Inteligencia artificial para la moderación de contenidos: ¿solución o problema?

La IA generativa está en todos lados, en la prensa, en la televisión, en las redes sociales, todo el mundo habla de ella. En cierta manera, también la IA predictiva. Pero lo cierto es que poco o nada se habla sobre la IA para la moderación de contenidos, utilizada por las grandes plataformas de redes sociales para fiscalizar el cumplimiento (o, mejor dicho, el incumplimiento) de la normativa y sus políticas sobre contenido ilícito como: desnudos, violencia, acoso, discurso de odio, actividades ilegales, etc.

¿Cómo funcionan los sistemas de moderación de contenido?

Un usuario postea su contenido. Acto seguido, la IA verifica la conformidad con su política. De hecho, esta revisión, en algunos casos, ocurre antes de que el contenido se "postee", es decir, antes de que se publique. Algunas categorías de contenido que pueden ser consideradas como ilícitas son claras, otras son más grises, por lo que deben ser remitidas a un humano para su revisión y/o, en su caso, etiquetado. De hecho, de conformidad con lo establecido en el art. 16 de la DSA, las plataformas tienen la obligación de poner a disposición de los usuarios mecanismos de notificación de contenido ilícito. Si el contenido se califica como ilícito, la plataforma puede adoptar distintas acciones, por ejemplo, restricciones de visibilidad, suspensiones del servicio o de los pagos o suspensiones o supresiones de la cuenta del usuario.

El contexto, siempre el contexto.

Lo que puede ser considerado contenido ilícito o no depende, como no, del contexto. La incapacidad de tener en cuenta el contexto se acrecienta con los mecanismos de moderación de contenido asistidos por IA.

Que se lo digan sino a aquel padre que envío fotos de los genitales de su hijo a su doctor y Google, al subirse automáticamente a su nube, las calificó como contenido de abuso sexual. Fue investigado por la policía y todo se aclaró, pero Google se negó a reactivar su cuenta. Hay más ejemplos en plataformas como Twitter (ahora "X") o Youtube.

Y ello ocurre por la incapacidad de los sistemas de IA de tener en cuenta el contexto, al interpretar literalmente el texto o las imágenes. Piénsese en como Facebook contribuyó a amplificar la violencia en Myanmar contra los Rohingyas. Los locales reportaron numerosas veces a la plataforma la ilicitud del contenido sin que Facebook tomara ninguna acción y, cuando lo hizo, tardó más de 48 horas.

Seguramente, un humano no tendría demasiado problema en tomar la decisión correcta, si no fuera porque las normas que las compañías facilitan a los moderadores de contenido les fuerzan a tomar las decisiones en unos pocos segundos, siguiendo las políticas predeterminadas de la compañía que se trate.

En otros casos, tener en cuenta el contexto es imposible, puesto que no hay información adicional en el contenido o la persona que debe tomar la decisión no tiene capacidad real para ello. En estos casos, tomar una decisión, incluso por un humano, no deja de ser una tarea prácticamente inviable ya que en ciertos países no existen moderadores de contenido y se centralizan las tareas en lugares que no pueden entender lo que está pasando a miles de kilómetros (sobre todo, en países de lengua no inglesa).

Al contexto, evidentemente, se suma el idioma. Cuando los moderadores de contenido no conocen el idioma local, las plataformas utilizan herramientas de traducción automática, con IA integrada, claro. El problema es que los textos que forman el corpus del modelo no contienen información o disponen de poca, de todos los idiomas del mundo. Incluso si las traducciones fueran perfectas, se necesitaría información cultural sobre el contexto.

Una consecuencia de esta falta de conocimiento de la cultura local es la homogeneidad de las políticas de las plataformas sobre lo que se puede considerar contenido ilícito, normalmente, al localizarse su sede en USA, estas suelen ser “USA-centric”.

Mecanismos de IA para la moderación de contenido

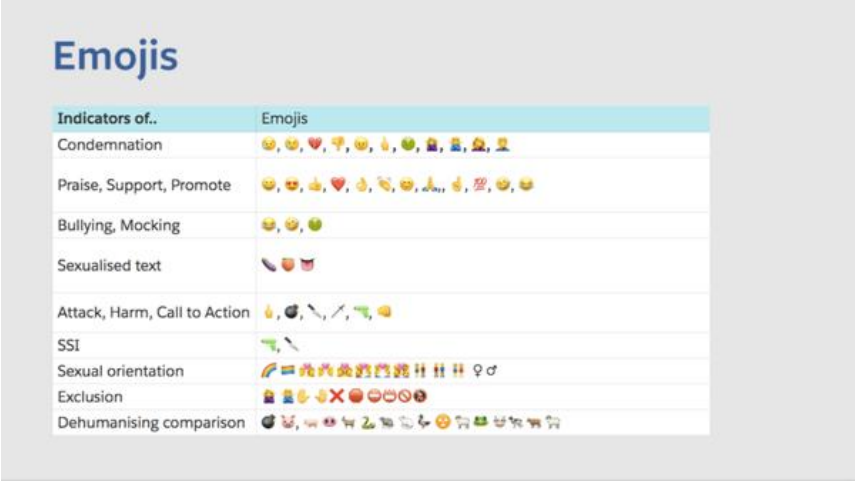
Existen dos mecanismos que utilizan IA para moderar contenidos. Uno es el "fingerprinting matching" y el otro es el aprendizaje automático.

El primero es utilizado para detectar contenido como fotos o videos dentro de las categorías prohibidas de las que el sistema dispone en sus bases de datos. Cuando un usuario sube contenido el mismo es comparado con la base de datos para verificar si hay alguna coincidencia con material catalogado como ilícito. Dicha base de datos está conformada, en su mayoría, por contenido que previamente ha sido etiquetado como ilícito de forma manual. Existen herramientas automatizadas, claro, como la que desarrolló en 2018 Google, si bien cierto tipo de contenido (CSAM[1] -abuso sexual infantil-) es mucho más complejo de etiquetar como ilícito que otro (un gato).

Para otro tipo de contenido ilícito como el discurso de odio, la desinformación o la incitación a la violencia, las herramientas de machine learning (ML) son predominantes. No obstante, una vez dichas herramientas han aprendido los patrones para distinguir, por ejemplo, lo que es desinformación de lo que es información veraz, los modelos deben ser reentrenados constantemente para mantenerse actualizados. La sociedad cambia, las políticas cambian (Facebook ha modificado sus políticas internas en muchísimas ocasiones), los hábitos culturales cambian, por ello es necesario ese reentrenamiento continuo de los modelos.

En 2018 se filtraron las normas internas de Facebook sobre discurso político. Los documentos establecían la interpretación que debía darse, por ejemplo, a determinados emojis. En general, los documentos planteaban serias dudas sobre el enfoque de Facebook para la moderación del contenido.

[1] https://support.google.com/transparencyreport/answer/10330933?hl=es-419#zippy=%2Cqu%C3%A9-es-csam



Fuente: https://www.nytimes.com/2018/12/27/world/facebook-moderators.html.

Es posible que en el futuro se utilicen modelos de lenguaje que se mantengan actualizados a través de información disponible en la web o mediante búsquedas en tiempo real a los efectos de detectar este tipo de contenido ilícito. No obstante, en el estado actual de la técnica tampoco parece que sea una buena idea, estos modelos todavía "alucinan", además, podrían potenciar cierto tipo de creencias mayoritarias (científicas, culturales, etc.) que se encuentran disponibles online, acallando aquellas más minoritarias, pero no por ello inciertas o erróneas.

Hay que tener en cuenta, a su vez, lo fácil que es para los usuarios sortear la moderación de contenido. ¿Cómo? Se llama "*algospeak*"[2], un fenómeno reciente, nacido de la necesidad de comunicarse libremente o de evitar la censura en redes sociales, sin ser penalizado por los algoritmos de moderación de contenido. Como los algoritmos no entienden del todo bien el contexto de las palabras, los usuarios modifican términos sensibles para evitar bloqueos o reducciones de visibilidad, por ejemplo, en inglés, "corn" se refiere a "porn".

[2] https://en.wikipedia.org/wiki/Algospeak.

Un ejemplo del problema de automatizar los mecanismos de moderación de contenido fue el caso de Kim Phuc. Kim Phuc tenía nueve años cuando fue fotografiada corriendo tras un ataque de napalm al norte de Saigón en junio de 1972, durante la guerra de Vietnam. Sufrió quemaduras horribles. Un fotógrafo y un corresponsal de guerra la llevaron al hospital. Les dijeron que no se esperaba que sobreviviera. La fotografía, en 2016, se subió a Facebook por parte de un medio Noruego como parte de un reportaje sobre imágenes de guerra. La fotografía fue suprimida por Facebook por considerarla pornografía infantil y desnudez.

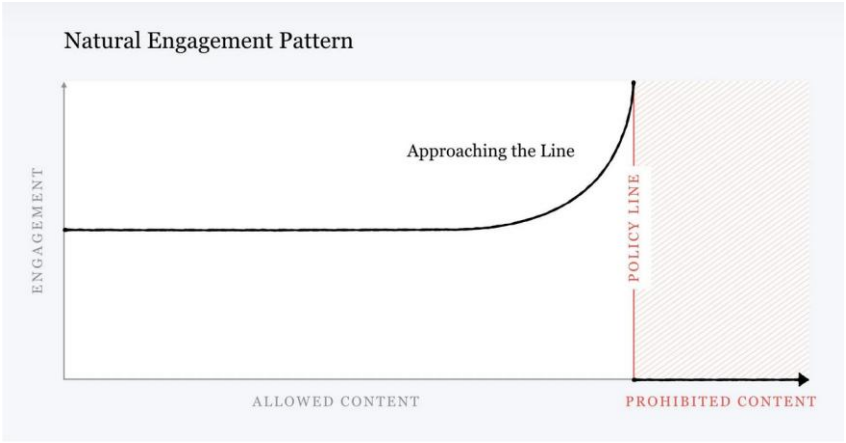
Phan Thi Kim Phuc, de nueve años (centro), al huir de su aldea en el sur de Vietnam en 1972. Se había despojado de su ropa quemada luego de que su comunidad fue bombardeada con napalm. Nick Ut.

La cuestión de la moderación de contenido no es baladí, como vemos. No olvidemos que incluso algunas personas buscan ayuda a través de las redes sociales momentos antes de quitarse la vida e incluso, lamentablemente, algunas de ellas lo han retransmitido en directo. O piénsese en retos virales como por ejemplo el "*blackout challenge*"[3], que consiste en grabarse mientras se practica la sofocación intencional hasta desmayarse, privando al

[3] https://es.wikipedia.org/wiki/Desaf%C3%ADo_del_apag%C3%B3n_(blackout_challenge).

cerebro de oxígeno, que pueden acaban resultando en la muerte de menores, y en los que los algoritmos recompensan a los usuarios con mayor atención por parte del resto, ya que, claro, más tiempo es igual a más anuncios y, consiguientemente, a más ingresos.

No podemos tampoco pasar por alto que, como el propio Zuckerberg indicó, el contenido que está próximo a violar las políticas de la plataforma o se ser considero ilícito, es más probable que genere mayor *engagement* con los usuarios. Por ejemplo, la desinformación, la violencia o el contenido extremo son lo que él mismo llamó "natural engagement pattern".



La razón por la que los algoritmos de las redes sociales amplifican el contenido dañino o la desinformación es que estos no operan en base al contenido subido sino en base a la interacción y al comportamiento de los usuarios con el mismo. Cuando un usuario realiza una publicación que es invisible a los algoritmos de moderación de contenido la misma es también invisible a los algoritmos de recomendación de contenido. Y, claro, una vez el usuario ha interactuado con el contenido amplificado, ilícito o rozando el ilícito, le es servido otro contenido similar por la plataforma. El pez que se muerde la cola.

Moderación de contenido y anuncios fraudulentos.

Además de contribuir a la desinformación, las redes sociales como Facebook, Instagram o TikTok son utilizadas como canales para difundir anuncios fraudulentos, eso no es ninguna novedad. Soluciones milagrosas para la pérdida de peso, el crecimiento del cabello, remedios médicos sin respaldo científico o inversiones millonarias en criptomonedas son solo algunos ejemplos.

Aunque las plataformas implementan medidas para detectar anuncios fraudulentos, estos sistemas no son infalibles. Los anunciantes evitan palabras o frases o utilizan expresiones ambiguas para eludir los sistemas de moderación automática. Los anunciantes también cambian ligeramente el diseño de los anuncios, lo que hace que los filtros, que generalmente dependen de la comparación con bases de datos previas de anuncios fraudulentos, no puedan detectarlos con precisión. Los cambios pueden ser sutiles, como una variación en la imagen o una ligera modificación del texto, lo que permite que el anuncio pase desapercibido. También es perfectamente posible que un anuncio de lo más burdo, con numerosas faltas de ortografía, pase los filtros de las plataformas.

Gran parte de la pasividad de las plataformas recae en su modelo de negocio, basado los ingresos de la publicidad derivados del pago por clic o por impresión, independientemente de la legitimidad del anuncio.

Si este tema te interesa, no dejes de ver el documental "The Clik Trap", que *"expone las alarmantes verdades que se esconden tras la economía de la publicidad digital, valorada en cientos de miles de millones de dólares, y muestra el impacto trascendental que esta tiene en nuestras vidas, tanto dentro como fuera de Internet. La nueva tecnología publicitaria, monopolizada por las grandes empresas tecnológicas, ha convertido el odio y la desinformación en una oportunidad de negocio rentable, de la que también se aprovechan ciberestafadores de todo el mundo. Al revelar cómo funciona esta compleja industria, quién la controla y quién está tomando medidas en su contra, "The Click Trap" cuestiona si es posible construir en el futuro un modelo publicitario*

digital más seguro, justo y transparente".

Moderación de contenido: RGPD y DSA.

El 12 de septiembre de 2015 el EDPB publicó sus *Guidelines 3/2025 on the interplay between the DSA and the GDPR*[4] sujetas a consulta pública. En las directrices se prevé un apartado específico respecto al tratamiento de datos personales en el uso de herramientas automatizadas para combatir el contenido ilícito, de conformidad con lo previsto en el art. 16 de la Ley de Servicios Digitales (DSA).

En la medida en que las plataformas deban tratar datos personales para identificar el contenido ilícito, la DSA no especifica en qué supuestos dicho tratamiento, contemplado en el artículo 7, sería legítimo o proporcionado. Por lo tanto, y considerando también la amplia definición de «contenido ilícito» de la DSA, es importante describir cómo los proveedores de servicios intermediarios pueden llevar a cabo acciones voluntarias, por iniciativa propia, en virtud del artículo 7 de la DSA, de conformidad con el RGPD.

El primer escenario contemplado es aquel en el que los proveedores de servicios intermediarios realizan el tratamiento en el contexto de sus investigaciones voluntarias por iniciativa propia para detectar, identificar y eliminar (o deshabilitar el acceso a) contenido ilícito. Para cumplir con el RGPD, este tratamiento debe llevarse a cabo de forma lícita, justa y transparente hacia los interesados, observando los principios generales del artículo 5 del RGPD, así como las obligaciones que este impone a los responsables del tratamiento.

En primer lugar, los proveedores de servicios intermediarios, como responsables del tratamiento, deben identificar una base jurídica conforme al artículo 6.1 del RGPD para llevar a cabo las acciones correspondientes sobre la detección, identificación o eliminación del contenido ilícito. Dado que los responsables del tratamiento no están legalmente obligados a realizar el tratamiento para estos fines, la base jurídica más

[4] www.edpb.europa.eu/our-work-tools/documents/public-consultations/2025/guidelines-32025-interplay-between-dsa-and-gdpr_en.

adecuada en este caso sería el artículo 6.1 f) del RGPD (intereses legítimos), siempre que se supere el triple juicio de necesidad, licitud o legitimidad y balance de intereses.

El segundo escenario contemplado por el artículo 7 de la DSA es aquel en el que los proveedores de servicios intermediarios realizan el tratamiento de datos de buena fe y con diligencia para cumplir con los requisitos del Derecho de la Unión y del Derecho nacional. Si bien el artículo 8 de la DSA impide a la UE o a los Estados imponer a los proveedores de servicios intermediarios obligaciones generales de supervisar la información que transmiten o almacenan, o de investigar hechos o circunstancias que puedan suponer una actividad ilícita, puede haber obligaciones legales específicas, en virtud del Derecho de la UE o de los Estados miembros, para que dichos proveedores detecten y aborden el contenido ilícito. Un ejemplo es el supuesto en que un usuario ejerce un derecho de supresión (art. 17 RGPD) sobre sus datos y la plataforma debe localizar el contenido en el que aparecen.

Así, en la medida en que dichas acciones requieran el tratamiento de datos personales, el cumplimiento de las obligaciones legales puede constituir una base jurídica para el tratamiento en virtud del Artículo 6.1 c) del RGPD.

Si el tratamiento se lleva a cabo en el contexto del cumplimiento de una orden emitida por una autoridad competente para actuar contra contenidos ilícitos o para proporcionar información sobre uno o más destinatarios individuales específicos del servicio, los prestadores de servicios intermediarios también deben comprobar que la orden cumple los requisitos establecidos en los artículos 9 o 10 de la DSA para justificar el tratamiento en virtud del artículo 6.1 c) del RGPD.

Otra cuestión importante es que las decisiones tomadas en virtud del artículo 7 pueden considerarse decisiones individuales automatizadas del art. 22 RGPD. Pensemos a estos efectos, como hemos visto, en los algoritmos que buscan contenido ilícito y, sin intervención humana, toman decisiones sobre el mismo con posible afectación jurídica o similar a los interesados. A estos efectos, debe evaluarse, si existe, el grado de participación humana en un sistema que implica el tratamiento automatizado de

datos personales para la detección y eliminación de contenido ilegal: si no hay participación humana, si la participación humana no es significativa, o si el humano se basa en gran medida en la recomendación algorítmica generada por el sistema al decidir si eliminar el contenido, la decisión aún se consideraría basada únicamente en el procesamiento automatizado del art. 22 RGPD. Consiguientemente, los proveedores deberán contar con una excepción del 22.2 RGPD y facilitar la información de los artículos 13 y 14 RGPD.

Por otro lado, para combatir el contenido ilícito en línea, la DSA establece que todos los proveedores de servicios de alojamiento (entre ellos las plataformas en línea, sin importar su tamaño) deben implementar mecanismos de "notificación y acción". Estos mecanismos permiten que individuos o entidades informen, por medios electrónicos, sobre contenidos que consideran ilegales.

Al recibir una notificación, el proveedor debe decidir si el contenido es efectivamente ilícito y puede tomar medidas como eliminarlo, reducir su visibilidad, restringir pagos al usuario, o suprimir o bloquear la cuenta.

Además, entidades calificadas como "informantes de confianza" (*trusted flaggers*), con experiencia en detectar contenido ilegal, pueden usar estos mecanismos de forma prioritaria.

El proceso puede implicar el tratamiento de datos personales del denunciante, del usuario afectado o incluso de terceros, si dichos datos aparecen en el contenido denunciado. En este caso, los proveedores actuarán como responsables del tratamiento de datos. De conformidad con el artículo 16.2 de la DSA, los proveedores deben facilitar y contar la inclusión del nombre y correo electrónico del informante, excepto en casos de delitos relacionados con abuso infantil (CSAM) en que no podrá requerirse la identificación del informante.

Los datos recabados se pueden para:

- Confirmar la recepción de la notificación.
- Comunicar la decisión tomada.
- Informar sobre la imposibilidad técnica u operativa de actuar.
- Notificar al usuario afectado.

El artículo 16.6 DSA permite el uso de medios automatizados para procesar notificaciones y tomar decisiones, pero los denunciantes deben ser informados de ello. Si la decisión automatizada produce efectos legales o similares, como se ha dicho, se aplican las previsiones del artículo 22 y concordantes del RGPD, que protege a las personas contra decisiones únicamente basadas en tratamiento automatizado de datos y garantiza, de alguna manera, la intervención humana en fase posterior.

Asimismo, el artículo 17 de la DSA obliga a los proveedores a emitir una declaración motivada, clara y específica cuando eliminen o restrinjan contenido, salvo que fuera ordenada por una autoridad judicial o administrativa (artículo 9 DSA) o se trate de un contenido engañoso de volumen considerable.

Conclusión. ¿Por qué la IA no es la solución a la moderación de contenido ilícito?

Existen razones intrínsecas al estado del arte actual de la tecnología, otras que tienen que ver con los modelos de negocio de las grandes plataformas y otras con las redes sociales en sí mismas.

La IA no es demasiado buena entendiendo el contexto y los detalles, cosa que, en principio, sí pueden hacer los humanos adecuados. Si bien ello puede mejorar en el futuro, será complicado que cuando se evalúe la licitud de un contenido se tenga en cuenta el contexto adecuado. Por otro lado, no hay ninguna razón, más allá de los motivos económicos que llevan a no invertir a las plataformas, para explicar porque tanto los mecanismos automatizados como los humanos se desempeñan mejor o con más dedicación en algunos países que en otros.

El hecho de que el mundo cambie cada cierto tiempo y la sucesión de determinados hechos (por ejemplo, la pandemia de la COVID-19) es una barrera importante para la IA en cuanto a la moderación de contenido, tanto para los mecanismos de etiquetado como para el ML.

No pueden tampoco dejarse de lado aquellos casos en los que la moderación de contenido ilícito pueda tener consecuencias en el mundo real, puesto que es difícil predecir las consecuencias de dichos actos (cuando se silencia contenido violando derechos fundamentales o se permite cierto contenido que supone incitación al odio contra minorías).

También es posible que la regulación, como la DSA, tienda a resultar en una suerte de censura colateral por el que las plataformas moderan contenido de manera mucho más agresiva que si no estuvieran obligadas a ello para guardarse las espaldas. En este sentido, la regulación supondrá un reto importante para la moderación de contenido ya que, por otro lado, las redes sociales, en la actualidad, son un lugar de disputas políticas en el que la IA poco puede decir respecto al contexto cultural, ético o legal, cuestión que corresponde, en esencia, aunque no sin dificultades, a los humanos.

Gerard Espuga Torné
Abogado.
gerardespuga@betalegal.com