Libertad de expresión en línea

Redes sociales, algoritmos y moderación automatizada de contenidos

Micaela Mantegna
Diplomatura en Gobernanza de Internet
Universidad de San Andrés
Septiembre 2017

"the Internet interprets censorship as damage and routes around it." - John Gilmore

La potencialidad de los desarrollos de machine learning para el reconocimiento de lenguaje natural permiten su creciente uso en aplicaciones de análisis y filtrado de contenido online.
Por un lado, las redes sociales lo utilizan como herramienta de moderación de comentarios o para "proteger" a sus usuarios del hostigamiento o ciberacoso.
Dado el grado de dispersión de las noticias falsas y el creciente rol que tuvieron en procesos electorales democráticos para manipular la opinión pública a través de la creación de tendencias y propagación de rumores, informaciones falsas o mal documentadas, la inteligencia artificial se plantea como una herramienta eficaz para poder manejar un fenómeno de otra forma ingobernable e incontrolable.
Sin embargo, la utilización de algoritmos para filtros y sugerencias conllevan el peligro de crear burbujas (*filter bubbles*) de información que no nos expongan a otras líneas de pensamiento ni al debate, lo que en el fondo acarrea el peligro de relacionarnos discursivamente solo con aquellos de pensamiento similar, y por ende reforzar la razón o certeza sobre las propias ideas y convicciones (*confirmation bias).*
Este debate se inserta en el marco de un problema mayor sobre el ejercicio de la libertad de expresión y sus límites en el discurso de odio (*hate speech*), así como los parámetros de la remoción de los contenidos en línea, directamente ligados a la frontera que se trace en relación a lo primero.

Reacciones contra el discurso racista, el surgimiento de la "tecnomoral"?

Los acontecimientos en Charlottesville, Virginia tuvieron el impacto en poner en el frente del debate público los límites sobre la libertad de expresión, el discurso de odio y la censura.
En este thread de Twitter (https://twitter.com/GoDaddy/status/896935462622957573), una usuaria con el peso de una cuenta con muchos seguidores (lo que se conoce como un *influencer* en la jerga) inquirió a Go Daddy, proveedor de DNS, acerca de porqué continuaban *hosteando* a The Daily Stormer tras haber publicado una foto de la victima del atentado de Charlosttesville rodeada de epitetos injuriosos.

> *Amy SiskindCuenta verificada @Amy_Siskind 14 ago. @GoDaddy you host The Daily Stormer - they posted this on their site. Please retweet if you think this hate should be taken down & banned.*

The Daily Stormer es un website de noticias neonazi y de supremacía blanca estadounidense que se considera parte del movimiento denominado *alt-right*, que alojaba su sitio hasta ese momento a través de la compañía Go Daddy.
Go Daddy respondió a ese tweet informando que se le otorgaba a The Daily Stormer un plazo de 24 horas para mover su sitio a otro proveedor, por haber violado los términos de servicio.

> *We informed The Daily Stormer that they have 24 hours to move the domain to another provider, as they have violated our terms of service. 20:24 - 13 ago. 2017*

Este movimiento, sean sus intenciones de genuino repudio u oportunamente publicitarias para aprovechar la tracción del *momentum* político que se vive en Norteamérica, generó un efecto dominó sobre otros respecto de sitios de igual tenor. Stormfront.org, uno de los sitios más antiguos en la red con propaganda sobre supremacía blanca, también sufrió la misma suerte, siendo suspendido por su proveedor. Las decisiones corporativas basadas en estos argumentos se extendieron de los DNS a otros ámbitos, Airbnb impidió que usuarios asociados a los movimientos racistas hicieran uso de su plataforma para alojarse y concentrarse previo a las manifestaciones. Uber anunció a sus usuarios que no estaban obligados a transportar racistas, PayPal Inc. bloqueo el uso de la plataforma para procesar donaciones a sitios que promuevan la intolerancia racial (Paasche 2017), Apple bloqueó el uso de Apple Pay para sitios que vendieran mercadería vinculada a propaganda supremacista (Morris 2017), Spotify removió bandas asociadas a estos grupos, y Facebook cerró páginas sobre el tema ("Right Wing Death Squad" y "White Nationalists United), basándose en su política sobre discurso de odio.
Por su parte, Cloudflare revirtió para este caso su política de larga data de mantenerse neutral al contenido y removió a The Daily Stormer de su servicio de protección contra DDoS (lo cual en términos de protección para la la libre expresión online equivalente a alguien atado en un campo de batalla y con un blanco pintado en la frente).
El caso de Cloudflare es interesante porque su CEO, Matthew Prince, expresamente recalcó en una entrevista con Gizmodo que se trataba de una decisión sobre el caso particular, y no de un cambio en la política general del servicio a futuro, remarcando que cree que debe discutirse seriamente sobre qué parte de la infraestructura de Internet es la adecuada para decidir sobre contenidos (*registrars, browsers, o social networks*). Además, que este trabajo corresponde a los mecanismos gubernamentales si el contenido viola la ley, y no a las decisiones personales o corporativas de los CEO de las empresas tecnológicas (Conger 2017). En una carta interna y un post subsequente explica los fundamentos personales de su decisión, admitiendo expresamente que importó literalmente “echar a alguien de internet” y que nadie debería tener ese poder (Prince 2017).
Cuestiones similares se habían planteado en cuanto al rol que los servicios contra denegaciones distribuidas de servicio (DDoS) tienen en la protección de la libre expresión, cuando sitios como Krebs On Security (“The Democratization of Censorship — Krebs on Security” 2017) demostraron que no podían mantenerse online sin ellos. En el caso de Krebs, la escala del ataque al que estaba siendo sometido hacía que no fuera rentable protegerlo, lo que en definitiva significó que el sitio fue silenciado al quedar offline. Los ataques de denegación de servicio muestran una doble cara, por un lado son una herramienta del

hacktivismo para ejercer una forma de protesta civil (Sauter 2014), mientras que por el otro operan una censura de facto por el poder de la fuerza bruta de las botnets.
Afortunadamente, existen iniciativas como "Project Shield", que utiliza la infraestructura de Google para brindar protección a organizaciones periodísticas o activistas que sufren ataques de DDoS cuando publican contenido controversial o que cuestiona instituciones o gobiernos, y que no pueden afrontar el costo de servicios privados.
Dado que el efecto de estos ataques es justamente invisibilizar contenidos y sitios, el proyecto cuenta con un mapa para visualizar los ataques DDoS más poderosos en tiempo real, con lo cual se puede mapear su coincidencia con conflictos sociales o disputas políticas.
Por su parte, el Berkman Klein Center for Internet and Society auspicia el proyecto herdict.org

Rol de las plataformas privadas que funcionan como un foro público. La noción de lo público en Internet.
Podemos poner como guardianes de la libre expresión a privados? Este esquema funciona en la medida en que todos esos intereses se alinean, los del discurso libre y una internet abierta con los valores que estas empresas quieren sustentar. Pero cuando las antorchas digitales claman en sus puertas, y el peso de un desastre de relaciones públicas es inminente, los términos y condiciones de servicio funcionan como una válvula reivindicatoria del carácter privado de sus emprendimientos.
En una nota The Guardian tituló muy acertadamente que The Daily Stormer fue "expulsado de Internet" (kicked off the internet), lo que señala como una muestra de la irrelevancia de la primer enmienda frente a las autorregulaciones privadas de los términos de servicio.
Un fenómeno similar ocurre en otras áreas del derecho, donde la pérdida de jurisdicción de los estados para hacer cumplir las decisiones basadas en su soberanía y normas de alcance local, se enfrenta frente al carácter global, distribuido y abierto de internet, regido informalmente por un compendio supranacional de normas informales, los "términos y condiciones de servicio". Irónicamente, el *enforcement* o la fuerza de validez jurídica de estos cuerpos para el caso individual descansa en su remisión a normativas contractuales regionales.

En el tratamiento de su responsabilidad concurren ideas similares a las que forjaron la de los medios de prensa tradicionales, donde la normativa constitucional y la jurisprudencia (desde (New York Times Co. v. Sullivan 1964)en adelante) ponderó el mayor valor social de su rol como intermediarios de la información que permite la libre circulación de las ideas. Castigar a través de sanciones económicas de responsabilidad por informaciones falsas o incorrectas podía conducir a la autocensura (al evitar publicar aquello sobre lo que no se tenía certeza) o a demoras (mientras se cotejaba) que en definitiva malograban el acceso al mercado de las ideas y al debate público de la información. Los medios cumplian así un rol de custodios de uno de los pilares de un gobierno republicano, en tanto permitían mediatamente la publicidad de los actos de gobierno.
De similar manera, en el ecosistema online, los intermediarios y las plataformas tienen una posición muy poderosa como dueños de los espacios virtuales de comunicación e interacción social, a la par que un estatus protegido por normas de irresponsabilidad o *safe harbors* en relación a los contenidos.

La idea de las redes sociales y los espacios virtuales como representaciones modernas de las “plazas públicas” donde se ejerce expresión protegida por la primer enmienda fue respaldada por el reciente precedente de la Corte Suprema de los Estados Unidos en (Packingham v. North Carolina 2017)En el caso, una ley del Estado de Carolina del Norte prohibía a los convictos por delitos sexuales acceder a sitios y redes sociales donde sepa que menores pueden ser miembros o tener páginas personales, lo cual implicaba en la práctica, vetarlos de la mayoría de las redes sociales.

En este precedente la Corte señaló como principio fundamental derivado de la Primera Enmienda, que todas las personas tengan acceso a lugares de intercambio de ideas, foros para ejercer la libre expresión a través de reuniones públicas donde puedan hablar, escuchar o protestar. Ese rol, cubierto otrora por los parques y lugares públicos, se ha trasladado al ciberespacio, y en particular a las redes sociales, que ofrecen capacidades relativamente ilimitada y de bajo costo para realizar todo tipo de comunicaciones (*Reno v. American Civil Liberties Union, 521 U. S. 844, 870).*

El tribunal destaca que las redes sociales funcionan como una “moderna plaza pública” en tanto permiten a los usuarios el acceso a la información y la intercomunicación, siendo principales fuentes para conocer los acontecimientos actuales o buscar empleo, además de que proveen uno de los más poderosos vehículos para ejercer el derecho ser oído, citando los intercambios políticos que se desarrollan en Twitter. Aunque no lo explicita con ese ejemplo, no cabe duda del rol crucial que Facebook y Twitter tuvieron en la elección presidencial, ni de la potente plataforma que es la cuenta de Donald Trump. A tal punto, que las menciones de determinadas empresas en tweets iracundos repercutieron significativamente en el precio de las acciones de estas firmas en la bolsa (Ingram 2017), gracias a la velocidad de algoritmos bursátiles automáticos (conocidos como *high frequency trading*) que capturan instantáneamente las observaciones de Trump en Twitter y luego compran o venden inmediatamente las acciones afectadas (Peltz 2017).

Por otro lado, la cuenta misma funciona como una herramienta comunicacional *sui generis,* en la medida en que la utiliza para diseminar tanto ideas propias como de gobierno, fuera de los canales oficiales. En una ironía reivindicatoria de la posmodernidad online, una de las figuras públicas con mayor peso global se vale de una plataforma privada como principal vía de expresión. Sus tweets son analizados y replicados por los medios, naturalizandolos como un canal más de la expresión de su investidura, a veces con mayor impacto que los mismos discursos presidenciales.

Esta dinámica da mayor peso a la demanda entablada por un grupo de usuarios bloqueados por Trump, con fundamento en la violación a la Primer Enmienda, alegando que la cuenta funciona como un foro público del cual, el Presidente como figura pública, no puede ni debe excluirlos (Savage 2017).

En este sentido, el caso de las plataformas online no es totalmente asimilable al ejemplo de la plaza, por cuanto en éstas tanto la función de intercambio social como el espacio en que se realiza, son ambas de carácter público. Las redes sociales son en cambio espacios privados, en los que se lleva adelante la satisfacción de un rol público. Aunque puede parecer a primera vista irrestricto, ya que cualquiera puede abrir una cuenta en Facebook o Twitter teniendo una cuenta de mail, en realidad el acceso de los individuos y su “permanencia” en las mismas se mantiene en la medida en que se acaten los términos y condiciones de servicio. En este

sentido, las plataformas de redes sociales son más similares al concepto de “privately owned public spaces” (POPS), que son espacios construidos y mantenidos por un propietario/desarrollador privado para uso público, a cambio de la concesión de alguna ventaja por parte de los gobiernos.

Regulación vs. gatekeeping

A diferencia de los medios tradicionales, los contenidos no circulan en un sentido unilateral frente a receptores pasivos, sino que la web permitió la participación e interacción, así como el *engagement* de los usuarios en la generación de contenidos. Justamente, una de las características de Internet a partir de la llamada web 2.0 es la participación de los usuarios en la producción de contenido, lo que ha transformado a la red en un bastión donde cualquiera podía verse tener su voz.

En los medios tradicionales la formación de opinión estaba intermediada por custodios que curaban celosamente el acceso del ciudadano común, mientras que las plataformas digitales fomentan la generación de contenidos por los usuarios (*user generated content, UGC*) como parte estructural de sus modelos de negocios. Desde la moda al periodismo ciudadano, a la fama viral de los youtubers, instagramers y tuitstars, internet ha demostrado el poder subversivo para socavar el poder de esos intermediarios.

Internet, a través de fenómenos como el blogging, redes sociales y seguidores, se posicionó como un foro donde cualquiera con iniciativa, ideas e inventiva podría acceder tanto a poder plasman sus ideas como hacerlas llegar a otros.

Una de las banderas para mantener los safe harbor que los eximían de responsabilidad era básicamente la posición de su rol como intermediarios sin injerencia en el contenido, sobre el cual solo eran responsables los usuarios. Una aserción de la neutralidad tecnológica de las plataformas frente a terceros, que los eximia de obrar *a priori* como guardianes de los contenidos. Sin embargo los términos y condiciones cuentan otra historia, ya que implican un *guideline* de conductas admisibles para los usuarios, tanto en el terreno de lo ilícito como de lo que cada plataforma considera como inmoral (algunas son más permisivas por ejemplo al desnudo). Denunciada o detectada la conducta pasan a ejercer un poder de veto *ex post* con la remoción del contenido, y un rol disciplinario sobre las cuentas de los usuarios (desde la suspensión hasta la eliminación permanente). Para el usuario común puede no pasar de la molestia de abrir una nueva cuenta, pero en cuentas con una audiencia cuidadosamente construida y donde los *followers* son el capital inmaterial de la marca, estas sanciones funcionan como una suerte de *capitis diminutio* digital. La pretensa libertad de contenidos está en realidad limitada al campo cuidadosamente demarcado por los términos y condiciones, que pueden pasar de funcionar como elemento regulador de conducta a un instrumento que fomenta la autocensura.

Bajo el rótulo de “*corporate censorship*”, Rebecca McKinnon narra distintos episodios de control de contenidos por parte de Apple en su mercado de aplicaciones, donde se censuran expresiones legítimas de disenso político y parodia, concluyendo que en el gobierno de nuestro acceso a las aplicaciones las empresas muestran poderes soberanos y un preocupante desdén por los derechos de expresión de los ciudadanos (MacKinnon 2013).

Tensiones entre libre expresión y los intereses de los afectados. Limites. Harmful speech, definición.

Aunque detrás se encuentren los intereses corporativos no expresados, las justificaciones para el ejercicio de facultades de vigilancia de contenido enarbolan la bandera de la protección de los usuarios.

Ciertamente, este argumento apela a la tensión subyacente entre la libre expresión y los intereses de aquellos que se sientan afectados por el ejercicio que terceros hagan de esta. En la búsqueda de equilibrio entre estos dos valores tenemos toda una escala de matices, que va desde sobreproteccionismo o extrema sensibilidad frente a determinados tipos de discurso (el cual es regularmente una excusa para silenciar un pensamiento opuesto a la moral personal), a verdaderos ataques individuales de acoso, con fenómenos como los trolls. Decidir que es expresión protegida dentro de esa escala es una tarea difícil y delicada (entendiendo que esta demarcación implicara qué contenidos pueden ser removidos y por ende, silenciados), el límite que se suele encontrar está en el uso de la expresión como discurso de odio.

En los sucesos puestos en marcha luego de Charlottesville se trata de un caso muy cerca de este límite. Se trata de un caso difícil de defensa de la libertad de expresión, porque aunque las reivindicaciones sobre supremacía blanca despiertan el repudio generalizado, los sucesos y mecanismos que pusieron en marcha puede ser tácticas usadas el día de mañana para silenciar a otros grupos de opinión.

Aunque resulte antipático en el caso concreto de los sitios de alt-right, el principal principio en juego es que Internet siga siendo una zona de libertad de expresión. Como indica la Electronic Frontier Foundation "*Proteger la libertad de expresión no es algo que hacemos porque estamos de acuerdo con todo el discurso que se protege. Lo hacemos porque creemos que nadie -ni el gobierno, ni las empresas comerciales privadas- deben decidir quién puede hablar y quién no…"("Fighting Neo-Nazis and the Future of Free Expression" 2017)*

No hay que olvidar que el discurso impopular también tiene protección dentro de la libertad de expresión de la primer enmienda, esto por cuanto el discurso que es ofensivo a la moral de hoy, puede abrir el camino para las reivindicaciones de derechos futuros, como fuera el caso de las reivindicaciones de igualdad de género, raza o el matrimonio igualitario. También se encuentra protegido el discurso ofensivo, máxime cuando tiene ribetes políticos. En (Cohen v. California 1971) la Suprema Corte de Estados Unidos claramente indicó que no puede castigarse el mero hecho de la comunicación, por la sola circunstancia de que el medio elegido para transmitir el mensaje al público haya sido ofensivo (en el caso una chaqueta protestando contra la guerra de Vietnam con las palabras "Fuck the Draft").

La posibilidad de expresarse sobre ideas no populares o en términos que resulten ofensivos sin temor a ser castigados es una precondición para evitar la autocensura. En una era donde nuestras comunicaciones, hábitos e interacciones online son fuertemente vigilados, no solo se necesita fortalecer la libre expresión, sino también la "privacidad intelectual" que la precede y le permite germinar. Neil Richards la define como ese ámbito de reserva donde se puede pensar sobre las ideas, aún las más heréticas, y discutirlas en el fuero interno o con algunos confidentes cercanos, sin miedo a ser descubiertos y como paso previo a su presentación en público (Richards 2015).

El límite del cobijo de este discurso impopular dentro de la libre expresión cede cuando se trata del llamado “discurso de odio”, que comúnmente se refiere a expresiones que atacan o degradan a una persona o personas como miembros de un grupo con características compartidas como la raza, el género, la religión, la orientación sexual o la discapacidad.(Sellars 2016)

El discurso de odio es uno de los tipos de expresiones nocivas o dañinas comprendidos en la noción más amplia del llamado “*harmful speech*”, que incluye una variedad de fenómenos como el acoso online, doxxing, o revenge porn, que impactan en las víctimas causandoles perjuicios legales, físicos o emocionales, y que pueden provenir de un solo individuo, grupos u ataques orquestados.La Comisión Europea vienen tratando estos problemas desde 1996, categorizandolos bajo dos rótulos, contenido ilegal y contenido dañoso (Akdeniz 2013)

Dada la dificultad conceptual de definirlo y catalogarlo, se han utilizado distintos enfoques. Unos priorizan el enfoque desde las víctimas, mensurando el daño producido, otros apelan a analizar la intención del emisor, mientras que una tercer perspectiva enfatiza en la necesidad de evaluarlo de manera holística en su contexto: daño, intención y contenido.

Dentro del ecosistema de usos delimitado por sus propios términos y condiciones, las empresas de Internet contribuyen a dar forma a las expresiones toleradas, sea a través de la palabra, imagenes o video. La industria sostiene una presión creciente para remover y regular el contenido, por lo cual, dada la masividad de la información producida por sus usuarios, se han volcado a la investigación y uso de algoritmos para una automatizar los procesos de supervisión.

Los desafíos de los grandes volúmenes de contenidos, moderación a través de algoritmos, problemas y soluciones.

Como se explicara, la arquitectura institucional respecto de la responsabilidad de los intermediarios no suele ubicarlos en un rol de control previo a la divulgación de los contenidos. Sin embargo, esto no los exime de la presión de intervención para remover contenidos que resulten ilegales o no cumplan con los términos de uso. Muchas veces, esto se realiza a través de mecanismos de denuncia habilitados para que los particulares que se sientan afectados por los mismos (sea porque por ej. los contenidos afecten derechos de autor de su titularidad o simplemente les resulten amorales u ofensivos) funcionen como un control distribuido, masivo y, por sobre todo, de bajo costo para las empresas.

El alto impacto económico de tener un equipo humano moderando contenidos hacía inviable la moderación para muchos sitios, que implicaba que cerraran los comentarios, o bien que existiera una demora creciente entre el posteo y la aprobación, ralentizando la circulación de la información.

The New York Times anunció en Septiembre del año pasado su asociación con el proyecto Jigsaw y la tecnología de Google para aplicar machine learning de manera masiva a la moderación de comentarios. El trato implicaba la cesión de los datos anonimizados de los comentarios a cambio de la construcción de un algoritmo automatizado que supervise las conversaciones en línea en esa sección. El modelo se basa en el aprendizaje sobre la base de datos de los comentarios moderados por su equipo humano, 14 personas que revisan aproximadamente 11000 comentarios diariamente. Las notas abiertas a comentario representaban hasta ese momento solo el 10% de las publicadas. Las reglas de moderación de

contenidos requieren interpretación y contextualización, como puede comprobarse en el experimento online que invita a los lectores a moderar cinco comentarios según las reglas y comparar sus resultados con los de los editores (Etim 2016).

Casi un año después, se anunció la implementación de “Moderator”, un sistema que permitirá abrir a comentarios todas las historias principales por un término de 8 horas durante la semana. Todo esto representa un gran avance para la libre expresión, porque importa que exista un lugar abierto para ejercerla donde antes no lo había. Sin embargo, ver el uso de algoritmos como una panacea es una falacia tecnosolucionista, porque no hay que olvidar que este es uno de los usos posibles, mientras que existen otros implican la *remoción de contenidos o su invisibilización.*

La tecnología detrás de los algoritmos es todavía imperfecta, hay que recordar que todas estas aplicaciones requieren de las computadoras reconocimiento de lenguaje natural y procesamiento inteligente de textos (*computational linguistics, intelligent text processing)*. Las aplicaciones de *machine learning* se basan mayormente en técnicas estadísticas de clasificación y predicción sobre los datos, de ahí que algunos autores cuestionen la “inteligencia” de estos sistemas, apelando a la parábola del “*chinese room*”[1].

Asimismo, bajo una pátina de eficiencia y neutralidad que en el ideario colectivo atribuimos a la tecnología, los algoritmos tienen dos problemas inherentes: el sesgo (*machine bias*) y la opacidad, que si no son evaluados pueden conducir a la discriminación y la falta de responsabilidad por los procesos (*unaccountability).*

Remoción e invisibilización de la información, fragmentación del discurso conforme a los usuarios (filter bubbles)

Por el contexto descrito, las plataformas están mutando su rol hacia uno más activo en la supervisión de los contenidos. Sea por la presión de la legislación, o de los propios usuarios que reclaman protección frente a los fenómenos de abusos y ataques, las empresas están recurriendo a desplegar algoritmos que automaticen estos procesos, lo cual puede terminar en una censura incontrolada. En determinados países de Europa, hay formas de discurso que son ilegales, como en Alemania la negación del Holocausto tiene penas de prisión. Este país aprobó recientemente una ley que impone a las redes sociales el deber de eliminar comentarios ilegales o racistas dentro de las 24 horas de ser notificados, bajo apercibimiento de cuantiosas multas. Este tipo de regulaciones crean incentivos que fomenta el rápido despliegue e implementación de formas de control algorítmico, que frente a la duda, optarán por decantarse hacia la remoción y evitar las sanciones, o que en definitiva es un incentivo a la censura.

Por otro lado, la invisibilización de la información implica que la misma sigue disponible, pero o bien se retacean las herramientas que permiten acceder a la misma (como ocurre con las desindexaciones de los motores de búsqueda), o bien queda sepultada bajo otra tonelada de información que la torna irrelevante. A efectos prácticos, el resultado es el mismo: el mensaje se pierde, aunque grite a todo pulmón, una voz no va a poder ser oída dentro de la cacofonía de un millar de voces.

---

[1] Este argumento que un programa no hace que una computadora sea inteligente, porque el hecho de que pueda realizar determinadas tareas no implica que sepa que las está haciendo, o sea que tenga una mente, o conciencia. El argumento fue presentado por primera vez por el filósofo John Searle en su artículo "Minds, Brains, and Programs", publicado en Behavioral and Brain Sciences en 1980.

En este punto se replica la discusión sobre los algoritmos responsables de posicionar las "tendencias" o que reflejan un "newsfeed". Los trending topics no se crean por una acumulacion meramente numérica de hashtags sobre un mismo tema, sino que atienden a distintos parámetros declarados vagamente por las redes sociales para definir lo que consideran como un tema popular[2]. Algunos de estos parámetros apelan a la personalización (en base a cuentas que se siguen, intereses, etc), lo que deja en descubierto que lo que consideramos tendencia, es en realidad lo que el algoritmo cree que es tendencia para nosotros. Esta clasificación de los contenidos nos es invisible, como usuarios no sabemos qué información está dejando afuera ni porqué, ni qué categoría nos ha adjudicado el algoritmo.

Una técnica empleada en machine learning es el *clustering,* se trata de un procedimiento que permite agrupar a una serie de data points de acuerdo a determinados criterios o variables, sea de similitud o distancia, que se emplea muy frecuentemente en algoritmos de recomendación o marketing de segmentación de usuarios. Puede implicar técnicas de reducción dimensional (aunar variables asignándoles una correlación causa efecto), que pueden asentarse sobre hipótesis erróneas (por ejemplo, tomar como parámetro de alto poder adquisitivo que se compran entradas para el teatro los días sábados, -el día más caro- ,cuando en realidad responde a que se trata del único día de franco, etc.)

Esta "personalización", que en la jerga se presenta como un valor de eficacia deseado bajo el eufemismo de "resultados relevantes para el usuario", conlleva la fragmentación de la información y puede ser un factor más coadyuvante a la fragmentación de Internet. La web contextual no nos va presentar todos los resultados, sino que para encontrarlos tendremos que salir de esa zona en la que el algoritmo nos han encasillado.

En efecto, a nivel de contenidos el peligro de estos algoritmos es que nos "emparejan" con quienes teóricamente se parecen a nosotros y piensan de manera similar, lo que puede crear una burbuja (*filter bubble)* que lleve a confirmar permanentemente nuestro propio sesgo cognitivo, al no exponernos a voces disidentes. Por ello es que una de las soluciones propuestas para estos espacios online es el "counterspeech", esto es presentar visiones diferentes aquella que responde, por ejemplo, al resultado de la búsqueda de un keyword.

En la radicalización de los supremacistas blancos algunos autores apuntan la contribución de los algoritmos de sugerencia de contenidos, que -con una eficaz aplicación de criterios de clasificación exentos de una valoración axiológica del contenido- recomendaban videos que no hacían sino reforzar las posturas de odio y la confirmacion como verdad revelada de hechos que en realidad son una interpretación sobre la historia. En definitiva, los algoritmos funcionan como una especie de escalera de descenso conformada por los pasos de todos los que los precedieron, y que refuerza el camino para los que vienen detrás.

---

[2] Las tendencias se determinan mediante un algoritmo y, de forma predeterminada, se personalizan de acuerdo con las cuentas que sigues, tus intereses y tu ubicación. Este algoritmo identifica los temas que gozan de popularidad en un momento dado, en lugar de los temas que han sido populares durante un tiempo o diariamente, para ayudarte a descubrir los últimos temas de discusión que van surgiendo en Twitter.
Nota: El número de Tweets relacionados con las tendencias es solo uno de los factores que el algoritmo tiene en cuenta a la hora de clasificar y determinar las tendencias. Si hay tendencias y hashtags que se relacionan con un mismo tema, el algoritmo los agrupa. Por ejemplo, es posible que tanto #MondayMotivation como #MotivationMonday se agrupen bajo #MondayMotivation.
Preguntas frecuentes sobre las tendencias de Twitter, https://support.twitter.com/articles/349215

En este sentido, los algoritmos refuerzan la polarización online y la creación de cámaras de resonancia (*echo chambers)* contrarias a la discusión que una sociedad democrática necesita.

Machine Bias y Opacidad

Los algoritmos son tan neutros como los datos con los que se entrenan, si los datos están sesgados o son discriminatorios, estos se van a reproducir y perpetuar en los modelos.

Recientemente Facebook sufrió fuertes críticas a raíz de una investigación de Propublica,en la que se advirtió que sus algoritmos de filtrado de contenidos (dividiendo expresión política de discurso dañino) protegían al sector de hombres blancos pero no así a los niños de color (Angwin 2017). La aplicación que estaban haciendo de las reglas comunitarias de moderación de contenidos estaba sesgada, reproduciendo modelos de discriminación.

Dada la dificultad de poder evaluar correctamente la expresión dentro de su contexto, incluyendo matices semánticos sutiles como la ironía o la parodia, lo recomendable seria que humanos revisen el contenido “flagueado” por los algoritmos como un mecanismo de reserva a prueba de fallos.

Otro componente problemático es la opacidad con la que estos procesos se conducen, tanto en sus reglas (enunciadas generalmente al público de manera muy vaga y laxa) como en la aplicación en cada caso concreto. Las decisiones de remoción de contenidos se realizan a puertas cerradas y en secreto, comunicando solo los resultados. Esta forma de operar perpetua los modelos “black box”, para evitar resultados de censura indebida, necesitan ser transparentes e instituir algún mecanismo de supervisión o auditoría de la bajada de contenidos, particularmente por personas ajenas a las plataformas. Así como se instituyeron los captcha como esfuerzo colectivo para reconocer imágenes, podría pensarse en un mecanismo que imponga ese toque humano de revisión de contextos para colaborar en la moderación de comentarios, como paso previo a la navegación o acceso a determinados sitios.

Bibliografía

Akdeniz, Yaman. 2013. “To Block or Not to Block: European Approaches to Content Regulation, and Implications for Freedom of Expression.” In *New Technologies and Human Rights: Challenges to Regulation*. Routlegde.

Angwin, Julia. 2017. “Facebook’s Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children — ProPublica.” *ProPublica*. June 28. https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms.

Cohen v. California. 1971. Supreme Court of the United States.

Conger, Kate. 2017. “Cloudflare CEO on Terminating Service to Neo-Nazi Site: ‘The Daily Stormer Are Assholes.’” *Gizmodo*. gizmodo.com. August 16. http://gizmodo.com/cloudflare-ceo-on-terminating-service-to-neo-nazi-site-1797915295.

Etim, Bassey. 2016. “Approve or Reject: Can You Moderate Five New York Times Comments?” *The New York Times*, September 20. https://www.nytimes.com/interactive/2016/09/20/insider/approve-or-reject-moderation-quiz.html.

“Fighting Neo-Nazis and the Future of Free Expression.” 2017. *Electronic Frontier Foundation*. August 17. https://www.eff.org/deeplinks/2017/08/fighting-neo-nazis-future-free-expression.

Ingram, Mathew. 2017. “Here’s What a Trump Tweet Does to a Company's Share Price.” *Fortune*, February 24. http://fortune.com/2017/02/24/trump-tweet-stocks/.

MacKinnon, Rebecca. 2013. *Consent of the Networked: The Worldwide Struggle for Internet Freedom*. Basic Books (AZ).

Morris, Chris. 2017. “These Companies Have Banned Hate Groups After Charlottesville.” *Fortune*. Fortune. August 17. http://fortune.com/2017/08/17/hate-groups-google-godaddy-apple-paypal/.

New York Times Co. v. Sullivan. 1964. Supreme Court of the United States.

Paasche, Franz. 2017. “PayPal’s AUP - Remaining Vigilant on Hate, Violence & Intolerance.” *PayPal*. August 15. https://www.paypal.com/stories/us/paypals-aup-remaining-vigilant-on-hate-violence-intolerance.

Packingham v. North Carolina. 2017. Supreme Court of the United States .

Peltz, James F. 2017. “When Trump Tweets, Wall Street Trades — Instantly.” *Los Angeles Times*, January 16. http://www.latimes.com/business/la-fi-agenda-trump-tweets-stocks-20170116-story.html.

Prince, Matthew. 2017. “Why We Terminated Daily Stormer.” *Cloudflare Blog*. Cloudflare Blog. August 16. http://blog.cloudflare.com/why-we-terminated-daily-stormer/.

Richards, Neil. 2015. *Intellectual Privacy: Rethinking Civil Liberties in the Digital Age*. Oxford University Press.

Sauter, Molly. 2014. *The Coming Swarm: DDOS Actions, Hacktivism, and Civil Disobedience on the Internet*. Bloomsbury Publishing USA.

Savage, Charlie. 2017. “Twitter Users Blocked by Trump File Lawsuit.” *The New York Times*, July 11. https://www.nytimes.com/2017/07/11/us/politics/trump-twitter-users-lawsuit.html.

Sellars, Andrew F. 2016. “Defining Hate Speech.” Berkman Klein Center for Internet & Society Research. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2882244.

“The Democratization of Censorship — Krebs on Security.” 2017. Accessed August 30. https://krebsonsecurity.com/2016/09/the-democratization-of-censorship/.