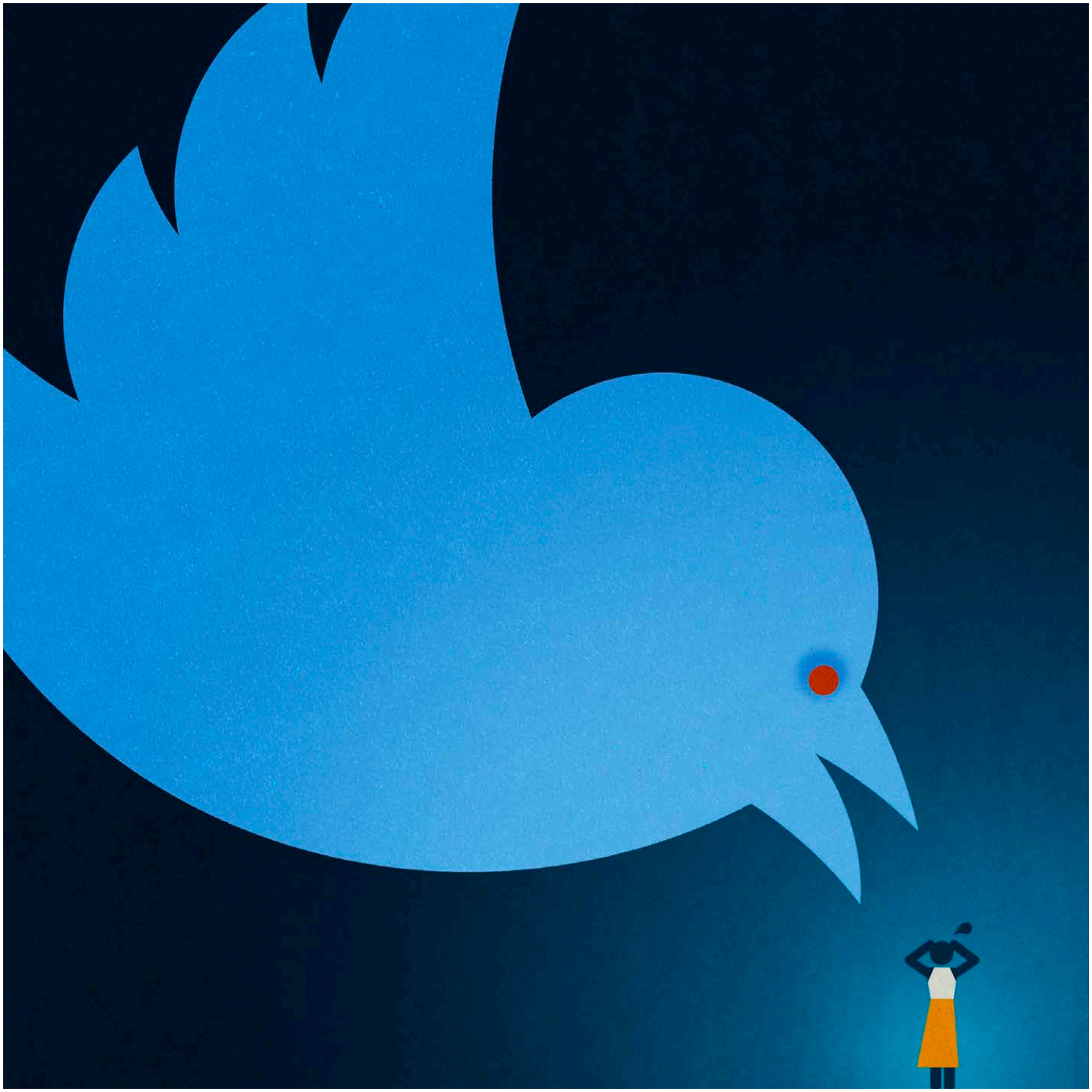

TABLA DE PUNTUACIÓN DE TWITTER

SEGUIMIENTO DE LOS PROGRESOS DE TWITTER CONTRA LA VIOLENCIA Y LOS ABUSOS ONLINE QUE SUFREN LAS MUJERES EN ARGENTINA

AMNISTÍA INTERNACIONAL

Amnistía Internacional es un movimiento global de más de 7 millones de personas que trabajan en favor del respeto y la protección de los derechos humanos.

Nuestra visión es la de un mundo en el que todas las personas disfrutan de todos los derechos humanos proclamados en la Declaración Universal de Derechos Humanos y en otras normas internacionales.

Somos independientes de todo gobierno, ideología política, interés económico y credo religioso. Nuestro trabajo se financia principalmente con las contribuciones de nuestra membresía y con donativos.

Publicado por primera vez en 2020
por Amnesty International Ltd
Peter Benson House, 1 Easton Street
London WC1X ODW, Reino Unido

Índice: AMR 13/4721/2021
Idioma original: Inglés

amnesty.org

Cover photo: © www.NickPurserDesign.com

AMNISTÍA INTERNACIONAL

INTRODUCCIÓN

Twitter es una red social que usan millones de personas en todo el mundo para debatir, conectar y compartir información entre sí. Como tal, puede ser una potente herramienta para hacer conexiones y expresarse. Pero, para muchas mujeres y personas no binarias, Twitter es una plataforma donde abundan la violencia y los abusos contra ellas, a menudo impunemente.[1]

En 2017, Amnistía Internacional encargó una encuesta online a mujeres de ocho países sobre sus experiencias de abusos en las redes sociales y usó la ciencia de datos para analizar los abusos que sufrieron las parlamentarias en Twitter antes de las elecciones anticipadas de 2017 en Reino Unido.[2] En marzo de 2018, Amnistía Internacional publicó *Toxic Twitter: Violence and abuse against women online*, un informe que denuncia la magnitud, la naturaleza y el impacto de la violencia y los abusos dirigidos a las mujeres en Estados Unidos y Reino Unido en Twitter.[3] Nuestra investigación concluyó que la plataforma no había asumido su responsabilidad de proteger los derechos de las mujeres en Internet al no investigar debidamente las denuncias de violencia y abuso ni responder a ellas de forma transparente, por lo que muchas mujeres guardan silencio o se autocensuran en la plataforma. Aunque Twitter ha hecho progresos a la hora de abordar este problema desde 2018, la empresa sigue incumpliendo sus responsabilidades en materia de derechos humanos y debe tomar más medidas para proteger los derechos de las mujeres en Internet.

La persistencia de este tipo de abusos menoscaba el derecho de las mujeres y de las personas no binarias a expresarse en condiciones de igualdad, libremente y sin temor. Como dice Amnistía Internacional en *Toxic Twitter*: "En lugar de fortalecer la voz de las mujeres, la violencia y los abusos que muchas de mujeres y personas no binarias experimentan en Twitter las obligan a autocensurar sus mensajes, limitar su interacción e incluso abandonar por completo la plataforma". Por otra parte, como pone de relieve nuestra investigación, los abusos experimentados tienen un carácter muy interseccional, al dirigirse a mujeres de color, mujeres de minorías étnicas o religiosas, mujeres pertenecientes a castas marginadas, mujeres lesbianas, bisexuales o transgénero y mujeres con discapacidad.

Desde la publicación de *Toxic Twitter* en marzo de 2018, Amnistía Internacional ha publicado una serie de informes adicionales, como el estudio Troll Patrol (Patrulla Antitroles) en diciembre de 2018, en el que Amnistía Internacional y Element AI colaboraron para analizar millones de tuits recibidos durante 2017 por 778 mujeres periodistas y políticas de Reino Unido y Estados Unidos de diversas posiciones políticas de todo el espectro ideológico.[4] Mediante el uso de herramientas punteras de la ciencia de datos y de técnicas de aprendizaje automático, pudimos ofrecer un análisis cuantitativo de la magnitud sin precedentes de abusos en Internet contra mujeres en Reino Unido y Estados Unidos.

En noviembre de 2019, Amnistía Internacional publicó una investigación sobre la violencia y los abusos contra mujeres en varias redes sociales, Twitter entre ellas, en Argentina en el periodo previo a los debates sobre la legalización del aborto en el país y durante éstos.[5] En enero de 2020, Amnistía Internacional publicó una nueva investigación que medía la magnitud y la naturaleza de los abusos que sufrieron las

1. Abusos similares suelen producirse también en otras plataformas como Facebook and Instagram. Esta tabla se centra específicamente en Twitter, a raíz de las investigaciones previas de Amnistía Internacional sobre esta plataforma, como se describe a continuación.

2. Amnistía Internacional, *Amnistía revela alarmante impacto de los abusos contra las mujeres en Internet,* comunicado de prensa, 20 de noviembre de 2017, https://www.amnesty.org/es/latest/news/2017/11/amnesty-reveals-alarming-impact-of-online-abuse-against-women/ (consultado por última vez el 24 de agosto de 2020); también, Amnesty Global Insights, *Unsocial Media: Tracking Twitter Abuse against Women MPs,* 4 de septiembre de 2017, https:// medium.com/@AmnestyInsights/unsocial-media-tracking-twitter-abuse-against-women-mps-fc28aeca498a (consultado por última vez el 24 de agosto de 2020).

3. Amnistía Internacional, *Toxic Twitter: A Toxic Place for Women,* Índice: ACT 30/8070/2018, marzo de 2018, https://www.amnesty.org/en/ latest/research/2018/03/online-violence-against-women-chapter-1/#topanchor (consultado por última vez el 24 de agosto de 2020).

4. Amnistía Internacional, *Troll Patrol Report,* diciembre de 2018, https://decoders.amnesty.org/projects/troll-patrol/findings (consultado por última vez el 24 de agosto de 2020).

5. Amnistía Internacional, *Corazones Verdes: Violencia online contra las mujeres durante el debate por la legalización del aborto en Argentina,* noviembre de 2019, https://amnistia.org.ar/corazonesverdes/files/2019/11/corazones_verdes_violencia_online.pdf (consultado por última vez el 24 de agosto de 2020).

mujeres políticas de India en Internet durante las elecciones generales celebradas en ese país en 2019.[6] La investigación de Amnistía Internacional detalló nuevos casos de violencia y abusos contra las mujeres en la plataforma, en esta ocasión en contextos geográficos y lingüísticos diversos, lo que motivó que se volviera a pedir a Twitter que abordase este problema urgente y continuado. Todos estos informes concluían con medidas concretas que debía tomar Twitter para cumplir con su responsabilidad de respetar los derechos humanos en el contexto de la violencia y los abusos contra las mujeres en la plataforma.

En septiembre de 2020, Amnistía Internacional publicó la primera Tabla de puntuación de Twitter. Esta Tabla de puntuación estaba concebida para hacer el seguimiento de los progresos globales de Twitter a la hora de abordar el lenguaje insultante según 10 indicadores que abarcan la transparencia, los mecanismos de denuncia, el proceso de revisión de las denuncias de abusos y las características de privacidad y seguridad mejoradas. Estos indicadores se elaboraron a partir de recomendaciones que Amnistía Internacional había formulado con anterioridad sobre la mejor forma en que Twitter puede abordar contenidos abusivos y problemáticos.

Esta es la segunda edición de la Tabla de puntuación, y hace el seguimiento de los progresos que, en su caso, ha hecho Twitter en el último año en relación con estos diez indicadores.[7] Aunque Twitter ha hecho algunos progresos, distan mucho de ser suficientes. La plataforma ha aumentado la cantidad de información disponible a través de su Centro de ayuda[8] y sus Informes de transparencia,[9] y al mismo lanza nuevas campañas de sensibilización, amplía el alcance de su política sobre conducta de odio para incluir el lenguaje que deshumanice a las personas por motivos de religión, edad, discapacidad o enfermedad, y mejora sus mecanismos de denuncia y sus características de privacidad y seguridad. Estos pasos son importantes; dicho esto, el problema sigue sin resolverse. Twitter debe tomar más medidas para que las mujeres y las personas no binarias —así como todas las personas usuarias en todos los idiomas— puedan usar la plataforma sin temor a sufrir abusos.

6. Amnistía Internacional, *Troll Patrol India: Exposing the Online Abuse Faced by Women Politicians in India,* 16 de enero de 2020, https://decoders.amnesty.org/projects/troll-patrol-india (consultado por última vez el 24 de agosto de 2020).

7. Además de este informe centrado en Argentina, Amnistía Internacional está publicando simultáneamente informes sobre esta cuestión en relación con Sudáfrica y Estados Unidos.

8. Twitter, Centro de ayuda, https://help.twitter.com/es (consultado por última vez el 6 de julio de 2021).

9. Twitter, Twitter Transparency Center, https://transparency.twitter.com (consultado por última vez el 6 de julio de 2021).

2. ¿QUÉ SON LA VIOLENCIA Y LOS ABUSOS CONTRA LAS MUJERES Y LAS PERSONAS NO BINARIAS EN INTERNET?

Según el Comité para la Eliminación de la Discriminación contra la Mujer, la violencia de género incluye "la violencia dirigida contra la mujer porque es mujer o que la afecta en forma desproporcionada" y, como tal, constituye una violación de sus derechos humanos.[10] El Comité establece asimismo que la violencia de género contra las mujeres incluye (entre otros aspectos) actos que infligen daños o sufrimientos de índole física, mental o sexual a las mujeres y amenazas de cometer esos actos.[11] Esto podría ser facilitado por medios en línea.

El Comité para la Eliminación de la Discriminación contra la Mujer (CEDAW) usa la expresión "violencia por razón de género contra la mujer" para reconocer expresamente las causas y los efectos relacionados con el género de la violencia.[12] La expresión violencia por razón de género refuerza además la noción de la violencia como problema social más que individual que exige respuestas integrales. Además, el CEDAW afirma que el derecho de las mujeres a una vida libre de violencia por razón de género es indivisible e interdependiente respecto de otros derechos humanos, como los relativos a la libertad de expresión, de participación, de reunión y de asociación.[13] La relatora especial sobre la violencia contra la mujer ha afirmado: "[L]a definición de violencia en línea contra la mujer se aplica a todo acto de violencia por razón de género contra la mujer cometido, con la asistencia, en parte o en su totalidad, del uso de las TIC, o agravado por este, como los teléfonos móviles y los teléfonos inteligentes, Internet, plataformas de medios sociales o correo electrónico, dirigida contra una mujer porque es mujer o que la afecta en forma desproporcionada".[14]

La violencia y los comportamientos abusivos contra las mujeres en las redes sociales, como Twitter, incluyen diversas experiencias: amenazas directas o indirectas de violencia física o sexual; insultos dirigidos a uno o varios aspectos de la identidad de una mujer (como los de carácter racista, transfóbico, etc.); acoso selectivo; atentados contra la intimidad como el doxeo (divulgación en Internet de datos privados que revelan la identidad de una persona con el fin de causar alarma o malestar); y la divulgación de imágenes sexuales o íntimas de una mujer sin su consentimiento.[15] En ocasiones, una o más formas de esa violencia y esos abusos se utilizarán conjuntamente como parte de un ataque coordinado contra una persona, lo que a menudo se designa con el término *"pile-on"*. Las personas que llevan a cabo una constante de acoso selectivo contra una persona suelen recibir el nombre de "trolls".[16]

---

10. ONU Mujeres, Recomendaciones Generales adoptadas por el Comité para la Eliminación de la Discriminación contra la Mujer, Recomendación general Nº 19, 11º periodo de sesiones, párr. 6, 1992, https://www.un.org/womenwatch/daw/cedaw/recommendations/recomm-sp.htm (consultado por última vez el 22 de agosto de 2020).

11. Comité para la Eliminación de la Discriminación contra la Mujer, Recomendación general núm. 35 sobre la violencia por razón de género contra la mujer, por la que se actualiza la recomendación general núm. 19, párr. 14, 26 de julio de 2017, doc. ONU CEDAW/C/GC/35, https://undocs.org/es/CEDAW/C/GC/35 (consultado por última vez el 22 de agosto de 2020).

12. Comité para la Eliminación de la Discriminación contra la Mujer, Recomendación general núm. 35 sobre la violencia por razón de género contra la mujer, por la que se actualiza la recomendación general núm. 19, párr. 14, 26 de julio de 2017, doc. ONU CEDAW/C/GC/35, https://undocs.org/es/CEDAW/C/GC/35 (consultado por última vez el 22 de agosto de 2020).

13. Comité para la Eliminación de la Discriminación contra la Mujer, Recomendación general núm. 35 sobre la violencia por razón de género contra la mujer, por la que se actualiza la recomendación general núm. 19, párr. 14, 26 de julio de 2017, doc. ONU CEDAW/C/GC/35, https://undocs.org/es/CEDAW/C/GC/35 (consultado por última vez el 20 de agosto de 2020).

14. Consejo de Derechos Humanos de la ONU, *Informe de la Relatora Especial sobre la violencia contra la mujer, sus causas y consecuencias acerca de la violencia en línea contra las mujeres y las niñas desde la perspectiva de los derechos humanos,* 18 de junio a 6 de julio de 2018, doc. ONU A/HRC/38/47, https://undocs.org/es/A/HRC/38/47.

15. Amnistía Internacional, *¿Qué son la violencia y los abusos contra las mujeres en Internet?,* 20 de noviembre de 2017, https://www.amnesty.org/es/latest/campaigns/2017/11/what-is-online-violence-and-abuse-against-women/ (consultado por última vez el 20 de agosto de 2020).

16. Amnistía Internacional, *¿Qué son la violencia y los abusos contra las mujeres en Internet?,* 20 de noviembre de 2017, https://www.amnesty.org/es/latest/campaigns/2017/11/what-is-online-violence-and-abuse-against-women/ (consultado por última vez el 20 de agosto de 2020).

3. LAS RESPONSABILIDADES DE TWITTER EN MATERIA DE DERECHOS HUMANOS

Las empresas, cualquier que sea el lugar del mundo donde lleven a cabo su actividad, tienen la responsabilidad de respetar todos los derechos humanos. Esta es una norma de conducta que cuenta con respaldo internacional.[17] La responsabilidad de las empresas de respetar los derechos humanos exige que Twitter tome medidas concretas para evitar causar abusos contra los derechos humanos o contribuir a ellos y para abordar los efectos en los derechos humanos en los que están implicadas, lo que incluye proporcionar recursos efectivos para cualquier efecto real. También les exige tratar de prevenir o mitigar las consecuencias negativas sobre los derechos humanos directamente vinculadas a sus operaciones o a servicios de sus relaciones comerciales, incluso si no han contribuido a que se produzcan. En la práctica, esto significa que Twitter debería evaluar, de forma continua y proactiva, la manera en que sus políticas y prácticas afectan a los derechos de quienes usan la plataforma respecto a la no discriminación, la libertad de expresión y de opinión, la libertad de reunión y de asociación, así como a otros derechos, y tomar medidas para mitigar o prevenir las posibles repercusiones negativas.

4. DEFINICIÓN DE CONTENIDO ABUSIVO Y PROBLEMÁTICO

CONTENIDO ABUSIVO. Tuits que promueven la violencia contra alguien por razón de su raza, origen étnico, casta, origen nacional, orientación sexual, género, identidad de género, filiación religiosa, edad, discapacidad o enfermedad grave. Algunos ejemplos son las amenazas físicas o sexuales, los deseos de daños físicos o muerte, la referencia a actos violentos, el comportamiento que causa temor o la difamación, los epítetos, los símiles racistas y sexistas reiterados, u otros contenidos que sean degradantes para una persona.[18]

CONTENIDO PROBLEMÁTICO. Los tuits con contenidos hirientes u hostiles, especialmente si se dirigen reiteradamente a la misma persona en múltiples ocasiones aunque no lleguen a la consideración de abusivos. Los tuits problemáticos pueden reforzar estereotipos negativos o perjudiciales contra un grupo de personas (por ejemplo, estereotipos negativos sobre una raza o pueblo que sigue una determinada religión). Creemos que esos tuits pueden seguir surtiendo el efecto de silenciar a una persona o un grupo de personas. Sin embargo, reconocemos que los tuits problemáticos pueden ser expresión protegida y no serían objeto necesariamente de eliminación de la plataforma.[19]

---

17. Principios Rectores sobre las Empresas y los Derechos Humanos, 2011, https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_sp.pdf (consultado por última vez el 22 de agosto de 2020).

18. Amnistía Internacional, *Troll Patrol,* https://decoders.amnesty.org/projects/troll-patrol/findings#abusive_tweet/abusive_sidebar.

19. Amnistía Internacional, *Troll Patrol,* https://decoders.amnesty.org/projects/troll-patrol/findings#inf_12/problematic_sidebar.

5. VIOLENCIA Y ABUSOS CONTRA LAS MUJERES Y LAS PERSONAS NO BINARIAS EN TWITTER EN ARGENTINA

En Argentina, Twitter tiene más de 5,4 millones de usuarios en actividad.[20] Es una plataforma en la que se difunde y se busca información, se comparten opiniones y se promueve el debate.

Ser parte de la conversación en Twitter tiene múltiples implicaciones. La inmediatez en el acceso a la información y la posibilidad de difundir datos, opiniones y noticias sin intermediarios conviven con la hostilidad y los mensajes de odio, los ataques coordinados, el *doxing* y la difusión de información falaz que no logra ser desmentida con la misma rapidez e impacto con que se instala. Esto repercute no sólo en la calidad del debate público sino también en la libertad de expresión de las personas, en su salud mental e incluso en otras esferas de su vida que trascienden el universo digital.

En Argentina, la plataforma de Twitter es un espacio que muchas mujeres y personas LGTBI+, reconocen como un lugar hostil y virulento. Los testimonios de les usuaries dan cuenta que hay quienes se han visto forzadas a modificar su conducta en las redes o a autocensurarse como forma de preservarse. Incluso, hay quienes han resuelto abandonar por completo la red social por el impacto causado por las agresiones y mensajes abusivos recibidos a través de la plataforma.[21]

Amnistía Internacional Argentina ha relevado que los ataques tienen un sesgo marcado hacía aspectos relacionados con el género o con otras características de la identidad de las mujeres y personas LGTBI+.[22] Según las investigaciones llevadas a cabo en Argentina, una de cada tres mujeres ha sufrido violencia en las redes sociales.[23]

En este sentido, la encuesta que formó parte del informe *Corazones Verdes* de 2019, en el contexto previo a la legalización del aborto en Argentina en 2018, dio cuenta del aumento de la violencia que padecen numerosas activistas y defensoras de los derechos humanos que luchan por la ampliación de los derechos de las mujeres y del colectivo LGTBI+. Durante el debate, el lenguaje abusivo se incrementó en un 42%; las amenazas psicológicas de violencia sexual, un 12%; los comentarios racistas, un 14%; y los comentarios homofóbicos o transfóbicos, un 15%.[24] Se evidencia, asimismo, el efecto de la violencia online en distintas esferas de la vida de las mujeres: el 39% de las mujeres que sufrieron este tipo de violencia sintió que su seguridad física estaba amenazada. Algunas también manifestaron el impacto que tuvo sobre su salud física y psicológica. Un 36% tuvo ataques de pánico, estrés o ansiedad y un 35%, pérdida de autoestima o de confianza. Un 34% dijo haber sentido miedo a salir y un 33% identificó haber atravesado un periodo de aislamiento psicológico.[25]

Durante los meses de septiembre a noviembre de 2021, Amnistía Internacional Argentina condujo una serie de entrevistas con periodistas y actrices, activistas y defensoras de los derechos humanos de las mujeres y personas LGTBI+, que destacaron a Twitter como una plataforma que propicia la violencia de género en la conversación. La proliferación de acciones concertadas y de cuentas mecanizadas —a

20. Statista.com: https://es.statista.com/estadisticas/1219122/numero-de-usuarios-activos-mensuales-twitter-argentina-sistema-operativo/.

21. Amnistía Internacional, *Corazones Verdes: Violencia online contra las mujeres durante el debate por la legalización del aborto en Argentina,* noviembre de 2019, https://amnistia.org.ar/corazonesverdes/files/2019/11/corazones_verdes_violencia_online.pdf

22. En 2018 y 2019, Amnistía Internacional Argentina ha realizado investigaciones sobre la violencia y los abusos contra mujeres y personas no binarias en las redes sociales y en Twitter en particular, en el contexto de los debates por la legalización del aborto. Amnistía Internacional, *Pañuelos Verdes: Relatos de la violencia durante el debate por la legalización de la interrupción legal del embarazo, diciembre de 2018,* https://amnistia.org.ar/atacadas-por-usar-panuelos-verdes-casos-de-violencia-en-el-contexto-del-debate-por-el-aborto-legal/. Amnistía Internacional, *Corazones Verdes: Violencia online contra las mujeres durante el debate por la legalización del aborto en Argentina*, noviembre de 2019, https://amnistia.org.ar/corazonesverdes/files/2019/11/corazones_verdes_violencia_online.pdf

23. Amnistía Internacional, Corazones Verdes: Violencia online contra las mujeres durante el debate por la legalización del aborto en Argentina, noviembre de 2019, https://amnistia.org.ar/wp-content/uploads/delightful-downloads/2021/11/Corazones_verdes_Violencia_online.pdf

24. Amnistía Internacional, *Corazones Verdes: violencia online contra las mujeres durante el debate por la legalización del aborto en Argentina,* 2019 https://amnistia.org.ar/wp-content/uploads/delightful-downloads/2021/11/Corazones_verdes_Violencia_online.pdf

25. Amnistía Internacional, *Corazones Verdes: violencia online contra las mujeres durante el debate por la legalización del aborto en Argentina,* 2018, https://amnistia.org.ar/corazonesverdes/informe-corazones-verdes.

través de bots y trolls— hacían la experiencia más hostil, lo que potencian las insuficientes medidas adoptadas por la plataforma para morigerar los casos de violencia de género online y la falta de regulación por parte del Estado.

EXPERIENCIAS DE ABUSO EN TWITTER

Las personas entrevistadas han coincidido en que Twitter se ha vuelto una red que incentiva interacciones hostiles o ser "picante" o "amarillista": es decir, emplear expresiones que generen escándalo, y sean impactantes y sensacionalistas.

La violencia de género hacia las mujeres y personas LGBTI+ se combina con un contexto en el que, en cuestión de segundos y desde cualquier lugar del mundo, pueden surgir ataques anónimos e incluso organizados.

Thelma Fardin —actriz, cantante y activista feminista de Argentina— denunció en 2018 que, cuando tenía 16 años, un actor la había sometido a abusos sexuales. Participa en Twitter desde ese año y considera que la red es un espacio de debate político y el lugar donde se construyen opiniones. Sin embargo, cree que se ha vuelto una plataforma tan virulenta que lo usa sólo cuando tiene algo que decir . En este sentido, agrega: "Del otro lado puede aparecer cualquier cosa, con una impunidad total, por el anonimato de estar detrás de una pantalla". "La experiencia empeoró en el último tiempo […] por el anonimato y la idea de lo impersonal que se genera. Creo que tiene que ver con un momento social que se está viviendo en América Latina específicamente […] y estamos viviendo la reacción a un avance a la conquista de tantos derechos".[26]

Marlene Wayar, activista travesti, directora del periódico travesti *El Teje* y coordinadora de Futuro Transgenérico, coincide con el testimonio anterior: "Fue un obstáculo [para el uso de Twitter] que mi comunicación sea honesta, desde mi propio ser y no hacer el esfuerzo por ser canchera, por ser violenta, efectista o por ser amarillista y agrega que ser genuina no provoca repercusión".[27]

Otra entrevistada, Marina Abiuso, periodista especializada en género y editora de Género en los canales televisivos TN y Canal 13, usuaria de Twitter desde 2013, da cuenta de los sesgos de género que tiene la violencia en Twitter: "Hay una hostilidad muy marcada, hay un hostigamiento muy marcado hacia las mujeres [...] Nosotras muchas veces creemos que tiene que ver con los climas de micro mundo... Y colegas de otras partes del mundo experimentan lo mismo. Twitter se volvió un terreno de mucha hostilidad, dejó de ser un espacio de conversación".[28] Agrega: "Las agresiones que recibimos las mujeres tienen una cuota de sexismo que es imposible de ignorar".[29]

En Argentina se repite el patrón que se da en otras partes del mundo, según lo relevado por el informe de Amnistía Internacional en *Toxic Twitter,*[30] en cuanto a las modalidades que adquieren las agresiones a mujeres y personas LGBTI+, asociadas a su género o identidad. "A lo largo de los últimos años, y con mucho más énfasis desde que me volví una cara más pública relacionada al feminismo, recibí desde imágenes de genitales, por privado y en público, mensajes de hombres pidiéndome que les enviara desde fotos desnudas hasta que les enviara de alguna parte del cuerpo por algún fetiche, agresiones, comentarios sobre mi cuerpo, muchísimos y de todo tipo, mucho comentario gordofóbico, homofóbico […] memes con mi imagen, he recibido amenazas, fotos de armas, mi foto intervenida con el bigote de Hitler, he recibido imágenes de niños muertos en el contexto de la discusión por aborto", dice Marina Abiuso.[31]

---

26. Entrevista realizada el 1 de octubre de 2021, Ciudad Autónoma de Buenos Aires.

27. Entrevista realizada el 1 de octubre de 2021, Ciudad Autónoma de Buenos Aires.

28. Entrevista realizada el 17 de septiembre de 2021. Ciudad Autónoma de Buenos Aires.

29. Entrevista realizada el 17 de septiembre de 2021. Ciudad Autónoma de Buenos Aires.

30. Amnistía Internacional, #ToxicTwitter, *Violencia y Abuso contra las Mujeres en Internet,* Disponible en: https://amnistia.org.ar/wp-content/uploads/delightful-downloads/2018/05/TOXICTWITTER-report_SP.pdf

31. Entrevista realizada el 17 de septiembre de 2021. Ciudad Autónoma de Buenos Aires.

Los ataques se profundizan cuando las mujeres y personas LGBTI+ adquieren roles más protagónicos en la esfera pública y en particular cuando embanderan causas feministas. Así fue la experiencia de Manu Mireles, persona trans no binaria, migrante y activista por los derechos del colectivo LGBTIQ+, secretaria general de la Asociación Civil Mocha Celis. Usa Twitter desde hace nueve años: "Comencé a usarla porque seguía cuentas que me daban información relevante; te conecta con un montón de información de forma inmediata". Además, rescata que, para los activismos, "es como un aprendizaje poder usar las plataformas para poner en agenda los temas que nos interesan, movilizar la opinión pública en ese sentido. Pero el anonimato que permiten las redes hace que haya mucha información que circula que promueve el odio".

Agrega que para ella la violencia ha repuntado en los últimos años. "Creo que mi experiencia en Twitter empeoró. Empezó a ser más conflictiva […] comencé a recibir mensajes de odio, amenazas, mucha violencia, porque además yo comencé a tener un rol de mayor visibilidad en el activismo, porque comencé a tener vocería de un espacio. Eso hizo que de pronto, en muy pocos meses, viva un montón de situaciones de violencia, incluso ataques directos. Me han llegado a enviar tuits de los negativos con 'ya vas a ver si te encuentro en la calle, lo que te va a pasar' y directamente mucha violencia como 'puto de mierda', 'maricón', 'que se mueran todas las travestis', cosas directamente asociadas a mi trabajo, a mi identidad y muchas amenazas […] y con un nivel de impunidad muy grande […]".[32]

Por su parte, Georgina Orellano, trabajadora sexual y activista por los derechos laborales de las trabajadoras y trabajadores sexuales, subraya que cuando la discusión sobre el trabajo sexual adquirió mayor visibilidad, aumentaron los ataques y debió cambiar su comportamiento: "Creo que el límite lo marqué cuando se empezó a utilizar mensajes donde se me deseaba la muerte, o donde incitaban mucho al odio hacia mi persona". "Cualquier persona se hace un perfil y destila odio y violencia y nadie controla eso y tampoco el daño que eso produce hacia las personas". En este sentido, indica que la escalada en la violencia la llevó a limitar ampliamente su participación en Twitter.

LOS MECANISMOS DE DENUNCIA Y MONITOREO SON INSUFICIENTES

Las personas entrevistadas coinciden en que las acciones tomadas por Twitter son insuficientes, poco efectivas y difusas, y que los procesos de denuncia muchas veces quedan sin respuesta. Además, advierten de otros inconvenientes que presentan los mecanismos actuales.

Un punto relevante es la necesidad de transversalizar la perspectiva de género en los mecanismos de monitoreo de la conversación en Twitter. "La perspectiva de género y diversidad sexual es una perspectiva que desafía constantemente un montón de mandatos […] hay muchas formas sutiles de violentar. Ya solamente reproducir estereotipos de género es una forma de violencia que es invisible para muchas personas",[33] afirma Manu Mireles, activista trans no binario.

Sobre este punto, Marlene Wayar, activista travesti, comentó que había vivido la falta de respuesta en ambas direcciones: "Yo de las redes nunca tuve respuesta alguna, ni cuando yo denuncié […] ni cuando fui denunciada. Eventualmente 'se [me] canceló alguna publicación porque infringía estas normas' y no tuve respuesta sobre qué norma infringía ni quién lo hizo [la denuncia] o si podíamos dialogar. Esto implicaría que aprenda. […]. Y si no mi derecho a defensa, que es básico, nodal, de contraargumentar".[34]

Otra de las entrevistadas, la periodista Marina Abiuso, plantea un aspecto relevante a considerar en el momento de evaluar y actualizar las políticas de seguridad de Twitter: la preservación de la evidencia de las agresiones y los abusos en redes sociales. "Muchas veces cuando aparece una agresión por redes sociales tendemos a tratarla en el marco de la red social y hay cosas que exceden al universo de las redes sociales. Hay amenazas y tipos de amenazas que pueden ser denunciables por fuera de las

32. Entrevista realizada el 17 de septiembre de 2021. Ciudad Autónoma de Buenos Aires.

33. Entrevista realizada el 17 de septiembre de 2021. Ciudad Autónoma de Buenos Aires.

34. Entrevista realizada el 1 de octubre de 2021, Ciudad Autónoma de Buenos Aires.

redes sociales. Cuando nosotras accionamos dentro de Twitter nos quedamos sin herramientas para poder accionar penalmente por fuera".[35]

Por otro lado, sobre la calidad y modalidad del monitoreo, agregó: "No siento que Twitter haga un control efectivo. Es muy fácil detectar cuándo está ocurriendo una situación de acoso selectivo y muchas veces se encuentra que las denuncias que se hacen, Twitter las revisa sin filtro humano […] como no está escrita 'la palabra', no se leen en su contexto. Una foto de un falcón verde puede no significar nada para Twitter pero en Argentina, yo sé perfectamente qué significa […]. Muchas veces ese tipo de cuestiones no aparecen contempladas",[36] afirma Marina Abiuso.

En este sentido, se sigue evidenciando la necesidad de tener controles que puedan analizar los tuits y las denuncias en su contexto local. Por ejemplo, cuando la entrevistada se refiere al envío de fotos de "un falcón verde", se trata del modelo de automóvil asociado a la desaparición forzada de personas durante la dictadura militar argentina (1976-1983). Un control automatizado de denuncias difícilmente identificaría esta imagen como un tuit abusivo.

Sobre los cambios en los últimos dos años, la mirada de las personas entrevistadas es crítica; pese a reconocer que hay avances, éstos son insuficientes. Thelma Fardin sostiene: "Creo que las herramientas que da Twitter para denunciar la violencia mejoraron porque antes incluso, cuando yo empecé con la red, no había figuras para ciertas cosas que querías denunciar […]. De todas maneras, creo que ya está tan naturalizado el tono, es una red que como identidad tiene una cosa violenta, de a ver quién es más picante, y en esa propia identidad que construyó la marca […] a pesar de que te dan herramientas para denunciarlo, se sigue fomentando que a ver quién es más canchero y más picante y en eso hay un matiz de odio".[37] En este sentido, señaló que, como árbitros del debate en la plataforma, podrían proponer otro tipo de reglas que promuevan la conversación en otros términos. Al mismo tiempo que se reconoce el poder de Twitter para movilizar conversaciones abiertas, la sensación generalizada indica que los avances para que sea un espacio seguro son escasos. La periodista Marina Abiuso dijo que "Twitter tiene en esto un problema. Porque fue la libertad y la posibilidad de decirlo todo lo que hizo popular a la plataforma. Fue uno de sus puntos fuertes. Y ahora no me da la impresión de que estén preocupados por morigerar las agresiones a mujeres de a pie. Lo hemos visto actuar muy rápidamente en casos muy altisonantes, en cosas que tienen que ver con figuras políticas, pero no lo hacen de la misma manera cuando se trata de particulares. No siento que haya habido un avance en esas herramientas".[38]

EFECTO SILENCIADOR

La avalancha de abusos y agresiones online contra mujeres y personas LGBTI+ genera una autocensura que lleva incluso a abandonar la conversación o, directamente, la plataforma y tiene un carácter aleccionador para otras usuarias. Referentes políticos, deportistas, periodistas, activistas, manifestaron haber cerrado su cuenta personal y haberse alejado de la red social para cuidar su salud mental.[39]

Muchas de las mujeres que hablaron con Amnistía Internacional en el marco de sus investigaciones resaltaron que no vale la pena correr el riesgo de vivir situaciones de violencia y abuso a cambio de expresarse libremente en Twitter. Cuando no abandonan su cuenta, destinan tiempo y esfuerzo a moldear sus mensajes para evitar ser atacadas, limitan lo que dicen y dejan de participar en debates de interés público. Las mujeres y las personas LGBTI+, vienen luchando históricamente por una participación igualitaria, libre y sin violencias en el espacio público y en el acceso a los derechos humanos. Esta lucha incansable se ha trasladado ahora a las redes sociales.

---

35. Entrevista realizada el 17 de septiembre de 2021. Ciudad Autónoma de Buenos Aires.

36. Entrevista realizada el 17 de septiembre de 2021. Ciudad Autónoma de Buenos Aires.

37. Entrevista realizada el 1 de octubre de 2021, Ciudad Autónoma de Buenos Aires.

38. Entrevista realizada el 17 de septiembre de 2021. Ciudad Autónoma de Buenos Aires.

39. La Nación, 3 de abril de 2021, "Inesperado: Ofelia Fernández cerró su cuenta personal de Twitter". Nota periodística disponible en: https://www.lanacion.com.ar/politica/inesperado-ofelia-fernandez-cerro-su-cuenta-personal-de-twitter-nid03042021/. *El País,* 3 de agosto de 2021, "La nadadora argentina Delfina Pignatiello se aleja de las redes sociales para cuidar su salud mental". Nota periodística disponible en: https://elpais.com/deportes/2021-08-03/la-nadadora-argentina-delfina-pignatiello-se-aleja-de-las-redes-sociales-para-cuidar-su-salud-mental.html.

6. METODOLOGÍA

Esta Tabla sintetiza todas las recomendaciones que hemos formulado a Twitter desde 2018 y las condensa en 10 recomendaciones fundamentales para evaluar la empresa.[40] Estas 10 recomendaciones se resumen en cuatro categorías generales: Transparencia, Mecanismos de denuncia, Proceso de revisión de denuncias de abusos, y Características de privacidad y seguridad Hemos optado por centrar la atención en estas cuatro categorías de cambio debido al impacto positivo que creemos que cada una de ellas puede tener en las experiencias de las mujeres en Twitter. El aumento de la transparencia es la medida más importante que Twitter puede tomar para identificar y abordar de forma adecuada los problemas derivados de su tratamiento de los abusos en su plataforma. Facilitar al máximo la denuncia de los abusos por parte de las personas usuarias y las decisiones de apelación ayuda a Twitter a colaborar directamente con quienes usan la plataforma para hacerla más segura. Mejorar sus procesos para examinar los informes de abusos permite a Twitter ser más eficiente a escala, al mismo tiempo que mantiene unos niveles más elevados de exactitud e integridad libres de sesgos. Desarrollar más características de privacidad y seguridad permite a Twitter empoderar directamente a quienes usan la plataforma para que se protejan.

Cada recomendación consta de entre uno y cuatro subindicadores distintos. A continuación determinamos si Twitter ha hecho algún progreso respecto a cada subindicador, y calificamos cada indicador como No aplicado, Trabajo en curso o Aplicado. No aplicado significa que Twitter no ha hecho ningún progreso para aplicar nuestras recomendaciones. Trabajo en curso significa que Twitter ha hecho algún progreso pero no ha aplicado plenamente nuestra recomendación. Aplicado significa que la empresa ha aplicado íntegramente nuestra recomendación. Hemos basado nuestra valoración en el examen de dos fuentes fundamentales: primero, las afirmaciones efectuadas por Twitter en correspondencia escrita con nosotros desde 2018; y en segundo lugar, la información disponible públicamente en el sitio web de Twitter, incluidas sus políticas, Informes de transparencia, blogs, tuits y páginas del Centro de ayuda. Antes de hacer pública la Tabla de puntuación, Amnistía Internacional escribió a Twitter para solicitar una actualización sobre los progresos en la aplicación de nuestras recomendaciones, y se ha reflejado la respuesta de la empresa.

Usamos subindicadores para generar una puntuación compuesta para cada recomendación. Si Twitter no ha hecho ningún progreso respecto a ninguno de los subindicadores para una recomendación concreta, calificamos esa recomendación como *No aplicado*. Si Twitter ha hecho progresos respecto a alguno de los subindicadores, calificamos esa recomendación como *Trabajo en curso*. Si Twitter ha aplicado plenamente cada subindicador, calificamos esa recomendación como *Aplicado* plenamente. Si Twitter ha hecho algún progreso respecto a algunos subindicadores pero no respecto a otros, calificamos esa recomendación como Trabajo en curso. En el contexto de las campañas públicas de sensibilización en curso, hemos examinado si estas campañas habían abordado todas las cuestiones que habíamos planteado, además de si estas campañas y los materiales relacionados estaban disponibles en idiomas distintos del inglés.

En el apartado *Explicación detallada de los indicadores* se incluye una descripción completa de cada recomendación y cada subindicador y del razonamiento en que se basa nuestra puntuación.

Nuestra intención es que estas puntuaciones sean dinámicas a medida que Twitter desarrolla su tratamiento de la violencia y los abusos contra las mujeres en su plataforma. Haremos el seguimiento de los progresos de Twitter mediante la supervisión de los Informes de transparencia, las actualizaciones de políticas, los lanzamientos de características y otros anuncios públicos, además de seguir interactuando directamente con Twitter.

También recibiríamos con agrado cualquier aportación adicional pertinente de organizaciones de la sociedad civil y personalidades académicas que trabajan en este asunto. Si desean aportar esa información, pónganse en contacto con Michael Kleinman, director de la Iniciativa Silicon Valley de Amnistía Internacional y Amnistía Internacional Estados Unidos, en mkleinman@aiusa.org; michael.kleinman@amnesty.org.

---

40. La Tabla tiene en cuenta las recomendaciones formuladas por Amnistía Internacional a Twitter en cuatro informes: *Toxic Twitter, Troll Patrol US/UK, Troll Patrol India y Corazones verdes Argentina.*

TABLA DE PUNTUACIÓN DE TWITTER

CATEGORÍA	SUBCATEGORÍA	RECOMENDACIÓN	PUNTUACIÓN
TRANSPARENCIA	Desglose	Mejorar la calidad y la eficacia de los Informes de transparencia mediante el desglose de los datos por tipos de abuso, región geográfica y situación de la cuenta verificada.	TRABAJO EN CURSO
	Moderadores de contenido	Aumentar la transparencia en cuanto al proceso de moderación de contenido mediante la publicación de datos sobre el número de moderadores empleados, los tipos de formación necesarios y el tiempo medio que tardan esas personas en responder a los informes.	NO APLICADO
	Apelaciones	Aumentar la transparencia del proceso de apelación mediante la publicación del volumen de apelaciones recibidas y los resultados de las apelaciones.	NO APLICADO
MECANISMOS DE DENUNCIA	Solicitud de característica	Desarrollar más características para reunir e incorporar aportaciones de las personas usuarias en todas las etapas del proceso de denuncia de abusos, desde el informe inicial hasta la decisión.	TRABAJO EN CURSO
	Apelaciones	Mejorar el proceso de apelación ofreciendo más orientación a las personas usuarias obre cómo funciona el proceso y cómo se toman las decisiones.	APLICADO
	Campaña pública	Seguir educando a las personas que usan la plataforma sobre los perjuicios causados a quienes son víctimas de abusos mediante campañas públicas y otras actividades de divulgación. Esto debería incluir el envío de una notificación/un mensaje a las personas usuarias que estén violando las Reglas de Twitter en relación con los efectos silenciadores y el riesgo de daños para la salud mental causados por el envío de violencia y abusos a otro usuario o usuaria en Internet.	TRABAJO EN CURSO
PROCESO DE EXAMEN DE LOS INFORMES DE ABUSOS	Transparencia	Ofrecer ejemplos más claros de qué tipos de comportamiento alcanzan el nivel de violencia y abuso y cómo evalúa Twitter las sanciones para estos tipos distintos de comportamiento.	TRABAJO EN CURSO
	Automatización	La automatización debe usarse en la moderación de contenido únicamente con estrictas salvaguardias, y siempre sujeta a criterio humano. Por tanto, Twitter debe informar de forma clara sobre cómo diseña y aplica los procesos automatizados para identificar abusos.	TRABAJO EN CURSO
CARACTERÍSTICAS DE PRIVACIDAD Y SEGURIDAD	Solicitud de característica	Proporcionar herramientas que faciliten que quienes usen la red social eviten la violencia y los abusos en la plataforma, incluidas listas compartibles de términos ofensivos y otras características adaptadas a tipos concretos de abuso que una persona denuncie.	TRABAJO EN CURSO
	Campaña pública	Educar a las personas que usan la plataforma sobre las características de privacidad y seguridad de que disponen mediante campañas públicas y otros canales de divulgación y facilitar al máximo el proceso para habilitar estas características.	TRABAJO EN CURSO

TRANSPARENCIA

1. Mejorar la calidad y la eficacia de los Informes de transparencia mediante el desglose de los datos por tipos de abuso, región geográfica y situación de la cuenta verificada.

Amnistía Internacional tuvo en cuenta cuatro indicadores distintos para evaluar los progresos de Twitter:

- Publicar el número de denuncias de conducta abusiva o perjudicial que Twitter recibe anualmente. Esto debe incluir cuántas de estas denuncias se refieren a dirigir "odio contra una raza, religión, género, casta u orientación", "acoso selectivo" y "amenaza de violencia o daño físico". En concreto, Twitter debe compartir también estas cifras para cuentas verificadas en la plataforma.[41] – **TRABAJO EN CURSO** [42]
- De los informes desglosados de abusos, publicar el número de informes que infringen —y que no infringen— las directrices de la comunidad de Twitter, por año y por categoría de abuso. En concreto, Twitter debe compartir también estas cifras para cuentas verificadas en la plataforma.[43] – **TRABAJO EN CURSO** [44]
- Publicar el número de informes de abusos que Twitter recibe anualmente que no recibieron respuesta de la empresa, desglosados por categoría de abuso denunciado y por país.[45] – **TRABAJO EN CURSO** [46]
- Publicar la proporción de personas usuarias que han formulado quejas contra cuentas en la plataforma y la proporción de estas personas que han sido objeto de quejas en la plataforma, desglosado por categorías de abuso.[47] – **NO APLICADO** [48]

Para determinar si Twitter había implementado alguno de estos cambios, examinamos su ultimo Informe de transparencia.[49] El Informe de transparencia más reciente —el Informe 18, que abarca el periodo comprendido entre julio y diciembre de 2020— sigue incluyendo la información publicada en el Informe 17, incluidos el total de cuentas objeto de acciones por abusos o acoso y conducta de odio (entre otras categorías), el número de cuentas suspendidas y el número de piezas de contenido eliminadas.[50] El Informe de transparencia 18 incluye también algunos parámetros nuevos, como las "impresiones", que captura el número de visitas que recibieron los tuits infractores antes de ser retirados,[51] y nuevos datos sobre la adopción de la autenticación de doble factor.[52] Twitter también está tomando más medidas de aplicación sobre contenidos infractores antes incluso de que reciban visitas.[53] Además, ha dado pasos recientemente

41. Amnistía Internacional, *Toxic Twitter,* cap. 8; Amnistía Internacional, Corazones Verdes, pp. 40, 44; Amnistía Internacional, *Troll Patrol India,* p. 49.

42. Este indicador no ha cambiado desde la publicación de la *Tabla de puntuación de Twitter 2020.*

43. Amnistía Internacional, *Toxic Twitter,* cap. 8.

44. Este indicador no ha cambiado desde la publicación de la *Tabla de puntuación de Twitter 2020.*

45. Amnistía Internacional, Toxic Twitter, cap. 8; Amnistía Internacional, *Troll Patrol India,* p. 49.

46. Este indicador no ha cambiado desde la publicación de la *Tabla de puntuación de Twitter 2020.*

47. Amnistía Internacional, *Toxic Twitter,* cap. 8.

48. Este indicador no ha cambiado desde la publicación de la *Tabla de puntuación de Twitter 2020.*

49. Twitter, *Informe de transparencia* 19, julio-diciembre de 2020, https://transparency.twitter.com/es_es.html (consultado por última vez el 16 de julio de 2021).

50. Véase Carta de Twitter India a Amnistía, 29 de noviembre de 2019 ("A petición de Amnistía, el Informe de transparencia incluye ahora datos desglosados sobre una serie de políticas clave y detalla el número de informes que recibimos y el número de cuentas sobre las que actuamos"); Twitter Argentina, Carta a Amnistía, 16 de enero de 2020.

51. "En total, las impresiones de tuits infractores representaron menos del 0,1% de todas las impresiones para todos los tuits globalmente desde el 1 de julio hasta el 31 de diciembre. Durante este periodo, Twitter eliminó 3,8 millones de tuits que habían violado las Reglas de Twitter, el 77% de los cuales recibieron menos de 100 impresiones antes de la eliminación, y un 17% adicional que recibió entre 100 y 1.000 impresiones. Sólo el 6% de los tuits eliminados tuvo más de 1.000 impresiones". Carta de Twitter a Amnistía, 27 de septiembre de 2021.

52. Twitter, Blog, *An update to the Twitter Transparency Center,* https://blog.twitter.com/en_us/topics/company/2021/an-update-to-the-twitter-transparency-center (consultado por última vez el 16 de julio de 2021).

53. Carta de Twitter a Amnistía, 27 de septiembre de 2021.

para establecer un proceso de verificación[54] basado en un proceso de consulta[55] sobre el borrador original de la política de verificación.[56]

Según el Informe de transparencia, un 82% más de cuentas fueron objeto de acciones, un 9% más de cuentas fueron suspendidas y un 132% más de contenidos fueron eliminados en comparación con el periodo del informe anterior. En cuanto a los abusos y el acoso, Twitter afirma que ha desplegado un aprendizaje automático más preciso y que había mejorado la detección y emprendido acciones sobre contenidos infractores, lo desembocó en un aumento del 142% de las cuentas objeto de acciones en comparación con el periodo del informe anterior.[57] En cuanto a la aplicación de la política sobre conducta de odio (que incluye los contenidos que incitan al miedo y/o los estereotipos que inspiran miedo a categorías protegidas como el género, la orientación sexual, la raza, la etnia o el origen nacional), se tomaron medidas en un 77% más de cuentas.

Estos datos son importantes e indican el progreso de Twitter hacia un mejor seguimiento y presentación de informes sobre contenido abusivo, y la información sobre medidas adoptadas para abordarlo. Ahora bien, el informe sigue sin ofrecer datos desglosados por subcategorías de tipos de abuso, no ofrece datos desglosados según el país, no proporciona datos sobre el número de informes de abusos que no recibieron respuesta de la empresa, y no ofrece datos sobre la proporción de personas usuarias que han formulado quejas. Twitter tampoco distingue las cuentas verificadas y las no verificadas.

En su respuesta de 2020, Amnistía Internacional, Twitter afirmaba: "Aunque comprendemos el valor y los motivos de los datos de país, hay matices que podrían interpretarse erróneamente, en primer lugar, que ciberdelincuentes oculten su ubicación y de ese modo puedan dar impresiones muy engañosas de cómo se manifiesta un problema, y que personas ubicadas en un país denuncien a una persona de otro país, lo cual no queda claro en los datos desglosados".[58] La respuesta completa de Twitter a este informe se incluye como anexo *infra*.

Como se explicaba en nuestra edición anterior de la Tabla, aunque la respuesta de Twitter muestra algunas de las consideraciones en liza, Amnistía Internacional no pide que Twitter proporcione datos por país sobre personas usuarias acusadas de abusos, sino que considera que Twitter debe proporcionar datos por país sobre personas usuarias que denuncian abusos, lo cual evita el problema que se plantea *supra*. Disponer de datos sobre cuántos usuarios/as de un país determinado denuncian abusos, y sobre cómo este número cambia con el tiempo, es un indicador fundamental para ayudar a determinar si las iniciativas de Twitter para abordar este problema tienen éxito en un país determinado. Twitter podría proporcionar también información contextual para corregir posibles interpretaciones erróneas de los datos.

Twitter subraya que: "A petición de Amnistía, el Informe de transparencia incluye ahora datos desglosados sobre una serie de políticas clave y detalla el número de informes que recibimos y el número de cuentas sobre las que actuamos".[59] Sin embargo, el informe de transparencia no facilita ningún dato sobre contenidos denunciados que no recibieron respuesta ni sobre el número de denuncias que se estudiaron, pero se concluyó que no violaban las directrices de la comunidad. Y así, no especifica cuántas denuncias se estudiaron realmente frente a las que fueron ignoradas.

En su carta de respuesta, dijeron que, en la fecha en que se publique la Tabla, la página de normas estará disponible en otros idiomas; reflejamos esto en nuestro análisis del indicador 10 *infra*.

---

54. Twitter, Centro de ayuda, *Acerca de las cuentas verificadas,* https://help.twitter.com/es/managing-your-account/about-twitter-verified-accounts (consultado por última vez el 16 de julio de 2021).

55. https://blog.twitter.com/es_es/topics/product/2020/ayudanos-a-dar-forma-a-nuestro-nuevo-enfoque-de-verificacion-de-

56. https://help.twitter.com/es/managing-your-account/about-twitter-verified-accounts

57. Twitter, Blog, *An update to the Twitter Transparency Center,* https://blog.twitter.com/en_us/topics/company/2021/an-update-to-the-twitter-transparency-center (consultado por última vez el 16 de julio de 2021).

58. Carta de Twitter a Amnistía, 26 de agosto de 2020.

59. Carta de Twitter India a Amnistía, 29 de noviembre de 2019; carta de Twitter Argentina a Amnistía, 16 de enero de 2020.

2. Aumentar la transparencia en cuanto al proceso de moderación de contenido mediante la publicación de datos sobre el número de moderadores empleados, los tipos de formación necesarios y el tiempo medio que tardan esas personas en responder a los informes.

Amnistía Internacional tuvo en cuenta tres indicadores distintos para evaluar los progresos de Twitter:

- Publicar el tiempo medio que tardan los moderadores en responder a los informes de abusos en la plataforma, desglosado por categoría del abuso denunciado. En concreto, Twitter debe compartir también estas cifras para cuentas verificadas en la plataforma.[60] – **NO APLICADO** [61]
- Compartir y publicar el número de moderadores de contenido que emplea Twitter, incluido el número de moderadores empleados por región y por idioma.[62] – **NO APLICADO** [63]
- Compartir qué formación reciben los moderadores para identificar la violencia de género y otras formas de violencia por razón de identidad y los abusos contra personas usuarias, así como qué formación reciben los moderadores sobre las normas internacionales de derechos humanos y la responsabilidad de Twitter de respetar los derechos de las personas usuarias en su plataforma, incluido el derecho de las mujeres a expresarse en Twitter libremente y sin miedo a sufrir violencia y abusos.[64] – **NO APLICADO** [65]

Para determinar si Twitter había implementado alguno de estos cambios, Amnistía Internacional examinó su último Informe de transparencia.[66] El informe no incluye datos sobre el tiempo medio de respuesta a los informes de abusos ni el número de moderadores de contenido empleados desglosados por región e idioma. El informe tampoco ofrece información alguna sobre la formación recibida por moderadores de contenido en relación con los abusos y la violencia por razón de género e identidad. Otras páginas de Twitter disponibles públicamente, como el Centro de ayuda, tampoco ofrecen información alguna sobre esta capacitación.

En su respuesta al anterior informe sobre la Tabla de puntuación, Twitter alegó que "medir el progreso o la inversión de una empresa en estas cuestiones tan importantes y complejas con una medición del número de personas empleadas no es un parámetro informativo ni útil y sólo sirve para afianzar aún más a las mayores empresas con los mayores recursos".[67] Pero Twitter también reconoció que sus "operaciones se vieron gravemente afectadas por la pandemia de COVID-19 durante la segunda mitad de 2020, como fue el caso del anterior periodo objeto de informe. Las diversas restricciones de cada país y los ajustes dentro de nuestros equipos debido a la COVID-19 afectaron a la eficiencia de nuestra labor de moderación de contenido y a la velocidad con la que aplicamos nuestras políticas. Aumentamos el uso del aprendizaje automático y de la automatización para tomar toda una serie de medidas sobre contenidos potencialmente engañoso o manipulador. Como muchas organizaciones —tanto públicas como privadas de todo el mundo— las alteraciones causadas por la COVID-19 repercutieron en nuestra empresa y se reflejan en algunos de los datos difundidos hoy".[68]

Amnistía Internacional cree firmemente que el número de moderadores de contenido es un indicador fundamental de la capacidad general de Twitter para responder a los informes de contenido abusivo y problemático, sobre todo en lo referente a mostrar la capacidad de Twitter —o la ausencia de ella— para

60. Amnistía Internacional, *Troll Patrol India,* p. 49.

61. Este indicador no ha cambiado desde la publicación de la *Tabla de puntuación de Twitter 2020.*

62. Twitter, Aplicación de las Reglas, enero a junio de 2020, https://transparency.twitter.com/es_es/reports/rules-enforcement.html#2020-jan-jun (consultado por última vez el 6 de julio de 2021).

63. Este indicador no ha cambiado desde la publicación de la *Tabla de puntuación de Twitter 2020.*

64. Amnistía Internacional, *Toxic Twitter,* cap. 8; Amnistía Internacional, *Troll Patrol India,* p. 49.

65. Este indicador no ha cambiado desde la publicación de la *Tabla de puntuación de Twitter 2020.*

66. Twitter, *Informe de transparencia* 19, julio-diciembre de 2021, https://transparency.twitter.com/es_es.html (consultado por última vez el 16 de julio de 2021).

67. Carta de Twitter a Amnistía, 27 de septiembre de 2021.

68. Twitter, Blog, *An update to the Twitter Transparency Center,* https://blog.twitter.com/en_us/topics/company/2021/an-update-to-the-twitter-transparency-center (consultado por última vez el 16 de julio de 2021).

abarcar informes de abusos en diferentes países e idiomas y la manera en que esto cambia con el tiempo. Incluso con inversiones en aprendizaje automático para detectar abusos en Internet, es importante tener una idea del número de moderadores humanos que examinen las decisiones automáticas. Esto es especialmente importante durante épocas de trastornos como la pandemia de COVID-19.

La tendencia a usar el aprendizaje automático para automatizar la moderación de contenido en Internet entraña también riesgos para los derechos humanos. Por ejemplo, David Kaye, ex relator especial de la ONU sobre la libertad de expresión, ha señalado: "La automatización puede aportar valor a las empresas que tienen que valorar enormes volúmenes de contenido generado por los usuarios".[69] Sin embargo, el relator especial advierte que en áreas relacionadas con asuntos que requieren un análisis del contexto, estas herramientas pueden ser de menor utilidad, o incluso problemáticas, de ahí la importancia de contar con un número suficiente de moderadores humanos. Un informe de mayo de 2021 del Center for Democracy and Technology expone con más detalle las limitaciones de los sistemas basados en algoritmos para la moderación de contenido.[70]

Twitter reconoció recientemente que "muchas personas expresaron preocupación por nuestra capacidad para aplicar nuestras normas de forma imparcial y coherente, por lo que desarrollamos un proceso de formación más largo y exhaustivo con nuestros equipos a fin de asegurarnos de que estaban mejor preparados para estudiar informes".[71] No obstante, no se ha divulgado ningún dato aún sobre cómo se forma a los moderadores de contenido, lo que confirma la necesidad de hacer pública esta información, pues de otro modo es imposible evaluar la calidad y los criterios de dicha formación.

3. Aumentar la transparencia del proceso de apelación mediante la publicación del volumen de apelaciones recibidas y los resultados de las apelaciones.

Amnistía Internacional tuvo en cuenta dos indicadores distintos para evaluar los progresos de Twitter:

- Compartir y publicar el número de apelaciones recibidas sobre informes de abusos, y la proporción de informes desestimados en este proceso, desglosados por categoría de abuso.[72] – **NO APLICADO** [73]
- Publicar información relativa a los criterios y la decisión para admitir (o no) apelaciones, año y número concreto por país de apelaciones recibidas, con resultados.[74] – **NO APLICADO** [75]

Para determinar si Twitter había implementado alguno de estos cambios, examinamos su último Informe de transparencia,[76] las páginas pertinentes del Centro de ayuda, tuits y varias cartas. El informe no ofrece ningún dato acerca de apelaciones ni de los criterios utilizados para tomar decisiones sobre las apelaciones, a pesar de que Twitter garantizaba que "seguimos estando comprometidos con la ampliación de nuestros futuros informes de transparencia con más datos detallados, datos sobre apelaciones incluidos, y esa meta sigue siendo un trabajo en curso".[77] Dicho eso, Twitter ha dado pasos para mejorar la transparencia

69. Consejo de Derechos Humanos de las Naciones Unidas, *Informe del Relator Especial sobre la promoción y protección del derecho a la libertad de opinión y de expresión,* A/HRC/38/35, 6 de abril de 2018, doc. ONU A/HRC/38/35, párr. 33, https://documents-dds-ny.un.org/doc/UNDOC/GEN/G18/096/75/PDF/G1809675.pdf?OpenElement.

70. Center for Democracy and Technology, Dhanaraj Thakur, Emma Llansó, *Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis,* https://cdt.org/insights/do-you-see-what-i-see-capabilities-and-limits-of-automated-multimedia-content-analysis/

71. Twitter, Twitter Seguro, *Actualizamos nuestras normas contra la conducta de odio,* https://blog.twitter.com/es_es/topics/company/2019/ConductasOdio (consultado por última vez el 6 de julio de 2021).

72. Amnistía Internacional, *Toxic Twitter,* cap. 8.

73. Este indicador no ha cambiado desde la publicación de la *Tabla de puntuación de Twitter 2020.*

74. Amnistía Internacional, *Troll Patrol India,* p. 49.

75. Este indicador no ha cambiado desde la publicación de la *Tabla de puntuación de Twitter 2020.*

76. Twitter, *Informe de transparencia 19*, julio-diciembre de 2021, https://transparency.twitter.com/es_es.html (consultado por última vez el 16 de julio de 2021).

77. Carta de Twitter a Amnistía, 27 de septiembre de 2021.

dentro del propio producto, en cuya aplicación las personas usuarias pueden recibir información sobre apelaciones.[78] Twitter también ha empezado a preguntar a las personas usuarias si desean pedir que se etiquete un contenido de delicado o de información errónea, así como a implementar anuncios de suspensión en la aplicación.[79] Sin embargo, esta información es específica de usuario y no proporciona información desglosada de toda la plataforma.

MECANISMOS DE DENUNCIA

4. Desarrollar más características para reunir e incorporar aportaciones de las personas usuarias en todas las etapas del proceso de denuncia de abusos, desde el informe inicial hasta la decisión.

Amnistía Internacional tuvo en cuenta tres indicadores distintos para evaluar los progresos de Twitter:

- Añadir una pregunta opcional para las personas usuarias que reciben una notificación sobre los resultados de cualquier informe sobre si están satisfechas o no con la decisión de Twitter. Twitter debe compartir y publicar anualmente estas cifras, desglosadas por categoría de abuso.[80] – **NO APLICADO** [81]
- Brindar a las personas usuarias la opción de proporcionar un número de caracteres limitado de contexto al formular denuncias de violencia o abusos para ayudar a los moderadores a comprender por qué se ha presentado una denuncia. Twitter debe someter a prueba finalmente la satisfacción de la persona usuaria respecto a los informes con contexto añadido y a los informes sin contexto añadido.[82] – **APLICADO** [83]
- Compartir información con las personas usuarias que han presentado un informe de violencia o abuso con enlaces y recursos de apoyo y sugerencias sobre la manera de afrontar cualquier efecto negativo o perjudicial.[84] – **TRABAJO EN CURSO** [85]

Para determinar si Twitter había implementado alguno de estos cambios, Amnistía Internacional examinó su último Informe de transparencia,[86] las páginas pertinentes del Centro de ayuda y varias cartas enviadas en los últimos dos años por la empresa en respuesta a nuestras solicitudes de actualizaciones.

El Centro de ayuda de Twitter sugiere que quienes denuncian abusos reciben notificaciones después del abuso, aunque no está del todo claro qué incluyen estas notificaciones además de recomendaciones de "otras medidas que [puede el usuario/a] tomar para mejorar [si] experiencia en Twitter".[87] Twitter ha anunciado recientemente que está explorando la idea de un "Centro de Seguridad: Una ventanilla única de herramientas de seguridad. Un espacio donde [las personas usuarias] puedan ver la situación de [sus] denuncias, bloqueos y actividad con el Servicio de Twitter (incluidos informaciones en su contra y si están a punto de ser suspendidas)".[88]

Sin embargo, esta información no es suficiente para determinar si Twitter permite que la persona usuaria haga comentarios directos o si esta información está personalizada para abordar adecuadamente la

78. Carta de Twitter a Amnistía, 27 de septiembre de 2021.

79. Carta de Twitter a Amnistía, 27 de septiembre de 2021.

80. Amnistía Internacional, *Toxic Twitter,* cap. 8; Amnistía Internacional, *Troll Patrol India,* p. 49.

81. Este indicador no ha cambiado desde la publicación de la *Tabla de puntuación de Twitter 2020.*

82. Amnistía Internacional, *Toxic Twitter,* cap. 8.

83. Este indicador no ha cambiado desde la publicación de la *Tabla de puntuación de Twitter 2020.*

84. Amnistía Internacional, *Toxic Twitter,* cap. 8.

85. Este indicador no ha cambiado desde la publicación de la *Tabla de puntuación de Twitter 2020.*

86. Twitter, *Informe de transparencia 19*, julio-diciembre de 2021, https://transparency.twitter.com/es_es.html (consultado por última vez el 16 de julio de 2021).

87. https://help.twitter.com/es/safety-and-security/report-abusive-behavior

88. https://twitter.com/tapatinah/status/1375224390430961666?s=20

preocupación individual de la persona usuaria. Aun cuando la plataforma sí recopile estos datos, la información no aparece en el último Informe de transparencia.[89]

En sendas cartas que nos remitió el 29 de noviembre de 2019 y el 16 de enero de 2020, Twitter afirmó que había mejorado su flujo de presentación de denuncias concediendo a las personas usuarias la opción de añadir contexto adicional antes de presentar una denuncia. La página correspondiente del Centro de ayuda confirma que Twitter permite a los usuarios publicar tuits adicionales.[90] Twitter permite también que quienes usan la plataforma aporten contexto adicional mediante la elección entre varias opciones preseleccionadas (por ejemplo, se les pregunta: "¿De qué manera es abusivo o perjudicial este tuit?", y las personas usuarias pueden elegir entonces opciones como "Es poco respetuoso u ofensivo", "Incluye información privada", "Incluye acoso selectivo" y "dirige odio a una categoría protegida (por ejemplo, raza, religión, género, orientación, discapacidad" etc.).[91] Además, Twitter proporciona ahora "aviso en tiempo de la medida adoptada contra los tuits denunciados".

Una página del Centro de ayuda facilita también información adicional sobre la denuncia de contenido sensible.[92] Para la pregunta "¿Qué problema tienes?" se pueden seleccionar las opciones: "Una cuenta me acosa a mí o a otra persona", "Una cuenta incita el odio contra una categoría protegida, como raza, religión, orientación, sexo, discapacidad u otra categoría" o "Una cuenta amenaza con violencia o daño físico",[93] entre otras.

En una carta remitida el 12 de diciembre de 2018,[94] Twitter nos informó de que ahora enviaba "notificaciones de seguimiento a las personas que denuncian abusos" y "recomendaciones de acciones adicionales que se pueden emprender para mejorar la experiencia, como usar la característica bloquear o silenciar". En la carta que nos envió el 29 de noviembre de 2019,[95] Twitter informó de que las personas usuarias ya no veían los tuits que habían denunciado. El Centro de Ayuda tiene una "Guía de estilo para la selección de contenidos"[96] que ofrece opciones para personalizar la experiencia de la persona usuaria en Twitter. Aunque esto indica algún progreso, creemos que Twitter debe hacer más para proporcionar a quienes lo usan enlaces y recursos sobre la mejor manera de afrontar los efectos de experimentar violencia y abusos en la plataforma.

En su respuesta al informe anterior, Twitter señaló: "Aunque apoyamos el espíritu de esta propuesta y así lo hemos hecho en relación con el apoyo a las víctimas que tienen un solo correo electrónico con los recursos necesarios para trasladar las denuncias de amenazas violentas a agentes de la ley, no está claro cómo puede implementarse esto en gran escala, en todas las políticas de Twitter. En el caso de una única política, podría haber disponible una gran variedad de diferentes cuestiones, potencialmente con cientos de organizaciones asociadas pertinentes". Twitter aclaró también que su "flujo de informes y notificaciones internas se traducen a 42 idiomas principales".[97]

En su respuesta al informe actual, Twitter señaló: "Mejorar la experiencia de la denuncia es un esfuerzo en curso. Como han indicado ustedes en su carta, estamos trabajando en un centro de denuncias y esperamos tener más que compartir muy pronto. Hemos relanzado recientemente nuestro Centro de Ayuda en todos los idiomas que ofrecemos para facilitar que personas de todo el mundo denuncien contenido. En el Centro de

89. Twitter, *Aplicación de las Reglas*, enero a junio de 2020, https://transparency.twitter.com/es_es/reports/rules-enforcement.html#2020-jan-jun (consultado por última vez el 6 de julio de 2021).

90. Twitter, Centro de ayuda, *Denunciar comportamientos abusivos*, https://help.twitter.com/es/safety-and-security/report-abusive-behavior (consultado por última vez el 6 de julio de 2021).

91. Twitter, *Denunciar comportamientos abusivos*, https://help.twitter.com/es/safety-and-security/report-abusive-behavior (consultado por última vez el 24 de agosto de 2020).

92. Twitter, Centro de ayuda, *La seguridad en Twitter y el contenido sensible*, https://help.twitter.com/es/forms/safety-and-sensitive-content/abuse (consultado por última vez el 6 de julio de 2021).

93. https://help.twitter.com/es/forms/safety-and-sensitive-content/abuse

94. Carta de Twitter Estados Unidos a Amnistía, 12 de diciembre de 2018.

95. Carta de Twitter India a Amnistía, 29 de noviembre de 2019.

96. Twitter, Centro de ayuda, *Guía de estilo para la selección de contenidos,* https://help.twitter.com/es/rules-and-policies/curationstyleguide (consultado por última vez el 6 de julio de 2021).

97. Correo electrónico de Twitter a Amnistía, 25 de agosto de 2020.

ayuda también exponemos con claridad nuestras opciones de cumplimiento, que ofrecen una orientación detallada sobre el cumplimiento y cómo que se evalúan las sanciones".[98] Twitter también indicó que "apoya a organizaciones que proporcionan asistencia a personas y organizaciones que buscan ayuda de emergencia rápida".[99]

5. Mejorar el proceso de apelación ofreciendo más orientación a las personas usuarias sobre cómo funciona el proceso y cómo se toman las decisiones.

Amnistía Internacional tuvo en cuenta un indicador para evaluar los progresos de Twitter:

- Proporcionar orientaciones claras a todas las personas que usan Twitter para que apelen contra cualquier decisión relativa a informes de abusos y estipular claramente en sus políticas cómo funcionará este proceso.[100] – **APLICADO** [101]
- En la actualidad, Twitter alerta a las personas usuarias de que "Nuestro equipo de soporte está experimentando algunos retrasos en las revisiones y respuestas en este momento, pero le recomendamos que informe todos los problemas potenciales".[102] Sin embargo, antes de la pandemia, un tuit publicado por @TwitterSafety el 2 de abril de 2019 confirmaba que Twitter había mejorado enormemente su proceso de apelación con el lanzamiento de un proceso de apelación interno y la mejora en un 60% de su tiempo de respuesta a las apelaciones. Twitter confirmó también esta característica en la carta que nos remitió el 29 de noviembre de 2019.[103] Twitter describe su proceso de apelación en su Centro de ayuda, en el apartado "Ayuda con las cuentas bloqueadas o limitadas".[104]

6. Seguir educando a las personas que usan la plataforma sobre los perjuicios causados a quienes son víctimas de abusos mediante campañas públicas y otras actividades de divulgación.

Amnistía Internacional tuvo en cuenta dos indicadores distintos para evaluar los progresos de Twitter:

- Llevar a cabo campañas y sensibilización públicas entre las personas usuarias sobre los efectos nocivos para los derechos humanos de experimentar violencia y abusos en la plataforma, especialmente la violencia y los abusos dirigidos a mujeres y/o grupos marginados. Esto debería incluir el envío de una notificación/un mensaje a quienes estén violando las Reglas de Twitter sobre el efecto silenciador y el riesgo de daños para la salud mental causados por el envío de violencia y abusos a otra persona usuaria.[105] – **TRABAJO EN CURSO** [106]
- Crear campañas públicas sobre Twitter en las que se anime a quienes usan la plataforma a utilizar los mecanismos de denuncia en nombre de otras personas que experimentan violencia y abusos. Esto puede ayudar a promover y reiterar el compromiso de Twitter de poner fin a la violencia y los abusos en la plataforma y reconocer la carga emocional que el proceso de denuncia puede tener para quienes experimentan abusos en la plataforma.[107] – **TRABAJO EN CURSO** [108]

---

98. Carta de Twitter a Amnistía, 27 de septiembre de 2021.
99. Carta de Twitter a Amnistía, 27 de septiembre de 2021.
100. Amnistía Internacional, *Toxic Twitter,* cap. 8.
101. Este indicador no ha cambiado desde la publicación de la *Tabla de puntuación de Twitter 2020.*
102. https://help.twitter.com/forms/general
103. Carta de Twitter India a Amnistía, 29 de noviembre de 2019.
104. Twitter, *Ayuda con las cuentas bloqueadas o limitadas,* https://help.twitter.com/es/managing-your-account/locked-and-limited-accounts (consultado por última vez el 6 de julio de 2021).
105. Amnistía Internacional, *Toxic Twitter,* cap. 8.
106. Este indicador no ha cambiado desde la publicación de la *Tabla de puntuación de Twitter 2020.*
107. Amnistía Internacional, *Toxic Twitter,* cap. 8.
108. Este indicador no ha cambiado desde la publicación de la *Tabla de puntuación de Twitter 2020.*

En noviembre de 2019, Twitter lanzó la campaña Twitter Safety Program. Twitter también ha lanzado recientemente el sitio rules.twitter.com para ofrecer información adicional sobre el cumplimiento de sus reglas. En su respuesta a este informe, Twitter afirmó: "Este nuevo recurso está incluido en los correos electrónicos enviados a las personas que se unen a Twitter así como enlaces a nuestro enfoque de la elaboración y aplicación de políticas que detalla factores considerados por los equipos de examen al determinar las acciones de cumplimiento".

En una carta de fecha 16 de enero de 2020, Twitter hacía referencia a la firma reciente en México de un pacto con varias partes interesadas del ámbito académico, la sociedad civil, UNESCO y otras alianzas internacionales, para abordar la violencia por razón de género en México.[109] En agosto de 2020, Twitter afirmó en otra carta que había "lanzado un comando de búsqueda específico de violencia por razón de género para líneas telefónicas de emergencia y apoyo en lenguas locales en ocho mercados de Asia: India, Indonesia, Malasia, Filipinas, Tailandia, Singapur, Corea del Sur y Vietnam".[110] Twitter ha publicado también vídeos en los que explica a las personas usuarias cómo denunciar contenidos problemáticos.[111]

En el Informe de Impacto Global 2020 (publicado en abril de 2021), Twitter ofrece un panorama de sus actividades relativas a la responsabilidad social corporativa.[112] Cabe destacar que la empresa anunció que daría subvenciones a socios sin fines de lucro para "sensibilizar sobre la violencia de género ante los casos aparecidos en lo que se conoce como 'pandemia en la sombra'".[113] Twitter también se asoció con partes interesadas globales y locales en nuevos mercados para ampliar los avisos del servicio de notificación de Twitter #ThereisHelp, que, según informes, ahora facilita información sobre violencia de género en 24 mercados.[114] De forma similar, Twitter lanzó eventos globales el #DíaDelInternetSeguro 2021 que incluyeron formación y presentaciones sobre seguridad.

Aunque estas iniciativas que reconocen específicamente los perjuicios por motivos de género son loables, en general se limitan a los esfuerzos sobre diversidad e inclusión, el apoyo económico a campañas u organizaciones dirigidas por mujeres y a tratar de problemas de violencia de género sin reconocer que esta violencia está extendida en la propia plataforma. Todas estas iniciativas pueden servir sin duda para sensibilizar sobre los perjuicios de los abusos y la violencia en la plataforma, pero creemos que Twitter debe hacer aún más, especialmente para abordar los perjuicios por motivos de género. Por ejemplo, Twitter no ha implementado todavía una característica para notificar a las personas usuarias que estén violando las Reglas de Twitter el efecto silenciador y el riesgo de daños para la salud mental causados por el envío de contenidos violentos o abusivos a otra persona que usa la red. En una publicación del blog de noviembre de 2020, el equipo de Política pública de Twitter se enorgullecía de que "los derechos de las mujeres han dominado las conversaciones en Twitter [en 2019] con 40 millones de tuits hasta ahora y suma y sigue".[115] Sin embargo, en la publicación no se hacía ningún reconocimiento o referencia a los abusos por motivos de género ni al discurso de odio en la propia plataforma de Twitter.

Otra página del Centro de ayuda de Twitter ofrece algunas orientaciones sobre cómo ayudar a alguien de quien el usuario o usuaria sabe que está sufriendo abusos en línea.[116] No obstante, Twitter debería tomar más medidas para alentar a las personas usuarias a que denuncien contenido perjudicial en nombre de

109. Carta de Twitter Argentina a Amnistía, 16 de enero de 2020.

110. Twitter confirmó esto en su carta a Amnistía de 27 de septiembre de 2021.

111. Twitter, *How to use Twitter | Reporting Abusive Behavior,* https://www.youtube.com/watch?v=HUEjPiCDaDk (consultado por última vez el 6 de julio de 2021).

112. https://about.twitter.com/content/dam/about-twitter/en/company/global-impact-2020.pdf

113. https://blog.twitter.com/en_us/topics/company/2020/our-work-to-combat-the-shadow-pandemic.html

114. https://blog.twitter.com/en_us/topics/company/2020/our-work-to-combat-the-shadow-pandemic.html

115. Twitter, Blog, Twitter Public Policy, *Our work to combat the 'shadow pandemic',* https://blog.twitter.com/en_us/topics/company/2020/our-work-to-combat-the-shadow-pandemic.html (consultado por última vez el 6 de julio de 2021).

116. Twitter, Centro de ayuda, Cómo ayudar a alguien que sufre abuso en línea, https://help.twitter.com/es/safety-and-security/helping-with-online-abuse (consultado por última vez el 6 de julio de 2021).

otras personas que sufren violencia y abusos, lo que incluye alentar expresamente a las personas usuarias a que denuncien abusos en nombre de otra persona.

Twitter también se ha comprometido en el marco de la World Wide Web Foundation a poner fin a la violencia de género en Internet, y ha afirmado en su carta más reciente que ofrecería más actualizaciones sobre esta iniciativa en los próximos meses.[117]

PROCESO DE EXAMEN DE LOS INFORMES DE ABUSOS

7. Ofrecer ejemplos más claros de qué tipos de comportamiento alcanzan el nivel de violencia y abuso y cómo valora Twitter las sanciones para estos tipos distintos de comportamiento.

Amnistía Internacional tuvo en cuenta dos indicadores distintos para evaluar los progresos de Twitter:

- Compartir ejemplos específicos de violencia y abusos que Twitter no tolerará en su plataforma para demostrar y comunicar a las personas usuarias cómo pone en práctica sus políticas.[118] – **APLICADO** [119]
- Compartir con las personas usuarias la manera en que los moderadores deciden las sanciones adecuadas cuando las personas usuarias de cuentas han violado las Reglas de Twitter.[120] – **TRABAJO EN CURSO** [121]

Para determinar si Twitter había implementado alguno de estos cambios, Amnistía Internacional recurrió a cartas de Twitter y a anuncios públicos de actualizaciones de política y prácticas recientes.

En una carta de fecha 29 de noviembre de 2019, Twitter nos notificó que había actualizado su flujo de denuncias "para ofrecer más detalle de lo que Twitter define como 'categoría protegida'", y que había renovado las Reglas de Twitter en junio de 2019 para simplificarlas y para añadir "detalles como ejemplos, instrucciones paso a paso sobre la manera de denunciar, y . lo que ocurre cuando Twitter toma medidas".[122] Un tuit de @TwitterSafety el 6 de junio de 2019 confirma que esta renovación de las Reglas tuvo lugar realmente. En marzo de 2020, Twitter amplió la política para incluir la edad, la discapacidad y la enfermedad.[123] En diciembre de 2020, Twitter anunció que había revisado de nuevo su política sobre odio para incluir la casta, la religión, la raza, la etnia y el origen nacional. Sin embargo, no incluía un análisis interseccional del impacto desproporcionado que sufren las mujeres y las castas y religiones con representación insuficiente.[124]

Twitter ha comenzado también a proporcionar información adicional en relación con la manera en que los moderadores deciden las sanciones adecuadas, explicando los cinco factores que estas personas tienen en cuenta.[125] Son los siguientes: "El comportamiento se dirige a un individuo, grupo o categoría protegida de personas; la denuncia ha sido presentada por la persona destinataria del abuso o por alguien que lo ha detectado; el usuario tiene antecedentes de violación de nuestras políticas; la gravedad de la violación; el contenido puede ser un tema de interés público legítimo".[126]

---

117. Carta de Twitter a Amnistía, 27 de septiembre de 2021. Para el marco de la Web Foundation, consulten: webfoundation.org/2021/07/generation-equality-commitments/

118. Amnistía Internacional, *Toxic Twitter*, cap. 8; Amnistía Internacional, Corazones verdes, p. 44.

119. Este indicador no ha cambiado desde la publicación de la *Tabla de puntuación de Twitter 2020.*

120. Amnistía Internacional, *Toxic Twitter,* cap. 8.

121. Este indicador no ha cambiado desde la publicación de la *Tabla de puntuación de Twitter 2020.*

122. Carta de Twitter India a Amnistía, 29 de noviembre de 2019.

123. Twitter, Blog, Twitter Seguro, *Actualizamos nuestras reglas contra las conductas de odio,* https://blog.twitter.com/es_es/topics/company/2019/ConductasOdio

124. Twitter, Blog, Twitter Seguro, *Actualizamos nuestras normas contra la conducta de odio,* https://blog.twitter.com/es_es/topics/company/2019/ConductasOdio (consultado por última vez el 6 de julio de 2021).

125. Twitter, Centro de ayuda, *Nuestro enfoque para el desarrollo de políticas y nuestra filosofía de control del cumplimiento,* https://help.twitter.com/es/rules-and-policies/enforcement-philosophy (consultado por última vez el 6 de julio de 2021).

126. Twitter, Centro de ayuda, *Nuestro enfoque para el desarrollo de políticas y nuestra filosofía de control del cumplimiento,* https://help.twitter.com/es/rules-and-policies/enforcement-philosophy (consultado por última vez el 6 de julio de 2021).

Además, el Centro de Ayuda expone las opciones de control del cumplimiento para los tuits que violan las normas de la comunidad:[127] desde limitar la visibilidad del tuit a solicitar su eliminación o que se oculte un tuit infractor mientras se espera a que se elimine.[128] Twitter viene explorando con razón opciones alternativas al enfoque binario a la moderación de contenido de eliminar/dejar Las opciones de control del cumplimiento incluyen ahora la posibilidad de poner una etiqueta al tuit y/o incluir un mensaje de advertencia, mostrar un aviso a las personas usuarias antes de que lo compartan o pulsen "me gusta", silenciar los "Me gusta", las respuestas y los retuits; y/o facilitar un enlace a información adicional.[129] En junio de 2021, Twitter aclaró, además, que "no permitimos la negación de hechos violentos, incluidas las referencias insultantes a sucesos concretos en los que las víctimas principales fueron categorías protegidas. Esta política abarca ahora el contenido personalizados y no personalizados".[130]

En general, el Centro de ayuda es una plataforma donde las personas usuarias pueden obtener información valiosa sobre cómo mejorar su experiencia en Twitter. Twitter ha actualizado recientemente el Centro de ayuda con información adicional en inglés y se prevé que ofrezca traducciones a otros idiomas. Hay que señalar que las personas usuarias pueden saber más sobre "Cómo ayudar a alguien que sufre abuso en línea", "Qué hacer con respecto a situaciones de suicidio y daño autoinfligido en Twitter" y cómo "denunciar comportamientos abusivos".[131] También nos satisface que Twitter tenga más noticias que compartir pronto sobre su centro de denuncias.[132]

Dicho esto, Twitter debe publicar más información sobre la importancia que concede a los factores antes expuestos. También debería explicar cómo deciden los moderadores de contenido cuál de las sanciones imponen. De hecho, la información compartida actualmente es poco precisa y no da suficientes detalles sobre cómo evalúan los moderadores los criterios para decidir sobre contenidos. Lo que es importante: no hay un análisis específico de género en las políticas sobre discurso de odio en la plataforma.

8. La automatización debe usarse en la moderación de contenido únicamente con estrictas salvaguardias, y siempre sujeta a criterio humano. En consecuencia, Twitter debe informar con claridad de cómo diseña e implementa los procesos automatizados para identificar abusos.

Amnistía Internacional tuvo en cuenta un indicador para evaluar los progresos de Twitter:

- Proporcionar detalles sobre cualquier proceso automatizado que se utilice para identificar abusos en Internet contra las mujeres, detallar las tecnologías utilizadas, los niveles de exactitud, cualquier sesgo identificado en los resultados y la información acerca de cómo (si) los algoritmos están actualmente en la plataforma.[133] – **TRABAJO EN CURSO** [134]

Para determinar si Twitter había implementado alguno de estos cambios, Amnistía Internacional examinó el último Informe de transparencia[135] y otros blogs, tuits y páginas del Centro de ayuda disponibles públicamente de Twitter sobre el uso de tecnología y automatización para moderar contenidos.

Encontramos debates sobre la forma en que Twitter está usando la tecnología basada en algoritmos para tomar medidas sobre contenido problemático en mayor escala y con mayor rapidez, por ejemplo, para

127. Twitter, Centro de ayuda, *Nuestras opciones de control del cumplimiento,* https://help.twitter.com/es/rules-and-policies/enforcement-options (consultado por última vez el 6 de julio de 2021).

128. https://help.twitter.com/es/rules-and-policies/enforcement-options

129. https://techcrunch.com/2021/07/01/twitter-colorful-misinformation-labels/?guccounter=1

130. https://twitter.com/TwitterSafety/status/1399863969246957568?s=20

131. https://twitter.com/TwitterSupport/status/1407392249097318402?s=20

132. Carta de Twitter a Amnistía, 27 de septiembre de 2021.

133. Amnistía Internacional, *Troll Patrol India,* p. 49; Amnistía Internacional, *Corazones Verdes,* pp. 33, 44.

134. Esto se actualizó de "No aplicado" en la *Tabla de puntuación de Twitter 2020.*

135. Twitter, *Informe de transparencia 18,* julio-diciembre de 2021, https://transparency.twitter.com/es_es.html (consultado por última vez el 16 de julio de 2021).

combatir la información engañosa durante la actual pandemia de COVID-19.[136] Como ya se ha mencionado, Twitter afirma que ha desplegado un aprendizaje automático más preciso y que había mejorado la detección y emprendido acciones sobre los abusos y el acoso en su plataforma, lo había desembocado en un aumento del 142% de las cuentas objeto de acciones en comparación con el periodo del informe anterior.[137] En su carta más reciente a Amnistía, Twitter decía que, actualmente, "el 65% del contenido abusivo sobre el que actúa aflora proactivamente para revisión humana en lugar de basarse en informes de personas que usan Twitter".[138]

Con anterioridad, en su respuesta a la primera versión de este informe, Twitter había afirmado que se basaba en el "cumplimiento automático cuando la violación de la política es de índole más seria (por ejemplo, explotación sexual de niños y niñas, contenidos extremistas violentos)" y cuando ha valorado que puede hacerlo "con gran exactitud". También dijo que no "suspende de forma permanente las cuentas basándose únicamente en nuestros sistemas automáticos de cumplimiento y seguirá buscando oportunidades de incorporar controles de revisión humanos cuando sean los de mayor efecto".

Twitter anunció recientemente la ampliación y crecimiento de su equipo de Ética, Transparencia y Responsabilidad del Aprendizaje Automático (META).[139] Actualmente, la información disponible públicamente en relación con el trabajo de este equipo sigue siendo muy poco precisa. Expone tres metas: investigar y comprender el impacto de las decisiones del Aprendizaje Automático; aplicar lo aprendido para mejorar Twitter y solicitando comentarios.[140] Un producto disponible públicamente es el análisis de sesgo racial y de género del algoritmo de recorte de imágenes de Twitter, y la evaluación de su compatibilidad con la capacidad de las personas usuarias para hacer sus propias elecciones.[141]

Sin embargo, Twitter no ha proporcionado aún suficiente transparencia respecto de la moderación de contenido a través de algoritmos. Cuando se redactan estas líneas, Twitter ha ofrecido sólo información limitada sobre la forma en que selecciona el contenido y la forma en que las personas usuarias pueden seleccionar su propia página de inicio en estas Preguntas frecuentes sobre los resultados de la búsqueda.[142] Twitter sólo comparte información básica sobre las sugerencias de cuentas y los "momentos" de Twitter.[143] No facilita información muy necesaria ni conjuntos de datos ni modelos, ni hay información pública sobre los esfuerzos de Twitter para monitorear la precisión y los sesgos al abordar los abusos contra mujeres.

En su carta más reciente a Amnistía Internacional, Twitter reconocía "los riesgos potenciales de la automatización" y aseguraba que "seguirá equilibrando esto con salvaguardias y revisión humana como parte de su estrategia general".[144] Twitter también afirma que "apoya el espíritu de los Principios de Santa Clara sobre la transparencia y la responsabilidad en la moderación de contenidos y que se compromete a compartir información más detallada en informes futuros sobre cómo hace cumplir las Reglas de Twitter".[145] Sin embargo, no hay evidencia en la actualidad de que se estén aplicando internamente estos principios.

---

136. Carta de Twitter India a Amnistía, 29 de noviembre de 2019 ("Más del 50% de los tuits objeto de acciones por abusos afloraron gracias a la tecnología, con lo que se redujo la responsabilidad de las personas que pueden experimentar abusos y acoso de informarnos").

137. Twitter, Blog, *An update to the Twitter Transparency Center,* https://blog.twitter.com/en_us/topics/company/2021/an-update-to-the-twitter-transparency-center (consultado por última vez el 16 de julio de 2021).

138. Carta de Twitter a Amnistía, 27 de septiembre de 2021.

139. Twitter también se comprometió a "compartir más resultados públicamente" en su carta a Amnistía de 27 de septiembre de 2021.

140. Twitter, Blog, *Presentamos nuestra iniciativa de aprendizaje automático responsable,* https://blog.twitter.com/es_es/topics/product/2021/presentamos-nuestra-iniciativa-de-aprendizaje-automatico-respons (consultado por última vez el 6 de julio de 2021).

141. Twitter, Blog, *Presentamos nuestra iniciativa de aprendizaje automático responsable,* https://blog.twitter.com/es_es/topics/product/2021/presentamos-nuestra-iniciativa-de-aprendizaje-automatico-respons (consultado por última vez el 6 de julio de 2021).

142. Véase también Twitter, Centro de ayuda, *Sobre las sugerencias de cuentas de Twitter,* https://help.twitter.com/es/using-twitter/account-suggestions (consultado por última vez el 6 de julio de 2021).

143. Twitter, Centro de ayuda, *Directrices y principios de los Momentos de Twitter,* https://help.twitter.com/es/rules-and-policies/twitter-moments-guidelines-and-principles (consultado por última vez el 6 de julio de 2021).

144. Carta de Twitter a Amnistía, 27 de septiembre de 2021.

145. Twitter, Transparencia, *Aplicación de las Reglas,* https://transparency.twitter.com/es_es/reports/rules-enforcement.html#2020-jan-jun (consultado por última vez el 6 de julio de 2021).

CARACTERÍSTICAS DE PRIVACIDAD Y SEGURIDAD

9. Proporcionar herramientas que faciliten que las personas usuarias eviten la violencia y los abusos en la plataforma, incluidas listas compartibles de términos ofensivos y otras características adaptadas a tipos concretos de abuso que esas personas denuncien.

Amnistía Internacional tuvo en cuenta tres indicadores distintos para evaluar los progresos de Twitter:

- • Proporcionar herramientas que faciliten que las mujeres eviten la violencia y los abusos, como una lista de términos ofensivos clave asociados al género y otras obscenidades o difamaciones basadas en la identidad entre las cuales las personas usuarias pueden elegir al habilitar la función de filtro. Una característica adicional podría permitir que se compartan fácilmente términos clave de su lista de silenciar con otras cuentas de Twitter.[146] – **TRABAJO EN CURSO** [147]
- Ofrecer información y asesoramiento personalizados basados en la actividad personal en la plataforma. Por ejemplo, compartir consejos y orientaciones útiles sobre configuración de privacidad y seguridad cuando las personas usuarias denuncian un caso de violencia o abusos. Esto debería adaptarse a la categoría específica de abuso denunciado por esas personas. Por ejemplo, a una persona que denuncia acoso selectivo se le podría asesorar sobre la manera de protegerse contra las cuentas falsas.[148] – **TRABAJO EN CURSO** [149]
- Comunicar claramente cualquier riesgo asociado a la utilización de características de seguridad junto con formas sencillas de mitigar esos riesgos. Por ejemplo, si se enseña a las personas usuarias a silenciar las notificaciones de cuentas que no siguen, debería explicarse el riesgo de no tener conocimiento de ninguna amenaza que se dirija contra ellas desde esas cuentas, junto con formas prácticas de mitigar esos riesgos (por ejemplo, que una persona amiga vigile la cuenta de Twitter).[150] – **TRABAJO EN CURSO** [151]

Para determinar si Twitter había implementado alguno de estos cambios, Amnistía Internacional examinó las cartas recibidas de Twitter y anuncios públicos de lanzamientos de nuevas características. Además de sus características de seguridad más antiguas, como bloquear y silenciar cuentas o garantizar que la comunicación entre Twitter y las personas usuarias está encriptada, Twitter ha lanzado una serie de nuevas características de seguridad en el último par de años, como la posibilidad de ocultar respuestas a los tuits; limitar las respuestas,[152] por ejemplo, cambiando quién puede responder incluso después de haber enviado el tuit;[153] desactivar la opción de que la gente envíe reacciones con emojis y respuestas de texto a fleets con mensajes directos;[154] mejorar la posibilidad de silenciar palabras,[155] eliminar seguidores sin tener que bloquear una cuenta,[156] nuevas configuraciones de conversación que permiten a la persona usuaria

146. Amnistía Internacional, *Toxic Twitter,* cap. 8

147. Este indicador no ha cambiado desde la publicación de la *Tabla de puntuación de Twitter 2020.*

148. Amnistía Internacional, *Toxic Twitter,* cap. 8

149. Este indicador no ha cambiado desde la publicación de la *Tabla de puntuación de Twitter 2020.*

150. Amnistía Internacional, *Toxic Twitter,* cap. 8

151. Este indicador no ha cambiado desde la publicación de la *Tabla de puntuación de Twitter 2020.*

152. TechCrunch, *Twitter considers new features for tweeting only to friends, under different personas and more*, https://techcrunch.com/2021/07/01/twitter-considers-new-features-for-tweeting-only-to-friends-under-different-personas-and-more/?guccounter=1) https://techcrunch.com/2020/08/11/twitter-now-lets-everyone-limit-replies-to-their-tweets/ (consultado por última vez el 6 de julio de 2021).

153. Twitter, Twitter Support, https://twitter.com/TwitterSafety/status/1415025551773892608?s=20 (consultado por última vez el 16 de julio de 2021).

154. Twitter Support, https://twitter.com/TwitterSupport/status/1370120178919477249?s=20 (consultado por última vez el 6 de julio de 2021). Los fleets de Twitter son una característica que "permite a las personas usuarias de Twitter publicar fotos, vídeos, su reacción a tuits o texto sencillo a pantalla completa que desaparece a las 24 horas". S. Rodriguez, CNBC, "Twitter to kill Fleets feature, its competitor to Facebook Stories", junio de 2021, https://www.cnbc.com/2021/07/14/twitter-to-kill-fleets-feature-its-competitor-to-facebook-snapchat-stories.html#:~:t ext=Twitter%20introduced%20Fleets%20in%20November,t hat%20disappears%20after%2024%20hours.

155. Twitter Support, https://twitter.com/TwitterSupport/status/1407051178689585163?s=20 (consultado por última vez el 6 de julio de 2021).

156. Carta de Twitter a Amnistía, 27 de septiembre de 2021.

escoger quién puede responder a las conversaciones que inicia[157] y una característica "sugerencia de hacer una pausa" que pregunta a la persona usuaria si quiere revisar una respuesta que incluye un lenguaje potencialmente perjudicial u ofensivo antes de enviarlo.[158] Twitter está desplegando actualmente el Modo Seguro, que es una característica que bloquea temporalmente cuentas por usar un lenguaje potencialmente dañino o enviar respuestas o menciones repetitivas y no deseadas.[159] También ha dado prioridad a personas de comunidades marginadas y a mujeres periodistas al someter a prueba el Modo Seguro, y colaboró con organizaciones de la sociedad civil en la fase de desarrollo de este producto.[160]

En su respuesta al informe anterior, Twitter señala: "En los últimos años hemos ampliado la capacidad de las personas para controlar sus conversaciones. Además de Silenciar y Bloquear, lanzamos la posibilidad de ocultar respuestas en noviembre de 2019, y más recientemente, en agosto de 2020, lanzamos una nueva configuración de las conversaciones que permite que las personas que usan Twitter, especialmente quienes han experimentado abusos, decidan quién puede responder a las conversaciones que inician.[161] Durante el experimento inicial, Amnistía Internacional constató que esta configuración impedía un promedio de tres respuestas potencialmente abusivas, y al mismo tiempo añadía sólo un retuit potencialmente abusivo con comentario y no experimentaba un aumento de los mensajes directos no deseados. Investigaciones públicas han revelado que las personas que sufren abusos encuentran útil esta configuración".[162]

Twitter ha hecho algún progreso en la personalización de la información que ofrece a las personas usuarias que denuncian abusos. En una carta remitida el 12 de diciembre de 2018, Twitter nos comunicó que ahora proporciona "notificaciones de seguimiento a las personas que denuncian abusos, así como recomendaciones de acciones adicionales que se pueden emprender para mejorar la experiencia, como usar la característica Bloquear o Silenciar".[163] Twitter debe ir un paso más allá para adaptar este asesoramiento a la categoría específica de abuso que la persona usuaria denuncia. Por ejemplo, Twitter se ha asociado con organizaciones como Glitch, una entidad benéfica de Reino Unido que hace campaña para poner fin a los abusos en Internet contra las mujeres y promueve la ciudadanía digital, para brindar asesoramiento a activistas de Black Lives Matter.[164] Estas iniciativas deben ampliarse.

Twitter también ha hecho algunos progresos en la mejora de sus procesos de gestión del acceso y sistemas de autenticación, y de sus capacidades de detección y supervisión. La plataforma afirma que "de forma similar a cómo detectamos proactivamente y te alertamos de una conducta sospechosa en tu cuenta para ayudarte a mantenerla segura, tenemos herramientas internas de detección y supervisión que ayudan a alertarnos de una conducta inusual o posibles intentos no autorizados de acceder a nuestras herramientas internas".[165] Además, se ha comprometido a invertir en herramientas de privacidad y seguridad, y en formación para quienes trabajan para la empresa y contratistas.[166] Desde junio de 2021, las personas usuarias tienen la opción de usar claves de seguridad como única forma de autenticación de doble factor (2FA).[167] Sin embargo, aparentemente, ninguna de estas características adopta un enfoque sensible al género.

---

157. Carta de Twitter a Amnistía, 27 de septiembre de 2021.

158. Carta de Twitter a Amnistía, 27 de septiembre de 2021.

159. Twitter, Blog, *Twitter introduce el Modo Seguro,* 1 de septiembre de 2021, https://blog.twitter.com/es_es/topics/product/2021/safety-mode (consultado por última vez el 3 de septiembre de 2021).

160. Twitter confirmó esto en su carta a Amnistía de 27 de septiembre de 2021.

161. Twitter confirmó esto en su carta a Amnistía de 27 de septiembre de 2021.

162. Carta de Twitter a Amnistía, 26 de agosto de 2020.

163. Carta de Twitter Estados Unidos a Amnistía, 12 de diciembre de 2018.

164. Twitter Reino Unido, https://twitter.com/TwitterUK/status/1277519085014847490?s=20 (consultado por última vez el 6 de julio de 2021).

165. Twitter, Blog, *Our continued work to keep Twitter secure,* https://blog.twitter.com/en_us/topics/company/2020/our-continued-work-to-keep-twitter-secure.html (consultado por última vez el 6 de julio de 2021).

166. Twitter, Blog, *Our continued work to keep Twitter secure,* https://blog.twitter.com/en_us/topics/company/2020/our-continued-work-to-keep-twitter-secure.html (consultado por última vez el 6 de julio de 2021).

167. Twitter, Blog, *Stronger security for your Twitter account,* https://blog.twitter.com/en_us/topics/product/2020/stronger-security-for-your-twitter-account (consultado por última vez el 6 de julio de 2021).

Entre las características recientes figura también el nuevo proceso para solicitar dentro de la aplicación la insignia azul de verificación que se está implementando actualmente. A fecha 20 de mayo de 2021, Twitter ofrece más claridad respecto de la eligibilidad de la verificación en nueva política basada en comentarios públicos.[168] Hay que señalar que el equipo de diseño de Twitter también anunció que está explorando características adicionales para aumentar el control y la seguridad de quienes usan la aplicación. Entre ellas figuran avisos que dan a las personas usuarias la opción de revisar su respuesta antes de publicarla si usa un lenguaje que podría ser dañino,[169] la posibilidad de que la persona usuaria se "desetiquete" y restrinja las menciones de ciertas cuentas, y otros ajustes para controlar las notificaciones y evitar que sigan escalando las menciones masivas.[170] Twitter ha anunciado también planes para "hacer experimentos en un futuro próximo para dar a las personas usuarias formas más proactivas de seleccionar su experiencia, como ofrecer un aviso anticipado sobre el tono de una conversación en la que puedan intervenir y [explorar] nuevas formas para que una persona salga de la conversación controlando quién puede @mencionarla, así como nuevas formas en que la gente puede filtrar palabras no deseadas en sus respuestas".[171]

Twitter comunica por fin los riesgos asociados a sus características de seguridad. En la carta de Twitter de agosto de 2020, señala: "Respecto a los riesgos asociados al uso de características de seguridad, decimos a la gente lo que ocurre cuando usa nuestras herramientas de seguridad, como Bloquear, Silenciar, Silenciar avanzado para palabras y hashtags, y lo que ocurre cuando una persona es bloqueada".[172] Sin embargo, Twitter no expone con claridad los riesgos o consecuencias de seleccionar determinadas opciones ni sugiere acciones específicas que pueden emprender las personas usuarias para mitigar estos riesgos.

Lamentablemente, a pesar de las características y mencionadas y de los progresos, Twitter no ha lanzado todavía las características propuestas por Amnistía Internacional en el pasado, como listas compartibles de palabras clave asociadas con obscenidades por razón de género o de otros tipos de identidad.

10. Educar a las personas que usan la plataforma sobre las características de privacidad y seguridad de que disponen mediante campañas públicas y otros canales de divulgación y facilitar al máximo el proceso para habilitar estas características.

Amnistía Internacional tuvo en cuenta un indicador para evaluar los progresos de Twitter:

- Llevar a cabo campañas y sensibilización públicas en Twitter sobre las diferentes características de seguridad que las personas usuarias pueden habilitar en la plataforma. Estas campañas podrían promoverse para las personas usuarias a través de varios canales, como posts promovidos en cuentas de Twitter, correos electrónicos y notificaciones internas de la aplicación animándolas a que aprendan a usar con confianza varias herramientas de seguridad.[173] – **TRABAJO EN CURSO** [174]

Para determinar si Twitter había implementado alguno de estos cambios, Amnistía Internacional examinó sus blogs, tuits y otros anuncios públicos recientes. Por ejemplo, la página correspondiente del Centro de

168. Twitter, Blog, *Relanzamos la verificación y lo que está por venir,* https://blog.twitter.com/es_es/topics/product/2021/relanzamos-la-verificacion-y-lo-que-esta-por-venir; y Twitter Support, https://twitter.com/TwitterSupport/status/1395403954377404417?s=20 (consultado por última vez el 6 de julio de 2021).

169. Twitter, Blog, *Tweeting with consideration*, https://blog.twitter.com/en_us/topics/product/2021/tweeting-with-consideration; and Twitter Support, https://twitter.com/TwitterSupport/status/1257717113705414658?s=20 (consultado por última vez el 6 de julio de 2021). Twitter confirmó esto en su carta a Amnistía de 27 de septiembre de 2021.

170. Twitter, Dominic Camozzi, https://twitter.com/_dcrc_/status/1404578211309056006?s=20 (consultado por última vez el 6 de julio de 2021).

171. Carta de Twitter a Amnistía, 27 de septiembre de 2021.

172. Carta de Twitter a Amnistía, 26 de agosto de 2020.

173. Amnistía Internacional, *Toxic Twitter*, cap. 8; Amnistía Internacional, *Corazones Verdes*, p. 44; Amnistía Internacional, *Troll Patrol India*, p. 49.

174. Este indicador no ha cambiado desde la publicación de la *Tabla de puntuación de Twitter 2020*.

ayuda ofrece una perspectiva de las características de seguridad claves de Twitter en breves vídeos de explicaciones y tutoriales.[175]

Twitter señaló en su respuesta a este informe que sigue "invirtiendo en campañas públicas y en sensibilización en Twitter sobre las diferentes características de seguridad". Explicó también que en julio de 2020 había concluido "una serie de experimentos para notificar a las personas dentro de la aplicación sobre herramientas de seguridad y lanzado un filtro de calidad de las notificaciones para informar sobre esta opción". Más recientemente, Twitter lanzó el podcast "I Wish I Knew", auspiciado conjuntamente por investigadores de Twitter, que comparten su experiencia de investigación, hablan de la colaboración interfuncional en la empresa y exploran cuestiones relacionadas con la investigación.[176] Twitter lanzó también un nuevo blog llamado "Common Thread."[177]

En general, Twitter debe seguir llevando a cabo este tipo de campañas y ampliando los canales a través de los cuales las promueve, lo que incluye realizar campañas en idiomas locales en los países donde aumentan los abusos contra las mujeres en la plataforma. Twitter debe seguir también encontrando nuevas formas de facilitar al máximo que las personas usuarias habiliten características de seguridad, lo que incluye ofrecer estos recursos en otros idiomas. Desde noviembre de 2021, la página de Reglas de Twitter está disponible en 42 idiomas.[178]

Twitter también ha desarrollado asociaciones con organizaciones orientadas a la justicia de género y ha creado el Consejo de Confianza y Seguridad para asesorarlas sobre asuntos relacionados con política de contenidos.[179] En asociación con ONU Mujeres y la Oficina de Derechos Humanos de la ONU, ha lanzado recientemente emojis a medida que aparecerán junto con determinados hashtags para ayudar a concienciar el Día Internacional de la Eliminación de la Violencia contra la Mujer y el Día de los Derechos Humanos.[180] Como se ha reconocido más arriba, Twitter creó el hashtag #ThereIsHelp, que sugería información de utilidad para las personas usuarias que buscaran ciertos términos relacionados con la violencia en el ámbito familiar y la violencia de género.[181] Recientemente ha ampliado la característica a 5 países más, por lo que está disponible en un total de 24 mercados y 17 idiomas. Aunque no es específico sobre el género, Twitter también ha expresado su compromiso de luchar contra el racismo y la xenofobia hacia las personas asiáticas,[182] y ha creado recursos sobre buenas prácticas para ONG acerca de protección de cuentas y herramientas de seguridad, entre otras características. Twitter también se ha comprometido con el marco de la World Wide Web Foundation para poner fin a la violencia de género online, como parte del Foro Generación Igualdad de ONU Mujeres.[183]

Sin embargo, todas las campañas mencionadas consideran la violencia basada en el género desde un ángulo externo, sin reflexionar sobre el papel del propio Twitter a la hora de facilitar los abusos online contra las mujeres. Y, lo que es importante, estas campañas no educan a las personas usuarias sobre lo que pueden hacer para impedir o reducir la violencia de género online en Twitter. Twitter ha manifestado

175. Twitter, Centro de ayuda, *Cómo estamos logrando que Twitter sea más seguro,* https://help.twitter.com/es/resources/a-safer-twitter (consultado por última vez el 6 de julio de 2021).

176. https://blog.twitter.com/en_us/topics/company/2021/i-wish-i-knew-podcast

177. Carta de Twitter a Amnistía, 27 de septiembre de 2021.

178. Véase https://help.twitter.com/es/rules-and-policies/twitter-rules

179. Twitter, Consejo de Confianza y Seguridad, https://about.twitter.com/es/our-priorities/healthy-conversations/trust-and-safety-council (consultado por última vez el 6 de julio de 2021).

180. Twitter, Blog, *Our work to combat the 'shadow pandemic',* https://blog.twitter.com/en_us/topics/company/2020/our-work-to-combat-the-shadow-pandemic.html (consultado por última vez el 6 de julio de 2021).

181. Twitter, Blog, *Our work to combat the 'shadow pandemic',* https://blog.twitter.com/en_us/topics/company/2020/our-work-to-combat-the-shadow-pandemic.html (consultado por última vez el 6 de julio de 2021).

182. Twitter, Blog, *Allyship right now: #StandForAsians,* https://blog.twitter.com/en_us/topics/company/2021/allyship-right-now-stand-for-asians.html (consultado por última vez el 6 de julio de 2021).

183. Carta de Twitter a Amnistía, 27 de septiembre de 2021.

su compromiso con “aumentar la educación y la concienciación sobre estos tipos de herramientas para las personas usuarias” y prevé “tener actualizaciones más detalladas en los próximos meses”,[184] especialmente en relación con el *Safety Playbook* en el que está trabajando actualmente para las personas usuarias que son víctimas de abusos con más frecuencia, como las mujeres.[185]

184. Carta de Twitter a Amnistía, 27 de septiembre de 2021.

185. Carta de Twitter a Amnistía, 27 de septiembre de 2021.

ANNEX: AMNESTY'S LETTER TO TWITTER

Nick Pickles, Public Policy Strategy

Cynthia Wong, Legal Director, Human Rights

Twitter, Inc.
1355 Market Street, Suite 900
San Francisco, CA 94103
United States

AMNESTY INTERNATIONAL
INTERNATIONAL SECRETARIAT
Peter Benenson House, 1 Easton Street
London WC1X 0DW, United Kingdom
T: +44 (0)20 7413 5500 F: +44 (0)20 7956 1157
E: amnestyis@amnesty.org W: www.amnesty.org

7 September 2021

Dear Nick and Cynthia,

Re: Tracking Twitter's Progress on Addressing Abuse and Violence Against Women

I am writing to provide Twitter an opportunity to respond to the findings of a forthcoming report by Amnesty International.

In March 2018, Amnesty International released *Toxic Twitter*, exposing experiences of violence and abuse experienced by women on Twitter and failures of the social media platform to uphold its responsibility to protect this group of users.

Such abuse undermines the right of women to express themselves equally, freely and without fear. As Amnesty International described in *Toxic Twitter*: "Instead of strengthening women's voices, the violence and abuse many women experience on the platform leads women to self-censor what they post, limit their interactions, and even drives women off Twitter completely." Moreover, as highlighted in *Toxic Twitter*, the abuse experienced is highly intersectional, touching women of color, women from ethnic or religious minorities, lesbian, bisexual or transgender women – as well as non-binary individuals – and women with disabilities.

Since the release of *Toxic Twitter*, Amnesty International has published a series of other reports – including the *Troll Patrol* report in December 2019, measuring violence and abuse against women on Twitter, as well as reports looking at violence and abuse against women on Twitter in India and Argentina – detailing further instances of violence and abuse against women on the platform and renewing calls for Twitter to address this urgent and ongoing issue. All of these reports concluded with concrete steps Twitter should take to fulfil its human rights responsibilities moving forward.

Company Registration: 01606776 Registered in England and Wales

In September 2020 Amnesty International published the first Twitter Scorecard. This Scorecard was designed to track Twitter's global progress in addressing abusive speech against ten indicators, covering transparency, reporting mechanisms, the abuse report review process, and enhanced privacy and security features. These indicators were developed based on recommendations that Amnesty International has made in the past regarding how Twitter can best address abusive and problematic content.

According to the 2020 Scorecard, we found that Twitter had made no progress in implementing three of the indicators, had made some progress implementing six of the indicators, and had fully implemented one of the indicators.

As you know, companies, wherever they operate in the world, have a responsibility to respect all human rights. This is an internationally endorsed standard of expected conduct. The corporate responsibility to respect requires Twitter to take concrete steps to avoid causing or contributing to human rights abuses and to address human rights impacts with which they are involved, including by providing effective remedy for any actual impacts. It also requires them to seek to prevent or mitigate adverse human rights impacts directly linked to their operations or services by their business relationships, even if they have not contributed to those impacts. In practice, this means Twitter should be assessing – on an ongoing and proactive basis – how its policies and practices impact on users' rights to non-discrimination, freedom of expression and opinion, as well other rights, and taking steps to mitigate or prevent any possible negative impacts.

We are currently preparing the second Twitter Scorecard Card, gauging Twitter's progress in addressing violence and abuse experienced by women on the platform. We have found that, compared to last year, Twitter has made relatively little progress.

Please find attached an Annex that details our analysis and findings. We would welcome any further information from Twitter to help inform this report. We would be grateful to receive your response to these points and to the analysis below by close of business September 21st; if we receive your response at a later date we may not be able to fully reflect it in the report. We will use your response in our report and campaigning materials, including using verbatim quotes. We will also publish your response on our website.

Please respond by email to mkleinman@aiusa.org.

We welcome the opportunity to continue the dialogue with Twitter on these questions.

Yours sincerely,

Michael Kleinman
Director, Silicon Valley Initiative

ANEXO: RESPUESTA DE TWITTER (disponible sólo en inglés)

Nick Pickles
Global Head of
Public Policy
Strategy,
Development &
Partnerships
@nickpickles

Twitter, Inc.
1355 Market St #900
San Francisco, CA
94103

27 September 2021

Dear Michael,

Thank you for once again sharing the findings of your Twitter Scorecard assessment regarding abuse and violence against women on Twitter.

We continue to pursue our mission to protect the health of the public conversation on Twitter and we've invested considerable resources devoted to this space. As your report highlighted, we believe we have made progress, but know that much of our work continues. We thank you for your detailed review and for this opportunity to provide you with an update on our efforts.

We maintain the belief that a one-size-fits all approach fails to take into account important distinctions between services, while solutions and investments that fall outside of your categories do not translate to an accurate representation of our progress to date. At Twitter we're committed to experimenting in public with product solutions that help address the fundamental problems our users are facing, and empowering them with controls to set their own experience. While many of these changes are not directly captured in your report scorecard, we believe these improvements will ultimately enable our most vulnerable communities to better engage in free expression without fear, a goal we share with Amnesty.

Transparency

On July 14, 2021 we launched our Twitter Transparency Report 18. As noted previously, in our new Transparency Center, we've expanded our Rules Enforcement metrics to include an increased range of policies and a more granular look at the actions we take, breaking down the total accounts actioned, the number of accounts suspended, and the number of pieces of content removed. We believe these metrics provide meaningful transparency and insight into how many accounts were actioned and which policies they violated. The most recent update illustrates that there

was a 77% increase in the number of accounts actioned for violations of our hateful conduct policy.

We're always looking for ways to share more context about our enforcement of the Twitter Rules. As you captured in your letter, we've also added new metrics, including the number of "impressions" or views, violative Tweets received prior to removal, as well as information about the adoption of two-factor authentication. In total, impressions on violative Tweets accounted for less than 0.1% of all impressions for all Tweets globally, from July 1 through December 31. During this time period, Twitter removed 3.8 million Tweets that violated the Twitter Rules; 77% of which received fewer than 100 impressions prior to removal, with an additional 17% receiving between 100 and 1,000 impressions. Only 6% of removed Tweets had more than 1,000 impressions. More broadly, as we work to remove harmful, violative content quickly and at scale, these numbers represent both our present efficiency and where improvement is needed. Our goal is to improve these numbers over time, taking enforcement action on violative content before it's even viewed.

It is important to note that we often action content for a different rule violation than that which was reported, which could lead to some of the asks in your report leading to confusion about the basis of our actions. As we have previously discussed, there is a wider issue about how we could quantify country-level data and how accurate these different calculations would be; for example, individuals can be located in one country and report Tweets sent by someone in a different country. As we stated previously, providing insight into how many accounts were actioned and which policies they violated is a cleaner and more descriptive way of documenting all known instances of abuse on Twitter.

In the area of content moderation, we have previously noted our disagreement with Amnesty's recommendation and outlined that our strategy is one that combines human moderation capacity with technology. Measuring a company's progress or investment on these important and complex issues with a measure of how many people are employed is neither an informative or useful metric, and only serves to further entrench the largest companies with the greatest resources.

Regarding appeals data, we remain committed to expanding our future transparency reports with more granular data, including appeals data, and

that goal remains a work in progress. However, in the meantime, we are striving to be more transparent with our users in other formats, such as timely transparency in the product itself. This is a key area we're investing in as we believe it is more valuable to our users to receive this information in-app rather than requiring them to reference our Transparency Report. We continue to experiment with our approach, such as prompting users if they want to appeal for a sensitive media or misinformation label directly in the product. In-app suspension banners are another way we're communicating with users when they log in, thereby ending reliance on email for these notifications; these banners also provide a link to appeal. Given these workstreams underway, we believe the Scorecard assessment for item 3 should be changed to *Work in Progress*.

Reporting Mechanisms and Abuse Report Process

Improving the experience of reporting is an ongoing effort. As you captured in your letter, we are working on a reporting center and hope to have more to share very soon.

We recently relaunched our Help Center in all supported languages to help make it easier for people globally to report content. In the Help Center we also clearly lay out our enforcement options which provide detailed guidance on enforcement and how penalties are assessed.

In addition, we support organizations that provide assistance to individuals and organizations seeking rapid response emergency help. As you noted, we have partnered with health authorities and nonprofit organizations in 27 markets to expand our #ThereIsHelp notification service. When people search terms associated with gender-based violence on Twitter, they will receive a notification with contact information for local hotlines and other resources to encourage them to reach out for help.

This June, as part of the UN Women Generation Equality Forum, we committed to the Web Foundation's framework to end online gender based violence. This pledge was the culmination of a year-long consultation process with over a hundred women focused NGOs to discuss solutions for online gender based violence. Many of the solutions proposed are projects we already have underway, and we are looking forward to sharing more updates in the coming months.

AMNISTÍA INTERNACIONAL ES UN MOVIMIENTO GLOBAL DE DERECHOS HUMANOS. LAS INJUSTICIAS QUE AFECTAN A UNA SOLA PERSONA NOS AFECTAN A TODAS LAS DEMÁS.

CONTÁCTANOS


info@amnesty.org


+44 (0)20 7413 5500

ÚNETE A LA CONVERSACIÓN


www.facebook.com/AmnestyGlobal


@AmnestyOnline

Índice: AMR 13/4721/2021
Novembre de 2021
amnesty.org

AMNISTÍA INTERNACIONAL