unesco

UNITED NATIONS OFFICE ON GENOCIDE PREVENTION AND THE RESPONSIBILITY TO PROTECT

Hacer frente al discurso de odio en las redes sociales: desafíos contemporáneos

Documento de debate

01

El presente documento ha sido elaborado por investigadores del Oxford Internet Institute, con el apoyo de la UNESCO, como contribución a la Estrategia y Plan de Acción de las Naciones Unidas para la Lucha contra el Discurso de Odio y en el marco del proyecto "#CoronavirusFacts: lucha contra la "desinfodemia" en materia de COVID 19 en entornos propensos a conflictos", financiado por la Unión Europea. Forma parte de la colaboración entre el Oxford Internet Institute y la UNESCO para desarrollar una herramienta que englobe los métodos y recursos existentes, así como los proyectos de investigación que se han creado para vigilar la existencia, la propagación y el impacto del discurso de odio en línea y evaluar las capacidades y prácticas para combatirlo. Se agradecerá toda observación en relación con este documento de discusión que contribuya a la ampliación de este estudio.

Para hacer frente y combatir el discurso de odio es necesario llevar a cabo un esfuerzo a diversos niveles que incluya abordar sus causas profundas y sus factores, impedir que desemboque en violencia y afrontar sus consecuencias sociales más amplias. Para generar respuestas eficaces al discurso de odio, por ejemplo mediante la educación, es esencial mejorar la vigilancia y el análisis del fenómeno a partir de datos claros y fiables. En esta era digital, también es necesario tener una mejor comprensión de la incidencia, la virulencia y el alcance del discurso de odio en línea.

Existen diversos desafíos a la hora de identificar el discurso de odio para fines de investigación. Por un lado, desde una perspectiva metodológica, tenemos las definiciones utilizadas para considerar la cuestión, los contextos históricos y sociales, las sutilezas lingüísticas, la variedad de las comunidades en línea y las formas del discurso de odio en línea (tipo de lenguaje, imágenes, etc.). Por otro lado, desde la perspectiva tecnológica, el discurso de odio es difícil de estudiar debido a la incoherente fiabilidad de los sistemas de detección, la naturaleza opaca de los algoritmos propietarios, la falta de acceso a los datos de que disponen las empresas, etc. Es indispensable tener una idea clara de cómo hacer frente a estos desafíos para poder comprender de qué manera aparece y se propaga el discurso de odio en línea y posteriormente formular respuestas eficaces.

La Estrategia y Plan de Acción de las Naciones Unidas identifica una serie de áreas prioritarias para vigilar y analizar el discurso de odio y estipula que las «entidades pertinentes de las Naciones Unidas deben estar en condiciones de reconocer, vigilar y analizar las tendencias relativas al discurso de odio y recopilar datos sobre ellas». En materia de discurso de odio en línea, se alienta a las entidades de las Naciones Unidas a «llevar a cabo más investigaciones sobre la relación entre el uso indebido de Internet y las redes sociales para difundir el discurso de odio y los factores que impulsan a las personas a cometer actos de violencia», así como a «señalar los riesgos y oportunidades relativos a la propagación del discurso de odio que comportan las nuevas tecnologías y las plataformas digitales». Y, por último, «a definir protocolos de actuación que tengan en cuenta las nuevas formas digitales del discurso de odio».

Durante el último año, la pandemia de COVID-19 ha puesto de manifiesto la pertinencia de la Estrategia y Plan de Acción de las Naciones Unidas para la Lucha contra el Discurso de Odio, dada la propagación del discurso de odio en todo el mundo que ha agravado la intolerancia y la discriminación hacia determinados grupos y ha desestabilizado sociedades y sistemas políticos.

Este documento de discusión pretende ofrecer una visión general de los aspectos clave que deben tenerse en cuenta para abordar la incidencia del discurso de odio en las redes sociales, ya sea mediante normativas específicas de las empresas de redes sociales, esfuerzos y leyes para combatirlo o medidas educativas preventivas. El documento está dividido en tres secciones: la parte 1 se centra en las definiciones del discurso de odio y los marcos legales conexos; la parte 2 revisa y analiza las herramientas y técnicas para vigilar el discurso de odio en línea y examina posibles maneras de medir la prevalencia de este último; y la parte 3 analiza posibles medidas para contrarrestar y prevenir este tipo de discurso.

1
DEFINICIÓN DEL DISCURSO DE ODIO

Las dificultades a la hora de abordar el discurso de odio y legislar al respecto empiezan con su definición, ya que no existe un acuerdo a nivel internacional sobre lo que significa el discurso de odio. Se evocan varias cuestiones legales, como la libertad de opinión y de expresión, la discriminación y la promoción de la discriminación o la incitación a ella, la hostilidad o la violencia.

El proyecto «**The Dangerous Speech**» de Susan Benesch[1] sugiere que existen dos dificultades principales con el término «discurso de odio». En primer lugar, «odio» es un término impreciso que puede presentar distintos niveles de intensidad y acarrear distintas consecuencias: el término «odio» de la expresión «discurso de odio» ¿significa que el hablante odia, que quiere persuadir a otros para que odien o que quiere que las personas se sientan odiadas?»[2] En segundo lugar, la expresión «discurso de odio» en esencia significa que se ataca a una persona o a un grupo a causa de su identidad o de su pertenencia a un grupo. Por ello, es necesario que la ley o la definición pertinentes especifiquen si se engloba a todas las identidades y grupos y, en caso de no ser así, a qué grupos incluiría. El proyecto «The Dangerous Speech» argumenta que las leyes demasiado generales pueden provocar un uso indebido contra grupos vulnerables o de la oposición cívica o política, lo cual puede perjudicar a esos mismos grupos que las leyes pretenden proteger. Sin embargo, también se puede argumentar que una definición que se centre casi exclusivamente en grupos e identidades específicos puede conducir a la exclusión legal o a la falta de herramientas jurídicas para abordar el problema.

Si bien el alcance del presente documento de discusión no permite examinar estos desafíos más a fondo, las distintas leyes nacionales e internacionales de todo el mundo dan fe de las complejidades e interpretaciones diversas que se pueden aplicar al discurso de odio.

A nivel global, además de la no vinculante *Declaración Universal de Derechos Humanos*, el *Pacto Internacional de Derechos Civiles y Políticos* (ICCPR) establece el derecho a la libertad de expresión en su artículo 19, seguido de la prohibición de toda apología del odio que constituya incitación a la discriminación, la hostilidad o la violencia (artículo 20). Los artículos 19 y 20 también establecen restricciones a la libertad expresión, que «deberán, sin embargo, ***estar expresamente fijadas por la ley*** y ser ***necesarias para***: **a) Asegurar el respeto a los derechos o a la reputación de los demás; b) La protección de la seguridad nacional, el orden público o la salud o la moral públicas**.»

Como complemento de estos principios, el Plan de Acción de Rabat propone «**una prueba de umbral que consta de seis parámetros**» para justificar las restricciones a la libertad de expresión, teniendo en cuenta **el contexto sociopolítico, la condición del hablante, la intención de incitar al antagonismo, el contenido del discurso, el alcance de su difusión y la probabilidad de causar daño**.

La *Convención Internacional sobre la Eliminación de Todas las Formas de Discriminación Racial* (ICERD) también trata el discurso de odio y establece una cláusula más estricta que el artículo 20 del ICCPR al no requerir intención o «apología del odio» e incluir la difusión en la lista de actos punibles. La *Convención para la Prevención y la Sanción del Delito de Genocidio* y la *Convención sobre la Eliminación de Todas las Formas de Discriminación contra las Mujer* (CEDAW) merecen también una mención a este respecto.

La organización ARTÍCULO 19, dedicada a la defensa de la libertad de expresión, ha elaborado los *Principios de Camden Sobre La Libertad de Expresión y la Igualdad*, basándose en debates con funcionarios de las Naciones Unidas y expertos académicos y de la sociedad civil. Estos Principios proporcionan una guía para interpretar los artículos del ICCPR y pretenden disuadir a los agentes de abusar del artículo 20, aclarando cuestiones relativas a la «incitación», así como aquello que constituye «discriminación», «hostilidad» y «violencia».

[1] The Dangerous Speech Project. 2021. https://dangerousspeech.org/
[2] Benesch, S. 2021. Dangerous Speech: A Practical Guide. The Dangerous Speech Project, 2021, p. 7. https://dangerousspeech.org/

Cuando las leyes y los principios internacionales se traducen en leyes nacionales, los enfoques de cada país sobre cómo definen el discurso de odio son ligeramente distintos en cuanto a la forma en que se expresa, a quién va dirigido y qué tipo de daño debe ocurrir para que el discurso se considere de odio. Y esta falta de unidad en la definición es precisamente uno de los mayores desafíos a la hora de luchar contra el discurso de odio en línea, que no se rige por las fronteras nacionales.

Las definiciones del discurso de odio también son importantes para los esfuerzos de investigación y promoción, en particular cuando se trata de identificar sus consecuencias en la sociedad. Los daños causados por el discurso de odio pueden afectar a las personas (en forma de daño psicológico), a los grupos y comunidades y a la sociedad (en forma de erosión de los derechos y los bienes públicos). Dado que el discurso de odio ataca a las personas basándose en las características del grupo al que pertenecen, el análisis del daño causado a la comunidad es de particular importancia. Los daños causados por el discurso de odio se distribuyen de manera desigual en la población en general, siendo los grupos marginados los que suelen llevarse la peor parte. Asimismo, se trata de daños cumulativos para aquellos que los sufren, y el hecho de haber experimentado anteriormente el discurso de odio se considera una variable clave para estimar el daño derivado de ser objeto de este tipo de discurso.[3]

DISCURSO DE ODIO EN LÍNEA

El discurso de odio en línea no es muy diferente del que tiene lugar fuera de línea. Su diferencia reside en las interacciones en las que sucede/se lleva a cabo, así como en el uso y la difusión de palabras, acusaciones y teorías conspiratorias específicas que pueden evolucionar, alcanzar máxima popularidad y desvanecerse muy rápidamente. Los mensajes de odio pueden hacerse virales en horas, incluso en minutos.

El informe de 2015 de la UNESCO *Countering Online Hate Speech* (Luchar contra el discurso de odio) subraya cómo se produce y propaga el discurso de odio en línea a bajo costo, sin necesidad de pasar por un proceso de edición, como sucede para otros trabajos escritos; los niveles muy distintos de exposición que experimenta en función de la popularidad de la publicación; y el hecho de poder publicarlo en distintos países, dado que los servidores y las sedes de las plataformas no tienen que estar en el mismo país que el usuario y que el público al que va destinado. El discurso de odio en línea también puede estar disponible durante más tiempo y experimentar olas de popularidad, conectar con nuevas redes o volver a aparecer, así como ser anónimo. Por ello, se ha debatido largo y tendido sobre la cuestión de quién modera los espacios en línea y de cuándo se debe eliminar un contenido en caso de que haya que hacerlo.

Este debate queda patente en leyes como **NetzDG** (relativa a la aplicación del derecho en la red) de Alemania, que se presentó en 2017 y establece que las plataformas de redes sociales con más de dos millones de usuarios deben aplicar un procedimiento transparente de moderación del contenido ilegal — en el que se incluye el discurso de odio—, eliminar contenido que se ha identificado como ilegal en un plazo de 24 horas e informar de manera periódica acerca de las medidas tomadas. Esta ley fue muy criticada por obligar a las plataformas a desempeñar un papel de «censura privatizada» respecto a decisiones que deberían depender de los juzgados, y se advirtió de que los plazos y las multas conducirían a las plataformas a una eliminación excesiva de contenidos para evitar el riesgo de recibir sanciones elevadas. En 2020 hubo una revisión de la ley en la que se requería a las plataformas de redes sociales que enviaran el contenido ilegal a la Policía Criminal Federal. En otra revisión que se llevó a cabo casi de manera simultánea, se reforzaron los derechos de los usuarios al requerir a las plataformas que el proceso de denuncia de contenidos ilegales fuera más intuitivo y al permitir la apelación de una decisión de eliminar o no una publicación.

En este sentido, la formulación de leyes para abordar el discurso de odio en línea y fuera de línea a menudo se ve plagada de complicados procesos de revisión relacionados tanto con los desafíos relativos a la definición como con la tarea de respetar la libertad de expresión en el marco de la ley. Ante estos desafíos, es imprescindible emplear métodos que vayan más allá de las medidas legales para hacer frente al discurso de odio.

[3] Gelber, K. and McNamara, L. 2016. Evidencing the harms of hate speech. Social Identities, Vol. 22, No. 3, pp. 324 341.

2
HERRAMIENTAS Y TÉCNICAS PARA VIGILAR Y MEDIR EL DISCURSO DE ODIO

Las políticas y herramientas en materia de detección, vigilancia y moderación del discurso de odio en línea varían según los contextos, los agentes y las plataformas.

Los métodos de detección se pueden agrupar, a grandes rasgos, en dos categorías: esfuerzos más integrales basados inicialmente en filtros de palabras clave y métodos de colaboración masiva, y aquellos que dependen de moderadores humanos que revisan el contenido señalado por los usuarios como discurso de odio y deciden si lo es o no. Si bien los enfoques manuales tienen la clara ventaja de captar el contexto y reaccionar con rapidez frente a nuevos acontecimientos, se trata de un proceso intensivo, prolongado y costoso que limita la escalabilidad y las soluciones rápidas. Como resultado de estos desafíos, del aumento del volumen de contenido producido en las redes sociales y de los avances en el aprendizaje automático y en las técnicas de procesamiento del lenguaje natural, plataformas y expertos han elaborado soluciones de detección automatizada que cada vez utilizan más. Muchas de las iniciativas más recientes utilizan una combinación de varios métodos. Algunas de las palabras clave de estas metodologías incluyen:

- **aprendizaje automático:** técnicas que utilizan algoritmos informáticos que pueden mejorar automáticamente a través de la experiencia y del uso de datos.
- **procesamiento del lenguaje natural:** técnicas que procesan y analizan grandes cantidades de datos de lenguaje natural.
- **enfoques basados en palabras clave:** métodos que utilizan una ontología o diccionario para identificar textos que pueden contener palabras de odio clave.
- **semántica distribucional:** métodos para cuantificar y categorizar similitudes entre palabras, expresiones y frases basándose en su distribución en grandes muestras de datos.
- **análisis de sentimientos:** métodos para explicar qué tipo de actitudes se transmiten en relación con un tema en un texto determinado.
- **metadatos de origen:** algunos métodos elaboran modelos a través de la metainformación de los datos como, por ejemplo, datos relativos a los usuarios asociados a los mensajes, que incluyen características específicas de la red como su número de seguidores.
- **aprendizaje profundo:** un tipo de algoritmo de aprendizaje automático que utiliza múltiples capas para extraer de manera progresiva características importantes de los datos brutos.

Las empresas de redes sociales han pasado de reaccionar a aquellas publicaciones señaladas por los usuarios como discurso de odio a detectar y abordar de manera proactiva este tipo de contenidos gracias a sus sistemas automatizados, incluso antes de que los usuarios lo

lleguen a ver. Si bien es necesario combatir el discurso de odio a escala, estos métodos también introducen complicaciones: la detección automatizada del discurso de odio puede resultar errónea y, por lo tanto, puede llegar a eliminar contenidos que no contengan ningún tipo de expresiones de odio. La eliminación excesiva de contenidos puede crear efectos intimidatorios y socavar la libertad de expresión.

Por ello, para mejorar los procesos de vigilancia, continuamente aparecen nuevas herramientas de detección del discurso de odio. **Perspective API**[4] , por ejemplo, una herramienta de código abierto de Jigsaw (una incubadora dentro de Google) y del equipo de tecnología de Google encargado de la tecnología para luchar contra los abusos, es utilizada por medios de comunicación y productos de Google. Emplea el aprendizaje automático para puntuar expresiones en función de su potencial de toxicidad en una conversación y está disponible en siete idiomas: alemán, español, francés, inglés, italiano, portugués y ruso. Por su parte, **Facebook** ha declarado que la última versión de su herramienta para vigilar y detectar discursos de odio en sus plataformas ha mejorado la comprensión semántica del lenguaje y la comprensión del contenido, incluido el análisis de imágenes, comentarios y otros elementos[5].

Investigadores y organizaciones de la sociedad civil también han trabajado para desarrollar herramientas que detecten el discurso de odio. He aquí algunos ejemplos:

- la plataforma keniana **Umati** fue una de las primeras en vigilar las publicaciones en línea escritas en idiomas pertinentes en Kenia.
- Davidson y otros (2017) desarrollaron **Hate Sonar** utilizando un enfoque de regresión logística basado en datos de foros de Internet y Twitter.
- ADL y D-Lab, de la Universidad de California Berkeley, desarrollaron el **Online Hate Index** (OHI), diseñado para transformar, mediante el aprendizaje automático, la comprensión humana del discurso de odio en una herramienta adaptable que se puede utilizar para el contenido de Internet con el fin de descubrir la magnitud y la propagación del discurso de odio en línea.
- el equipo del proyecto «Measures & Counter-measures» del Alan Turing Institute desarrolló una herramienta que utiliza métodos de aprendizaje profundo **para detectar prejuicios contra la cultura y las personas del este asiático** en las redes sociales.
- Moon y otros (2020) desarrollaron una herramienta de **detección del discurso de odio en coreano** en la que utilizaron una calificación de «prejuicio» además de la de «odio»;.
- **Hate Meter** detecta discursos de odio antimusulmanes mediante el uso del aprendizaje automático y de técnicas de procesamiento del lenguaje natural; la plataforma está disponible en francés, inglés e italiano.
- **COSMOS** recopila y analiza datos de Twitter en tiempo real mediante la especificación de palabras clave, utilizando análisis de sentimientos y procesamiento del lenguaje natural.
- **MANDOLA** detecta contenido de odio a través de una combinación de análisis de sentimientos, procesamiento del lenguaje natural, aprendizaje automático y aprendizaje profundo.

Es importante resaltar que la vigilancia del discurso de odio en línea depende del acceso a los datos, en especial por parte de las plataformas de redes sociales. Asimismo, muchas de las herramientas que existen en la actualidad son monolingües y suelen limitarse al inglés, por lo que es necesario llevar a cabo más investigaciones en relación con el rendimiento de métodos de detección multilingües. Por otro lado, la mayor parte de las investigaciones y de la vigilancia del discurso de odio en las plataformas de redes sociales se centra en los Estados Unidos y Europa, por lo que existe una brecha tanto en las herramientas y datos como en la comprensión de la magnitud y las dinámicas de propagación del discurso de odio en otras regiones. La naturaleza contextual del discurso de odio hace que sea de vital importancia cerrar esta brecha.

[4] Para obtener más información acerca de Perspective API puede dirigirse a: https://www.perspectiveapi.com/
[5] Fuente: https://ai.facebook.com/blog/ai-advances-to-better-detect-hate-speech/

PREVALENCIA DEL DISCURSO DE ODIO EN LAS PLATAFORMAS DE REDES SOCIALES

Gracias al uso de herramientas de detección automatizada basadas en los métodos disponibles, Twitter, Facebook, Instagram y YouTube consiguen señalar o eliminar cada vez más contenido. Entre enero y marzo de 2021, YouTube eliminó 85 247 vídeos que violaban su política relativa al discurso de odio. Sus dos informes anteriores muestran cifras similares. En ese mismo trimestre, Facebook denunció un total de 25,2 millones de elementos de contenido en relación con los cuales había tomado alguna medida, y en el caso de Instagram fueron 6,3 millones. Según el último informe de transparencia de Twitter, entre julio y diciembre de 2020 la empresa eliminó 1 628 281 elementos de contenido que se consideraba que violaban su política relativa al discurso de odio.

La prevalencia del discurso de odio en las plataformas de redes sociales se determina mediante una muestra de contenido que los usuarios visualizan. Dicho de otra manera, solamente captura (una estimación de) el discurso de odio que permanece en la plataforma después de que la empresa haya detectado y eliminado elementos de manera proactiva. Hasta el momento, Facebook es la única plataforma que publica datos sobre la prevalencia. La empresa informó de que, entre enero y marzo de 2021, se registró entre un 0,05% y un 0,06% de prevalencia del discurso de odio, una cifra ligeramente inferior a la de los dos informes anteriores. Algunos estudios indican que la prevalencia del discurso de odio en las plataformas principales, como Twitter y Wikipedia, es de menos del 1% del contenido total, mientras que en plataformas alternativas más especializadas, como Gab o 4chan, del 5% al 8% del contenido puede ser abusivo[6]. Sin embargo, la evidencia sobre la prevalencia del discurso de odio en las plataformas de redes sociales es incompleta, en parte a causa de la falta de transparencia y de acceso a los datos por parte de las plataformas.

[6] Zannettou et al. 2018; Mathew et al. 2018; Hine et al. 2017; Vidgen et al. 2019.

3
COMBATIR EL DISCURSO DE ODIO EN LÍNEA

Cabe resaltar que la lucha contra el discurso de odio —y, por extensión, la elección de las herramientas y estrategias adecuadas, así como de los esfuerzos para prevenirlo— se ve complicada por varios factores. Existe una disparidad de opiniones en las respuestas de los agentes a preguntas clave en una variedad de contextos. ¿De qué manera daña el discurso de odio y cuándo se considera que el daño es lo suficientemente grave como para justificar su regulación? Y, más en detalle, ¿qué tipo de daños vinculados a comportamientos en los que se utiliza el discurso de odio merecen normativas acordes al derecho internacional de los derechos humanos y a las regulaciones en pro de la libertad de expresión?

La arquitectura de Internet también añade desafíos singulares a la hora de combatir el discurso de odio. Entre ellos, cabe citar la permanencia, la itinerancia, el anonimato y el carácter transjurisdiccional del contenido en línea, la gran variedad de arquitecturas de las plataformas y un sistema heterogéneo de gobernanza de Internet de múltiples actores.

A pesar de estos desafíos, existe una gran cantidad de grupos e individuos que participan de distintas maneras en la lucha contra el discurso de odio en línea y refuerzan a modo preventivo la resiliencia de los usuarios de Internet en contra de este discurso.

RECURSO LEGAL-ESTATAL

El recurso legal es una vía importante para combatir el discurso de odio. A pesar de que las posiciones en relación con el discurso de odio y el discurso de odio en línea varían en función de las regiones y de que continúan evolucionando a medida que se profundiza más en el tema, existe una serie de principios internacionales, acuerdos regionales, leyes estatales y ejemplos de jurisprudencia dictada por los tribunales que están en consonancia con las normativas internacionales de derechos humanos y que incluyen cláusulas pertinentes al discurso de odio en línea y fuera de línea, tal como se ha resumido al principio de este documento de discusión.

Sin embargo, rápidamente surgen problemas relacionados con las respuestas estrictamente jurídicas para combatir el discurso de odio en línea. Por ejemplo, inquietudes relativas al equilibrio de los derechos, la posibilidad de que agentes poderosos abusen de las restricciones de derechos y el confiar en el umbral de incitación a la violencia, junto con una mala comprensión de la relación entre el discurso de odio y la violencia fuera de línea. Más importante aún para combatir el discurso de odio en línea, un problema clave para los recursos jurídicos es la limitación de la autoridad de los Estados individuales sobre los espacios digitales en línea. Hacer frente al discurso de odio en línea de manera eficaz no puede depender exclusivamente de los recursos jurídicos nacionales.[7]

En 2016, un conjunto de empresas informáticas acordó el Código de Conducta de la Comisión Europea para la Lucha contra la Incitación Ilegal al Odio en Internet, en el que se requiere a las empresas que revisen en el plazo de un día aquellos elementos que hayan sido denunciados. A pesar tratarse de un enfoque difícil, debido a la alta variabilidad en los términos de servicio y en las definiciones operacionales del discurso de odio, es también un esfuerzo importante para promover la colaboración y vincular los enfoques judiciales y extrajudiciales en el ámbito del discurso de odio.

[7] Desde 2013, la iniciativa de la UNESCO para la formación de jueces ha mejorado las competencias de los actores judiciales en materia de normas regionales e internacionales relativas a la libertad de expresión, el acceso a la información y la seguridad de los periodistas en regiones de todo el mundo, motivados en particular por el hecho de que una de las cuestiones que suscita más interés entre los operadores judiciales es la manera de tratar los casos de discurso de odio. Unos 23 000 actores judiciales han recibido capacitación sobre estas cuestiones, en particular mediante cursos en línea masivos y abiertos (MOOC), formación y talleres sobre el terreno y la publicación de una serie de manuales y directivas.

LAS RESPUESTAS DE LAS EMPRESAS INFORMÁTICAS

En 2021, tanto YouTube como Facebook informaron de un incremento del contenido encontrado y señalado por cada una de ellas, así como de una mayor proporción de contenido señalado por las empresas en comparación con el señalado por los usuarios. Esto se debe a un creciente uso de los sistemas automatizados de detección. Sin embargo, esta tendencia va acompañada de un aumento del contenido restaurado en comparación con los períodos anteriores. Entre enero y marzo de 2021, Facebook restableció 408 700 elementos de contenido e Instagram, 43 700. Si bien los informes sugieren que las plataformas toman cada vez más medidas contra el contenido de odio, no queda claro si esto se debe a un aumento del abuso, a un endurecimiento de las políticas de las plataformas o a un mayor número de falsos positivos.

Las empresas de redes sociales dependen de jurisdicciones nacionales, por lo que se ven directamente afectadas por las leyes nacionales y por ello son más receptivas a las solicitudes para contener el discurso de odio. Sin embargo, las plataformas de redes sociales no están vinculadas al territorio, por lo que solamente se les requiere que cumplan sus propias condiciones de servicio, que pueden o no ser más estrictas que las normas establecidas por los acuerdos internacionales mencionados en la sección anterior. Algunas de las medidas tomadas por las plataformas de redes sociales son la eliminación de material considerado de incitación al odio, el envío de advertencias a los usuarios que publican discursos de odio, la restricción de su actividad en la plataforma o su expulsión de la misma. Estas normas de la comunidad evolucionan de manera constante, en particular, en relación con el uso de métodos de moderación automatizados frente a los métodos humanos.

En vista de estos desafíos, en los últimos años ha cobrado fuerza un movimiento multisectorial que exige una mayor transparencia de las empresas de Internet como manera de fomentar su responsabilidad. Esto incluye la propuesta de medidas legislativas y regulatorias en unos 30 países y regiones, incluso mediante la Ley Europea de Servicios Digitales actualmente en elaboración. Asimismo, las empresas han tomado medidas para ser más transparentes. En 2021, la organización Access Now indexó unas 70 empresas que publicaban informes de transparencia de manera periódica , aunque todavía queda mucho por hacer a este respecto.

La nota breve de la UNESCO titulada «Dejar entrar el sol: transparencia y responsabilidad en la era digital» presenta la transparencia como una tercera vía entre un exceso de regulación estatal del contenido —que ha dado lugar a restricciones desproporcionadas de los derechos humanos— y un enfoque de laissez faire que no ha conseguido hacer frente de manera eficaz al contenido problemático, como el discurso de odio o la desinformación. Esta nota proporciona un conjunto de 26 principios de alto nivel que abarcan cuestiones relacionadas con el contenido y el proceso, la debida diligencia y la reparación, el empoderamiento, la dimensión comercial, la recopilación y el uso de datos personales y el acceso a los datos.

RESPUESTAS EXTRAJUDICIALES E INTERVENCIONES PREVENTIVAS

Existen otras respuestas extrajudiciales que provienen de esfuerzos de investigación y promoción por parte de la sociedad civil o que se centran en medidas preventivas para fortalecer la resiliencia de los usuarios de Internet ante el discurso de odio. Entre ellas figuran iniciativas que se centran directamente en las causas y consecuencias del discurso de odio en línea, en particular mediante la educación, así como iniciativas que exigen la aplicación de mejores medidas judiciales e informáticas.

Las iniciativas basadas en la educación ocupan un lugar central en estos esfuerzos y a menudo se centran en la prevención a largo plazo. Las intervenciones educativas pueden servir para concienciar acerca de las consecuencias nocivas del discurso de odio, abordar sus causas profundas y alertar de manera eficaz acerca de las técnicas retóricas y de manipulación que se utilizan para difundir el odio, tanto en línea como fuera de línea. En particular, se han elaborado y aplicado programas

[8] https://www.accessnow.org/transparency-reporting-index/

de capacitación básica en todo el mundo en materia de información y medios de comunicación, con el objetivo de proporcionar a los usuarios de Internet las competencias que les permitan examinar la información del contenido en línea e identificar la desinformación y el contenido perturbador o de incitación al odio. Asimismo, se han realizado campañas para contrarrestar el discurso de odio con narrativas alternativas, como la Iniciativa liderada por Facebook para alentar la valentía civil en línea, que se llevó a cabo en Alemania, el Reino Unido de Gran Bretaña e Irlanda del Norte y Francia en 2017. Otras iniciativas de la sociedad civil se centran en promover el cambio por parte de las plataformas. En julio de 2020, la campaña "Stop Hate for Profit" creó una coalición de unas 1200 empresas de todo el mundo que instaba a llevar a cabo un boicot publicitario contra las grandes plataformas para exigir la moderación del discurso de incitación al odio, así como una pausa publicitaria en aquellas cuentas que promovieran la discriminación hacia grupos determinados. Esta campaña se unió a un creciente número de voces que reclaman que se haga frente al discurso de incitación al odio, en particular en vista de la intensificación de la discriminación contra grupos marginales durante la pandemia de la COVID-19, y que exigen a varias empresas de redes sociales que cambien sus directivas comunitarias.

RECOMENDACIONES

Con el objetivo de elaborar políticas con base empírica que consigan contener el discurso de odio en línea - e impedir que se transforme en violencia, a la vez que garantizan la libertad de expresión - resulta fundamental reconocer, vigilar y analizar las tendencias del discurso de odio, así como recopilar datos al respecto, a fin de determinar las estrategias adecuadas para combatirlas. Las recomendaciones que figuran a continuación pretenden indicar las acciones clave encaminadas a hacer frente a los nuevos desafíos causados por el emergente discurso de odio viral y, en particular, abordar las consecuencias fuera de Internet para alcanzar la paz, la estabilidad y el disfrute de los derechos humanos por parte de todos.

1. Promover definiciones inclusivas del discurso de odio que respeten la libertad de expresión

- Garantizar que las definiciones están en consonancia con las normas internacionales, en particular las estipuladas por el ICCPR y el Plan de Acción de Rabat.

2. Crear coaliciones de múltiples partes interesadas

- Fomentar el intercambio de datos y de conocimientos especializados entre organizaciones de derechos humanos, intermediarios de Internet y el público.
- Empoderar a las partes interesadas y en particular a las comunidades locales para que vigilen y detecten el discurso de odio en las redes sociales adaptado a su contexto y a sus idiomas.

- Llevar a cabo diálogos entre múltiples interesados acerca de las tendencias del discurso de odio, su incidencia y la manera de combatirlo.
- Exigir a las plataformas que, en colaboración con grupos expertos y con el público, elaboren definiciones y rutinas operacionales que no se limiten a América del Norte y Europa Occidental para incluir más países de todo el mundo.

3. Recopilar datos y fomentar prácticas de datos abiertos en las que ya se han recopilado datos siempre respetando la protección de los datos personales

- Recopilar datos cualitativos con personas afectadas por el discurso de odio para comprender mejor el alcance y la naturaleza del daño.
- Abogar por que las empresas de plataformas de Internet mejoren sus prácticas de transparencia, en particular mediante la publicación abierta de datos sobre las denuncias relativas al discurso de odio y su resolución, así como sobre la precisión y el funcionamiento de sus sistemas de moderación de los contenidos, especialmente para fines de investigación.
- Apoyar la elaboración de herramientas y metodologías asequibles, accesibles y fáciles de utilizar que puedan emplearse para vigilar y detectar el discurso de odio en contextos multiculturales y plurilingües dentro de un plazo que permita la toma de medidas.

4. Alentar a las plataformas a que ofrezcan posibilidades firmes de reparación a aquellas personas cuyo contenido ha sido eliminado

- Facilitar la colaboración entre las empresas de redes sociales y los grupos de la sociedad civil que trabajan en el ámbito de los derechos digitales para garantizar que los procesos de moderación y eliminación del contenido se ajusten a las necesidades de la comunidad.

5. Capacitar acerca del discurso de odio, los medios de comunicación y la información y proporcionar competencias digitales a través de programas educativos

- Suministrar financiación y recursos para la elaboración de programas educativos que fomenten la resiliencia ante el discurso de odio, a partir de las tendencias actuales del discurso de odio, y que respondan a los desafíos conexos. Para ello es necesario que exista una colaboración estrecha entre las empresas de redes sociales, los institutos de investigación y las partes interesadas en la educación.
- Dar prioridad a los enfoques educativos de prevención que adviertan acerca de los efectos nocivos del discurso de odio en línea y fomenten la adquisición de nociones básicas acerca de la información y los medios de comunicación, además de llevar a cabo esfuerzos de mitigación y de solución.
- Establecer y apoyar asociaciones entre institutos educativos y empresas de redes sociales para aumentar el acceso a la información y a los recursos y hacer frente al discurso de odio en las plataformas de redes sociales mediante campañas encaminadas a la difusión o la reorientación de los usuarios hacia recursos externos.

6. Apoyar a las organizaciones que trabajan en el ámbito del discurso de odio en línea

- Apoyar a las organizaciones especializadas que se dedican a vigilar y combatir el discurso de odio y garantizar que dispongan de los recursos adecuados, en particular a aquellas que estén mejor posicionadas para tener en cuenta los contextos locales.

Este documento forma parte de una colección de documentos de debate, encargados y producidos por la UNESCO y la Oficina del Asesor Especial para la Prevención del Genocidio de las Naciones Unidas (OSAPG). Los documentos son una contribución directa a la Estrategia y el Plan de Acción de las Naciones Unidas y se publican en el contexto del Foro Multilateral y la Conferencia Ministerial sobre Combatir los discursos de odio a través de la educación en septiembre y octubre de 2021.

El estallido de la pandemia de COVID-19 ha subrayado la pertinencia de la Estrategia y el Plan de Acción de las Naciones Unidas, generando una ola de discursos de odio en todo el mundo, exacerbando aún más la intolerancia y la discriminación hacia determinados grupos y desestabilizando las sociedades y los sistemas políticos. Los documentos de debate tratan de desentrañar las cuestiones clave relacionadas con este reto mundial y proponer posibles respuestas y recomendaciones.

Este documento fue encargado por la Sección de Libertad de Expresión y Seguridad de los Periodistas de la UNESCO como parte del proyecto « #CoronavirusFacts, Addressing the 'Disinfodemic' on COVID-19 in conflict-prone environments » financiado por la Unión Europea. Fue redactado por Jonathan Bright, Antonella Perini, Anne Ploin y Reja Wyss, del Oxford Internet Institute de la Universidad de Oxford.

Publicado en 2021 por la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura, 7, place de Fontenoy, 75352 Paris 07 SP, France

© UNESCO 2021

Este documento está disponible en acceso abierto bajo la licencia Attribution-ShareAlike 3.0 IGO (CC-BY-SA 3.0 IGO) (http://creativecommons.org/licenses/by-sa/3.0/igo/).
Al utilizar el contenido del presente documento, los usuarios aceptan las condiciones de utilización del Repositorio UNESCO de acceso abierto (www.unesco.org/open-access/terms-use-ccbysa-sp).

Título original: *Addressing hate speech on social media: Contemporary challenges*
Publicado en 2021 por la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO).

Los términos empleados en este documento y la presentación de los datos que en ella aparecen no implican toma alguna de posición de parte de la UNESCO en cuanto al estatuto jurídico de los países, territorios, ciudades o regiones ni respecto de sus autoridades, fronteras o límites.

Las ideas y opiniones expresadas en esta obra son las de los autores y no reflejan necesariamente el punto de vista de la UNESCO ni comprometen a la Organización.

Este documento ha sido elaborado con el apoyo financiero de la Unión Europea. Su contenido es responsabilidad exclusiva de los autores y no refleja necesariamente la opinión de la Unión Europea.

Diseño gráfico: Dean Dorat

CI/FEJ/2021/DP/01

Financiado por la Unión Europea

Educar contra los #DiscursosDeOdio