unesco

Financiado por la Unión Europea

# LAS POLÍTICAS DE LAS GRANDES PLATAFORMAS SOBRE DISCURSO DE ODIO DURANTE EL COVID-19

Ana Laura Pérez

# LEGALES

Publicado en 2020 por la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura, 7, place de Fontenoy, 75352 París 07 SP, Francia y la Oficina Regional de Ciencias de la UNESCO para América Latina y el Caribe, UNESCO Montevideo, Luis Piera 1992, piso 2, 11200 Montevideo, Uruguay.


**MTD/CI/2021/PI/01**



Coordinación editorial: Sandra Sharman
Diseño gráfico: Trigeon.

Esta publicación contó con
el apoyo de OBSERVACOM.

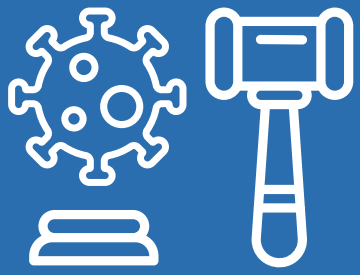
# CONTENIDO

# RESUMEN EJECUTIVO

Este documento da cuenta de un aumento de posteos considerados discurso de odio a partir de la llegada de la pandemia COVID-19 en Facebook, Twitter y YouTube. Aunque dispar, ese aumento puede establecerse a partir de los informes de transparencia de las distintas plataformas y los crecimientos registrados en la moderación de esos contenidos a partir de marzo de 2020.

Dado que en ese mismo período, y como consecuencia de las medidas de aislamiento tomadas ven la mayoría de los países del mundo, las plataformas tomaron la decisión de aumentar el uso de herramientas de inteligencia artificial en sus procesos de moderación no es posible asegurar a ciencia cierta que ese crecimiento se haya registrado por un aumento en la creación y publicación de mensajes o a partir de un cambio en los sistemas de detección que afectó los resultados entre un año y otro.

# INTRODUCCIÓN



La llegada de la pandemia de COVID-19 en 2020 tuvo impactos que van más allá de los servicios sanitarios y las poblaciones del mundo entero. Impactos que en algunos casos no hemos podido analizar a cabalidad aún y que probablemente lleven años lograr determinar a ciencia cierta.

Algunos estudios registran el aumento de contenido clasificable como discurso de odio dentro de las plataformas hacia algunos grupos específicos como resultado de la pandemia de COVID-19, así como existe evidencia acerca de un crecimiento en el número de contenidos eliminados de las redes sociales bajo esa etiqueta.

Desde 2020, las plataformas y redes sociales realizaron cambios sustanciales en sus criterios de moderación, añadieron nuevas cláusulas a sus normas comunitarias y debieron aumentar el peso de la moderación automática en sus procesos habituales debido a que muchos de sus trabajadores debieron irse a sus casas. A eso se sumó el ingreso definitivo al debate público de la preocupación por el impacto del discurso de odio en esas plataformas y sus posibles consecuencias en la ocurrencia de episodios de violencia.

Este año, Twitter, Facebook y YouTube -en distinta medida cada una-, pasaron de intentar mantenerse imparciales respecto del debate público que se daba en sus plataformas a, por ejemplo, bloquear las cuentas de un presidente en ejercicio durante los últimos días de su gobierno.

Si se analizan los reportes de transparencia de las plataformas sociales, es claro que durante 2020, el contenido categorizado como discurso de odio creció significativamente en las redes sociales, así como la eliminación de posteos por estas razones. La falta de información suficientemente desagregada acerca de lo que cada una de las plataformas analizadas entiende por discurso de odio, así como sobre los procesos de decisión y las tasas de error en la toma de decisiones hacen difícil determinar los motivos de este crecimiento.

Por su parte, las plataformas han declarado públicamente haber experimentado problemas en los procesos de moderación como resultado del envío de miles de trabajadores afectados a esas áreas a sus casas y la consiguiente decisión de aumentar el peso de la moderación automatizada y los sistemas de inteligencia artificial. También han reconocido la posibilidad de que esto haya redundado en un aumento de las tasas de error como consecuencia de los problemas en el software de machine learning para entender el contexto en el que muchos de los contenidos son creados y las diferencias entre esas capacidades y las de los moderadores humanos.

Facebook en particular, tanto en esa plataforma como en su hermana Instagram han registrado aumentos exponenciales de contenidos etiquetados como discurso de odio durante el período en el que el COVID-19 se instaló en el mundo. Es notorio que el aumento se registra de forma significativa a partir del segundo trimestre de 2020, cuando en la mayoría de los países del mundo los gobiernos comenzaron a implementar medidas de distanciamiento físico sostenido y cuarentenas o lockdowns.

Sin embargo, no existen datos que permitan establecer si entre las razones de este aumento también puede incluirse un cambio en los criterios de revisión de los contenidos y la ida hacia un modelo más agresivo de moderación de contenidos o a un aumento efectivo en el discurso de odio en redes sociales a partir de 2020.

Este trabajo busca analizar los aumentos en el discurso de odio en línea a partir de la llegada de la pandemia de COVID-19 al mundo y las acciones implementadas por Facebook, Twitter y YouTube, definir su alcance, efectos, motivos y posibles consecuencias.

# EL DISCURSO DE ODIO EN INTERNET

El discurso de odio es un término complejo y dispar en su definición, así como en el que no existe acuerdo tanto en las distintas plataformas como en gobiernos y sus regulaciones y leyes. En el **documento Estrategia y Plan de Acción de las Naciones Unidas para la Lucha contra el Discurso de Odio firmado por el secretario general de Naciones Unidas, António Guterres** se define al discurso de odio como "cualquier forma de comunicación de palabra, por escrito o a través del comportamiento, que sea un ataque o utilice lenguaje peyorativo o discriminatorio en relación con una persona o un grupo sobre la base de quiénes son o, en otras palabras, en razón de su religión, origen étnico, nacionalidad, raza, color, ascendencia, género u otro factor de identidad". Se añade que "en muchos casos, el discurso de odio tiene raíces en la intolerancia y el odio, o los genera y, en ciertos contextos, puede ser degradante y divisivo".

Desde el punto de vista legal, el derecho internacional no prohíbe el discurso de odio como tal sino "la incitación a la discriminación, la hostilidad o la violencia", la primera definida por Naciones Unidas como "una forma de expresión muy peligrosa, ya que tiene por objeto explícito y deliberado dar lugar a discriminación, hostilidad y violencia, que también podrían provocar o incluir actos de terrorismo o crímenes atroces". Debido a esto, se explica en el documento citado, el derecho internacional no exige que los Estados prohíban el discurso de odio que no alcanza el umbral de la incitación. Pero Naciones Unidas advierte sin embargo que "aún cuando no está prohibido el discurso de odio puede ser perjudicial".

"En todo el mundo, estamos presenciando una inquietante oleada de xenofobia, racismo e intolerancia, con un aumento del antisemitismo, el odio contra los musulmanes y la persecución de los cristianos. Se están explotando los medios sociales y otras formas de comunicación como plataformas para promover la intolerancia. Los movimientos neonazis y a favor de la supremacía blanca están avanzando, y el discurso público se está convirtiendo en un arma para cosechar ganancias políticas con una retórica incendiaria que estigmatiza y deshumaniza a las minorías, los migrantes, los refugiados, las mujeres y todos aquellos etiquetados como 'los otros'. Y no se trata de un fenómeno aislado, ni de las estridencias de cuatro individuos al margen de la sociedad. El odio se está generalizando, tanto en las democracias liberales como en los sistemas autoritarios y, con cada norma que se rompe, se debilitan los pilares de nuestra común humanidad. El discurso de odio constituye una amenaza para los valores democráticos, la estabilidad social y la paz", asegura Naciones Unidas en el documento.

Asimismo, en el **documento Recomendación General Nº 15 Relativa a la Lucha contra el Discurso de Odio y Memorándum Explicativo de la Comisión Europea contra el Racismo y la Intolerancia (ECRI) del Consejo de Europa**, define el discurso de

odio como “el uso de una o más formas de expresión específicas- por ejemplo, la defensa, promoción o instigación del odio, la humillación o el menosprecio de una persona o grupo de personas, así como el acoso, descrédito, difusión de estereotipos negativos o estigmatización o amenaza con respecto a dicha persona o grupo de personas y la justificación de esas manifestaciones basada en una lista no exhaustiva de características personales o estados que incluyen la raza, color, idioma, religión o creencias, nacionalidad u origen nacional o étnico al igual que la ascendencia, edad, discapacidad, sexo, género, identidad de género y orientación sexual”.

Según la definición, el discurso de odio “no solo tiene por objeto incitar a que se cometan actos de violencia, intimidación, hostilidad o discriminación, sino también actos que cabe esperar razonablemente que produzcan tal efecto” y “motivos que van más allá de la raza, color, idioma, religión o creencias, nacionalidad, origen étnico o nacional y ascendencia”. También se añade que el alcance del término “expresión” refiere “a los discursos orales y publicaciones en cualquiera de sus formas, incluyendo el uso de los medios electrónicos y su difusión y almacenamiento”. Así como al hecho de que el discurso de odio “puede tomar forma oral o escrita o cualquier otra forma como pinturas, señales, símbolos, dibujos, música, obras de teatro o videos” y “también abarca el uso de conductas específicas como gestos para comunicar una idea, mensaje u opinión”. La definición incluye “la negación, trivialización, justificación o condonación públicas de delitos de genocidio, delitos de lesa humanidad o delitos en caso de conflicto armado cuya comisión haya sido comprobada tras recaer sentencia los tribunales o el enaltecimiento de las personas condenadas por haberlos cometido”.

Varios países del mundo tienen legislación que prohíbe el discurso de odio y que en general se enfocan en la incitación al odio hacia personas basado en sus características identitarias.

En América Latina el enfoque ha tendido a ser muy centrado en lo legislativo y, como afirma Marianne Díaz Hernández en **su trabajo Discurso de Odio en América Latina: Tendencias de Regulación, Rol de los Intermediarios y Riesgos para la Libertad de Expresión** en la mayoría de los casos apostan a la sanción penal directa, la sanción penal accesoria (como agravante de un delito principal) y la prohibición, que sin crear sanciones de tipo criminal, establece medidas reparatorias. Díaz Hernández añade que en América Latina varios países (Costa Rica, El Salvador, Perú, Argentina, Bolivia y Uruguay, por mencionar algunos) “tienen tipificada la incitación al odio como delito en sus legislaciones penales generales”.

A su vez, entre aquellos países que han elegido el modelo sancionatorio, no todos caracterizan la incitación al odio bajo los mismos parámetros, siendo que algunos exigen la presencia real o potencial de un daño para configurar el delito. En ese sentido, la Comisión Interamericana de Derechos Humanos ha resaltado que “como principio, en vez de restringirlos, los Estados deben impulsar mecanismos preventivos y educativos y promover debates más amplios y profundos, como una medida para exponer y combatir los estereotipos negativos”.

Sin embargo, existe un cierto acuerdo acerca de que el discurso de odio podría tener un rol en generar las condiciones para la violencia hacia grupos específicos de la sociedad. El académico Alexander Tsesis **argumenta que la principal motivación del discurso de odio intimidatorio es perpetuar y aumentar las inequidades existentes**. "Aunque no siempre la circulación del discurso de odio intimidatorio deriva en la existencia de violencia discriminatoria, establece un racional para el ataque de grupos particularmente desfavorecidos", sostiene.

Los actos de violencia contra los Rohingya en Myanmar, por ejemplo, muestran el rol que posteos de Facebook con contenidos de discurso de odio han tenido en el proceso. En 2018, **una investigación de Reuters llevada a cabo en conjunto con el Human Rights Center de UC Berkeley School of Law** detectó la existencia de más de 1.000 posteos en los que se definía a los Rohingya u otros musulmanes como perros, gusanos y violadores.

**Este contenido fue creado y distribuido en el inicio de una campaña de limpieza étnica y crímenes contra la humanidad llevado adelante por las Fuerzas Armadas de Myanmar que derivó en que 740.000 personas de la etnia Rohingya huyeran a Bangladesh.**

# 2020: UN "TSUNAMI DE ODIO Y XENOFOBIA"

**En mayo de 2020, el secretario general de Naciones Unidas, Antonio Guterres, advirtió que la pandemia del COVID-19 desencadenó un "tsunami de odio y xenofobia" en el mundo "buscando chivos expiatorios y fomentando el miedo" y llamó a "actuar ahora para fortalecer la inmunidad de nuestras sociedades contra el virus del odio".**

"Se ha vilipendiado a los migrantes y refugiados como fuente del virus, y acto seguido se les ha denegado el acceso a tratamiento médico", dijo y añadió que se ha mostrado a las personas de la tercera edad como "caricaturas despreciables" que "sugieren que también son las más prescindibles", así como periodistas, profesionales de la salud, trabajadores humanitarios y defensores de los derechos humanos "están siendo atacados por el simple hecho de hacer su trabajo".

En ese sentido, la alta comisionada de Naciones Unidas para los Derechos Humanos, Michelle Bachelet, **planteó en la 13 Sesión del Foro sobre Cuestiones de Minorías en noviembre de 2020** que las redes sociales han significado nuevas "oportunidades para el ejercicio de libertades fundamentales como la expresión, asociación y participación, expandiéndolos hasta extensiones sin precedentes" sin embargo "esta expansión ha traído consigo nuevas y significativas amenazas al espacio cívico y los derechos de las personas".

"Uno de ellos es el discurso de odio, que está largamente extendido online a través de varias plataformas sociales. Las minorías han sido el objetivo de la incitación a la discriminación, la hostilidad y la violencia de forma desproporcionada. Esto puede llevar a tensiones, agitación y ataques contra individuos y grupos. También puede ser usado para servir a intereses políticos y contribuir a un clima de miedo entre comunidades minoritarias", sostuvo Bachelet.

La alta comisionada dijo que "los mismos derechos que las personas tienen offline deben ser protegidos también online" y añadió que las empresas propietarias de redes sociales "tienen la responsabilidad de prevenir, mitigar y remediar las violaciones a los derechos humanos que causan o contribuyen a ocurrir".

"Las empresas de redes sociales tienen la alternativa de retirar o dejar material en línea. También pueden marcar el contenido, agregar material compensatorio, advertir al difusor y sugerir que se modere. La eliminación sólo estaría justificada en los casos más graves. Cualquier solución propuesta para abordar el discurso de odio en las redes sociales debería trabajar para cerrar una enorme brecha en la transparencia y la responsabilidad democrática en la toma de decisiones de las plataformas. No solo debemos esperar que sigan las pautas

de derechos humanos, sino que también necesitamos mecanismos para monitorear y evaluar sus actos, aseguró Bachelet.

En esa misma línea, el relator Especial de las Naciones Unidas sobre la libertad de religión o de creencias, Ahmed Shaheed, denunció en abril de 2020 la propagación en plataformas sociales de una teoría de "conspiración" que afirma "que los judíos o Israel son los responsables de crear y propagar el virus de la COVID-19".

**"La lucha contra los discursos de odio en línea no tendrá éxito si los medios generales o redes sociales no se toman en serio los informes sobre odio cibernético dirigido contra los judíos y otras minorías (...) Los medios deben eliminar cualquier publicación que incite al odio o la violencia además de identificar y notificar las noticias falsas", señaló. Shaheed añadió que "en estos momentos extraordinariamente difíciles, es más necesario que nunca garantizar que todas las personas puedan ejercer, sin temor alguno y en la mayor medida posible, su derecho a la libertad de religión o creencias, a la vez que se protege la salud pública".**

Durante la pandemia se registró un aumento del discurso de odio en redes sociales. Primero, en febrero, el objetivo fue sobre todo la comunidad china porque fue en ese país donde nació el COVID-19. Luego, el odio se dirigió hacia el uso de mascarillas, hasta llegar incluso a culpar a las población LGBTIQ por ser el supuesto origen de un virus considerado castigo divino. Según **un estudio realizado por la empresa L1ght**, el discurso de odio hacia China o ciudadanos chinos creció 900% en Twitter y aumentó 200% el tráfico hacia sitios que propagan el discurso de odio o hacia posteos específicos contra la comunidad china o asiática.

En muchos casos esas expresiones también surgieron desde líderes políticos de diferentes partes del mundo, tanto en sus plataformas sociales (con millones de seguidores) como fuera de ellas. **El uso del término "virus chino"** en sus redes sociales por parte del entonces presidente de Estados Unidos, Donald Trump, y el uso de "virus de Wuhan" por parte del también entonces secretario de Estado Mike Pompeo pueden haber alentado **el uso del discurso de odio en EE.UU.**

En febrero de 2020, el gobernador de la región italiana de Véneto, Luca Zaia, uno de los primeros epicentros de la pandemia, dijo a periodistas que el país gestionaría mejor el virus que China debido a la **"higiene que tiene nuestro pueblo (...) los ciudadanos italianos, la formación cultural que tenemos, de ducharnos, lavarnos, lavarnos muy a menudo las manos (...), mientras que todos hemos visto los videos con chinos que comen ratas vivas".**

En abril de ese mismo año, el entonces ministro de educación de Brasil, **Abraham Weintraub sugirió** en un tuit que la pandemia era parte del "plan de dominación mundial" del gobierno chino.

Esa intensificación de la retórica racista en redes sociales y medios de comunicación, coincide con el aumento de ataques contra esos mismos grupos registrados en varias partes del mundo. **En el Reino Unido personas asiáticas han sido golpeadas** y se han convertido en **blanco de burlas** y de acusaciones de propagar el coronavirus. Dos mujeres atacaron a unas estudiantes chinas en Australia, las golpearon y dieron patadas a una de ellas y les gritaron **"vuelvan a China"** y **"malditas inmigrantes"**. En España, dos hombres golpearon a **un joven estadounidense de origen chino** hasta dejarlo en coma durante dos días. En el estado norteamericano de Texas, un hombre con un cuchillo **atacó a una familia birmana** acusándolos de ser factor de contagio del coronavirus.

En África se han reportado incidentes de discriminación y ataques contra personas asiáticas acusadas de ser portadoras de coronavirus, así como extranjeros en general, en **Kenia**, **Etiopía** y **Sudáfrica**.

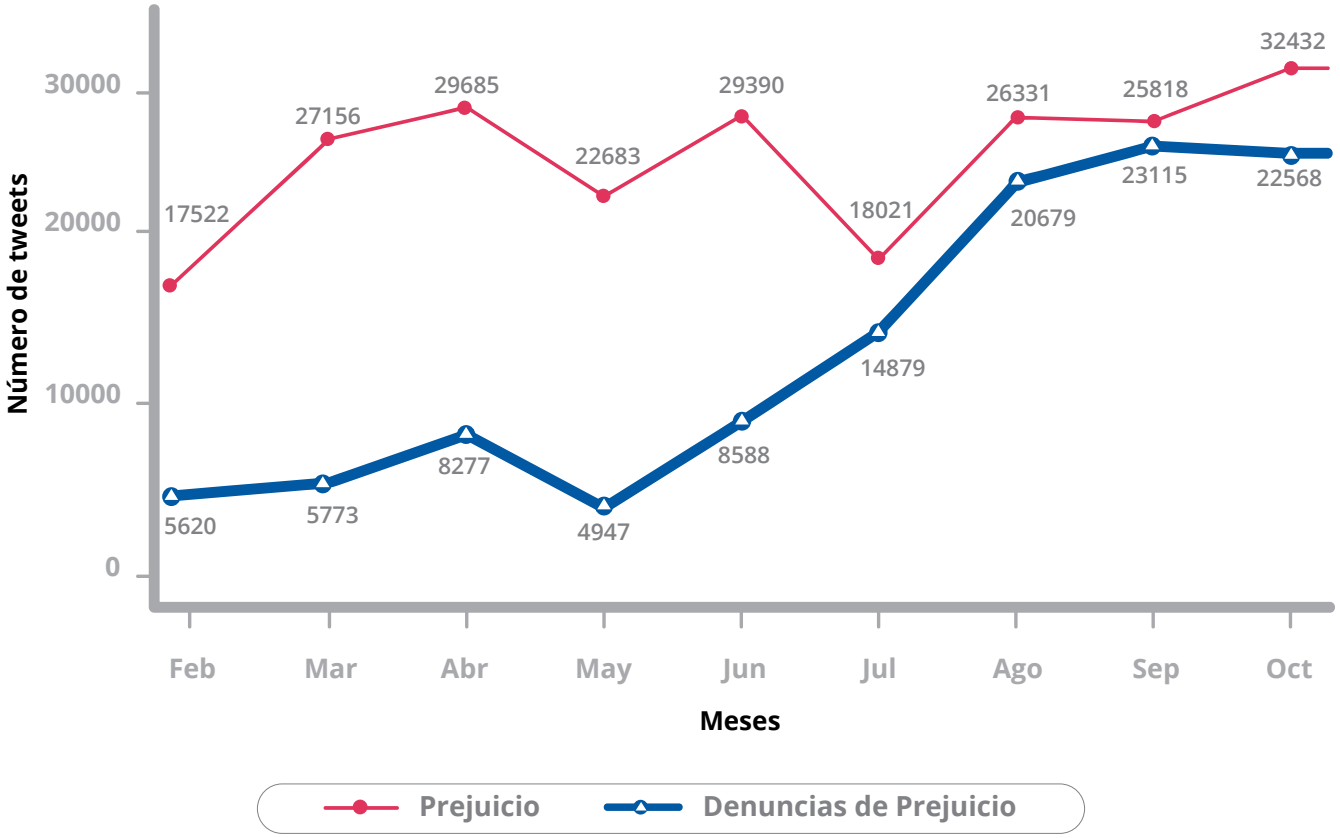
En América Latina también se han registrado casos. En Brasil, los medios de comunicación dieron cuenta de la ocurrencia de **casos de hostigamiento** y **rechazo**–a personas de origen asiático. En uno de esos episodios, una estudiante de Derecho denunció que fue víctima de racismo y xenofobia por una pasajera en el metro de Río de Janeiro. "Esta mujer esperó a que yo fuera a la puerta del vagón para gritar 'mira a la china que se va, cerda china', 'asquerosa' y 'se queda acá y nos enferma a todos'", publicó Marie Okabayashi en Twitter con un video de la agresora.

**La historiadora mexicana Yuriko Valdez**, de origen chino y autora del documental El legado de mi raza. Chinos y mestizos en Mexicali, advierte acerca de la proliferación de actitudes xenófobas por parte de la comunidad de esa localidad, así como de los numerosos comentarios racistas en redes sociales en las publicaciones que hablaban sobre festejos como el del Año Nuevo Chino el 25 de enero. A los habituales comentarios de "los chinos comen ratas y perros", se sumaron los de "chinos cochinos" o "nos van a contagiar porque en la China está el foco de infección del coronavirus". Reacciones en la misma línea, por parte de "personas orgullosas que presumen de ser verdaderamente de Mexicali" se presentaron cuando se promocionaba la inauguración de una exposición de la Asociación China en el Zoológico Bosque de la Ciudad: "Los chinos no se merecen un homenaje", "están enfermos por el coronavirus", entre otros mensajes que Valdez relata.

"Las expresiones de racismo y xenofobia relacionadas con COVID-19 en plataformas digitales han incluido acoso, discurso de odio, proliferación de estereotipos discriminatorios y teorías de conspiración. No es sorprendente que los líderes que intentan atribuir COVID-19 a ciertos grupos nacionales o étnicos sean los mismos líderes populistas nacionalistas que han convertido la retórica racista y xenófoba en el centro de sus plataformas políticas", dijo E. Tendayi Achiume, relatora Especial sobre **las Formas Contemporáneas de Racismo, Discriminación Racial, Xenofobia y Formas Conexas de Intolerancia.**

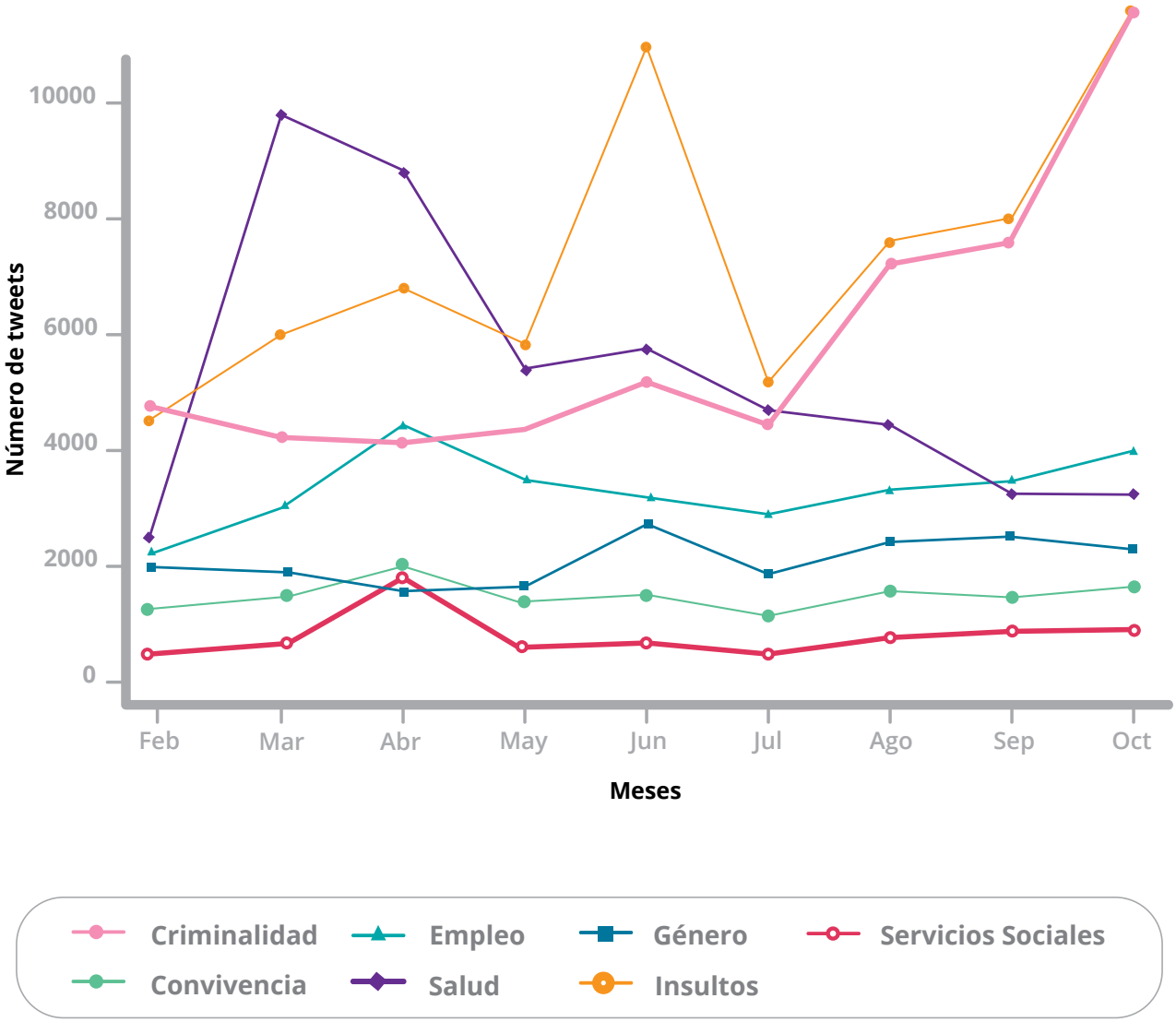
**La Unidad de Migración del Banco Interamericano de Desarrollo (BID)** llevó adelante un estudio entre febrero y diciembre de 2020 en el que dio seguimiento a conversaciones sobre los **inmigrantes en Twitter**. En esa investigación se monitorearon siete países de la región considerados importantes receptores de migrantes: Argentina, Chile, Colombia, Costa Rica, Ecuador, Panamá y Perú. En este monitoreo se recogieron tuits con términos como *asilo, xenofobia, migrante, inmigrante, refugiado, exiliado* y una vez realizada la recolección un algoritmo los clasifica en ocho categorías mutuamente excluyentes. Las primeras siete categorías engloban tuits que expresan prejuicios hacia los migrantes en las áreas de Criminalidad, Empleo, Género, Servicios Sociales, Convivencia, Salud e Insultos Generales. La octava categoría engloba tuits en los que se denuncian o se repudian estos prejuicios, según se explica en la investigación.

A partir de esto, y usando los tuits de febrero como línea base pre-pandemia, el estudio encontró un incremento de expresiones de prejuicio hacia los migrantes de 70% en dos meses, pasando de 17.522 tuits mensuales en febrero a 29.685 en abril.



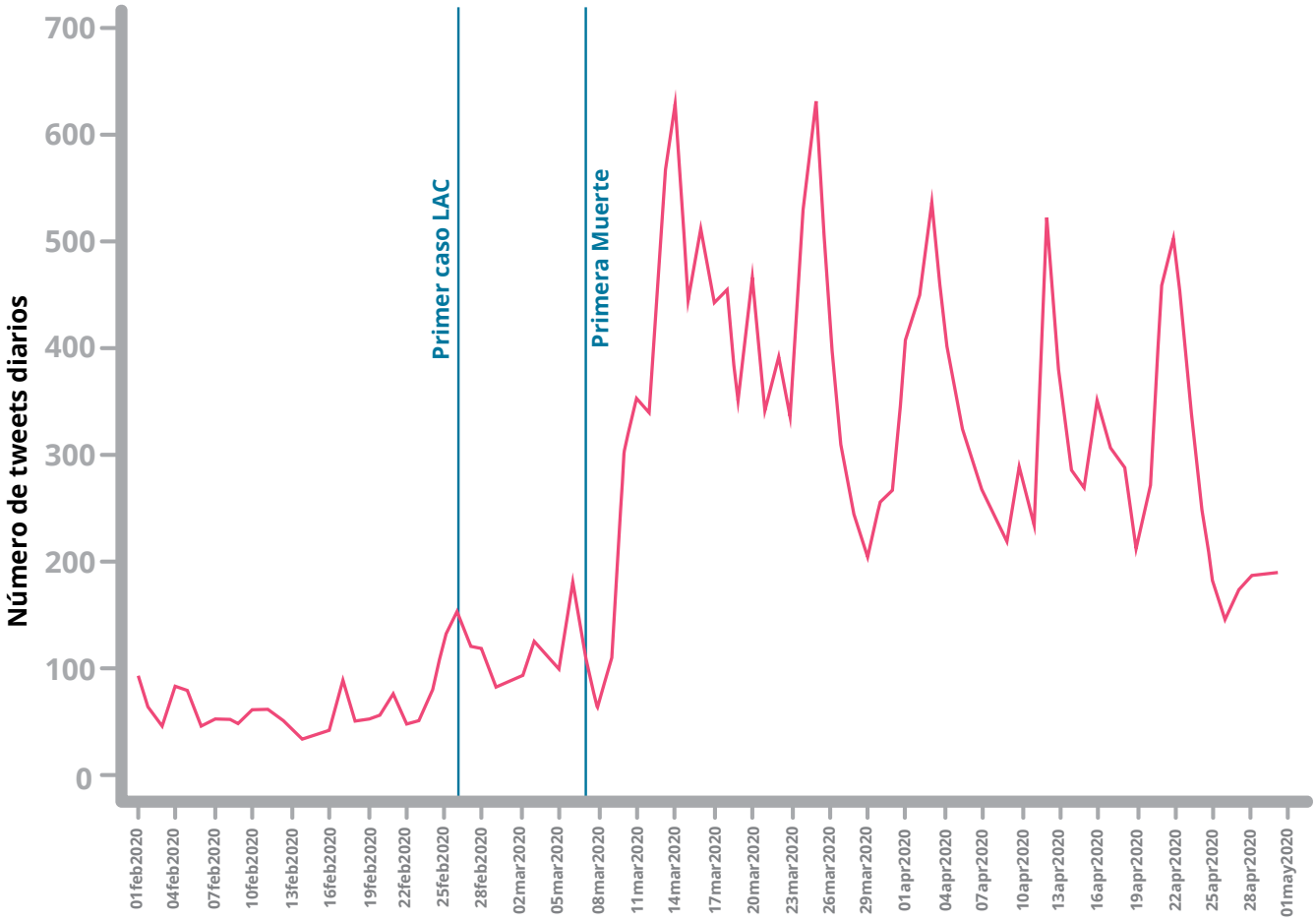

**Fuente: Unidad de Migración del BID en base a datos generados por Citibeats.**

Según el estudio, gran parte de este aumento registrado entre febrero y abril se explica por los prejuicios ligados a la salud, "principalmente explicados por el miedo a que los migrantes transmitan la enfermedad o colapsen los sistemas de salud".




**Fuente: Unidad de Migración del BID en base a datos generados por Citibeats.**

El estudio del BID asegura que "estos prejuicios se vieron disparados por la primera muerte de COVID-19 anunciada en la región", **ocurrida en marzo de 2020 en Argentina**.




**Fuente: Unidad de Migración del BID en base a datos generados por Citibeats.**

En los meses siguientes, los autores del estudio plantean que encontraron oscilaciones en los niveles de xenofobia o prejuicio pero siempre superiores a los de febrero pre-pandemia. En octubre, registran una subida, en este caso explicada por otros factores (como la criminalidad) que no guardan relación directa con la pandemia y una baja en los tuits referidos a razones sanitarias.

# LAS POLÍTICAS DE LAS PLATAFORMAS SOBRE DISCURSO DE ODIO DURANTE LA PANDEMIA

## “Creo firmemente que Facebook no debe ser el árbitro de la verdad de todo lo que la gente dice online”.

**Esa frase, dicha por Mark Zukerberg** de forma repetida durante años es un buen resumen de la actitud de las plataformas acerca de la moderación de contenidos hasta 2020. Incluso luego de las elecciones de 2016 en Estados Unidos, Facebook, Twitter y YouTube enfrentaron serias críticas por su rol en la diseminación de desinformación, odio y teorías conspirativas pero se mantuvieron muy resistentes a tomar acciones al respecto.

En 2020 eso cambió. Facebook, Twitter y YouTube hicieron cambios en sus normas comunitarias y términos de funcionamiento que durante años se habían resistido a hacer, desde etiquetar como falsa información de cuentas de personas públicas hasta eliminar posteos de un presidente en ejercicio de Estados Unidos y eliminar su cuenta.

En junio de 2020, la muerte del afroamericano George Floyd como resultado de su arresto a manos de cuatro policías de Mineápolis, generó una ola de protestas mundiales contra el racismo y la brutalidad policial. El en ese entonces presidente norteamericano Donald Trump realizó **una serie de posteos en sus plataformas sociales y en uno en particular escribió: “Cuando comience el saqueo, comienza el tiroteo”**. Esto fue interpretado por gran parte de la comunidad afroamericana como una amenaza a los manifestantes. Twitter resolvió ocultar el contenido. Facebook no.

En medio de las críticas, el CEO de Facebook, Mark Zuckerberg escribió un posteo en el que explicó sus razones para mantener el posteo de Trump publicado. “Estoy fuertemente en desacuerdo con lo que dijo el presidente acerca de esto, pero creo que la gente debe verlo por sí misma, porque en última instancia la responsabilidad de aquellos en posiciones de poder sólo puede cumplirse cuando su discurso es analizado de forma abierta”, escribió.

Semanas después, un grupo de empresas -entre las que figuran Unilever, Coca Cola, Verizon y Honda- anunció el inicio de la campaña *Stop Hate for Profit* y la suspensión por un mes de la compra de publicidad a la plataforma. El vicepresidente de medios de Unilever, Luis Di Como, dijo que seguir publicándose “en estas plataformas en este momento no daría valor añadido a la gente y

la sociedad". **"Dada la actual polarización y las elecciones que hay en Estados Unidos, tiene que haber mucho más cumplimiento en el área del discurso de odio", reclamó.**

"Respetamos profundamente la decisión de cualquier marca y seguimos enfocados en la importante labor de retirar el discurso de odio y entregar información crítica sobre la votación", fue la respuesta de Carolyn Everson, vicepresidenta del grupo de negocios globales de Facebook, el lunes. "Nuestras conversaciones con comercios y organizaciones de derechos civiles son sobre cómo podemos ser una fuerza para el bien juntos".

Sin embargo, en enero de 2021, Trump fue suspendido indefinidamente de Twitter, Facebook y algunos de sus videos eliminados de YouTube por difundir mensajes denunciando un supuesto fraude electoral en las últimas elecciones norteamericanas dirigidos a partidarios que irrumpieron en el Capitolio en Washington generando fuertes episodios de violencia y miedo entre legisladores y funcionarios, así como **la muerte de personas.**

"Los impactantes eventos de las últimas 24 horas demuestran claramente que el presidente Donald Trump tiene la intención de usar el tiempo que le queda en el cargo para socavar la transición pacífica y legal del poder a su sucesor electo, Joe Biden", escribió Zuckerberg en un posteo de Facebook en el que explicó la decisión del bloqueo.

# FACEBOOK Y LA REMOCIÓN DE CONTENIDO DE ODIO

En sus Normas Comunitarias, Facebook específicamente define el discurso de odio como "un ataque directo a las personas por lo que denominamos 'características protegidas': raza, etnia, nacionalidad, discapacidad, religión, clase, orientación sexual, sexo, identidad de género y enfermedad grave".

**"Definimos un ataque como lenguaje violento o deshumanizante, estereotipos dañinos, afirmaciones de inferioridad, expresiones de desprecio, repulsión o rechazo, insultos, o incitaciones de exclusión o segregación. Consideramos que la edad es una característica protegida cuando se menciona junto con otra característica protegida. También protegemos a los refugiados, migrantes, inmigrantes y solicitantes de asilo de ataques graves, aunque sí permitimos los comentarios y las críticas relacionadas con las políticas de inmigración. De manera similar, ofrecemos ciertas protecciones para características, como la profesión, cuando se mencionan junto con una característica protegida", explica Facebook.**

Y se añade: “Somos conscientes de que, a veces, las personas comparten contenido que incluye lenguaje que incita al odio emitido por otra persona con la intención de reprobarlo o concientizar a los demás. En otros casos, el lenguaje que, de otra manera, infringiría nuestras normas se puede usar de forma autorreferencial o motivadora . Nuestras políticas están diseñadas para dar espacio a estos tipos de lenguaje, pero exigimos que la intención quede clara. Si no es el caso, el contenido podría eliminarse”.

La compañía categoriza el discurso de odio en tres niveles según considera la gravedad de lo publicado en la red social. El Nivel 1 es todo aquel “contenido dirigido a una persona o un grupo de personas con “características protegidas” que incluya “lenguaje que incite a la violencia o que la apoye, tanto de forma escrita como visual” y “lenguaje o imágenes deshumanizantes en forma de comparaciones, generalizaciones o afirmaciones basadas en comportamientos no aptas (de forma escrita o visual) en relación con: insectos, animales que culturalmente se perciben como inferiores desde el punto de vista intelectual o físico, suciedad, bacterias, enfermedades y heces, depredadores sexuales, infrahumanidad, criminales sexuales y violentos, otros criminales (incluidos, por ejemplo, ‘ladrones’, ‘bandidos’), afirmaciones que niegan la existencia, burlas acerca del concepto de los crímenes de odio, de hechos de este tipo o de sus víctimas, incluso si no aparece una persona real en una imagen”.

También se consideran discursos de odio Nivel 1 “determinadas comparaciones, generalizaciones o afirmaciones basadas en comportamientos deshumanizantes (tanto de forma escrita como visual) que incluyen personas negras y simios o criaturas similares a simios, personas negras y maquinaria agrícola, caricaturas de personas negras con la cara pintada de negro, personas judías y ratas, personas judías controlando el mundo o instituciones importantes como cadenas de medios de comunicación, la economía o el gobierno, negación o distorsión de la información sobre el Holocausto, personas musulmanas y cerdos, personas musulmanas y relaciones sexuales con cabras o cerdos, personas mexicanas y criaturas similares a gusanos, mujeres como objetos domésticos o referirse a las mujeres como propiedades u ‘objetos’”, así como “referirse a personas transgénero o de género no binario como si no fueran seres humanos o a dalits o personas de castas registradas o ‘bajas’ como sirvientes”.

Los discursos de odio Nivel 2 para Facebook son aquellos que se refieren a grupos protegidos e incluyen “generalizaciones que denotan inferioridad (tanto de forma escrita como visual)” como “las deficiencias físicas” relacionadas con “higiene, incluidos, entre otros: ‘mugriento’, ‘sucio’”, de “apariencia física” como “‘feo’, ‘horrible’, a las “deficiencias mentales” como “tonto”, “estúpido”, “idiota”, referidas a la educación como “analfabeto”, “inculto”, a la salud mental como “enfermo mental”, “retrasado”, “loco”, “demente” y a las “deficiencias morales” relacionadas con “rasgos de la personalidad que se consideran negativos culturalmente, incluidos, entre otros: ‘cobarde’, ‘mentiroso’, ‘arrogante’, ‘ignorante’” y “términos despectivos relacionados con la actividad sexual” como “puta”, “zorra”,

"pervertido". También se incluyen en el Nivel 2 "expresiones que denotan insuficiencia" como "inútil", "inservible", "expresiones de superioridad o inferioridad respecto de otra característica protegida, "expresiones relacionadas con salirse de la norma como "anormal" y "expresiones de desprecio" como el "reconocimiento de intolerancia respecto de características protegidas como "homofóbico", "islamofóbico", "racista", así como "expresiones que indican que una característica protegida no debería existir" y "expresiones de odio", "rechazo" y "repulsión" como "odio" o "no respeto", "no me gusta", "no me importa", "vomitivo", "repugnante", "desagradable", etcétera. También se incluyen en esta categoría los insultos "relacionados con los genitales o el ano para hacer referencia a una persona", "frases o términos ofensivos con la intención de insultar", "términos o frases que incitan a participar en actividades sexuales o que hacen referencia al contacto con los genitales o el ano, o con heces u orina".

Por último, Facebook categoriza como discurso de odio Nivel 3 el contenido en imagen o texto que refiera a "segregación en forma de incitaciones, declaraciones de intención, defensa o apoyo, o declaraciones de aspiraciones o condiciones en relación con la segregación", "exclusión en forma de incitaciones, declaraciones de intención, defensa o apoyo, o declaraciones de aspiraciones o condiciones en las que se incluyen exclusión explícita, es decir, actos como expulsar a determinados grupos o indicarles que no tienen permiso, exclusión política, es decir, negar el derecho a la participación política, exclusión económica, es decir, negar el acceso a prestaciones económicas y limitar la participación en el mercado laboral, exclusión social, es decir, actos como negar el acceso a determinados espacios (físicos y online) y servicios sociales" y "contenido que describe o señala negativamente a personas mediante estigmas, donde estigmas se definen como palabras inherentemente ofensivas usadas como etiquetas insultantes".

En julio de 2020 y como consecuencia de **la campaña Stop Hate for Profit** en la que más de 1.200 compañías de todo el mundo se unieron en un boicot de publicidad contra las principales plataformas reclamando aumentar la moderación del discurso de odio así como la suspensión de la publicidad desde cuentas que promovían la discriminación hacia colectivos específicos. Uno de los principales reclamos de las organizaciones y compañías involucradas era la eliminación de todas las cuentas de Trump.

Entre las demandas de la coalición, se reclama que las plataformas eliminen a "grupos o páginas enfocados en la supremacía blanca, milicias, antisemitismo, islamofobia y conspiraciones violentas", se "aumenten los recursos destinados a monitorear grupos de discurso de odio y violencia", se "cambie la política de las plataformas para prohibir cualquier pagina de evento que llame a las armas", así como "comprometer 5% de sus ingresos anuales a la financiación de un fondo independiente que sostenga iniciativas, académicas y de organizaciones, que luchen contra el racismo, el odio y la división causada por la inacción de Facebook".

A partir de estos reclamos, Facebook publicó en junio de 2020 **un posteo en el que respondió a algunos de los pedidos de Stop Hate for Profit**. Respecto al pedido de la organización de "crear una moderación separada compuesta por expertos en odio basado en la identidad para los usuarios que expresen que han sido atacados", Facebook aseguró que "los informes de discurso de odio en Facebook ya se canalizan automáticamente a un conjunto de revisores con capacitación específica en nuestras políticas de odio basadas en la identidad en 50 mercados que cubren 30 idiomas" y además se realizan "consultas con expertos en odio basado en la identidad para desarrollar y desarrollar las políticas que estos revisores capacitados aplican". También anunciaron su "intención de incluir la prevalencia del discurso de odio en los futuros Informes de Cumplimiento de Normas Comunitarias (CSER), a la espera de que no haya más complicaciones de COVID-19".

Ese mismo mes, el vicepresidente de Políticas Públicas de Facebook, Richard Allan, escribió una columna en la que abordaba las diferencias en la definición de discurso de odio en los distintos lugares del mundo y las dificultades de la plataforma para **detectarlo de forma adecuada y tomar medidas al respecto**. "No existe una respuesta universalmente aceptada para cuando alguien cruza la línea. Aunque algunos países tienen leyes contra el discurso de odio, sus definiciones varían significativamente. En Alemania, por ejemplo, las leyes prohíben la incitación al odio, podrías ser objeto de una redada policial por publicar ese tipo de contenido online. En Estados Unidos, por otra parte, aún los más viles tipos de discurso están legalmente protegidos por la Constitución norteamericana", escribe Allen.

"Gente que vive en el mismo país -o en la puerta de al lado- a menudo tienen distintos niveles de tolerancia al discurso. Para algunos, el humor crudo sobre un líder religioso puede ser considerado blasfemo y discurso de odio contra todos los seguidores de esa fe. Para otros, una batalla basada en insultos de género puede ser una manera mutuamente disfrutable de compartir una risa. ¿Es OK para una persona postear cosas negativas sobre gente de una determinada nacionalidad mientras comparta esa nacionalidad? ¿Qué pasaría si un joven se refiere a un grupo étnico determinado usando insultos raciales está citando la letra de una canción?", se pregunta el ejecutivo de Facebook.

Allen también se refiere en el texto a los errores en la eliminación de contenido que se categorizó erróneamente como discurso de odio. "Si fracasamos en remover contenido que ustedes reportan como discurso de odio, sentimos que no estamos viviendo según los valores de nuestras Normas Comunitarias. Cuando removemos algo que ustedes postean y creen que es un punto de vista razonable, puede sentirse como una censura. Sabemos que tan mal se siente la gente cuando cometemos esos errores, por lo que estamos trabajando constantemente para mejorar nuestros procesos y explicar las cosas más en profundidad", señala el ejecutivo de Facebook.

Añade que los errores de Facebook en la moderación de contenido "han causado una gran preocupación en algunas comunidades, incluidos aquellos grupos que sienten que

actuamos -o dejamos de actuar- debido a sesgos". "El año pasado (2019), Shaun King, un prominente activista afroamericano posteó un mail con discurso de odio que había recibido e incluía insultos racistas. Eliminamos el posteo de King por error al no reconocer inicialmente que estaba siendo compartido para condenar el ataque", relató. En julio, Nick Clegg, vicepresidente de Asuntos Globales y Comunicaciones de Facebook, escribió **un artículo en el que aseguró que la compañía había tomado varias medidas a partir de las que había registrado significativos avances en la eliminación de discurso de odio en su plataforma**. "Un informe reciente de la Comisión Europea encontró que Facebook evaluó el 95,7% de los informes de discursos de odio en menos de 24 horas, más rápido que YouTube y Twitter", escribió Clegg. "El mes pasado, informamos que encontramos casi el 90% del discurso de odio que eliminamos antes de que alguien lo informe, en comparación con el 24% de hace poco más de dos años. Tomamos medidas contra 9,6 millones de piezas de contenido en el primer trimestre de 2020, frente a 5,7 millones en el trimestre anterior. Y el 99% del contenido de ISIS y Al Qaeda que eliminamos se retira antes de que alguien nos lo informe", aseguró.

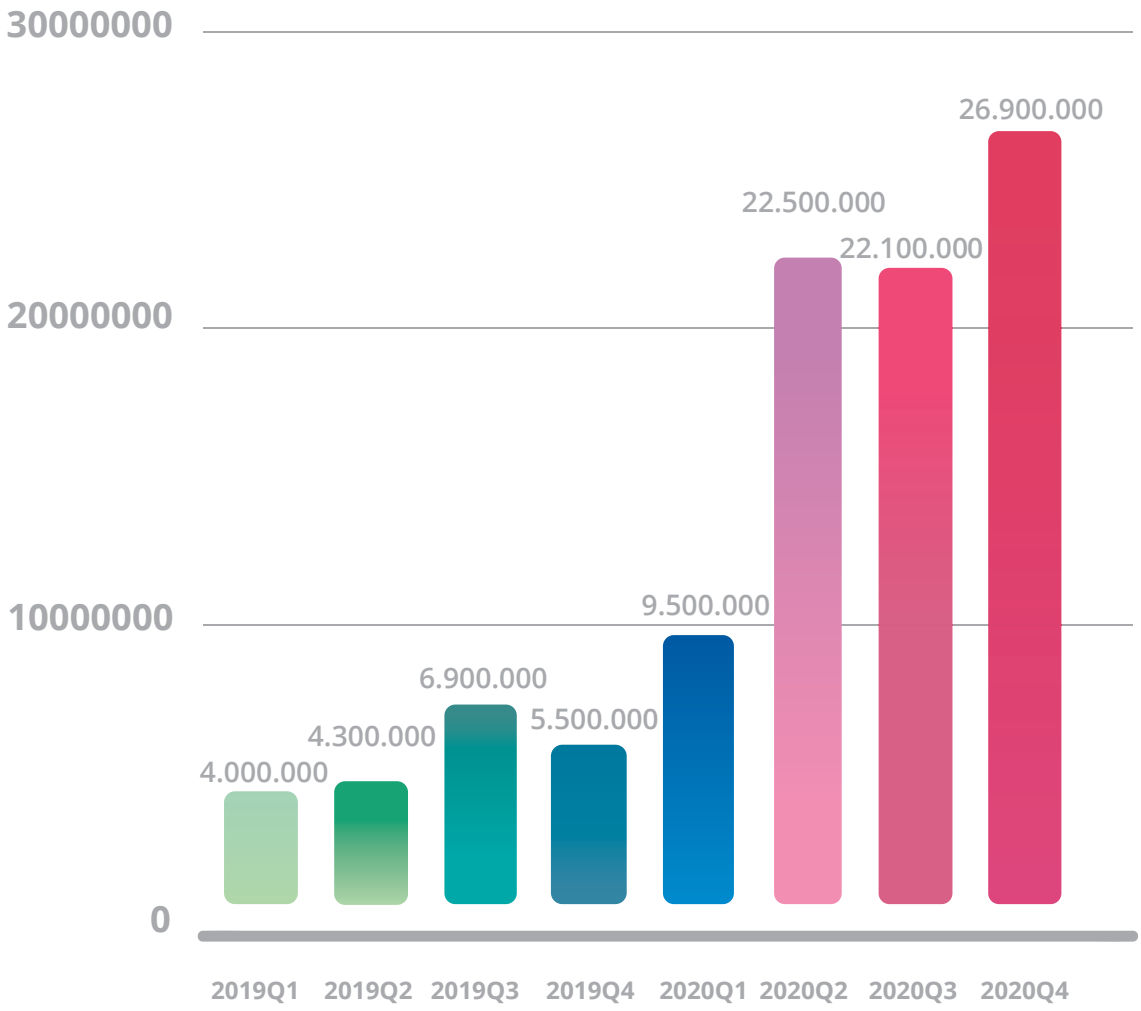
**Según el Community Standards Enforcement Report (CSER) publicado en febrero de 2021**, la cantidad de piezas de contenido sobre las que Facebook tomó acciones pasó de 20.700.000 en 2019 a 81.000.000, lo que significa un aumento de casi 300% en la cantidad de contenido categorizado como de discurso de odio entre un año y el siguiente.

En noviembre Facebook comenzó a medir la prevalencia de discurso de odio en la red social y detectó que entre julio y setiembre ese número se ubicó entre 0,10% y 0,11%. Esto significa que de cada 10.000 visualizaciones de publicaciones realizadas en la red social, entre 10 y 11 serían categorizadas como discurso de odio según Facebook. Entre octubre y diciembre de 2020 esa cifra descendió a 0,07% a 0,08% de prevalencia. En su informe, Facebook no aclara si este cambio se debe a un aumento de las publicaciones generales, un descenso efectivo de la categoría discurso de odio con relación al total o de un cambio en los criterios o procesos de detección.

Si se analiza 2020 en profundidad, puede detectarse un aumento muy significativo en la cantidad de piezas de contenido con discurso de odio sobre las que Facebook tomó acciones a partir del segundo cuarto de 2020. Entre enero y marzo, se accionó sobre 9.500.000, mientras que en los siguientes meses la cifra se duplicó pasando a 22.500.000 entre abril y junio, 22.100.000 entre julio y setiembre, y 26.900.000 entre octubre y diciembre.

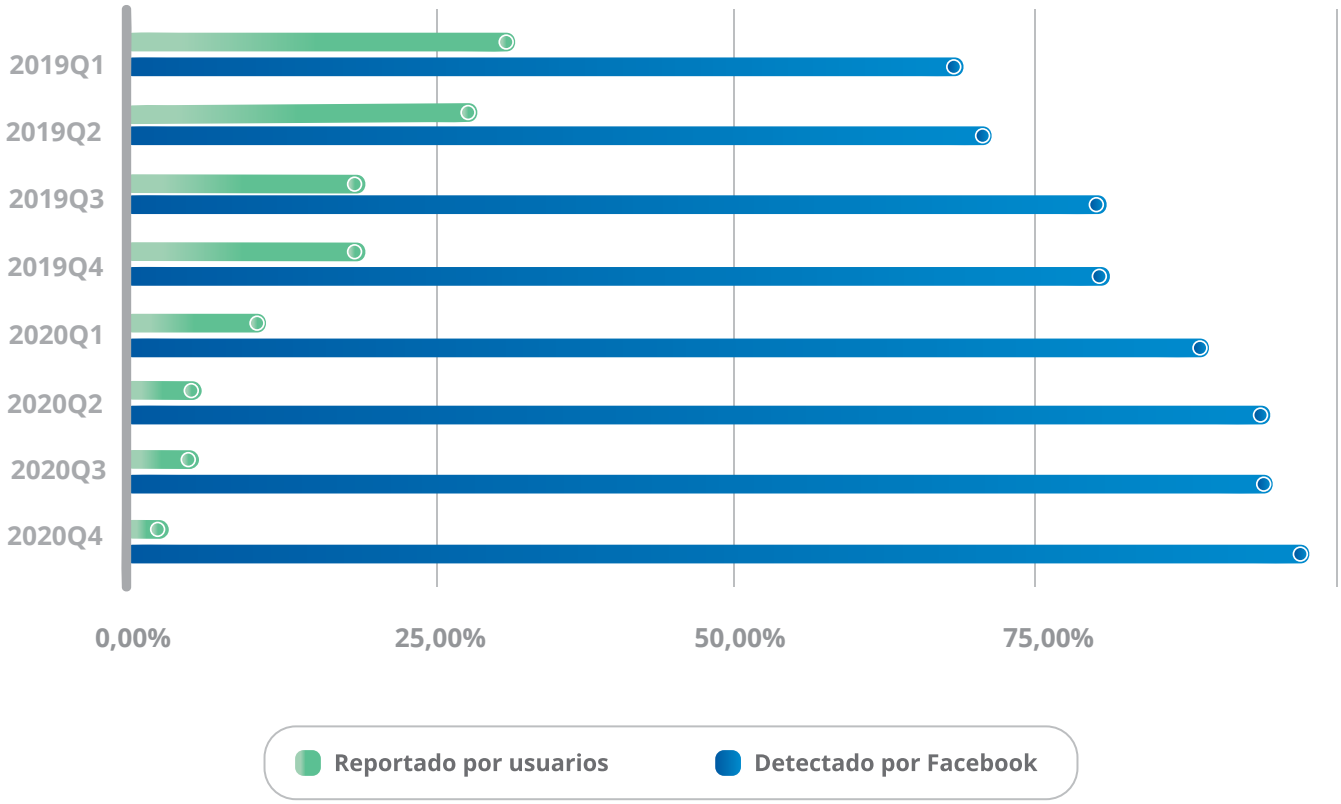
Según el informe del CSER, el crecimiento en el número de piezas detectadas así como en el porcentaje de proactividad de detección se debe "primariamente al mejoramiento de los sistemas tecnológicos de detección en los idiomas árabe y español" así como "a la expansión de la automatización en portugués".

## Contenido sobre el que se tomaron acciones por discurso de odio




Otro aspecto llamativo es el aumento del porcentaje detectado por Facebook en relación con aquel denunciado por usuarios sobre el total de los contenidos accionados debido a la categoría discurso de odio. Con pequeñas fluctuaciones, el peso de los sistemas internos de Facebook sobre el total de contenido con discurso de odio viene elevándose de modo sostenido desde 2018, llegando a ser prácticamente el total en el último cuarto de 2020. En el último cuarto de 2017, Facebook accionó sobre 1.700.000 por discurso de odio de las que 76,4% fueron detectadas a partir de denuncias de usuarios. En 2020 esa relación se invirtió. Entre enero y marzo 89,3% de los contenidos categorizados como discurso de odio provinieron de los sistemas de detección de Facebook y algo similar ocurrió entre abril y junio (94,7%), julio y setiembre (94,7%), y octubre y diciembre (97,1%).

**Del contenido "Discurso de Odio" cuanto fue denunciado**




Algo similar ocurrió con Instagram, también propiedad de Facebook, donde la acción sobre el contenido de odio se mide desde el último cuarto de 2019. Entre enero y marzo de 2020, Instagram detectó y tomó acciones sobre 578.000 piezas de contenido por entender que incurrían en su definición de discurso de odio, pasando a 3.200.000 entre abril y junio, 6.500.000 entre julio y setiembre, y 6.600.000 entre octubre y diciembre de 2020.

En el primer cuarto del año 57,1% del contenido fue detectado a partir de denuncias de usuarios, mientras que en el siguiente cuarto la relación cambió radicalmente y denuncias de los usuarios pasaron a ser solamente 15,1% del total de acciones tomadas sobre la categoría discurso de odio. Esta relación se mantuvo en los trimestres siguientes: 5,2% entre julio y setiembre, 4,9% entre octubre y diciembre.

A mediados de marzo de 2020, y luego de continuos pedidos de esos equipos a raíz de las medidas de aislamiento por la pandemia de COVID-19, Facebook decidió enviar a sus más de 15.000 moderadores de contenido en 20 lugares diferentes, a trabajar a sus casas.

El CEO de Facebook, Mark Zuckerberg dijo esa semana que Facebook se vería forzado durante la pandemia que azota a

la mayor parte del mundo a "apoyarse más activamente en software de inteligencia artificial para tomar las decisiones de moderación de contenido". La compañía también aseguró que iba a realizar entrenamientos full time para que prestaran "atención extra" al contenido "altamente sensible". Advirtió que los usuarios "deberían esperar más cantidad de errores mientras Facebook mejoraba el proceso, en parte porque solamente una fracción de los humanos seguirían siendo parte y debido a que el software toma decisiones más ingenuas que los humanos lo que podía generar "falsos positivos", incluyendo la remoción de contenido que no debería haber sido removido. "Creará un trade-off contra algunos tipos de contenido que no tienen riesgos físicos inminentes para las personas", **sostuvo Zuckerberg.**

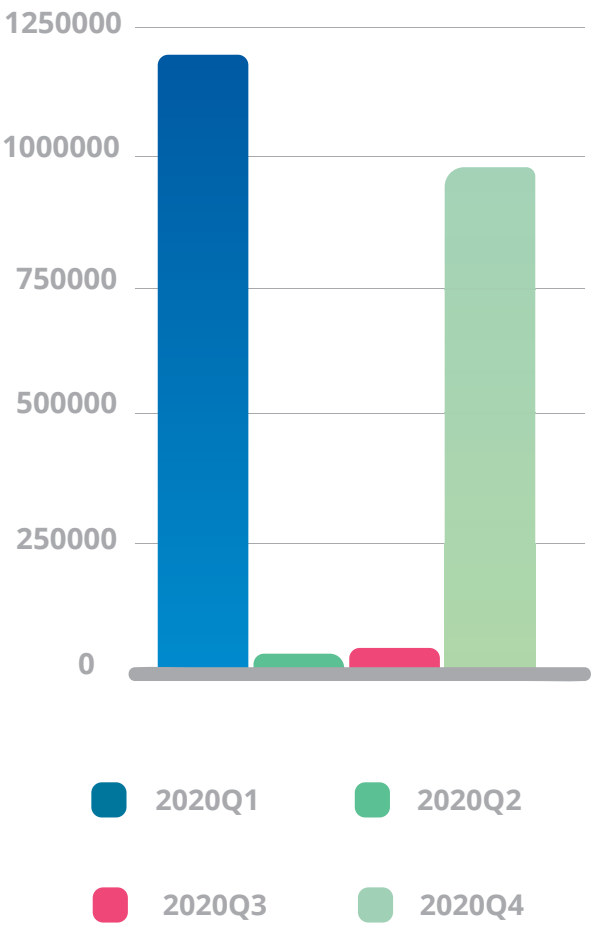
En noviembre de 2020, Facebook **anunció cambios en sus sistemas de moderación que implican** un aumento de la presencia de moderación automatizada en las primeras etapas de contacto con el contenido. Chris Palow, ingeniero y parte del equipo de Integridad de Facebook, admitió durante la conferencia de prensa que "la inteligencia artificial nunca será perfecta" y "tiene sus límites" para separar discurso de odio de aquel que no lo es, por ejemplo, por ser paródico o humorístico. "El sistema busca unir inteligencia artificial con revisión humana para hacer menor la cantidad de errores", explicó. Facebook no hace públicas las cifras porcentuales de contenidos que son categorizados erróneamente como contenido que debe ser eliminado.

Meses después, en febrero de 2021, el jefe de Políticas de Contenido Orgánico de Facebook, Varun Reddy, dijo que la plataforma estaba experimentando problemas por la ausencia de moderadores humanos en el proceso de moderación de gran parte de sus contenidos. La inteligencia artificial aprende de los moderadores humanos, explicó, esa reducción en la presencia humana ha cambiado **"cuán efectiva es la inteligencia artificial a lo largo del tiempo".**

"Estamos trabajando con proveedores para tener la mayor capacidad de vuelta online que nos sea posible (...) No estamos todavía en el punto de inicio pero desde que se inició el lock down el 25 de marzo. En los meses que vendrán esperamos que los sistemas vuelvan a su eficacia completa", dijo Reddy en febrero de este año.

Otro aspecto afectado por el aislamiento de los empleados de Facebook fue el proceso de apelación frente a contenidos que los usuarios entienden que fueron eliminados injustamente. "Debido a una reducción temporal de nuestra capacidad de revisión como resultado del COVID-19, no siempre podemos ofrecer a los usuarios la opción de apelar. Aún dimos a las personas la opción de decirnos que están en desacuerdo con nuestra decisión, lo que ha ayudado en la revisión en muchos de estos casos y a restaurar el contenido en los casos en los que fue apropiado", afirma Facebook en su **informe CSER**. Allí puede verse que entre abril y junio de 2020 las apelaciones casi no existieron, alcanzando solamente 70.000 en todo el mundo durante esos 6 meses, cuando en el trimestre anterior habían llegado a 1.200.000. En el período siguiente, entre octubre y diciembre, las apelaciones alcanzaron 984.200 casos.

**Apelaciones sobre el contenido accionado**




En 2020, Facebook también alcanzó cifras récord de contenido restaurado en relación con períodos anteriores, pasando de 483.400 contenidos en 2019 a 703.200 en 2020. De estos últimos, Facebook restableció 589.300 sin apelación alguna.

# LA MODERACIÓN DEL DISCURSO DE ODIO EN TWITTER

En diciembre de 2020, Twitter anunció una actualización de sus reglas para combatir la expansión del discurso de odio en su plataforma y apoyó su decisión en "investigaciones que vinculan el lenguaje deshumanizador con la violencia fuera de internet". En 2019, Twitter actualizó sus reglas respecto del discurso de odio para incluir la religión y la casta como grupos protegidos, en marzo de 2020 añadieron edad, discapacidad y enfermedad, y en diciembre de 2020, **anunciaron la prohibición de lenguaje de deshumanizara a las personas en razón de su raza, etnia o nacionalidad.**

En la publicación incluyeron una serie de ejemplos para ilustrar aquellos discursos que no serían permitidos a partir del anuncio:

**"Todos los (nacionalidad) son cucarachas quevivendelosbeneficiosdelEstadoydeben ser expulsados", "Las personas que son (raza) son sanguijuelas y solo buenas para una cosa", "Hay demasiados (nacionalidad, raza, etnia) gusanos en nuestro país y deben irse", "Todos los (grupo etario) son sanguijuelas y no merecen nuestro apoyo", "Las personas con (enfermedad) son ratas que contaminan todo a su alrededor", "(Grupo religioso) deben ser castigados. No estamos haciendo lo suficiente para deshacernos de esos apestosos animales".**

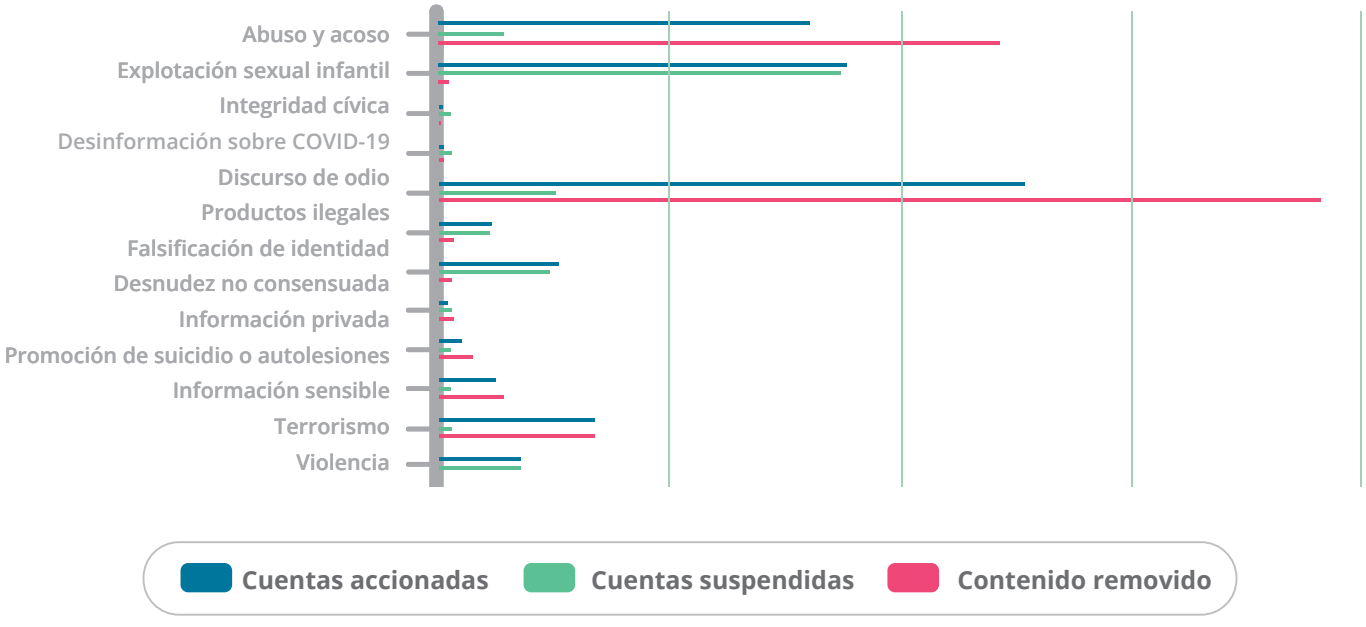
En octubre de 2019, la ahora vicepresidenta norteamericana, Kamala Harris, publicó **una carta abierta** al CEO de Twitter, Jack Dorsey en la que hacía un llamamiento a moderar algunas de las publicaciones del entonces presidente Donald Trump porque a su juicio violaban las normas comunitarias de la red social, entre ellas las referidas al discurso de odio. "Ningún usuario, sin importar su trabajo, riqueza o estatus debería estar exento de cumplir las reglas de uso de Twitter", sostenía Harris en la carta.

Ese mismo año, **un estudio de la Universidad de New York (NYU)** demostró la existencia de correlación entre el número de tuits racistas y la cantidad de crímenes de odio racistas ocurridos en 100 ciudades a través de Estados Unidos. "Pienso que existe un sentimiento en los tuits encontrados que está relacionado con el favorecimiento de un ambiente que favorece estos crímenes", sostiene Rumi Chunara, uno de los autores del estudio. Añade que en sentido contrario, **"tener conversaciones productivas mejora el medioambiente y los resultados".**

**"Actualmente el sistema hace super fácil acosar y abusar de otros", dijo Dorsey en 2019**, y añadió que "uno de los problemas es el tamaño del peso que adjudica a los seguidores y los me gusta".

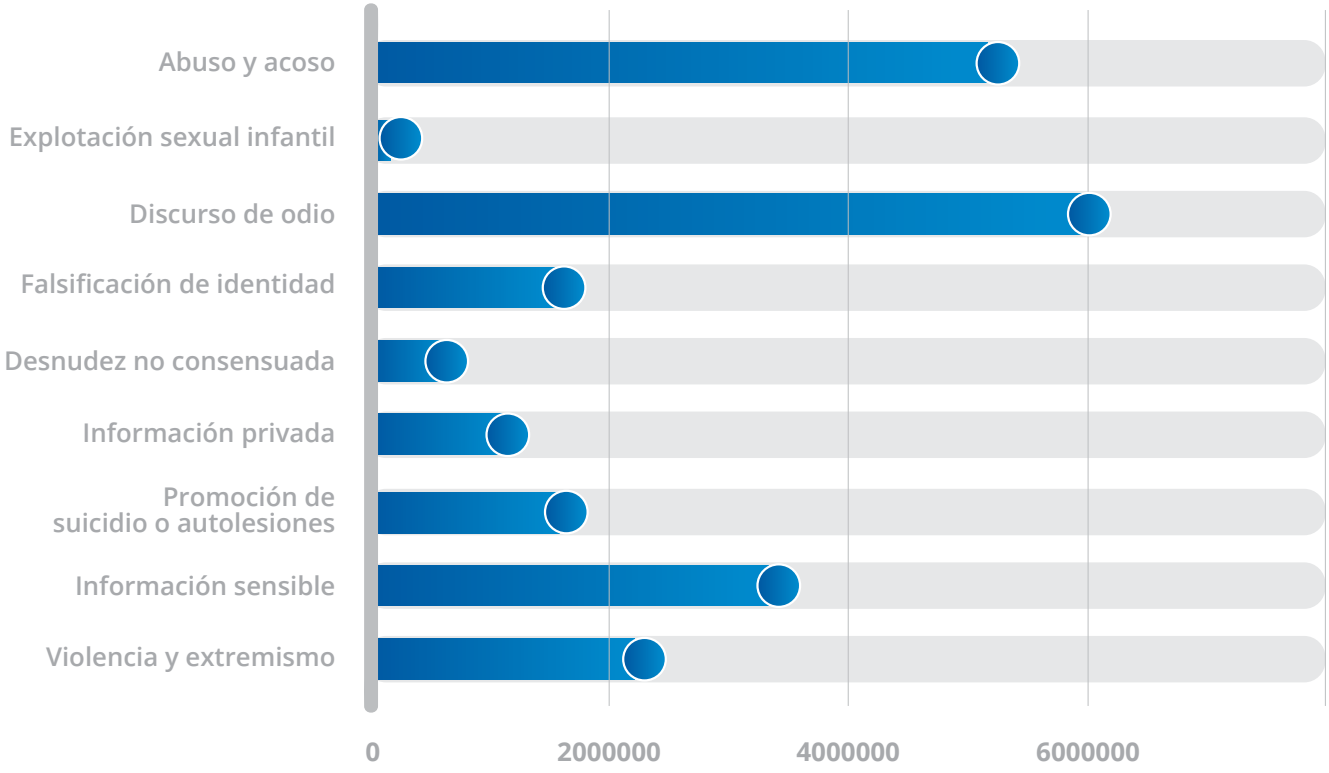
¿Qué pasó en 2020 y durante pandemia del COVID-19? Según **el último informe disponible** de Twitter Transparency Report, entre enero y junio de ese año se tomaron acciones sobre 1.940.082 cuentas, de las que 925.954 fueron suspendidas y 1.927.063 contenidos fueron removidos. Una cifra muy similar de contenido fue removido en el mismo período de 2019 (1.914.471) pero inferior en el caso de las cuentas suspendidas (687.397).

**Cuentas accionadas Enero - Junio por motivos**



Del total, se tomaron acciones sobre 645.416 cuentas a partir de contenidos etiquetados como discurso de odio, lo que representa 33,2% de las cuentas sobre las que se tomaron acciones. Se reportaron 12.400.000 cuentas en el período enero-junio, de las que cerca de la mitad (6.055.642) por razones de discurso de odio. Según el informe se reportó 30% más cuentas que en el mismo período del año anterior.

**Cuentas reportadas Enero - Junio**




Twitter reporta una reducción de 35% de las cuentas accionadas por razones de discurso de odio respecto del período anterior, aunque reconoce que dadas las circunstancias los equipos se concentraron en revisar contenidos que podían ocasionar daño o que se relacionaban con información errónea sobre COVID-19 y registraron "importantes atrasos en todas las restantes áreas".

**En abril de 2020, Twitter publicó un posteo en su blog informando sobre algunos cambios consecuencia de la decisión de enviar a gran parte de sus empleados a sus casas para observar las medidas de distanciamiento social impulsadas por gobiernos de todo el mundo.**

Parte de esas medidas fue el "aumento en el uso de machine learning y automatización para tomar un rango amplio de acciones en contenido potencialmente abusivo y manipulador". "Queremos ser claros: mientras trabajamos para que los sistemas sean consistentes, puede haber ocasiones en las que la falta de contexto produzca que cometamos errores. Como resultado, no suspenderemos de forma permanente basados solo en sistemas de moderación automatizada. En su lugar, continuaremos buscando oportunidades en las que los chequeos de moderación humana tengan el mayor impacto", se afirmaba en el texto.

Allí Twitter informó que durante la pandemia de COVID-19, la tecnología automatizada sería usada para "llamar la atención sobre contenido con más chances de causar daño y ese será revisado primero" y para "identificar proactivamente violaciones a las reglas antes de que sean reportados (los equipos aprenden en base a las decisiones pasadas, por lo que con el tiempo, la tecnología podrá ayudar a rankear el contenido o revisar cuentas de forma automática)". Para el contenido que "requiera contexto adicional, como la información engañosa sobre el COVID-19", Twitter asegura que sus equipos "continuarán revisando manualmente los reportes".

La red social aclara que los tiempos de respuesta ante los reportes se extenderán "más allá de los tiempos normales" y admite que "debido a que los sistemas automatizados no tienen todo el contexto ni el insight de los equipos humanos, se cometerán errores".

# YOUTUBE Y LA MODERACIÓN DEL DISCURSO DE ODIO EN PANDEMIA

**En YouTube la última actualización de las Normas Comunitarias** respecto del discurso de odio data de 2019. En la actualidad, la definición de lo la empresa propiedad de Google entiende como discurso de odio supone "contenido que promueve la violencia y el odio contra individuos o grupos basado en alguno de los siguientes atributos: edad, casta, discapacidad, etnia, identidad de género, nacionalidad, raza, estatus migratorio, religión, sexo o género, orientación sexual, víctimas de un evento violento o sus familias y veteranos".

En las Normas Comunitarias se añade que en YouTube no se permite que se "deshumanice a individuos o grupos con estas características, que afirme que son física o mentalmente inferiores, o que alabe o enaltezca la violencia contra ellos" ni "tampoco permitimos el uso de estereotipos que inciten o promuevan el odio basado en estas características, ni los insultos raciales, étnicos, religiosos o de otro tipo cuyo objetivo principal sea promover el odio", que "alegue la superioridad de un grupo sobre los que tienen cualquiera de las características mencionadas anteriormente para justificar la violencia, la discriminación, la segregación o la exclusión" o "niegue que han ocurrido sucesos violentos bien documentados".

En marzo de 2021, YouTube fue parte de un fuerte debate sobre sus políticas de moderación de discurso de odio cuando removió un video del comentarista Steve Crowder por considerar que violaba sus políticas relacionadas con la difusión de información errada sobre COVID-19. En ese video Crowder hacía una serie de comentarios acerca de la decisión del gobierno republicano de dar un subsidio a granjeros de minorías raciales por considerarlos históricamente excluidos de las políticas de ayuda a ese sector. **Los comentarios incluían caracterizaciones sobre las formas de hablar, moverse y pensar de los afroamericanos.**

Tras reclamos de distintas organizaciones de defensa de los derechos de las minorías raciales, YouTube emitió un comunicado en el que aseguró que sus "políticas prohíben el contenido que promueve el odio hacia grupos basado en su raza" pero "aunque es ofensivo, este video de Steven Crowder, no viola esas políticas".

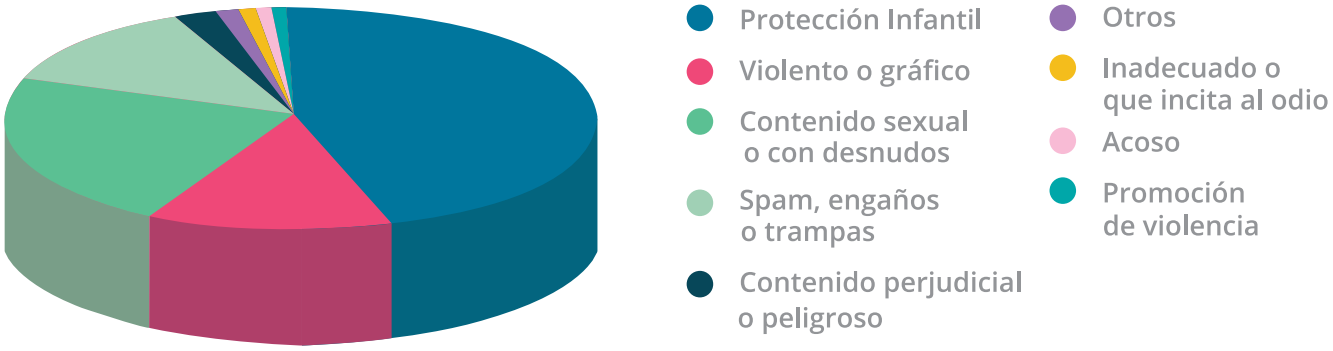
En abril de 2021, YouTube hizo pública información en la que aseguraba que había mejorado sus sistemas de detección de discurso de odio en la plataforma. "No queremos que YouTube sea una plataforma que pueda provocar daños en el mundo de manera atroz", **sostuvo el jefe de Producto de la plataforma, Neal Mohan.**

En YouTube el fenómeno parece ser más complejo de detectar. En los hechos no está claro a partir de los datos disponibles que se haya registrado un aumento significativo del discurso de odio en esa plataforma, aunque sí se registraron episodios aislados con destaque en los medios y la opinión pública.

Entre abril y junio de 2020, YouTube eliminó 11.401.696 videos, sin contar más de 30.000.000 millones de videos que fueron eliminados como consecuencia de la eliminación de 1.998.635 canales en ese mismo período. De esos más de once millones de videos, solamente 552.062 videos fueron eliminados sin utilizar sistemas de detección automática. Entre julio y setiembre fueron eliminados 7.872.684 videos y solamente 481.721 sin detección automática, y entre octubre y diciembre fue de 9.321.948 videos eliminados y solamente 521.866 sin el uso de sistemas automatizados de detección de infracciones a las reglas de YouTube.

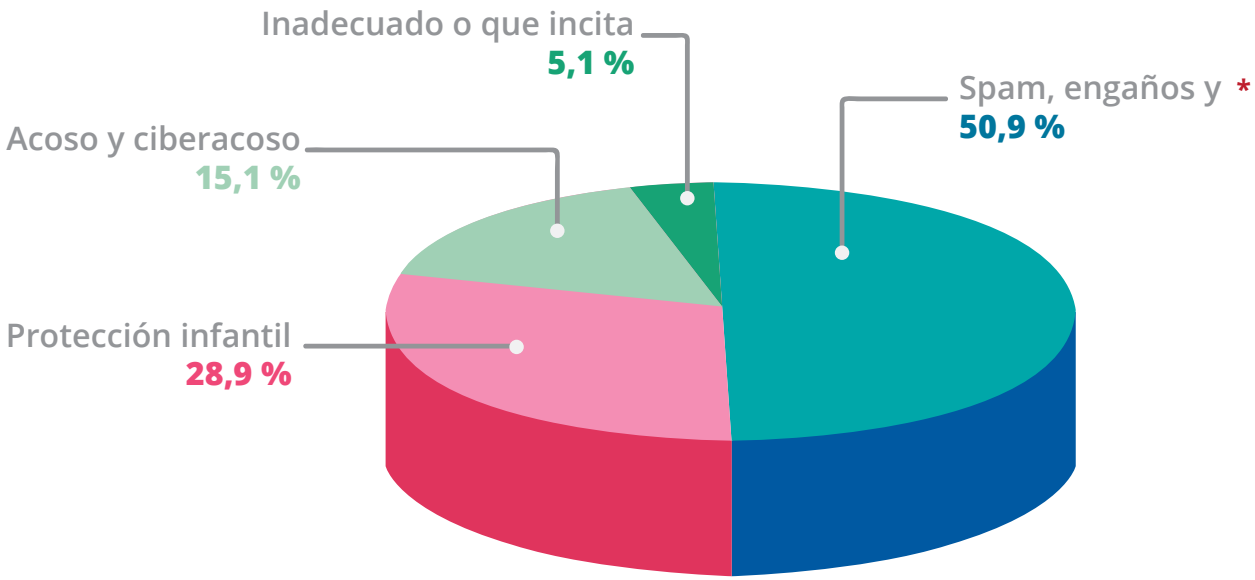
Respecto a los motivos, el discurso de odio no ocupó un espacio significativo alcanzando una cifra de 97.362 videos eliminados en el último trimestre de 2020, aunque sí registró un leve aumento entre el trimestre abril-junio y los dos siguientes, pasando de 0,7% de los videos eliminados a más de 1%.

## Videos retirados según motivo octubre - diciembre




Si se analiza la eliminación de comentarios en los videos, unos 906.196.160 retirados en el último trimestre de 2020, la motivación "incitación al odio" trepa a un 5% entre los motivos especificados para la eliminación de esos contenidos. Esto implica que en el último trimestre de 2020, fueron eliminados de YouTube más de 46 millones de comentarios porque los sistemas de moderación automatizada entendieron que infringían las normas de la plataforma referidas a la categorización "discurso de odio".

## Comentarios retirados, según motivo de retirada Oct- Dic



Cuando YouTube envió a sus moderadores de contenido a sus casas en marzo debido a la pandemia de COVID-19 y extendió dramáticamente el uso de sus filtros automáticos, esto derivó en la duplicación de los videos que fueron eliminados en el segundo trimestre de 2020. El crecimiento dejó abierto el debate en la red social propiedad de Alphabet sobre la precisión de los procesos de moderación automatizada.

"En respuesta a la situación del COVID-19, hemos tomado medidas para proteger a nuestro personal externo y reducir el personal presencial en las oficinas. Como resultado, y de forma temporal, estamos usando más la tecnología para llevar a cabo algunas de las tareas que normalmente hacen los revisores de carne y hueso, por lo que estamos eliminando más contenido que es posible que no infrinja nuestras políticas. Esto influye en algunas de las métricas de este informe y probablemente seguirá influyendo en las métricas de ahora en adelante", escribió la compañía en una publicación de blog que acompaña a **su informe de transparencia del último trimestre**. "Debido a que la responsabilidad es nuestra máxima prioridad, elegimos lo último: utilizar la tecnología para ayudar con parte del trabajo que normalmente realizan los revisores", explicó Google.

En el informe del segundo trimestre, YouTube admitió que el aumento en la eliminación de contenido se debió a que la compañía "aceptó un nivel más bajo de eficacia para estar seguro de que estaban eliminando todas las piezas de contenido que les era posible".

"Una de las decisiones que tomamos al inicio de la pandemia cuando estuvo claro que las máquinas no iban a ser tan precisas como los humanos, fue que íbamos a errar del lado de estar seguros que los usuarios estarían protegidos, aun cuando eso pudiera resultar en un número ligeramente superior de videos que fueron bajados de la plataforma", aseguró el jefe de producto de YouTube, Neil Mohan a **la publicación especializada norteamericana Mashable.**

En setiembre, YouTube anunció que los moderadores humanos comenzarían a retornar a las oficinas y trabajarían en la revisión de los sistemas de moderación para intentar volver a las cifras del inicio de 2020.

Como se pudo apreciar anteriormente, el uso de sistemas de detección automática, según sus propios ejecutivos, significó que YouTube eliminó numerosos contenidos que en los hechos no violaban sus Normas Comunitarias, alcanzando a duplicar la cantidad de apelaciones, que pasaron de 166.000 en el primer trimestre a 325.000 en el segundo de 2020.

A diferencia de Facebook, YouTube no redujo la atención a los procesos de apelación y continúo manteniendo los plazos de proceso anteriores al COVID-19. Esto implicó que la cantidad de videos restituidos tras las apelaciones también pasaron de 41.000 a 161.000 en ese período. **Esto significó un aumento de la tasa normal de restablecimiento de YouTube, normalmente en un 25% de las apelaciones a prácticamente la mitad.**

**En su informe de transparencia, YouTube detalla su proceso específico de moderación de discurso de odio** y aborda alguna de las dificultades específicas que este tipo de contenidos presentan con relación a otros tipos también prohibidos por las Normas Comunitarias.

"La política sobre la incitación al odio es compleja de aplicar a gran escala, ya que las decisiones que se toman requieren un análisis matizado del contexto y un entendimiento minucioso del idioma en cuestión. Para poder aplicar nuestra política de forma coherente, hemos ampliado nuestro equipo de revisión experto en la cuestión y en temas lingüísticos. Además, estamos implementando el aprendizaje automático para detectar posible contenido que incite al odio y enviarlo al equipo de revisión, y aplicamos las lecciones que hemos aprendido en otros tipos de contenido, como el extremismo violento. A veces nos equivocamos, así que disponemos de un proceso de apelación para los creadores que consideren que su contenido se ha retirado inadecuadamente. Evaluamos constantemente nuestras políticas y directrices de aplicación, y seguiremos colaborando con expertos y la comunidad para hacer cambios cuando sea necesario", afirman.

YouTube añade que además de "retirar contenido" que infringe sus Normas Comunitarias, trabajan en "reducir las recomendaciones de contenido que estén al límite" de infringir sus directrices. **"Desde hace tiempo, también tenemos directrices de contenido adecuado para anunciantes, en las que se prohíbe mostrar anuncios en vídeos que incluyan contenido que incite al odio", sostienen.**

# CONCLUSIONES

Bajo innumerables presiones políticas, sociales y mediáticas, Facebook, YouTube y Twitter han realizado en los últimos meses cambios en sus Normas Comunitarias relacionadas con el discurso de odio y han tomado decisiones a las que en los años anteriores parecían resistirse fuertemente y que implican un aumento sustancial en su rol de reguladores de lo que puede y no puede decirse en estos nuevos espacios públicos.

Es difícil saber cuán exitosos han sido estos cambios e incluso cómo definir el éxito frente a medidas que las propias plataformas admiten que no está claro que estén funcionando de manera adecuada durante la pandemia de COVID-19. Estas medidas incluyeron, en algunos casos como Facebook e Instagram, acciones altamente restrictivas e incluso poco garantistas como la virtual desaparición de los procesos de apelación durante varios meses. Esto supuso no sólo la eliminación de contenidos de interés público sino además la pérdida del derecho a reclamar la revisión de parte de miles de usuarios en América Latina.

Más allá de la falta de elementos para determinar a cabalidad todos y cada uno de los motivos que explican este cambio en los criterios, el hecho es que en 2020 las plataformas tomaron decisiones e hicieron cambios en la forma y los procesos en los que moderan contenido. Esos cambios en los procesos y en las Normas Comunitarias que los regulan, significaron un giro dramático respecto al modo en el que Facebook, Twitter y YouTube habían tratado el contenido creado por los usuarios hasta el momento.

Dos fenómenos parecen haber ocurrido este año, un aumento muy significativo de publicaciones con contenidos generalmente considerados “discurso de odio” a partir de la pandemia COVID-19 en las redes sociales. Facebook es la red social en la que -al menos a partir de los datos provistos por las propias plataformas-, se registró un crecimiento de mayor envergadura. Entre 2019 y 2020 los posteos intervenidos por esta red social considerados como discurso de odio crecieron casi 300%. Al profundizar el análisis de 2020, llama la atención que ese crecimiento se produce de manera mucho más significativa en el segundo cuarto del año. Como se destaca en el capítulo anterior, partir de marzo -momento de explosión de la pandemia mundial de COVID-19- la cantidad de posteos moderados por la plataforma por ser considerados discurso de odio, se duplicó y se mantuvo en esos guarismos durante el resto del año. En Twitter y YouTube también se registraron crecimientos aunque no tan considerables.

El segundo fenómeno que debe destacarse fue el hecho de que, a partir de los efectos que este aumento en el discurso de odio y los reclamos surgidos desde la sociedad civil al respecto, Facebook, Twitter y YouTube decidieron profundizar su fiscalización e intervención como expandir los tipos de contenidos que consideran fuera de sus Normas Comunitarias. Sin embargo no parece claro, a juicio de numerosos analistas en el mundo, que esas medidas sean suficientes ni adecuadas, además que existen importantes problemas en la forma en la que fueron implementadas, afectando derechos fundamentales.

A pesar de una creencia popular, aún muy arraigada, las redes sociales nunca fueron espacios de intercambio totalmente abiertos o "sin regulación". Las plataformas llevan años moderando contenidos que ellas consideran "ilegales" pero también aquellos que responden a caracterizaciones aún más vagas (y no prohibidas legalmente) como las que entienden como indecentes, obscenos y fuera de la moral de sus países de origen.

La llegada al mundo del COVID-19, una pandemia global que llevó a millones de personas a aislarse en sus casas, reducir sus contactos y trabajar de forma remota tuvo impactos de todo tipo. Uno de ellos, fue el aumento del discurso de odio en las plataformas sociales pero otro, tal vez menos detectable en una primera aproximación, fue el cambio en los procesos de moderación que se llevan adelante sobre los contenidos que los usuarios publican. Según una investigación realizada en la plataforma de búsqueda Crowdtangle (que permite rastrear el uso de hashtags o palabras en Facebook, Instagram y Twitter), entre febrero de 2020 y marzo de 2021 se generaron en Facebook 43.779 posteos que utilizaron la expresión "virus chino" y registraron un total de 3.535.409 interacciones. Los dos picos principales se registraron en marzo y abril de 2020.

Los gobiernos de todo el mundo llamaron a sus ciudadanos al distanciamiento social sostenido y con ello las plataformas debieron enviar a miles de moderadores humanos a sus casas. Esta decisión provocó un aumento por demás significativo del uso de herramientas automatizadas y de inteligencia artificial en la revisión de los millones de publicaciones que cada minuto se suben a las redes sociales. Aunque en constante proceso de mejora, esos sistemas automatizados aún no son capaces de entender las diferencias del lenguaje, idioma, idiosincrasia y cultura de los millones de usuarios en todo el mundo, así como la importancia del contexto para la definición de conceptos tan complejos como el discurso de odio.

De acuerdo con el estudio de Unesco Countering Online Hate Speech existen al menos cinco posibles aproximaciones no legislativas al problema del discurso de odio en línea y en ellas se alude directamente al rol de las plataformas como parte sustancial de la solución del problema. En el documento, Unesco propone la monitorización y análisis por parte de la sociedad civil, promoción de contra discurso de pares por

parte de individuos, acción organizada por parte de ONGs para reportar los casos a las autoridades, creación de campañas para promover acciones por parte de las compañías de internet que alojan el contenido específico y empoderamiento de usuarios a través de la educación y el entrenamiento con respecto al conocimiento, las capacidades y los aspectos éticos del ejercicio de la libertad de expresión en internet.

Es claro además, que los errores en la detección de discurso de odio en las plataformas pueden resultar en la eliminación de contenido no comprendido en esa definición y por ende en una afectación importante de la libertad de expresión como derecho humano fundamental.

Las plataformas han crecido de forma exponencial en el mundo entero y se han transformado en espacios de intercambio de ideas, por lo que allí ocurre afecta directamente (o tiene la potencialidad de afectar) cómo se procesa el debate público. Permitir que tanto los gobiernos como las plataformas se conviertan en reguladores del contenido puede traducirse en el silenciamiento de voces disidentes, en especial, en sociedades autoritarias.

Pero tal como afirma Díaz Hernández, el problema no es únicamente que las prohibiciones resulten en restricciones indebidas o desproporcionadas a la libertad de expresión, sino que también suelen ser ineficaces para abordar y solucionar el problema de fondo pues no cumplen el rol de contrarrestar el discurso de odio sino que, con frecuencia, agravan el clima de violencia y la polarización social que dio lugar al contenido de origen.

Es importante además tener en cuenta que los problemas derivados de la regulación de contenido en plataformas no solo involucran la regulación del contenido mismo sino la arquitectura de internet tal como la conocemos, así como sus características de espacios teóricamente extra-espaciales y extra-territoriales. A partir de esta estructura y del papel que en este ecosistema juegan las plataformas y redes sociales, cada uno de esos entornos tiene sus propias reglas de funcionamiento y generado sus propias definiciones de lo que está o no prohibido y permitido. En este sentido, parte del problema es que no se trata solo de lo que la legislación de cada Estado entiende por discurso de odio sino de lo que ese término significa para Facebook, Twitter o YouTube, cuando éstas no están sujetas a controles democráticos y no ofrecen garantías de debido proceso ni transparencia, entre otras.

La pandemia mundial trajo consigo impactos de todo tipo en la vida de las personas. Tal vez uno de ellos sea, también, el comienzo de una discusión acerca del rol de las plataformas como moderadores de contenido, los problemas derivados de permitirles o impulsarlos a ocupar ese rol de gatekeepers en Internet.

**ANA LAURA PÉREZ**

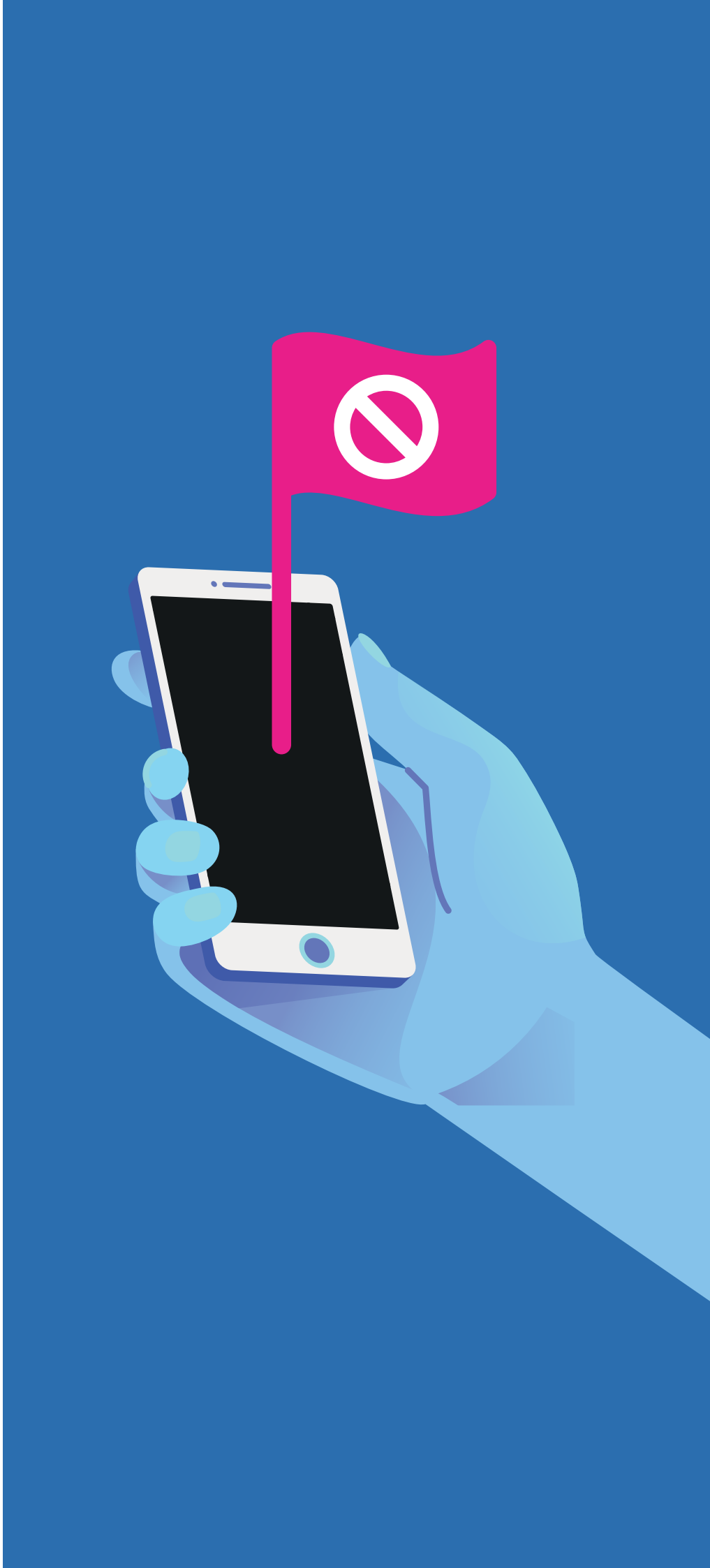
## SOBRE LA AUTORA

Es licenciada en Comunicación orientación Periodismo de la Universidad ORT, diplomada en Estudios Latinoamericanos de la Universidad de Montevideo y Master en Business Administration del Instituto de Estudios Empresariales de Montevideo.

Trabaja hace 20 años como periodista y editora en algunos de los medios de comunicación más influyentes de su país: los diarios El Observador y El País y el semanario Búsqueda, así como conduce y participa en programas de TV Ciudad, canal público de la Intendencia de Montevideo. Actualmente es además gerenta de Producto Digital del diario El País

Fue coordinadora de Periodismo y Contenidos Digitales de la Licenciatura de Comunicación de la Universidad ORT, donde es docente desde hace casi diez años.

Ha participado como conferencista, oradora y panelista en diversos eventos sobre periodismo, en particular sobre desinformación y plataformas digitales, temas en los que se ha especializado en los últimos años y sobre los que ha dado capacitaciones y cursos de formación a periodistas de Uruguay y de varios países de América Latina.



Financiado por la
Unión Europea