TABLA DE PUNTUACIÓN DE TWITTER:

SEGUIMIENTO DE LOS PROGRESOS DE TWITTER PARA ABORDAR LA VIOLENCIA Y LOS ABUSOS CONTRA LAS MUJERES EN INTERNET

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 51/2993/2020
Idioma original: Inglés

amnesty.org

Cover photo: © Getty

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, Twitter es una plataforma donde abundan la violencia y los abusos contra ellas, a menudo impunemente.

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.[1] 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.[2] 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 algunos 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 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 ellas 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, lesbianas, bisexuales o transgénero —así como a personas no binarias— 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*, 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 y medios de comunicación de todo el espectro ideológico.[3] 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.[4] En enero de 2020, Amnistía Internacional publicó una nueva investigación que medía la magnitud y la naturaleza de los abusos que sufrieron las mujeres políticas de India en Internet durante las elecciones generales celebradas en ese país en 2019.[5] La

1. 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).

2. 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).

3. 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).

4. 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).

5. 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).

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.

Amnistía Internacional publica la Tabla de Twitter en un intento de seguir pidiendo cuentas a Twitter sobre la protección de las mujeres de la violencia y los abusos en su plataforma. Esta Tabla de puntuación está 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 ha formulado con anterioridad sobre la mejor forma en que Twitter puede abordar contenidos abusivos y problemáticos.

¿QUÉ SON LA VIOLENCIA Y LOS ABUSOS CONTRA LAS MUJERES EN INTERNET

Según el Comité para la Eliminación de la Discriminación contra la Mujer, la violencia de género es "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.[6] 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.[7] 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.*[8] 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.[9] 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".[10]

6. 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).

7. Comité para la Eliminación de la Discriminación contra la Mujer, *Recomendación general num. 35 sobre la violencia por razón de género contra la mujer, por la que se actualiza la recomendación general num.* 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)

8. Comité para la Eliminación de la Discriminación contra la Mujer, *Recomendación general num. 35 sobre la violencia por razón de género contra la mujer, por la que se actualiza la recomendación general num.* 19, 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).

9. Comité para la Eliminación de la Discriminación contra la Mujer, *Recomendación general num. 35 sobre la violencia por razón de género contra la mujer*, por la que se actualiza la recomendación general num. 19, 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).

10. 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 de 2018, doc. ONU A/HRC/38/47, https://undocs.org/es/A/HRC/38/47

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.[11] 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".[12]

LAS RESPONSABILIDADES DE TWITTER EN MATERIA DE DERECHOS

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[13] La responsabilidad de las empresas de respetar 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, así como a otros derechos, y tomar medidas para mitigar o prevenir las posibles repercusiones negativas.

Tal como se refleja en la Tabla de puntuación *infra*, Twitter ha hecho algunos progresos a la hora de abordar este asunto. La plataforma ha aumentado la cantidad de información disponible a través de su Centro de ayuda[14] y sus Informes de transparencia,[15] 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. Se trata de medidas importantes, y reconocemos los esfuerzos realizados por Twitter hasta la fecha. Ahora bien, el problema persiste, y Twitter debe hacer más para que las mujeres —y todas las personas usuarias, en todas las lenguas— puedan usar la plataforma sin temor a abusos.

Actualizaremos esta Tabla de puntuación cada seis meses.

11. 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).

12. 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).

13. *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).

14. Twitter, Centro de ayuda, https://help.twitter.com/es (consultado por última vez el 24 de agosto de 2020).

15. Twitter, *Twitter Transparency Center*, https://transparency.twitter.com (consultado por última vez el 24 de agosto de 2020).

DEFINICIÓN DE CONTENIDO ABUSIVO Y PROBLEMÁTICO

CONTENIDO ABUSIVO. Tuits que promueven la violencia contra alguien por razón de su raza, origen étnico, 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.[16]

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.[17]

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.[18] 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 aplicad*o 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 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.

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

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

18. 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*.

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.

© Amnestía Internacional Australia

TWITTER'S TABLA DE PUNTUACIÓN

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.	NO APLICADO
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

EXPLICACIÓN DETALLADA DE LOS INDICADORES

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.[19] – **TRABAJO EN CURSO**
- 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.[20] – **TRABAJO EN CURSO**
- 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.[21] – **TRABAJO EN CURSO**
- 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.[22] – **NO APLICADO**

Para determinar si Twitter había implementado alguno de estos cambios, examinamos su ultimo Informe de transparencia.[23] Nos complace comprobar que el último Informe de transparencia —que abarca el periodo julio-diciembre de 2020— incluye más información que los informes anteriores, como el número total de cuentas objeto de acciones por abusos/acoso y conducta de odio (entre otras categorías), el número de informes suspendidos y el número de contenidos eliminados.[24]

Ahora bien, el informe no ofrece datos desglosados por subcategorías de tipos de abuso, no distingue entre cuentas verificadas y no verificadas, 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.

En su respuesta a 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". La respuesta completa de Twitter a este informe se incluye como anexo *infra*.

---

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

20. Amnistía Internacional, Toxic Twitter, cap. 8.

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

22. Amnistía Internacional, Toxic Twitter, cap. 8.

23. Twitter, Twitter Rules Enforcement, julio a diciembre de 2019, https://transparency.twitter.com/en/twitter-rules-enforcement.html (consultado por última vez el 25 de agosto de 2020).

24. 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.

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; en cambio, creemos 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. Asimismo, Twitter podría proporcionar también información contextual para corregir posibles interpretaciones erróneas de los datos.

Además, aunque la página de inicio del Informe de transparencia está disponible en otros idiomas como el español, los apartados específicos como Cumplimiento de las Reglas sólo están disponibles en inglés.

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 3los 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.[25] – **NO APLICADO**
- Compartir y publicar el número de moderadores de contenido que emplea Twitter, incluido el número de moderadores empleados por región y por idioma.[26] – **NO APLICADO**
- 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.[27] – **NO APLICADO**

Para determinar si Twitter había implementado alguno de estos cambios, examinamos su ultimo Informe de transparencia.[28] 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 a este informe, Twitter subrayó lo siguiente: "Nuestra estrategia combina la capacidad de moderación humana con la tecnología. Medir el progreso o la inversión de una empresa en estos asuntos importantes y complejos con una simple medición del número de personas a las que se emplea no es un elemento informativo ni útil, pues no tiene en cuenta las inversiones en aprendizaje automático, detección proactiva, avances en dotación de herramientas e infraestructuras... El uso de nuevas herramientas para abordar esta conducta desde una perspectiva de comportamiento nos permite identificar proactivamente cuentas y contenidos infractores a escala y reducir al mismo tiempo la carga sobre las personas que usan Twitter. Detectamos proactivamente 1 de cada 2 tuits que retiramos por abuso, frente a 1 de cada 5 tuits en

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

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

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

28. Twitter, *Twitter Rules Enforcement*, julio a diciembre de 2019, https://transparency.twitter.com/en/twitter-rules-enforcement.html (consultado por última vez el 25 de agosto de 2020).

2018. Esta cifra representa una considerable mejora para quienes sufren abusos, pero no se refleja en el número de moderadores empleados".

Amnistía Internacional no está de acuerdo con este análisis. 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 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.

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, 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".[29] 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.

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.[30] – **NO APLICADO**
- 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.[31] – **NO APLICADO**

Para determinar si Twitter había implementado alguno de estos cambios, examinamos su último Informe de transparencia, las páginas pertinentes del Centro de ayuda y varias cartas.[32] El informe no ofrece ningún dato acerca de apelaciones, ni de los criterios utilizados para tomar decisiones sobre las apelaciones.

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 cuatro 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.[33] – **NO APLICADO**

29. 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://www.undocs.org/es/A/HRC/38/35

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

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

32. Twitter, *Twitter Rules Enforcement*, julio a diciembre de 2019, https://transparency.twitter.com/en/twitter-rules-enforcement.html (consultado por última vez el 25 de agosto de 2020).

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

- 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.[34] – **APLICADO**
- 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.[35] – **TRABAJO EN CURSO**

Para determinar si Twitter había implementado alguno de estos cambios, examinamos su último Informe de transparencia,[36] 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 remite a quienes denuncian abusos diversas notificaciones una vez presentados los informes, pero Twitter no solicita respuestas directas de las personas usuarias para valorar su satisfacción con los resultados de los informes. Aun cuando la plataforma recopila en cierto modo estos datos, la información no aparece en el último Informe de transparencia.[37]

En sendas cartas que nos remitió el 29 de noviembre de 2019[38] y el 16 de enero de 2020,[39] Twitter afirma que ha 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. 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", etc.).[40] Además, Twitter proporciona ahora "aviso en tiempo de la medida adoptada contra los tuits denunciados". Sin embargo, Twitter sigue sin proporcionar un número de caracteres limitado para que las personas usuarias aporten contexto adicional sobre los motivos por los que presentan la denuncia.

En una carta remitida el 12 de diciembre de 2018,[41] Twitter nos informó de que ahora envía "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 otra carta enviada el 29 de noviembre de 2019,[42] nos informó de que las personas usuarias que informan de abusos reciben ahora "aviso puntual de la medida adoptada contra los tuits denunciados" y no ven ya los tuits de los que han informado. Aunque esto sugiere algún progreso, creemos que Twitter debe hacer más para proporcionar a quienes lo usan enlaces y recursos sobre la manera de afrontar los efectos de experimentar violencia y abusos en la plataforma.

En su respuesta a este informe, 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

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

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

36. Twitter, *Twitter Rules Enforcement*, julio a diciembre de 2019, https://transparency.twitter.com/en/twitter-rules-enforcement.html (consultado por última vez el 25 de agosto de 2020).

37. Twitter, *Twitter Rules Enforcement*, julio a diciembre de 2019, https://transparency.twitter.com/en/twitter-rules-enforcement.html (consultado por última vez el 25 de agosto de 2020).

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

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

40. 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).

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

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

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".[43]

5. Mejorar el proceso de apelación ofreciendo más orientación a las personas usuarias obre 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.[44] – **APLICADO**

Un tuit publicado por @TwitterSafety el 2 de abril de 2019 confirma que Twitter ha 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.[45] Twitter describe su proceso de apelación en su Centro de ayuda, en el apartado "Ayuda con las cuentas bloqueadas o limitadas".[46]

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.[47] – **TRABAJO EN CURSO**
- 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.[48] – **TRABAJO EN CURSO**

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".

Twitter ha detallado también algunas campañas específicas. En una carta de fecha 29 de noviembre de 2019, Twitter hablaba de sus iniciativas para lanzar diversas campañas centradas en la seguridad a lo largo de los años, como la campaña PositionOfStrength en India en 2016, dirigida a las mujeres; la colaboración

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

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

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

46. 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 27 de agosto 2020).

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

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

WebWonderWomen, también centrada en las mujeres; la campaña EduTweet, centrada en educadores y docentes; y "Tweesurfing", centrada en las personas "milleniales".[49] En otra 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.[50] Además, en su respuesta a este informe, Twitter afirmó 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: Corea del Sur, Filipinas, India, Indonesia, Malasia, Tailandia, Singapur y Vietnam". Twitter ha publicado también vídeos en los que explica a las personas usuarias a quién denunciar contenidos problemáticos.[51]

Todas estas iniciativas sirven 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. En concreto, 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.

Además, aunque esta página del Centro de ayuda ofrece alguna orientación sobre la manera de ayudar a una persona conocida que se ve afectada por abusos online, Twitter debe hacer más para animar a las personas usuarias a informar de los contenidos perjudiciales en nombre de otras personas que experimentan violencia y abusos, incluso animando explícitamente a quienes usan la plataforma a denunciar los abusos en nombre de otras personas.

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.[52] – **APLICADO**
- 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.[53] – **TRABAJO EN CURSO**

Para determinar si Twitter había implementado alguno de estos cambios, recurrimos a cartas de Twitter y a anuncios públicos de actualizaciones de política 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".[54] Un tuit de @TwitterSafety el 6 de junio de 2019 confirma que esta renovación de las Reglas tuvo lugar.

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. Son los siguientes: "El comportamiento se dirige a un individuo, grupo o categoría protegida

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

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

51. Twitter, *How to use Twitter | Reporting Abusive Behavior*, https://www.youtube.com/watch?v=HUEjPiCDaDk (consultado por última vez el 24 de agosto de 2020).

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

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

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

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".[55] Ahora bien, Twitter debe difundir más información sobre el grado de importancia que se concede a cada uno de estos factores, y explicar cómo los moderadores deciden entre diferentes sanciones como eliminar el tuit en cuestión y/o limitar temporalmente la capacidad del usuario para publicar nuevos tuits.

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.[56] – **NO APLICADO**

Para determinar si Twitter había implementado alguno de estos cambios, examinamos el último Informe de transparencia[57] y otros blogs y páginas del Centro de ayuda disponibles públicamente de Twitter sobre el uso de tecnología y automatización para moderar contenidos. Aunque hemos encontrado debates sobre maneras en las que Twitter utiliza la tecnología para actuar respecto a contenidos problemáticos en mayor escala y con mayor celeridad —por ejemplo, para combatir la información inexacta durante la actual pandemia de COVID-19[58] —, no hemos encontrado ningún debate público sobre el algoritmo utilizado ni sobre cómo vigila Twitter la exactitud y el sesgo, sobre todo a la hora de abordar los abusos contra las mujeres.

En su respuesta a este informe, Twitter afirmó que se basa 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". Twitter afirmó también que no "suspende de forma permanente las cuentas basándose únicamente en nuestros sistemas de cumplimiento automático y seguirá buscando oportunidades de incorporar controles de revisión humanos cuando sean los de mayor efecto".

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.[59] – **TRABAJO EN CURSO**

55. Twitter, *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 27 de agosto de 2020).

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

57. Twitter, *Twitter Rules Enforcement*, julio a diciembre de 2019, https://transparency.twitter.com/en/twitter-rules-enforcement.html (consultado por última vez el 25 de agosto de 2020).

58. 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").

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

- 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 contra ellas. 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.[60] – **TRABAJO EN CURSO**
- 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).[61] – **TRABAJO EN CURSO**

Para determinar si Twitter había implementado alguno de estos cambios, examinamos 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, Twitter ha lanzado una serie de nuevas características de seguridad en los últimos años, como la posibilidad de ocultar respuestas a los tuits. Sin embargo, 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.

En su respuesta a este informe, 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. Durante el experimento inicial constatamos 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".

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,[62] 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". 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.[63] Estas iniciativas deben ampliarse.

Twitter comunica también los riesgos asociados a sus características de seguridad. En su respuesta a este informe, Twitter 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". Sin embargo, Twitter no incluye información ni asesoramiento sobre la manera de mitigar los riesgos asociados a sus características de seguridad.

60. Amnistía Internacional, *Toxic Twitter*, cap. 8.
61. Amnistía Internacional, *Toxic Twitter*, cap. 8.
62. Carta de Twitter Estados Unidos a Amnistía, 12 de diciembre de 2018.
63. Twitter Reino Unido, https://twitter.com/TwitterUK/status/1277519085014847490?s=20 (consultado por última vez el 24 de agosto de 2020).

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..

- 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.[64] – **TRABAJO EN CURSO**

Para determinar si Twitter había implementado alguno de estos cambios, examinamos sus blogs, tuits y otros anuncios públicos recientes. Por ejemplo, el 8 de noviembre de 2019, @TwitterSafety tuiteó una campaña para educar a las personas usuarias sobre características como bloquear, silenciar y filtrar contenidos. El 5 de abril de 2020, @TwitterSupport tuiteó un hilo semejante.

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 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".

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. A tal fin, Twitter confirmó en un correo electrónico que la página rules.twitter.com se publicará en 17 idiomas adicionales a principios de septiembre, y que la página que explica su enfoque de la elaboración de políticas y su filosofía del cumplimiento está disponible actualmente en 18 idiomas.[65]

CONCLUSIÓN

Twitter sigue sin hacer lo suficiente para proteger a las mujeres de la violencia y los abusos en Internet.

Desde la publicación de *Toxic Twitter* en 2018, Amnistía Internacional ha seguido poniendo de relieve la magnitud de los abusos a los que se enfrentan las mujeres en Twitter, por ejemplo enArgentina, India, Estados Unidos y Reino Unido. Mientras tanto, las mujeres han seguido denunciando los abusos que experimentan en Twitter y la falta de respuesta adecuada de la empresa.

Los persistentes abusos que sufren las mujeres en la plataforma menoscaban su derecho a expresarse en condiciones de igualdad y libertad y sin temor. Estos abusos son muy interseccionales, y mujeres de minorías étnicas o religiosas, de castas marginadas, mujeres lesbianas, bisexuales o transgénero —así como las personas no binarias— y las mujeres con discapacidad son objeto de abusos.

Aunque la empresa ha hecho algunos progresos bien recibidos, la Tabla de puntuación de Twitter muestra cuánto queda por hacer. La finalidad de la Tabla de puntuación no es sólo hacer el seguimiento de los progresos de Twitter, sino también ofrecer recomendaciones concretas sobre medidas que Twitter debe tomar para abordar este asunto. De las 10 recomendaciones que se exponen *infra*, hasta la fecha Twitter sólo ha aplicado plenamente una. Mediante esta Tabla de puntuación, seguiremos haciendo el seguimiento de los progresos de Twitter en este asunto decisivo en el futuro.

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

65. Correo electrónico de Twitter a Amnistía, 27 de agosto de 2020.

ANNEX: TWITTER'S RESPONSE

Page 1

26 August 2020

Nick Pickles
Global Head of Public Policy
Strategy & Development

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

npickles@twitter.com
@nickpickles

Dear Michael,

Thank you for sharing the findings of your upcoming report and concerns about abuse and violence against women on Twitter. Protecting the health of the public conversation on Twitter is a priority and we continue to invest and make progress in this space.

We appreciate the detailed review citing prior correspondence and acknowledging the investment we have made to protect the health of the conversation. We've made progress in some areas but know we have more to do.

At a high level, we are concerned the Scorecard's framework does not fairly or fully capture our work. A number of items were proposed in regional Amnesty reports, and not in the global report on Twitter, so it is not clear where recommendations are for specific countries or in a global context. If Amnesty is proposing a single scorecard, we would request Amnesty similarly consolidate its recommendations accordingly.

This also has the effect of impacting the scoring where Twitter has fulfilled the request of the initial Amnesty report, but by modifying the recommendations in subsequent regional reports it is now assessed as incomplete.

In your letter, the score is not included for every section, nor is an overall framework of what the score is measured against included, or whether this analysis will extend to other services.

Finally, a number of the suggested approaches are neither relevant or appropriate for Twitter, but these are still scored. We remain concerned that a one-size-fits all approach fails to take into account important distinctions between services.

To your concerns about Twitter's increased reliance on automated content moderation during the pandemic, we rely on automated enforcement when the policy violation is of a more serious nature (e.g. child sexual exploitation, violent extremist content) and have assessed we can do so with high accuracy. We do not permanently suspend accounts based solely on our automated enforcement systems and will continue to look for opportunities to build in human review checks where they are most impactful. This process helps ensure we make the most of

available resources without changing how we evaluate and action on content as a result of COVID-19.

The fall edition of our Twitter Transparency Report (covering the January to June 2020 period) will include a number of improvements in how we define and present enforcement metrics. In the meantime, we will continue to post relevant updates on our policies and metrics in our Coronavirus response blog post.

We responded to a civil society coalition letter on this topic back in July, our response is attached for reference.

Transparency
We believe the Scorecard assessment for items 1.1 and 1.2 and some prescribed indicators are incorrect.

On August 19th we published our most recent Twitter Transparency Report (TTR) within the new Twitter Transparency Center. We now include expanded Rules Enforcement metrics that more closely align with the Twitter Rules including the Hateful Conduct policy, which covers the protected categories listed in your letter. This report expanded the number of policies covered and added more granularity on the actions we take, breaking down the total accounts actioned, the number of accounts suspended and the number of pieces of content removed.

The report card states '*not implemented*' however this data is available in the Transparency Report, which details that in the most recent reporting period July to December 2019:

- Twitter received 4,634,583 reports of hateful conduct, 3,906,683 reports of abuse and 1,722,576 reports of violent threats.
- Twitter took action on 970,109 accounts for violations of our hateful conduct policy, removing 1,445,469 pieces of content and suspending 170,994 accounts. This data is also available relating to our abuse and violent threat policies.

The new Transparency Center includes all our disclosed data in one place and allows for comparison over time. We remain committed to expanding the TTR in future with more granular data, including appeals data. We believe these metrics provide more meaningful transparency and insight into how many accounts were punitively actioned and which policies they violated.

While we have recently updated the Transparency Report, it should be noted that the previous version did include violations broken down across seven key policies (including violent threats and hateful conduct) and the number of reports received.

While we understand the value and rationale behind country-level data, there are nuances that could be open to misinterpretation, not least that bad actors hide their locations and so can give very misleading impressions of how a problem is manifesting, and individuals located in one country reporting an individual in a different country, which is not clear from aggregate data.

On content moderation, we have previously outlined to Amnesty 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 crude measure of how many people are employed is neither an informative or useful metric. It fails to take into account investments in machine learning, proactive detection, tooling and infrastructure advances, not to mention normalising a narrative that the only way to solve these challenges is to continually hire more people, an approach that risks entrenching an approach that benefits the largest and best resourced companies.

We have teams working around the world to provide timely responses and leverage technology to scale our efforts. Previously, our actions were largely predicated on people reporting accounts or content that violated the Twitter Rules before we could take action. By using new tools to address this conduct from a behavioral perspective, we're able to proactively identify violative accounts and content at scale while reducing the burden on people who use Twitter. We proactively detect 1 in 2 of the Tweets we take down for abuse, compared to one in five Tweets in 2018. This is a significant improvement for those facing abuse, but is not captured by the number of moderators employed.

Similarly, abstract measurements of time may seem useful, but how does that reflect the constant re-prioritisation of reports happening, in part based on the potential severity of harm, or our efforts to limit the impact of bad-faith reporters?

We agree on the need to increase training of moderators on hateful content, particularly identity-based hate, and regularly evaluate the efficacy of content moderation efforts. We will share more on our progress in this area in the future.

With regard to providing individuals with links and resources for support, while we support the spirit of this proposal and have done so with regards to supporting victims having a single email with the necessary resources to take reports of violent threats to law enforcement, it is unclear how this could be implemented at scale, across all of Twitter's policies. In the case of a single policy alone, there could be a vast range of different issues at hand, with potentially

hundreds of relevant partner organisations. We would welcome further discussion on how this could work in practice.

Reporting Mechanisms and Abuse Review Process
We believe the Scorecard assessment for items 6.2 and 7.2 should be revised.

We recently launched a new rules.twitter.com site on how we enforce our rules as part of the Twitter Safety Program campaign launched in November 2019 to educate people about our tools. This new resource is included in emails sent to individuals joining Twitter as well as links to our approach to policy development and enforcement which details factors considered by review teams when determining enforcement actions. When we communicate with people we serve, we do include links to find out more about the process.

Automation
As we have previously discussed, we do not take action for violations of our hateful conduct policy without human review.

Privacy and Security Features
We believe the Scorecard assessment for items 9.1, 9.3, and 10.1 should be revised.

Over the past few years we have expanded people's ability to control their conversations. Aside from Mute and Block, we launched the ability to Hide replies in November 2019 and more recently as of August 2020, we launched new conversation settings that allows people on Twitter, particularly those who have experienced abuse, to choose who can reply to the conversations they start. During the initial experiment we found that these settings prevented an average of three potentially abusive replies while only adding one potentially abusive Retweet with Comment and didn't experience a rise in unwanted Direct Messages. Public research revealed that people who face abuse find these settings helpful.

On risks associated with using safety features, we tell people what happens when they use our safety tools including Block, Mute, advanced Mute for words and hashtags, and what happens when individuals are blocked.

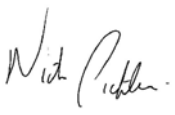
We are continuing to invest in public campaigns and awareness on Twitter about the different safety features. Last month we concluded a series of experiments that notify people in-app about our safety tools and we launched a notifications quality filter prompt to inform people about this option.

There is recent work and policy launches that speak to our commitment to reduce abuse and harassment on Twitter not included in the assessment. Some initiatives include:

- We are testing ways to prompt individuals and add a layer of friction when posting potentially hateful content and sharing articles without having accessed the content first.
- We created and disseminated a resource on the Twitter Rules on Safety and Guidelines on Abuse & Manipulation with best practices for NGOs on account protection and safety tools and will be updating to include the most recent conversation settings launch.
- We launched a dedicated gender-based violence search prompt for hotlines and support in local languages in eight Asia Pacific markets: India, Indonesia, Malaysia, Philippines, Thailand, Singapore, South Korea, and Vietnam.
- We expanded our rules against hateful conduct to include language that dehumanizes others on the basis of religion, age, disability or disease. We plan to expand this policy and are actively consulting with human rights groups to include race, ethnicity, and national origin later in the year.
- Just last month, we updated our URL policy to limit or prevent the spread of URL links to content outside Twitter that promotes violence against, threatens or harasses other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.
- When receiving Direct Messages, we now include the sender's profile information and indicate how the sender is connected to the receiver which can help people quickly identify potentially abusive content.

We respect the work that Amnesty International performs to bring awareness in the field of human rights and support towards vulnerable communities. We welcome further conversations on these issues to learn from your expertise and insights and would be happy to discuss these issues on a call with you and your colleagues.

Best wishes,



Nick Pickles
Global Head of Public Policy Strategy and Development

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 51/2993/2020
September 2020
amnesty.org

AMNISTÍA INTERNACIONAL