Universitas. Revista de Ciencias Sociales y Humanas
ISSN: 1390-3837
ISSN: 1390-8634
revistauniversitas@ups.edu.ec
Universidad Politécnica Salesiana
Ecuador

Larrondo, Manuel Ernesto; Grandi, Nicolás Mario
Inteligencia Artificial, algoritmos y libertad de expresión
Universitas. Revista de Ciencias Sociales y Humanas, núm. 34, 2021, Marzo-, pp. 177-194
Universidad Politécnica Salesiana
Ecuador

DOI: https://doi.org/10.17163/uni.n34.2021.08

Disponible en: https://www.redalyc.org/articulo.oa?id=476165932008

UNIVERSITAS
Revista de Ciencias Sociales y Humanas de la Universidad Politécnica Salesiana

https://doi.org/10.17163/uni.n34.2021.08

Inteligencia Artificial, algoritmos y libertad de expresión

Artificial Intelligence, algorithms and freedom of expression

Manuel Ernesto Larrondo
Universidad Nacional de La Plata
larrondomanuel@gmail.com
https://orcid.org/0000-0002-0569-502X

Nicolás Mario Grandi
Universidad Nacional de La Plata
drgrandinicolas@hotmail.com
https://orcid.org/0000-0003-4191-8849

Resumen
La Inteligencia Artificial puede presentarse como un aliado al momento de moderar contenidos violentos o de noticias aparentes, pero su utilizaciónsin intervención humana que contextualice y traduzca adecuadamente la expresión deja abierto el riesgo de que se genere censura previa
En la actualidad esto se encuentra en debate dentro del ámbito internacional dado que, al carecer la Inteligencia Artificial de la capacidad para contextualizar lo que modera, se ésta presentando más como una herramienta de censura previa indiscriminada, que como una moderación en busca de proteger la libertad de expresión.
Por ello luego de analizar la legislación internacional, informes de organismos internacionales y los términos y condiciones de Twitter y Facebook, sugerimos cinco propuesta tendientes a mejorar la moderación algorítmica de contenidos.
En primer término proponemos que los Estados compatibilicen sus legislaciones internas respetando los estándares internacionales de libertad de expresión. También instamos a que desarrollenpolíticas públicas consistentes en implementar legislaciones protectoras de las condiciones laborales de supervisores humanos sobre las decisiones automatizadas de remoción de contenido.
Por su parte, entendemos que las redes sociales deben presentar términos y condiciones claros y consistentes, adoptar políticas internas de transparencia y rendición de cuentas acerca de cómo opera la IA en la difusión y remoción de contenido en línea y, finalmente, deben realizar evaluaciones previas de impacto de su IA a los derechos humanos.

Palabras clave
Inteligencia Artificial, moderación automática de contenidos, *fakenews,* libertad de expresión, redes sociales.

Forma sugerida de citar: Grandi, N. (2021). Inteligencia Artificial, algoritmos y libertad de expresión. *Universitas, 34,* pp. 177-194.

Abstract
Artificial Intelligence can be presented as an ally when moderating violent content or apparent news, but its use without human intervention that contextualizes and adequately translates the expression leaves open the risk of prior censorship.
At present this is under debate within the international arena given that, since Artificial Intelligence lacks the ability to contextualize what it moderates, it is presented more as a tool for indiscriminate prior censorship, than as a moderation in order to protect the freedom of expression.
Therefore, after analyzing international legislation, reports from international organizations and the terms and conditions of Twitter and Facebook, we suggest five proposals aimed at improving algorithmic content moderation.
In the first place, we propose that the States reconcile their internal laws while respecting international standards of freedom of expression. We also urge that they develop public policies consistent with implementing legislation that protects the working conditions of human supervisors on automated content removal decisions.
For its part, we understand that social networks must present clear and consistent terms and conditions, adopt internal policies of transparency and accountability about how AI operates in the dissemination and removal of online content and, finally, they must carry out prior evaluations impact of your AI on human rights.

Keywords
Artificial Intelligence,automatic content moderation, fake news, freedom of expression, social networks.

Inteligencia Artificial y libertad de expresión. Planteo del problema y anticipo de propuesta

La libertad de pensamiento y expresión humana es la base fundamental de toda sociedad democrática.

Así lo reconocen el art. 19 de la Declaración Universal de Derechos Humanos al igual que el art 18 (libertad de pensamiento) y art 19, 1) y 2) del Pacto Internacional de Derechos Civiles y Políticos (PIDCP) al prever que "nadie podrá ser molestado a causa de sus opiniones" así como también que toda persona gozará del derecho a "buscar, recibir y difundir informaciones e ideas de toda índole" sin consideración de fronteras o "procedimiento" que elija (ONU, 1966).

Sin embargo, al mismo tiempo dicho artículo reconoce que ese derecho puede estar sujeto a restricciones que deben ser fijadas por ley necesaria para: a) asegurar el respeto a los derechos o a la reputación de los demás y b) la protección de la seguridad nacional, el orden público o la salud o la moral públicas.

Por su parte, el Sistema Interamericano de Derechos Humanos consagra en igual sentido y alcance de protección amplia a ese derecho en el artículo 13 de la Convención Americana, con la particularidad de que expresamente prohíbe la censura bajo cualquier modalidad y solo la contempla en forma previa para proteger a los derechos de la niñez y adolescencia. En igual sentido, destaca que quien ejerza este derecho se encuentra sujeto a las responsabilidades ulteriores que deben estar fijadas por ley respetando su necesidad, legitimidad y proporcionalidad.

Respecto a las posibles restricciones y responsabilidades posteriores al ejercicio del derecho, advertimos que el Comité de DDHH de la ONU —interpretando los alcances del art. 19 del PIDCP— se inclina por una postura aún más protectora al considerar que la libertad de opinión "no autoriza excepción ni restricción alguna" a su ejercicio, ya sea "por ley u otro poder" (ONU, 2011).

Es evidente que los límites y alcances del ejercicio de este derecho son centro de análisis e interpretación complementaria de parte de los principales organismos internacionales.

Sin ir más lejos, basta con citar la vigencia de la OC 5/85 dictada por la Corte Interamericana que señala que la libertad de expresión no se agota en su faz individual, sino que comprende además a la dimensión colectiva, subrayando que el libre pensamiento y su difusión son inseparables, de forma tal que una limitación previa —estatal o privada— a cualquiera de ellos, sería incompatible con los estándares interamericanos que protegen a este derecho (Sec. Gral. OEA, 2017).

En Argentina, la Corte Suprema de Justicia de la Nación (CSJN), ha seguido igual sintonía remarcando que la libertad de expresión es una de las libertades de mayor relevancia, en tanto que, sin su debida protección, el sistema democrático funcionaría solo de forma aparente.

En efecto, la libertad de expresión, en su faz colectiva, busca garantizar el debate público a fin de facilitar oportunidades de expresión alos diversos estamentos de la sociedad con el fin de que todas las personas puedan dar a conocer sus ideas y opiniones, evitándose así que predomine una o un grupo por sobre otros (CSJN, 2013).

Estos estándares intercontinentales fueron consensuados a lo largo de casi todo el siglo XX y principios del presente siglo XXI. Con la irrupción de las redes sociales y demás intermediarios en línea llegó el momento en que prácticamente la mitad de la población mundial ejerce la triple acción de difundir, investigar y difundir información a través de Internet por medio de las principales plataformas que emplean "Inteligencia Artificial" (en adelante IA) con las que interactuamos día a día.

Ahora bien, ¿qué es la IA? De momento no se ha logrado consensuar una única definición a nivel mundial. En esta ocasión nos inclinaremos por citar aquella que brinda la Relatoría Especial sobre la promoción y protección del derecho a la libertad de opinión y de expresión de la ONU al decir que "es una 'constelación' de procesos y tecnologías que permiten que las computadoras complementen o reemplacen tareas específicas que de otro modo serían ejecutadas por seres humanos, como tomar decisiones y resolver problemas" (ONU, 2018).

Agrega la Relatoría que en la base de la IA están los "algoritmos" que son códigos informáticos diseñados y escritos por seres humanos. Toda clase de dato que procesa un algoritmo es traducida y arroja un resultado determinado tales como inferencias, sugerencias o predicciones. Así, el caudal de infinitos datos que genera una persona por segundo al interactuar en la red, conlleva al necesario desarrollo de la IA ante la imposibilidad material de que una persona pueda hacerlo por sí sola en poco tiempo y de forma eficiente.

Bastará como muestra de ello advertir que el volumen de generación de datos en línea crece exponencialmente cada segundo, a punto tal que en solo un minuto de navegación en Internet en Google se llevan adelante más de tres millones de búsquedas, en Facebook se envían más de treinta millones de mensajes y se ven más de dos millones de videos, en Twitter se publican más de cuatrocientos cincuenta mil tuits, en Instagram se postean más de cuarenta y seis mil fotos, en YouTube se suben más de cuatro millones de horas de videos y casi el doble en Netflix. A esta gran cantidad de información se la ha denominado big data, y surge de la interrelación de nuestros aparatos electrónicos conectados en la Web. Ya la capacidad de almacenamiento no se mide más en kilobyte compuesto por un numero de cuatro cifras, sino que la evolución nos ha llevado alhella bytes, el cual tiene de veintisiete cifras(DAUS, 2019), es decir, la información es seiscientas veces mayor.

Este inmenso caudal de datos que generamos a través de las redes sociales e intermediarios conforman un excéntrico lugar virtual en el que conflui-

mos con otras personas, al igual que con "bots" y demás sistemas automatizados basados en IA.

Si bien estos últimos coadyuvan al ejercicio humano de la libre expresión, al mismo tiempo han surgido evidencias concretas que nos alertan sobre el serio riesgo de que, poco a poco, la IA "usurpe" el derecho humano a recibir, investigar y difundir contenido en línea al decidir, de forma automatizada qué contenido perdura y cuál es removido según sus "términos y condiciones".

Una primera aproximación al empleo de IA de parte de las plataformas podría considerarse adecuado para la remoción de contenido violento, de desinformación o de aquel que incite al odio, por ejemplo.

Sin embargo, anticipando el desarrollo y conclusión de nuestra propuesta, consideramos que sin moderación humana que contextualice y traduzca adecuadamente la expresión línea, puede existir un serio riesgo de que las plataformas otorguen prevalencia a la IA como moderadora automatizada de contenido en línea, incumpliendo así con los estándares internacionales anteriormente referidos en relación con que el derecho humano a recibir, investigar y difundir sería limitado no ya por una ley necesaria, con fin legítimo y proporcional sino por una "constelación" de facto conformada por algoritmos inhumanos.

Por lo tanto, nuestro trabajo consistirá en comenzar por explicar de qué manera los organismos internacionales conceptualizan y diagnostican el uso de la IA en la moderación automatizada de contenido en línea, así como también apuntaremos sus principales ventajas y desventajas. Seguidamente analizaremos cuáles son las implicancias de la IA en el ejercicio de la libertad de expresión a través de las plataformas como Facebook y Twitter en miras a evaluar si su implementación ha coadyuvado en los últimos tiempos a restringir o no a la expresión. Finalmente, y con base en el marco fáctico y jurídico analizado, postulamos la necesaria intervención de la supervisión humana cuando la IA sugiera la remoción de contenido en línea de interés público.

IA y moderación automatizada de contenido en línea. Razones de su implementación y la necesaria supervisión humana

La IA suele ser considerada como un conjunto de sistemas tecnológicos automáticos e imparciales tendientes a facilitar la eficacia en la moderación de contenidos en busca de mitigar posibles discursos de odio, discriminato-

rios, terroristas, etc., y así mejorar la experiencia de sus usuarios y la construcción de ciudadanía.

Sin embargo, la ONU ha remarcado que, en el campo de la moderación de contenido, si bien la IA tiene sus aristas positivas también se resaltan las negativas.

Entre los beneficios del uso de IA, destaca la ONU que la selección personalizada de contenido aumenta la experiencia en línea de cada persona permitiendo encontrar rápidamente la información solicitada, incluso en diversos idiomas. Sin embargo, esa virtud inicial tiene como elemento disvalioso la limitación de que cada persona pueda acceder adiversos puntos de vista, interfiriendo así con la posibilidad personal de ahondar y confrontar distintas ideas y opiniones con individuos que tengan otra posición ideológica. Política, religiosa o social. De esta manera, esa segmentación de contenido que aparenta ser muy útil y eficaz, podría al mismo tiempo reforzar las creencias individuales y llevar a la exacerbación de contenidos violentos o bien a la desinformación con el solo fin de mantener la participación en línea del usuario (ONU, 2018).

Sandra Álvaro explica que los algoritmos ya son parte de nuestra vida cotidiana poniendo como ejemplo a Facebook quien tiene un algoritmo llamadoEdgerankqueanaliza nuestros datos de navegación —los "me gusta" que ponemos, los amigos que tenemos y los comentarios que hacemos— y con ello nos perfila con el fin de mostrarnos aquellas historias que nos gustan y ocultarnos aquellos que nos aburran y mostrarnos nuevos amigos que concuerden con nuestro perfil e ideología (Álvaro, 2014).

Esta situación, que genera una suerte de burbuja informativa, ha despertado el interés de la Unión Europea en tanto alerta que los seres humanos que interactúan con los sistemas de IA deben ser capaces de mantener una autodeterminación plena y efectiva sobre sí mismos y de poder participar en el proceso democrático. Por eso insta a que los sistemas de IA no deben coaccionar, manipular, inferir o agrupar injustificadamente a los seres humanos.

A criterio del órgano europeo, la IA debe entonces diseñarse para aumentar, complementar y potenciar las habilidades cognitivas, sociales y culturales humanas, siguiendo así los principios de diseño centrados en el ser humano (Eur. Comm., 2019).

Ante esta nueva realidad en la cual la información de toda índole desborda en la red, ya en marzo de 2018 la Comisión Europea instaba a las plataformas de internet a que usasen filtros automáticos para verificar y en su

caso retirar contenido extremista, aunque —al mismo tiempo— sugería que se emplease la revisión humana con el fin de evitar los errores que provienen de los sistemas automatizados.

Ello así, ya que el uso de IA en la moderación automatizada de contenido puede llegar a afectar al ejercicio de la libertad de expresión dado que, por el momento, entre sus limitaciones se destaca la imposibilidad de que pueda evaluar el contexto, los usos idiomáticos y aspectos culturales de los seres humanos.

Si bien en los últimos tiempos la IA ha mejorado exponencialmente en el Procesamiento del Lenguaje Natural (PLN), aún no ha logrado un desarrollo tal que le permita comprender la totalidad de los matices lingüísticos y culturales por los cuales se expresan los humanos.

Ello ha llevado a que, al momento de moderar contenidos de manera automática, el algoritmo empleado por las plataformas haya eliminado también imágenes de desnudez con valor histórico, cultural o educativo, relatos históricos y documentales de conflictos, pruebas de crímenes de guerra, intervenciones en contra de los grupos que promueven el odio o esfuerzos por impugnar o denunciar el lenguaje racista, homófobo o xenófobo.

Esto demostraría que, en esta faz del desarrollo de la IA, aún nos encontramos con sistemas automatizados débiles que necesitan la supervisión humana para poder llevar adelante sus acciones sin que se afecten otros derechos.

Justamente es en ese contexto en el cual la IA pierde su "poder mágico" de solucionar la remoción de contenido injurioso, discurso de odio o bien la eventual desinformación en línea. Por tal razón, las empresas de internet han instado a los usuarios a que perfeccionen el contenido observado con diferentes elementos contextuales, aunque, cabe aclarar, la viabilidad y la eficacia de esas directrices no son claras (ONU, 2018).

En ese sentido, el Comité de DDHH de la ONU entiende que, a diferencia de las personas, los algoritmos carecen de corpus y mente, es decir, no son aún capaces de comprender cuando una expresión es irónica, o es una parodia o bien para confirmar con precisión si una determinada manifestación puede ser calificada de alabanza al "terrorismo". Por lo tanto, la automatización de su operatividad matemática tiende más a optar por un resultado rápido consistente en limitar o removerdeterminada expresión sin tener en cuenta que ello redunda en que se afecta considerablemente al derecho humano a recibir, investigar y difundir (ONU, 2018).

De igual forma, la utilización de la IA al momento de subir archivos en la web, con el fin de proteger los derechos de propiedad intelectual tanto de los videos ha generado dudas por la gran cantidad de bloqueos que se producen, lo cual, sumado a los posibles filtrados ante contenidos vinculados con el terrorismo u otro tipo de posiciones extremas, pueden llegar a obtener lo contrario, es decir, en vez de proteger derechos se puede establecer regímenes totalitarios, al aplicar una censura previa automatizada.

En efecto, si bien resulta de suma utilidad el empleo de algoritmos de comparación criptográfica para detectar imágenes de abusos sexuales a menores, por el contrario, su aplicación al contenido "extremista" —que por lo general requiere la evaluación del contexto— es difícil sin la existencia de normas claras que definan qué es el "extremismo" (ONU, 2018).

En este sentido la ONU entiende que las plataformas deberían transparentar la forma en cómo utilizan la IA, explicando de forma detallada con datos agregados que ilustren ejemplos de casos reales o casos hipotéticos a fin de clarificar cómo es su interpretación y la aplicación de normas concretas (ONU, 2018).

Asimismo, siendo responsabilidad de las empresas prevenir y eventualmente disminuir los efectos negativos en los derechos humanos con el uso de la IA, es claro que parte de su política de transparencia debería consistir en comenzar por reconocer las importantes limitaciones que adolece la automatización en la moderación de contenido, tales como aquellas dificultades ya referidas sobre la interpretación del contexto como así también la amplia variación de matices idiomáticos y el significado y las particularidades lingüísticas y culturales. Es por eso que, como mínimo, la tecnología actual y futura para abordar aspectos relacionados datos a gran escala debería estar sometida a una auditoría rigurosa y, desde luego, contar con aportes de la sociedad civil tendientes a enriquecer el análisis.

Para finalizar este acápite, queremos referirnos a un último aspecto vinculado con nuestra propuesta de que se garantice la supervisión humana ante la posible remoción automatizada de contenido en línea.

Nos referimos en concreto a aquel por el cual se insta a las plataformas a que fortalezcan y garanticen que la moderación automatizada de contenido en línea cuente con la posibilidad de revisión y supervisión por seres humanos capacitados en conocer los estándares internacionales de libertad de expresión.

Para tal fin, la ONU afirma que resulta imprescindible que se brinde una protección adecuada a las condiciones de trabajo en la que prestan tareas ya

que deben ser compatibles con las normas de derechos humanos aplicables a los derechos laborales (ONU, 2018).

Tal postulación tiene su basamento, por ejemplo, en un caso concreto de "precarización laboral" de moderadores que trabajaban para Facebook.

En efecto, esta empresa en 2015 contaba con menos de cuatro mil quinientas personas como moderadoras de contenido audiovisual, pero, con motivo del COVID-19, debió ampliar la plantilla contratando a unos quince mil moderadores, la mayoría de las cuales están bajo la modalidad de subcontratados/as en diversas ciudades de todo el mundo (Dublín, Berlín, Manila).

Así informa el magazine "The New Yorker" que a menudo los moderadores trabajan en horas impares por las diversas zonas horarias en el mundo a lo cual se agrega la falta de sueño y el fuerte impacto psicológico que padecen al absorber todo lo que visualizan en sus pantallas sin contar con un "protocolo" estandarizado a fin de indicar qué contenido debe permanecer en línea y cuál no.

A raíz de ello, en mayo de 2020 miles de moderadores se unieron a una demanda colectiva contra Facebook alegando trastornos psicológicos y, por tal razón, acordaron con la empresa un convenio de pago por USD 52 000 000 (Marantz, 2020).

La supervisión humana que postulamos para la revisión de la decisión automatizada de contenido si bien no lograría prevenir de forma absoluta la censura en línea, es posible anticipar que sí contribuiría a suplir los graves defectos de la IA que no logra —aún— interpretar contextos, términos lingüísticos, ironía, humor satírico, imágenes artísticas de desnudez, etc.

Veamos a continuación determinados casos puntuales que, según nuestra posición, acompañan esta propuesta de implementar la supervisión humana frente a la desinformación y remoción automatizada de contenido.

Cómo operan Twitter y Facebook

Reglas de Twitter

La red social Twitter cuenta con una serie de reglas tituladas "Políticas y pautas generales" que deben respetarse en miras a utilizar la plataforma. Un sector de esas pautas se vincula, en lo que respecta a nuestro análisis, al contenido en línea que se relaciona con temas de interés público.

Si bien esta red social anticipa que toma diversas clases de medidas sobre los tuits que incumplen sus reglas, al mismo tiempo reconoce que en ciertas ocasiones —sin precisar cuáles al menos a título ejemplificativo— mantienen en línea determinados tuits que pueden ser de utilidad a la sociedadpuesya que, de otra manera se borrarían. ¿Cuándo un tuit sería considerado de interés público? Informa la plataforma que lo califican así cuando se presenta como "un aporte directo para la comprensión o el debate de un asunto que le preocupa a todo el público" (Twitter, 2020).

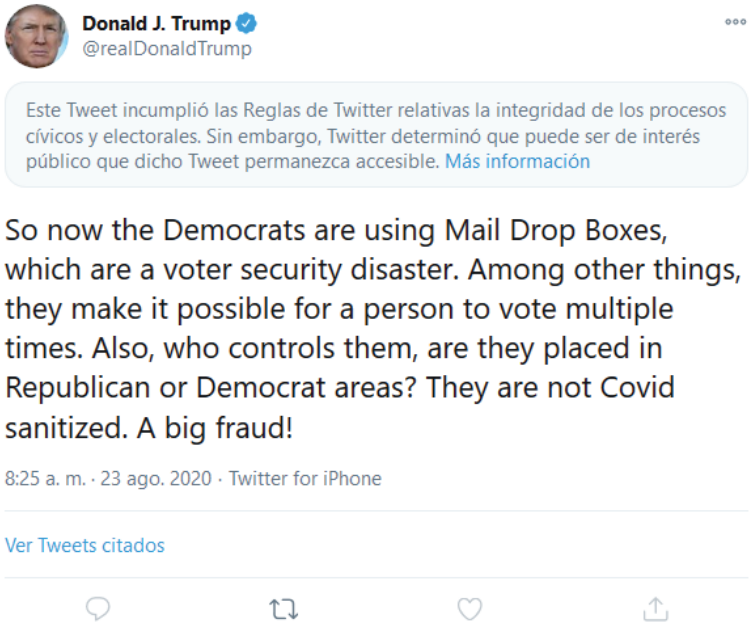
Así, destaca esta red social que son de interés público aquellos tuits que emiten funcionarios gubernamentalespues es importante conocer que hacen con el fin de debatir sus acciones u omisiones. Twitter anticipa así que dará prevalencia a la difusión de contenido de interés público basándose en los siguientes cuatro criterios que conforman una excepción a la remoción directa de contenido, a saber:

- El tuit incumple alguna o varias reglas de Twitter.
- El/la autor/a del tuit es una cuenta verificada.
- La cuenta tiene más de 100 000 seguidores/as.
- La cuenta representa a un integrante actual o potencial de Gobierno o Poder legislativo local, nacional o supranacional: i) titulares actuales de un cargo de liderazgo elegido o designado por un Órgano de Gobierno o legislativo; ocandidatos o nominados para cargos políticos.

Puede ocurrir, sin embargo, que un funcionario público publique un tuit vulnerando así los términos y condiciones de Twitter. En ese caso, como excepción, la plataforma informa que podrá optar por conservar el tuit que, de lo contrario, se eliminaría. Para tal fin, inserta detrás de este un aviso que tiene como fin contextualizar el incumplimiento de las reglas y permitir a las personas ingresar a verlo, en caso de desearlo.

Acudiendo al empleo de IA, expresa que, al colocarse ese aviso, también se está disminuyendo la posibilidad de interactuar con ese tuit, por medio de "Me gusta", "Retweet" o bien de compartirlo en esa misma red social para generar que el algoritmo de Twitter evite recomendarlo. Se advierte así que a través de estas acciones se intentaría restringir el alcance del tuity, al mismo tiempo, garantizar al público su posibilidad de visualizarlo y debatir sobre el tema que se trate.

Como primera observación a realizar, queremos resaltar el marco acotado y restrictivo que implementa Twitter cuando requiere que una cuenta posea 100 000 seguidores para poder entonces estar incluida dentro de las condiciones del estándar de interés público. La medición cuantitativa basada solo en cantidad de seguidores —que bien podrían ser conformadas mayoritariamente por cuentas bots— creemos que atentaría contra un análisis cualitativo del discurso que se trate en tanto para definir si un discurso es o no de interés público, pues debería acudirse a la supervisión humana siguiendo estándares jurisprudenciales tales como el de la Corte Interamericana de DDHH que define al interés público con aquellas opiniones o informaciones sobre asuntos en los cuales la sociedad tiene un legítimo interés de mantenerse informada sobre el funcionamiento del Estado o derechos e intereses generales (CIDH, 2009, 2011). Retomando entonces el análisis de las medidas que pone en práctica Twitter sobre este punto, un ejemplo de ello podemos verlo en concreto en uno de los tantos tuits que emitió el 23 de agosto 2020 el presidente Donald Trump con motivo de la contienda electoral presidencial.



Como puede observarse, en dicho tuit el presidente Trump hizo alusión al posible fraude electoral que podría cometerse a través del sistema de voto ciudadano por correo. En ese caso, Twitter insertó un aviso en el tuit que advirtió al público acerca del incumplimiento de las reglas relativas a la integridad de los procesos cívicos electorales, aunque igualmente se optó que el tuit permanezca accesible. Para más información, se anexó un enlace para que remite al usuario a la lectura de las políticas y pautas generales sobre interés público citadas anteriormente.

Para estos casos en particular, advertimos que Twitter informa que su "Equipo de Trust & Safety", el cual se conforma con profesionales expertos en diversos campos, implementará un segundo análisis, con el fin deanalizar el tuit y dar una opinión para conservar o no la visibilidaddelmismo con base en los criterios de interés público. Posteriormente, las primeras recomendaciones efectuadaspor este equipo se darán a conocer a un grupo de referentes internos de la red social con amplios conocimientos en la materia y en el contexto cultural en el cual se circunscribió el tuit paraque, luego de que ellos se expidan, los líderes Trust & Safety tomen finalmente la decisión de sicorresponde aplicar el aviso o eliminar el tuit.

Sin embargo, esta modalidad de revisión personalizada al parecer no sería aplicada por Twitter de manera uniforme para situaciones de interés público. Un ejemplo de ello puede apreciarse cuando en octubre 2020 Twitter impidió que los usuarios compartieran un artículo del diario *New York Post* vinculado al candidato presidencial JoeBiden y sus eventuales contactos con un empresario ucraniano. ¿Por qué lo impidió? El aviso indicó el siguiente fundamento: "Tu Tweet no se pudo enviar porque Twitter o nuestros socios identificaron este enlace como potencialmente dañino" (Cox, 2020). Ningún tipo de información adicional se brindó acerca de si en tal decisión pudo haber intervenido un equipo de profesionales de "Trust & Safety" tal como sí pareciera que lo hizo al remitir a sus políticas de casos de "interés público" en el aviso insertado en el tuit del presidente Trump.

Yendo al análisis de las políticas en general de esta plataforma, merece la pena referirnos al caso de difusión de contenido multimedia. Así Twitter anticipa que focalizará su atención en contenidos que se encuentren considerablemente alterados o falsificados con la intención deliberada de engañar. Sin embargo, no explica de qué manera arribaría a semejante conclusión, esto es, cómo determinaría que un cierto contenido audiovisual ha sido alterado o falsificado. Para tal fin, Twitter alerta que cuenta con la facultad de aplicar

su propia tecnología —no la especifica ni informa— o recogerdenunciapor medio de sus colaboradores o socios externos. Solo en aquellos supuestos en los cuales sea imposible determinar con certeza si lo expuesto en contenido multimedia fue modificado o es una copia, puede ser—no lo asegura— que no tome ninguna medida para restringirlo o referenciarlo (Twitter, 2020).

Asimismo, y siempre en relación con la difusión de contenido multimedia sobre el cual omite brindar detalles acerca de cómo concluye en que podría dar lugar a confusión o que sugiera una intención dolosa de engañar, informa que analiza el contexto del tuit para determinar si alerta acerca de que el contenido se encuentra modificado o falsificado, aunque no precisa si para tal fin intervienen profesionales tal como sí lo indica expresamente los contenidos de interés público. De esa manera, la falta de precisión nos inclina a inferir que Twitter emplearía IA a los fines de revisar:

- El texto del tuit que se anexa al elemento multimedia o incluido en él.
- Los metadatos asociados al elemento multimedia.
- La información del perfil de la cuenta que difunde el elemento multimedia.
- Los sitios web enlazados en el tuito en el perfil de la cuenta que difunde el elementomultimedia.

En tal sentido, observamos que las medidas automatizadas que Twitter adopta ante un contenido que la misma plataforma califica como falso o alterado, pues impide que se pueda compartir en Twitter y, en consecuencia, podría ser borrado al mismo tiempo que podrá suspenderse de forma permanente a la cuenta de la cual emana el referido contenido.

Facebook

Facebook informa en su plataforma que su estrategia para detener la información errónea consiste en tres acciones puntuales:

- Quitar cuentas y contenido que infrinjan sus normas comunitarias o políticas publicitarias.
- Reducir la distribución de noticias falsas y contenido no auténtico como “títulos anzuelo”.
- Informar a las personas brindándoles mayor contexto a las publicaciones que visualizan.

Esta triple acción tendería a eliminar a los "malos actores" que con frecuencia difunden historias falsas y, según indica, disminuiría drásticamente el alcance de esas historias ayudando a las personas a mantenerse informadas sin sofocar el discurso público.

Destaca también que para esta labor utiliza aprendizaje automático para ayudar a sus equipos a detectar el fraude, hacer cumplir sus políticas contra el spam y bloquear millones de cuentas falsas todos los días cuando intentan registrarse (Facebook, 2020).

Informa que toma "medidas" —aunque no explica en qué consistirían— contra páginas enteras y sitios web que comparten repetidamente noticias falsas, lo que reduciría su distribución general de noticias. Resaltan que como Facebook no tiene intención en ganar dinero con información errónea o ayudar a quienes la generan a obtener ganancias, se les impide a esos editores publicar anuncios y usar sus funciones de monetización como InstantArticles.

Asimismo, destaca que parte de su estrategia para combatir a la desinformación consiste en asociarse con varios países con verificadores de datos de terceros para revisar y calificar la precisión de los artículos y publicaciones en Facebook. Estos verificadores de datos serían independientes ya que, apunta, están certificados a través de la Red Internacional de Verificación de Datos no partidista. Así, cuando esas organizaciones califican algún contenido como falso, Facebook clasifica significativamente como más baja a esa historia en el News Feed. De esta manera, afirman que esto reduce las vistas futuras en más del 80% (Lions, 2018). En consonancia a lo que observa Agustina del Campo, se advierte que Facebook ha pasado de "un sistema que dependía casi enteramente de sus usuarios para las denuncias de contenido violatorio de sus normas, a un sistema de activación y de 'enforcement' proactivo de sus términos y condiciones de servicio".En lo que refiere a la denominada infodemia,este cambio implicó que esta red socialautomatice la moderación de contenidos que serían "posiblemente" falsos, para luego reenviar directamente ese mismo contenido a otros usuarios o a los denominados "verificadores", incluso antes de que alguien eleve una denuncia interna sobre dicho contenido (Del Campo,2020).

Para cerrar, resulta ilustrativo el siguiente gráfico elaborado por Facebook como muestra global de remoción de contenido de 2013 a 2019 en los últimos seis años (Facebook Transparency, 2019):

Conclusión: la necesaria supervisión humana como regla y no como excepción en la decisión final de remoción de contenido en línea

A lo largo del presente trabajo hemos analizado y descripto sucintamente el marco jurídico internacional relacionado a la protección de la libertad de pensamiento y expresión como derecho humano que se ejerce sin importar a través de qué medio o plataforma se haga. Con el crecimiento exponencial de las diversas plataformas en línea y el volumen de datos que crece segundo a segundo gracias a la interacción de los usuarios, hemos dado cuenta acerca de cómo los organismos internacionales destacan el uso y empleo de IA en la distribución y también en la restricción automatizada de contenidos en línea.

También hemos expuesto que, en términos generales, el uso inhumano de algoritmos predictivos en lo que refiere puntualmente a la remoción automática e irreflexiva de contenidos en línea, vulnera los estándares internacionales de libertad de expresión tal como la prohibición de censura previa.

Con el empleo único de IA en la decisión de remover contenido en línea, se visualiza la primera situación fáctica que contradice aquel estándar de prohibición de censura: que una serie de instrucciones programadas por humanos con funciones predictivas y con capacidad de lectura de lenguaje natural, dispone sin más qué información recibimos a través de las redes sociales con las que interactuamos a diario. Esta situación contradice abiertamente el estándar previsto por el artículo 13 de la Convención Americana de

DDHH, el artículo 19 del PIDCP, el art. 10 del Tratado Europeo de DDHH, entre otros, por cuanto en general "solo es admisible la restricción a la libre expresión mediante el dictado de una ley necesaria, que persiga un fin legítimo y que guarde proporcionalidad con el derecho que se intenta proteger".

En miras a que dicho estándar no se convierta en letra muerta y al mismo tiempo sin que deba afectarse el empleo de IA en la moderación de contenido en línea, a fin de lograr un equilibrio entre ambos es que postulamos que resulta trascendental la supervisión humana para la necesaria revisión sobre toda decisión automatizada de remoción de contenido. Los ejemplos ilustrativos referenciados a lo largo de este trabajo permiten inferir que, si bien la supervisión humana de la decisión adoptada por la IA no lograría prevenir de forma absoluta la censura en línea, es posible anticipar que sí contribuiría a suplir los graves defectos de esta última que no logra —aún— interpretar contextos, términos lingüísticos, ironía, humor satírico, imágenes artísticas de desnudez, etc.

Para tal fin, resulta imperioso que los organismos internacionales tales como la ONU, OEA, Comisión Europea, etc. continúen el estudio global de esta problemática y, a partir de allí, persistan en instar a los Estados a que:

Compatibilicen sus legislaciones internas respetando los estándares internacionales de libertad de expresión. Si bien cada país es soberano y tiene el poder de regular el discurso en las plataformas de Internet de manera más directa, existen casos puntuales como el que sucede en Alemania donde se encuentra vigente la Ley NetzDG desde 2018. Esta ley requiere que las redes sociales eliminen rápidamente el discurso ilegal, con un enfoque específico en el discurso de odio y los delitos de odio, de lo contrario deberían abonar multas de miles de euros. El loable fin que pudo tener el dictado de esa normativa confronta con un hecho innegable: que una pretensa rápida eliminación de contenido en línea supuestamente ilegal, pasa por alto relevantes garantías constitucionales como el debido proceso y derecho de defensa al delegarse en plataformas privadas la decisión de confirmar qué contenido merece o no permanecer en línea cuando, en su caso, tal resolución correspondería que sea adoptada por un Juez natural, al menos en lo que respecta a aquellos Estados democráticos.

También, es necesario que desarrollen políticas públicas consistentes en implementar legislaciones protectoras de las condiciones laborales (aspectos psicofísicos en particular) del personal dependiente que cumple labores de supervisión de toda decisión automatizadas de remoción de contenido en línea bajo las órdenes de plataformas.

Asimismo, sería pertinente requerir a las empresas a que sus términos y condiciones se expliquen claramente y sean consistentes con los estándares de derechos humanos establecidos para la libertad de expresión.

Por otra parte, también sería conveniente que se solicite a esas empresas que operan física o virtualmente en sus territorios a que igualmente adopten políticas internas de transparencia y rendición de cuentas acerca de cómo opera la IA en la difusión y remoción de cada contenido en línea que recibe toda persona al interactuar con su plataforma. Todo ello, claro está, junto con la necesaria colaboración que dichas empresas deberían brindar en perfeccionar los actuales mecanismos de apelación interna ante la eventual decisión automatizada y supervisada que disponga el bloqueo de una cuenta o remoción de un contenido en línea (ONU, 2018).

Y finalmente, complementando lo anterior, esas mismas empresas deben realizar la debida diligencia a través de evaluaciones de impacto sobre los derechos humanos, es decir, de cuáles son sus reglas, cómo se aplican y qué medidas toma para prevenir que sean vulnerados. Si bien es claro que los detalles de las acciones de cumplimiento individuales deben mantenerse privados, los informes de transparencia brindan a su vez información esencial sobre cómo la empresa está abordando los desafíos del día (New America,2020).

Bibliografía

Álvaro S. (2014). El poder de los algoritmos: cómo el software formatea la cultura. *CCCBLAB.Investigación e Innovación en Cultura*. https://bit.ly/3tiAGrO

Corte Interamericana de Derechos Humanos (2009). Caso Tristán Donoso vs Panamá. Serie C No. 193.

Corte Interamericana de Derechos Humanos (2011).Caso FontevecchiayD'amico Vs. Argentina. Serie C No. 238.

Corte Suprema de Justicia de la Nación Argentina (2013). Grupo Clarín S.A. y otros c/ Poder Ejecutivo Nacional y otros/ Acción meramente declarativa. Ver voto de la mayoría. Octubre 29.

Cox, J. (2020). TwitterSaysItBlocked NY Post Hunter BidenArticleBecauseItContainsHacked Data.https://bit.ly/2MNSDxB

Daus, G. (2019). Entrevista: Walter Sosa Escudero y la big data: el experto ante el diluvio de datos. *Clarín*.https://bit.ly/39AzxnC

Del Campo, A. (2020). *¿La desinformación en democracia o la democracia de la desinformación?*Univ. de Palermo. Facultad de Derecho. Centro de Estudios en Libertad de Expresión y Acceso a Información. Septiembre.

EuropeanCommission (2019). Ethicsguidelinesfortrustworthy AI. https://bit.ly/3alzqM1

Facebook (2020). Normas Comunitarias. Punto 21. https://bit.ly/3cwmyFm

Facebook Transparency (2019). Content RestrictionsBased on Local Law. https://bit.ly/3ameWCW

Lions, T. (2018). HardQuestions: What'sFacebook'sStrategyforStopping False News?https://bit.ly/3j6VGwZ

Marantz, A. (2020). Why Facebook Can'tFixItself. *The New Yorker.*https://bit.ly/3r5oMQa

New America (2020). So WhatShould Companies Do? *New America*https://bit.ly/3je4DET

Organización de las Naciones Unidas, ONU (1966). Pacto Internacional de Derechos Civiles y Políticos. Diciembre 16.

Organización de las Naciones Unidas, ONU (2011). Comentario General n 34 del Comité de Derechos Humanos. Párrafo 9. Septiembre 12.

Organización de las Naciones Unidas, ONU (2018). Informe del Relator Especial sobre la promoción y protección del derecho a la libertad de opinión y de expresión. Abril 6.

Organización de las Naciones Unidas, ONU (2018). Promoción y protección del derecho a la libertad de opinión y expresión. Agosto 29.

Secretaría General de la Organización de los Estados Americanos (2017). Libertad de expresión: a 30 años de la Opinión Consultiva sobre la colegiación obligatoria de periodistas: Estudios sobre el derecho a la libertad de expresión en la doctrina del Sistema Interamericano de Derechos Humanos. Bogotá.Colombia. Noviembre.

Twitter (2020). Política relativa a los contenidos multimedia falsos y alterados. https://bit.ly/3raXE2f

Twitter (2020). Acerca de las excepciones de interés público en Twitter. https://bit.ly/36zKXGy

Fecha de envío: 2020/10/31; Fecha de aceptación: 2021/01/31;
Fecha de publicación: 2021/03/01