. Entre 2019 e 2020, as postagens intervencionadas por essa rede social consideradas como discurso de ódio cresceram quase 300%. Ao aprofundar a análise de 2020, é notável que esse crescimento ocorra de forma muito mais expressiva no segundo trimestre do ano. Conforme destacado no capítulo anterior, a partir de março (momento da explosão global da pandemia de COVID-19), o número de postagens moderadas pela plataforma por serem consideradas discurso de ódio dobrou e manteve-se nesses números durante o resto do ano. O Twitter e o YouTube também registraram aumentos, embora não tão consideráveis.

O segundo fenômeno a ser destacado foi o fato de que, com base nos efeitos desse aumento do discurso de ódio e nas denúncias da sociedade civil a esse respeito, o Facebook, o Twitter e o YouTube decidiram aprofundar seu controle e sua intervenção no sentido de ampliar os tipos de conteúdo que eles consideram como estando fora de suas Normas Comunitárias. No entanto, não parece estar claro, na opinião de muitos analistas do mundo, que essas medidas sejam suficientes ou adequadas, além de haver problemas importantes na forma em que foram implementadas, afetando direitos fundamentais.

Apesar de uma crença popular, ainda profundamente enraizada, as redes sociais nunca foram espaços de intercâmbio totalmente abertos ou “desregulamentados”. As plataformas moderam há anos conteúdos que consideram “ilegais”, mas também aqueles que respondem a caracterizações ainda mais vagas (e não legalmente proibidas), como as que entendem como indecentes, obscenas e fora da moral dos seus países de origem.

A chegada ao mundo da COVID-19, uma pandemia global que levou milhões de pessoas a se isolarem em casa, reduzirem seus contatos e trabalharem remotamente, teve impactos de todos os tipos. Um deles foi o aumento do discurso de ódio nas plataformas sociais, mas outro, talvez menos detectável em uma primeira aproximação, foi a mudança nos processos de moderação que são realizados sobre o conteúdo que os usuários publicam. De acordo com uma investigação realizada na plataforma de busca Crowdtangle (que permite rastrear o uso de hashtags ou palavras no Facebook, Instagram e Twitter), entre fevereiro de 2020 e março de 2021, foram geradas 43.779 postagens no Facebook que utilizavam a expressão “vírus chinês”, e registrou-se um total de 3.535.409 interações. Os dois picos principais foram registrados em março e abril de 2020.

Governos de todo o mundo apelaram a seus cidadãos que fizessem um distanciamento social sustentado e, com isso, as plataformas tiveram que enviar milhares de moderadores humanos para suas casas. Essa decisão causou um aumento extremamente significativo do uso de ferramentas automatizadas e inteligência artificial na revisão das milhões de publicações que são enviadas às redes sociais a cada minuto. Embora em constante processo de aprimoramento, esses sistemas automatizados ainda não são capazes de compreender as diferenças de idioma, linguagem, idiosincrasia e cultura de milhões de usuários ao redor do mundo, bem como a importância do contexto para a definição de conceitos tão complexos quanto discurso de ódio.

De acordo com o estudo da Unesco Countering Online Hate Speech, existem pelo menos cinco abordagens não legislativas possíveis para o problema do discurso de ódio na Internet, e elas aludem diretamente ao papel das plataformas como uma parte substancial da solução do problema. No documento, a Unesco propõe o monitoramento e análise da sociedade civil, promoção do contradiscurso dos pares por parte dos indivíduos, ação organizada por parte das ONGs para denunciar os casos às autoridades, criação de campanhas de promoção de ações por parte das empresas de Internet que hospedam o conteúdo específico e empoderamento dos usuários por meio de educação e treinamento quanto aos conhecimentos, capacidades e aspectos éticos do exercício da liberdade de expressão na Internet

Também é claro que erros na detecção de discurso de ódio nas plataformas podem resultar na eliminação de conteúdo não incluso nessa definição e, portanto, em um impacto significativo sobre a liberdade de expressão como um direito humano fundamental.

As plataformas cresceram exponencialmente em todo o mundo e tornaram-se espaços de troca de ideias, então o que acontece lá afeta diretamente (ou tem o potencial de afetar) como o debate público é processado. Permitir que governos e plataformas se tornem reguladores de conteúdo pode resultar no silenciamento de vozes dissidentes, especialmente em sociedades autoritárias.

Porém, como afirma Díaz Hernández, o problema não é unicamente que as proibições resultam em restrições indevidas ou desproporcionais à liberdade de expressão, mas que também muitas vezes são ineficazes para abordar e resolver o problema subjacente, uma vez que não desempenham o papel de neutralizar o discurso de ódio. Em vez disso, muitas vezes acabam agravando o clima de violência e polarização social que deu origem ao conteúdo original.

É importante também ter em mente que os problemas derivados da regulação de conteúdo nas plataformas envolvem não só a regulação do conteúdo em si, mas também a arquitetura da Internet como a conhecemos, bem como suas características de caráter teoricamente extra-espacial e extraterritorial. Com base nessa estrutura e no papel que as plataformas e redes sociais desempenham nesse ecossistema, cada um desses ambientes possui suas próprias regras de funcionamento e gerou suas próprias definições do que é ou não proibido e permitido. Nesse sentido, parte do problema é que não se trata apenas do que a legislação de cada Estado entende por discurso de ódio, mas do que esse termo significa para o Facebook, o Twitter ou o YouTube, quando não estão sujeitos a controles democráticos e não oferecem garantias de devido processo legal ou transparência, entre outras.

A pandemia global trouxe consigo impactos de todos os tipos na vida das pessoas. Talvez um deles seja também o início de uma discussão sobre o papel das plataformas como moderadoras de conteúdo, os problemas derivados de permitir ou encorajá-las a ocupar o papel de gatekeepers na Internet.



## **ANA LAURA PÉREZ**

Uruguay

### **SOBRE A AUTORA**

É formada em Comunicação, com orientação em Jornalismo pela Universidade ORT, diploma em Estudos Latino-Americanos pela Universidade de Montevideu e mestrado em Business Administration pelo Instituto de Estudos Empresariais de Montevideu.

Há 20 anos, trabalha como jornalista e editora em alguns dos meios de comunicação mais influentes de seu país: os jornais El Observador e El País, e o semanário Búsqueda. Também apresenta e participa de programas em TV Ciudad, um canal de TV público da Prefeitura de Montevideu. Atualmente também é Gerente de Produto Digital do jornal El País.

Foi coordenadora de Jornalismo e Conteúdo Digital da Licenciatura em Comunicação da Universidade ORT, onde é professora há quase dez anos.

Tem participado como conferencista, palestrante e panelista em diferentes eventos sobre jornalismo, em particular sobre desinformação e plataformas digitais, temas nos quais tem se especializado nos últimos anos e sobre os quais tem ministrado capacitações e cursos de formação para jornalistas do Uruguai e de vários países da América Latina.



Financiado pela União Europeia

