Cuentas y narrativas adversas: Herramientas para identificar ataques en el debate digital y cómo reaccionar

Linterna Verde

Autores
Daniela Flórez
Cristina Vélez
Con el apoyo del equipo de Linterna Verde

Diseño y diagramación
Diego Villate

Abril de 2022

www.linternaverde.co

contacto@linternaverde.co
@linterna

Contenido

Introducción

En Linterna Verde, cada vez recibimos más solicitudes de organizaciones de la sociedad civil y periodistas para analizar la naturaleza y el alcance de cuentas que, a través de acciones coordinadas, técnicas de trolleo, narrativas adversas y posicionamiento de hashtags, intentan afectar sus mensajes y campañas en redes sociales. Desde hace cuatro años, cuando empezamos a documentar este tipo de alarmas, los equipos de comunicaciones tocan nuestra puerta con preguntas similares: ¿cómo diferenciar un par de publicaciones críticas en redes sociales de un ataque? ¿cuándo es mejor quedarse quieto para no amplificar las narrativas adversas y cuándo hay que reaccionar? ¿qué estrategias de acción tenemos disponibles a corto, mediano y largo plazo? Incluso a lo largo de 2021, uno de cada tres monitoreos que realizamos giró alrededor de este tipo de asuntos.

Si bien las redes sociales están llenas de agresiones, no todo este ruido digital debe identificarse como un ataque: antes, es importante clasificar la gravedad, la frecuencia y la dirección de las acciones digitales de cuentas adversas para saber cuándo y cómo actuar. Para ahorrar energía, es fundamental administrar bien las peleas y saber leer el contexto, no solo digital, sino también *offline,* de manera que podamos actuar de forma articulada. Por eso en Linterna Verde decidimos crear un insumo para la evaluación del riesgo con base en las preguntas más frecuentes que recibimos de las organizaciones de sociedad civil.

El resultado es esta guía, cuyo propósito es ayudar a las organizaciones a identificar los elementos centrales de un ataque, reconocer las señales de alerta y determinar los distintos mecanismos de respuesta con los que cuentan. El documento se divide en cuatro partes: en primer lugar, nos detenemos en una definición conceptual necesaria para avanzar en la identificación de un ataque, esto es la diferencia entre ruido digital y abuso en línea; en segundo lugar, profundizamos en los tipos de ataques más comunes que enfrenta la sociedad civil; posteriormente, recogemos algunas pautas para la medición del riesgo teniendo en cuenta las dinámicas y elementos que configuran el ataque así como la caracterización del atacante; finalmente, ahondamos en las respuestas de acción inmediata y abrimos un capítulo alrededor del trabajo narrativo como mecanismo de respuesta en el mediano y largo plazo.

1. Diferenciar el ruido del ataque

Aunque un bombardeo de críticas negativas puede encender las alarmas y llevar a una crisis reputacional, este en sí mismo no debe considerarse como un ataque, pues puede hacer parte de una expresión legítima de descontento o disenso.

Por otro lado, las organizaciones están en el núcleo del debate político, lo que significa, también, estar expuestas a la toxicidad. Hay espacios en las redes donde la agresión y la indignación son la regla. Las redes constituyen escenarios atractivos para promotores de contenidos hostiles, no solo por la posibilidad de reproducir rápidamente un mensaje y alcanzar audiencias masivas a un bajo costo, sino también por los incentivos propios de estas plataformas cuyos algoritmos han evolucionado para privilegiar el contenido más rentable en interacciones por encima del menos dañino. Esta forma de polución ya hace parte del sistema y es necesario aprender a navegarla.

No siempre es claro a quién va dirigido un comentario hostil o si tuvo la intención de agredir. En estos casos, se hace necesario identificar el contexto de emisión del comentario, bien sea la discusión en sí, o bien la situación más amplia de la conversación en el país de origen, lo cual implica un ejercicio analítico más amplio.

Para clasificar algo como un ataque, es clave identificar si los mensajes o contenidos van dirigidos a un destinatario particular y evaluar si hay una clara intención de hacer daño. Cuando el lenguaje abusivo, poco razonable o amenazante va dirigido directamente a la organización, a sus colaboradores o aliados, o cuando busca expulsar a una persona o grupo específico de las discusiones o desincentivar su participación, puede ser una señal de alarma.

Además, la frecuencia con que estas expresiones agresivas se producen debe cumplir con un patrón de continuidad en el tiempo o responder a momentos de coyuntura específicos que ya se tienen identificados. Cuando las agresiones dejan de ser esporádicas, se empieza a configurar lo que la teoría ha denominado como *abuso en línea*; es entonces cuando la situación empieza a requerir atención especial y seguimiento por parte de la organización.

2. Cuándo es necesario preocuparnos: abuso en línea

Si bien no existe una definición axiomática sobre abuso en línea y los términos para referirse a este fenómeno son diversos, esta guía acoge la definición hecha por Pen America según la cual el abuso en línea involucra todos aquellos "comportamientos hirientes persistentes o severos contra una persona o grupo en internet" cuyo efecto, en muchos casos, es sacar a los usuarios de la conversación. Esta definición resulta útil ya que recoge dos elementos centrales para que un comportamiento en la red constituya abuso: de un lado, la persistencia que conlleva repetición o coordinación, y, del otro, la severidad dadas las consecuencias para la reputación e integridad física o moral de la persona o grupo objeto del ataque.

El abuso en línea constituye una categoría amplia para referirse a los múltiples tipos de ataques, intimidación y violencia que tienen lugar en el ecosistema digital. Estos pueden ir desde compartir información privada sobre una persona u organización, hasta el hackeo, el hostigamiento, las amenazas y el discurso violento.

El abuso en línea puede manifestarse en cualquier rincón de internet en el que haya interacción. El anonimato y la posibilidad de escalar rápidamente el ataque sin mayores repercusiones es aprovechado por muchos usuarios para ejercer comportamientos abusivos sobre otros. Las redes sociales son un terreno fértil para este problema, pero no son el único: aplicaciones de mensajería, blogs, salas de chat, plataformas de correo electrónico y secciones de comentarios de sitios web, son algunos de los más utilizados.

A medida que internet se expande y nuevas plataformas emergen, la toxicidad en línea encuentra más espacio. Por ejemplo, durante la pandemia por el Covid-19, el uso extendido de Zoom para la realización de llamadas y conferencias en línea dio lugar al *'Zoombombing',* una forma de sabotaje digital por medio del cual se irrumpe en una reunión virtual sin autorización y se proyecta contenido ofensivo o inapropiado. Como esta, nuevas estrategias para ejercer violencia en línea surgen constantemente, lo que hace que el trabajo de identificación y documentación del abuso en línea deba hacerse de manera constante.

3. ¿Cómo identificar el abuso?

Uno de los mayores retos de la investigación sobre las violencias en línea es determinar cuándo se configura abuso y cuándo no. Ciertos usos del lenguaje como la ironía, el humor, el sarcasmo, la metáfora o las expresiones locales complejizan la interpretación del contenido, el tono y la intención de los emisores, exigiendo un amplio conocimiento del contexto en el que se desarrolla la conversación. No se trata solamente de un reto de la automatización de algoritmos de identificación; incluso las personas que moderan los contenidos de las redes sociales —quienes tienen la difícil labor de proteger a los usuarios frente al discurso violento y de paralelamente velar por la libertad de expresión— tienen dificultades para hacer una marcación precisa, aun estando apoyadas en herramientas computacionales[1]. En efecto, la percepción del abuso en línea suele ser subjetiva: un estudio del Pew Research Center sobre el estado del abuso en línea en 2020[2], muestra que en los Estados Unidos el 41% de usuarios de internet han experimentado ataques o comportamiento abusivo. De estos, el 36% no pensaban que eran víctimas de un ataque y el 21% no estaba seguro de considerar lo que les sucedió como "abuso en línea".

Esta dificultad para nombrar el abuso se debe a que la mayoría de ataques no se ubican en el extremo; en muchos casos la expresión abusiva está dada por el tono del mensaje y no por su contenido. Si bien hay un tipo de mensaje que tiene la intención expresa de hacer daño, insultar, trolear, amenazar o esparcir odio, hay otros más ambiguos que pueden llegar a ser mucho más prominentes, cuya intención es compartir una opinión o interactuar con otros pero lo hacen de una manera que agrede, antagoniza o disuade a los demás de seguir participando en la conversación[3].

El reto es identificar los elementos que subyacen en el mensaje y hacer una lectura amplia del contexto en el que fue emitido. Para ello, es importante comprender las distintas formas de ataque, sus dinámicas y las características del atacante. Esta guía ofrece una mirada a estos tres elementos con el fin de proporcionar un insumo útil para que periodistas, activistas y organizaciones de la sociedad civil desarrollen capacidades que les permitan saber cuándo son víctimas de algún tipo de abuso y cómo enfrentarlo.

1 Rhodes Artificial Intelligence Lab (RAIL). The challenge of identifying subtle forms of toxicity online. Disponible en: https://medium.com/jigsaw/the-challenge-of-identifying-subtle-forms-of-toxicity-online-465505b6c4c9

2 Pew Research Center. Online harassment 2020. Disponible en: https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/

3 Rhodes Artificial Intelligence Lab (RAIL). The challenge of identifying subtle forms of toxicity online. Disponible en: https://medium.com/jigsaw/the-challenge-of-identifying-subtle-forms-of-toxicity-online-465505b6c4c9

4. Formas de abuso en línea

El abuso en línea puede adoptar múltiples formas que no siempre resultan evidentes. En ocasiones, activistas y organizaciones pueden desestimar ataques que constituyen abuso al no saber cómo nombrarlos; en otras, pueden sentir abuso sin que en realidad se configure uno, o apresurarse a responder comentarios adversos que permanecen en rincones de internet, generando un efecto contraproducente al amplificarlos y llevarlos a escenarios *mainstream*.

A continuación se reseñan algunos de los tipos de ataques más comunes que enfrentan periodistas, activistas y organizaciones de la sociedad civil.

Tipo de ataque	Descripción
Troleo	Comportamiento que busca sembrar discordia, división y ridiculización en el ecosistema digital. Suele provenir de cuentas de humanos que ocultan su identidad pero también pueden ser cuentas semi-automatizadas. Los trolls acostumbran a publicar en una misma conversación repetidamente en vez de hacerlo en varias publicaciones.
Reportes falsos	Coordinación para reportar una cuenta o campaña de un usuario o grupo de usuarios haciendo creer a las plataformas que se trata de cuentas o contenidos inapropiados y prohibidos por estas, con el propósito de que sean sancionadas o suspendidas.
Ataques coordinados	Organización de campañas anonimizadas de desprestigio en contra de una persona, organización o sitio web. Estos ataques se coordinan principalmente a través de grupos cerrados y se ejecutan masivamente por medio de cuentas falsas.
Comportamiento inauténtico	La suma de varios actores, contenidos y estrategias que indican una actividad inorgánica y tienen la intención de contaminar el ecosistema.
Dog-whistling	Mensaje codificado para comunicarse entre un grupo específico de personas. Utiliza lenguaje o contenido simbólico que solo algunos pueden entender. Aunque se asocia principalmente al discurso político, también es común entre grupos que buscan coordinarse para atacar a una persona, grupo u organización.
Envenenamiento de hashtags	Creación o utilización de un hashtag existente con un mensaje abusivo para iniciar un ataque coordinado en las redes sociales con el ánimo de alcanzar viralidad. Por su parte, técnicas como el #typosquatting buscan alterar la ortografía de un hashtag abusivo para que no sea reconocido por los algoritmos que intentan bloquearlo.
Deepfakes	Contenidos –video, audio o imagen– modificados con tecnología de inteligencia artificial conocida como 'aprendizaje profundo' que alteran con un alto grado de veracidad lo que las personas dicen o hacen. Los deepfakes son difíciles de identificar para un usuario común y pueden usarse con el propósito de generar dudas y socavar la confianza que se tiene sobre una persona u organización.
Publicaciones difamatorias	Publicaciones que buscan desacreditar a un grupo o persona por medio de la difusión de información falsa o manipulada que afecta su buena reputación y fama.
Amenazas	Mensajes de intimidación y hostigamiento hacia un grupo o persona que expresan la intención de hacer daño o mal y buscan generar miedo y angustia. Pueden apelar a la muerte, a la violencia sexual y a otras formas de violencia.
Hacking	Conjunto de técnicas utilizadas para violar la seguridad de un dispositivo o red con el fin de acceder a este sin autorización. En contextos de abuso, tiene la intención de atacar, incriminar o afectar a una persona u organización.
Doxing	Publicación en internet de documentos o información privada de otra persona sin autorización. El *doxing* se utiliza como forma de intimidación y puede incluir números de teléfono personales, dirección de la vivienda de una persona y de su grupo familiar, fotografías, entre otros.
Espionaje	Mecanismo de intimidación que afecta principalmente a periodistas, activistas y defensores de derechos humanos. Esta práctica busca hacerse con el control de información privada de los afectados y vigilar sus acciones a través de software y tecnología especializada.
Discurso de odio	Cualquier forma de manifestación verbal y no verbal que incita, promueve y justifica la intolerancia y la violencia contra ciertas personas por factores de su identidad, ya sea por su género, nacionalidad, identidad sexual, o por su pertenencia a un determinado grupo étnico, religioso o racial.

5. Pautas para evaluar el riesgo de un ataque

El ecosistema digital es un espacio en disputa que mezcla múltiples incentivos, actores y motivaciones. De la misma manera en que es importante tener en cuenta cuáles organizaciones, instituciones, actores y voces son aliadas, también es imprescindible entender quiénes son nuestros detractores y qué dicen. Para ello, es preciso hacer un ejercicio de análisis de proporciones: ¿qué tanta es la actividad en comparación a ataques previos que ha sufrido? ¿está creciendo el número de menciones a su cuenta o el uso de una etiqueta con su marca a un nivel en que ya es visible en la capa de contenidos públicos en algunas redes? Es importante tener documentados casos anteriores y conocer algunas cifras que ayuden a entender el tamaño de la conversación.

Cuando una conversación marginal se introduce a la conversación *mainstream*, se debe tener especial cuidado. Cualquier actor puede introducirla, pero por lo general son medios de comunicación, influenciadores o las mismas organizaciones quienes lo hacen. Así mismo, cuando los miembros del equipo o servicios específicos son blanco de ataque, es necesario desplegar protocolos o estrategias de protección como denunciar las cuentas de los atacantes o volver las cuentas privadas. Tener varios voceros en cuentas y canales diferentes permite distribuir el riesgo y lograr que nuestras comunidades digitales sean menos porosas.

De otro lado, que algunas conversaciones, temas o etiquetas negativas tengan un volumen alto de interacciones, no implica necesariamente algo preocupante, especialmente si esos ataques se limitan a burbujas de conversación específicas y no están llegando a audiencias estratégicas. En esos casos, el riesgo disminuye sustancialmente.

Para evaluar el riesgo de un ataque, es recomendable responder a las siguientes preguntas:

- ¿El ataque representa una amenaza física o digital que genera miedo, inseguridad o una sensación de peligro?
- ¿El ataque utiliza información sensible, confidencial o privada que pone en riesgo su labor o su integridad física o moral?
- ¿El ataque tiene potencial de ser intensificado si no se actúa de manera oportuna?
- ¿El ataque está siendo replicado por influenciadores, figuras públicas o personas con poder de influencia?
- ¿En el ataque se evidencia el uso de recursos financieros, técnicos o de otro tipo?
- ¿El ataque está teniendo un impacto negativo en su labor y reputación?
- ¿El ataque evidencia una intención explícita de hacer daño, insultar o provocar?
- ¿El ataque se está discutiendo en medios mainstream y está alcanzando una escala de menciones significativa?

Estas preguntas aportan información valiosa para la toma de decisiones estratégicas. Sin embargo, otros factores deben ser tenidos en cuenta en la evaluación del riesgo: de un lado, las dinámicas del abuso en línea y, del otro, las características del atacante.

5.1 Dinámicas del abuso en línea

Aunque un bombardeo de críticas negativas puede encender las alarmas y llevar a una crisis reputacional, la crítica negativa en sí misma no es una forma de abuso pues puede hacer parte de una expresión legítima de descontento o disenso. El abuso, por su parte, responde a unas dinámicas específicas que permiten determinar el nivel de severidad y toxicidad de un mensaje. El siguiente cuadro muestra las dinámicas que deben ser consideradas en el análisis del abuso en línea y algunas preguntas guía para evaluarlas.

	Descripción	Preguntas
Frecuencia	Presencia del ataque en el pasado	¿Qué tan recurrente es el ataque?
Atributos[4]	Características del mensaje que ayudan a identificar su toxicidad	¿El mensaje es innecesariamente hostil?
		¿El mensaje tiene la intención explícita de hacer daño, insultar o provocar?
		¿El mensaje hace una generalización injusta acerca de un grupo específico?
		¿El mensaje es condescendiente?
		¿El mensaje es sarcástico?
		¿El mensaje es despectivo?
Sofisticación	Evidencia de coordinación, vigilancia o comportamiento automatizado e inauténtico	¿El contenido del ataque incluye información que pudo ser obtenida por medio de vigilancia o escucha?
		¿El ataque requirió de habilidades técnicas específicas o del uso de datos personales?
		¿El mensaje se repite muchas veces desde una misma cuenta?
Alcance	Exposición de un contenido a una audiencia en una o múltiples plataformas	¿Cuál es el nivel de exposición del contenido abusivo?
		¿El contenido abusivo aparece en distintas plataformas?
		¿Hay una conversación alrededor del ataque? ¿cuál es su tamaño?
Amplificación	Proceso, planeado o espontáneo, mediante el cual un contenido aumenta su alcance y audiencia	¿El contenido abusivo alcanzó medios *mainstream*?
		¿Influenciadores o figuras públicas han replicado el contenido abusivo?
		¿Cuántas veces ha sido mencionado el ataque?
Canal	Medio a través del cuál se manifiesta el ataque	¿El ataque se hizo por medio de un canal privado o público?
		¿Cómo influye la elección del canal en el ataque?
Contexto	Circunstancias que rodean el ataque	¿El ataque responde a una situación de coyuntura?
		¿El mensaje utiliza términos de doble sentido?

5.2 Características del agresor

Es común que muchas de las formas de abuso en línea operen bajo el anonimato, lo que dificulta la identificación de los agresores. Sin embargo, preguntarse por la autoridad, influencia, identidad y motivación del agresor, así como por el tono del mensaje, proporciona insumos valiosos para determinar formas de actuar frente al abuso.

4 Estos atributos fueron definidos por el equipo de Rhodes Artificial Intelligence Lab.

	Descripción	Preguntas
Autoridad	Nivel de legitimidad y credibilidad del atacante y de su red de contactos en el entorno digital	¿El ataque se hace desde cuentas reales y con credibilidad?
		¿Es una figura pública o está respaldado por una?
Influencia	Capacidad del atacante de influenciar la manera de pensar y actuar de un grupo o persona	¿Cuenta con credibilidad en una base de seguidores?
		¿Tiene capacidad para reproducir el ataque en múltiples plataformas?
		¿Sus mensajes son replicados por cuentas de terceros?
Identidad	Tipo de cuenta o perfil desde donde se produce el ataque	¿Puede reconocer la identidad del atacante?
		¿El atacante utiliza perfiles falsos o esconde su identidad por medio de seudónimos?
		¿La cuenta desde donde se ejecuta el ataque utiliza nombres alfanuméricos u oculta información relevante?
Motivación	Razones que inducen al agresor al ataque	Diferencias políticas
		Rencillas personales
		Ideologías conspirativas
		Pensamientos racistas, xenófobos, misóginos o discriminatorios en razón de un rasgo de identidad.
Tono	Forma en que se manifiesta el ataque	¿Los mensajes contienen datos personales o nombres propios?
		¿Los mensajes utilizan signos de exclamación, mayúsculas o elementos que enfatizan el abuso?
		¿El contenido abusivo tiene referencias explícitas a la violencia?

5.3 Acciones coordinadas, comportamiento inauténtico y automatización

La coordinación es una estrategia que todo tipo de actores utiliza en redes sociales, desde las organizaciones de sociedad civil que trabajan en derechos humanos, hasta operaciones políticas u oficiales que buscan sembrar discordia, división o silenciar voces. Algunas de estas operaciones coordinadas, que violan las normas comunitarias de las plataformas, pueden resultar en suspensiones masivas de cuentas o eliminación de contenidos, lo cual evita la amplificación de ese material[5]. Dada la proliferación de casos en los que se detectaron intentos coordinados con malas intenciones, Facebook acuñó el término "contenido inauténtico coordinado".

El comportamiento inauténtico tiene que ver con aspectos como la veracidad de los contenidos, la autenticidad de las cuentas y la estrategia para posicionar una conversación aprovechando los ofrecimientos del algoritmo y del ecosistema de la red. En particular, estos son algunos de los actores que suelen participar en este tipo de operaciones:

Actores de acciones coordinadas	Descripción
Bot	Cuentas o servicios completamente automatizados. No todos los bots se programan con malas intenciones; las alertas meteorológicas, alertas de tráfico, alertas de accidente y atención de servicio al cliente son, en muchos casos, automatizados.
Semi-bot	Cuentas semi-automatizadas, parcialmente controladas o supervisadas por humanos.
Troll	Actores que buscan sembrar discordia, división y ridiculización en el ecosistema digital. Pueden ser semi-automatizados.
Cuentas falsas	Usuarios que proveen información falsa al crear sus cuentas, generalmente con fines de influencia. Cabe resaltar, sin embargo, que algunas plataformas permiten cuentas parodia o el uso de seudónimos.

5 Este documento no se ocupa de abordar las tensiones entre la moderación de contenidos y la acción de los usuarios, que hace parte de un debate más amplio sobre la responsabilidad de los intermediarios de internet. Sobre este tema, visite los proyectos de Linterna Verde circuito.digital y letrachica.digital.

Algunas estrategias utilizadas por estas cuentas son cambiar los *handlers* (el nombre de una cuenta seguido del signo @), esconderse en el anonimato y proveer datos falsos en la cuenta o tener herramientas de coordinación ocultas para mover tendencias, como un grupo de WhatsApp. Es importante tener estos fenómenos en el radar, pues es por medio de operaciones de influencia que se pueden sabotear conversaciones o etiquetas e instalar nuevos encuadres que van en contravía de la agenda o campaña que las organizaciones quieran promover.

Aquí hay algunos indicadores que pueden ser utilizados para identificar este tipo de estrategías de manipulación en línea:

Matriz de indicadores[6]		
Indicadores de acciones coordinadas Interacciones que podrían evidenciar formas de trabajo en equipo de varias cuentas.	1.1	Existencia de un grupo de cuentas que publican una secuencia similar de etiquetas a través de múltiples tuits.
	1.2	Patrones de picos de actividad u otros patrones temporales de actividad.
	1.3	Acciones de co-retuiteo: donde las mismas fuentes son citadas y amplificadas por varias cuentas en períodos cortos de tiempo.
	1.4	Repetición de las mismas imágenes o memes en múltiples cuentas.
Indicadores de actividad inauténtica Comportamientos poco transparentes o posiblemente engañosos de cuentas que buscan ocultar su naturaleza para confundir a otros usuarios.	2.1	Actividad semi-automatizada o presencia de bots.
	2.2	La repetición de textos, aunque estos sean parafraseados.
	2.3	Publicaciones que no incluyan contenido orgánico y solo contengan ruido digital diseñadas para inflar una tendencia a través de etiquetas, menciones o enlaces.
	2.4	Usuarios con cambios de nombres de usuario o handlers.
	2.5	Un número importante de cuentas suspendidas o restringidas.

6 Estos indicadores hacen parte de una metodología estándar desarrollada por Linterna Verde para este tipo de análisis de coordinación y comportamiento inauténtico basada en el trabajo de Gleicher (2018) y Pacheco et al. (2020).

6. ¿Cómo responder a un ataque?

Si bien es común escuchar la frase “no alimentes el troll”, responder al abuso es una opción válida y en algunos casos poderosa para hacer frente a los ataques. Determinar cómo actuar frente al abuso en línea puede resultar difícil, especialmente si el ataque exige una reacción temprana debido a su severidad y a su potencial para causar daño. No hay reglas absolutas: una respuesta apresurada puede ayudar a amplificar el ataque o a desactivarlo, así como la inacción puede hacer que el ataque pase desapercibido o que escale.

Existen distintos tipos de respuesta para el abuso, algunos ofrecen un marco de acción inmediata mientras que otros pueden activar procesos de largo alcance como lo es el trabajo narrativo. La elección dependerá en gran medida de la evaluación del riesgo. Factores como las motivaciones del atacante, el alcance, la sofisticación, la frecuencia y las posibles consecuencias para la integridad física o moral del sujeto u organización objeto del ataque, influyen en la elección de la respuesta.

6.1 Respuestas de acción inmediata

No hay una sola forma de aproximarse al abuso en línea y un solo ataque puede requerir varias acciones, pero es importante considerar distintos caminos en lugar de dejarse llevar por la rabia, el estrés o el miedo. Lo primero que se debe considerar es el resultado que se espera obtener con la respuesta; este deberá ser razonable y factible, detener el abuso, conseguir que se elimine un contenido abusivo, lograr evidenciar la identidad del acosador, lograr que se suspenda una cuenta o evitar que escale. Con esto en mente, son varias las cosas que se pueden hacer:

Documentar

Documentar el ataque es importante para denunciar el abuso ante las plataformas, la policía, los entes legales o cualquier otra instancia. También será útil para identificar algunos rasgos de la identidad del atacante y su modus operandi en casos en los que el ataque escale. Hacer capturas de pantalla de mensajes en redes sociales y guardar copia de mensajes de texto, correos y comentarios, así como llevar un registro de las fechas y plataformas utilizadas para el ataque, es la mejor manera de documentar el abuso y de prepararse para tomar medidas frente a él. En esta guía de PEN America, encontrará información detallada sobre las formas de documentación.

Reportar

Las redes sociales tienen reglas comunitarias para arbitrar el comportamiento inadecuado en sus plataformas. Los ataques que configuran abuso en línea generalmente violan estos términos de servicio. Es importante saber que reportar no garantiza que las plataformas tomarán medidas para eliminar el contenido o sancionar a los responsables, pero sin duda aumenta las posibilidades de que se activen los mecanismos de moderación de contenidos y permite dejar constancia del abuso.

Familiarizarse con las normas comunitarias es un primer paso para saber cómo actuar frente a los ataques en cada plataforma. En Letra Chica y Circuito podrá encontrar los cambios más recientes en las políticas de moderación de contenidos de Facebook, Twitter y YouTube.

¿Para saber cómo reportar el abuso en las distintas plataformas consulte la guía de HeartMob sobre seguridad en las redes sociales.

Considerar la respuesta

Aunque el abuso en línea puede escalar rápidamente y desatar una tormenta, es recomendable tomarse un momento para analizar la severidad, frecuencia, alcance y circunstancias del comportamiento abusivo con el fin de determinar la respuesta. Varios caminos son posibles:

a. **Ignorar**

Es una opción cuando el riesgo es bajo y los mensajes abusivos o violentos tienen un alcance reducido. Hay olas de menciones y notificaciones que al estar asociadas a una burbuja específica de conversación o a una cuenta sin mayor tracción, no ameritan respuesta. Es más, responder puede llegar a amplificar mensajes que hasta entonces tenían un alcance limitado. Esto se puede identificar gracias al monitoreo regular de ciertas conversaciones.

b. **Silenciar**

Algunas plataformas permiten silenciar ciertos contenidos para que dejen de aparecer en nuestro *feed*. Esto significa restringir el contenido que vemos en las notificaciones o muro de noticias de manera predeterminada. Sin embargo, esta medida solo oculta los contenidos en el *feed* propio, para el resto de la red seguirán siendo visibles. Pese a que es una medida útil en algunos contextos, silenciar también conlleva el riesgo que los ataques continúen esparciéndose sin nuestro conocimiento, que no se detecten amenazas que es importante que conozcamos y sobre las que se deben tomar otro tipo de acciones, entre otros.

c. **Bloquear**

Esta acción evita que algunos usuarios interactúen con nosotros. Al bloquear a una persona en la red limitamos su capacidad para contactarnos y enviarnos mensajes. El bloqueo cumple dos funciones: de un lado, restringe el acceso que tiene una cuenta a nuestro contenido y perfil; del otro, hace que el contenido de la cuenta bloqueada no sea visible para nosotros[7]. En ocasiones el bloqueo puede llevar a una intensificación del abuso y a la aparición de nuevas cuentas desde donde se siguen efectuando ataques ya que los usuarios pueden ver que han sido bloqueados[8].

d. **Exponer**

El abuso en línea encuentra resguardo en el anonimato, por eso, exponer los mensajes de abuso o la identidad de quienes están detrás de estos —cuando se tiene acceso a ella—, es una manera de cambiar el foco de atención y ponerlo en los acosadores. Cuando se expone un contenido abusivo debe evitarse compartir el contenido original, pues esto contribuye a su amplificación. En este caso es recomendable hacer una captura de pantalla y publicarla junto a un mensaje proactivo y preciso que utilice un lenguaje distinto al del acosador.

Por ejemplo, para dar cuenta de que la libertad de expresión puede tener consecuencias, el Research Group en Suecia expuso en redes públicamente la identidad de una serie de trolls —entre los que había políticos de derecha— y esto llevó a varias personas a retractarse. Aunque hay opiniones encontradas sobre esta aproximación, exponer es una estrategia válida para hacer responsables a los trolls por sus publicaciones.

Responder sin amplificar

Antes de responder es crucial hacer un análisis de riesgo para determinar la pertinencia de una confrontación, si se trata de una amenaza a su integridad física evite responder y tome medidas para reportar a las autoridades,

7 Meedan. Content Moderation Toolkit.
Disponible en: https://meedan.com/reports/content-moderation-toolkit/

8 Pen America. Manual contra el abuso en línea.
Disponible en: https://onlineharassmentfieldmanual.pen.org/es/bloquear-silenciar-restringir/

si se trata de otra forma de abuso considere factores adicionales como su estabilidad emocional para iniciar una confrontación y el resultado que se espera obtener con la respuesta.

Las respuestas pueden ser variadas, desde enviar un mensaje directo, convocar a aliados y personas del círculo de confianza para responder en conjunto o hacer una declaración general sobre el tipo de abuso que está sufriendo, hasta publicar capturas de pantalla y evidencias de los contenidos abusivos que le han llegado. En todos los casos el lenguaje empleado deberá desligarse del utilizado por el agresor.

Basados en el trabajo de Pen America, compartimos algunas recomendaciones para responder al abuso en línea.

- **Centrar la respuesta en el contenido de la agresión.** Condenar el contenido del abuso en lugar de centrarse en el atacante puede dar lugar a una interacción más productiva.

- **Cambiar el enmarcado del mensaje.** Esto dirige la atención de la audiencia a otros temas dejando sin tracción el ataque.

- **Repetir el mensaje de respuesta.** La repetición se traduce en familiaridad, y la familiaridad, a su vez, en confianza.

- **No compartir sus publicaciones.** Compartir las publicaciones originales para denunciarlas solo contribuye a amplificar el mensaje agresor. En ese caso, haga una captura de pantalla y cambie el enmarcado del mensaje.

- **Hacer explícitas las consecuencias del abuso.** Mencionar los perjuicios que el abuso genera para usted o su comunidad le da una dimensión más clara al abuso y permite visualizar los daños e impactos concretos de este tipo de comportamientos.

- **Evitar el lenguaje hostil.** El abuso busca provocar reacciones emocionales fuertes y a menudo irreflexivas. Esto sin duda no ayuda a desescalar el abuso, por el contrario, puede contribuir a su intensificación.

6.2 Qué hacer a mediano y a largo plazo

¿Qué pasa si después de documentar los ataques vemos que, más allá de su frecuencia y severidad, estos implican la promoción de una narrativa adversarial que pone en riesgo la credibilidad y confianza en el trabajo de la organización? En este punto hay que empezar a pensar en el diseño de contranarrativas o narrativas alternas.

Las contranarrativas son estrategias discursivas cuyo propósito es contener la dispersión de las expresiones de odio y narrativas adversariales ampliamente difundidas. Su principal función es debilitar la retórica divisiva que induce al miedo y justifica la violencia, posicionando un mensaje incluyente que utilice un lenguaje distinto al usado por quienes difunden la narrativa adversarial. De este modo, el trabajo contranarrativo busca aumentar la capacidad de respuesta de las organizaciones y activistas a mediano plazo frente a narrativas que promueven la intolerancia y el discurso de odio.

Desarrollar un trabajo sobre narrativas implica tener claro ese punto de partida. Es necesario establecer si los contenidos vinculados a los ataques tienen un volumen y una continuidad suficiente para identificar patrones y tendencias. Por ejemplo, en una muestra de cuatro post de Facebook y cinco tuits publicados por una cuenta opositora durante un periodo de un año, no hay suficiente material para identificar una narrativa adversarial y los caminos para enfrentarla. Por eso es necesario identificar los valores, polaridades e indignaciones que intentan ser alimentados con estas agresiones y cuáles son los discursos de odio subyacentes. Igualmente, es importante relacionarlo con el contexto *offline* para la organización y el tema que trabaja.

Esta guía se centra en la construcción de contranarrativas desde la comunicación estratégica como medio de contención del extremismo en línea.

7. Pasos para la construcción de la contranarrativa

Las contranarrativas son principalmente estrategias de reacción y contención. Su objetivo es neutralizar el impacto de una narrativa adversarial al oponerse, denunciar y detener la propagación de las expresiones extremistas que la componen.

De acuerdo con Rachel Brown[9], estas intervenciones para contrarrestar el discurso de odio deben tener como objetivos:

- Reducir la probabilidad de que las audiencias acepten y difundan discursos peligrosos.
- Reducir la probabilidad de que las audiencias aprueben o participen en daños dirigidos a otros grupos.
- Aumentar la disposición de los miembros de la audiencia a hablar en contra de los esfuerzos para fomentar el odio.

Para lograrlo, es importante (i) comprender el contexto en el que se enmarca la narrativa adversarial o violenta, (ii) diseñar la contranarrativa y, (iii) establecer un marco para la evaluación de resultados.

i. Deconstruir la narrativa violenta

Antes de desarrollar una estrategia contranarrativa, es necesario hacer un análisis en profundidad del contexto, esto significa desagregar los componentes de la narrativa que se busca confrontar. Identificar los elementos constitutivos de la narrativa adversarial es el punto de partida del trabajo narrativo y permite comprender los factores que posibilitan la dispersión de expresiones de odio y contenido violento; conocer los actores, motivaciones e incentivos detrás del mismo; recopilar información útil acerca de la audiencia, y adoptar estrategias más efectivas para contrarrestar su impacto.

Dangerous Speech Project[10] elaboró un marco para determinar cómo se configuran las expresiones de odio en un determinado contexto y qué tan dañinas pueden llegar a ser, a partir de cinco elementos: el mensaje, el contexto social e histórico, el medio utilizado para difundir el mensaje, la audiencia y el mensajero. Estos cinco elementos deberán ser analizados uno por uno con el fin de establecer en qué medida contribuyen a la narrativa violenta y determinar formas de contención para reducir el impacto del mismo.

El mensaje

En el discurso de odio es común encontrar marcas distintivas como la deshumanización, la representación del otro como una amenaza, el reforzamiento del miedo, entre otros. Deconstruir el mensaje pasa por comprender el lengua-

9 Brown, R. (2016). Defusing Hate: A Strategic Communication Guide to Counteract Dangerous Speech. Disponible en: https://www.ushmm.org/genocide-prevention/reports-and-resources/defusing-hate-a-guide-to-counteract-dangerous-speech

10 Dangerous Speech Project (2021). Dangerous Speech. A practical guide. Disponible en: https://dangerousspeech.org/guide/

je, los encuadres, los valores subyacentes, las historias que lo componen y el tono.

- **Lenguaje:** el lenguaje es el bloque básico de la narrativa. Muchos de los promotores de odio utilizan un sistema de códigos para comunicarse, creando así una jerga compartida y al mismo tiempo excluyente.[11]

  En el contexto de una narrativa adversarial, es común que se utilice el lenguaje para deshumanizar al grupo objetivo, por ejemplo, al compararlo con cosas, animales, insectos o microorganismos, generalmente vistos como prescindibles, repulsivos y merecedores de la violencia, lo que en últimas contribuye a la aceptación o justificación de las expresiones violentas contra este. Por ejemplo, el uso que se hace en Colombia de la palabra *"desechable"* para hacer referencia a los habitantes de calle, deja a esta población en condición de *"algo no-humano"* que puede ser tirado, rechazado o eliminado por resultar *"inútil, incómodo o molesto"*[12].

Comprender el lenguaje que utiliza la narrativa adversarial permite demarcar los límites de nuestro propio trabajo narrativo.

Preguntas clave:

¿Cuál es la narrativa adversarial?	
¿Qué términos, estereotipos o lenguaje deshumanizante está siendo utilizado?	
¿Qué palabras, conceptos o imágenes se reiteran en los mensajes incendiarios?	
¿De qué manera se justifica la violencia contra el grupo objetivo?	

11 Ibid.

12 Real Academia Española. (2020). Diccionario de la lengua española (23.4.a ed.). Consultado en http://www.rae.es

- **Encuadre:** los encuadres hacen referencia a la forma en que se organiza y presenta una información, qué se incluye y qué se excluye del relato, qué metáforas se utilizan y cuál es su carga valorativa. Los encuadres pueden contener una descripción del problema, una interpretación causal, una evaluación moral y una recomendación o tratamiento del asunto[13]. Responder estas preguntas permite entender el encuadre que alberga la narrativa adversarial y, posteriormente, en el diseño de la contranarrativa pensar en un encuadre distinto favorable para nuestro objetivo.

Preguntas clave:

¿En la narrativa se responsabiliza al grupo objetivo de algún problema?	
¿La narrativa enmarca al grupo objetivo como una amenaza? ¿Qué argumentos ofrece?	
¿En la narrativa se identifica una evaluación moral sobre quién es bueno y quién es malo?	
¿La narrativa adversarial hace parte de un enmarcado más amplio? ¿Qué dice ese enmarcado?	
¿Cómo se conecta ese encuadre con otros temas relevantes y sensibles?	
¿Organizaciones, gobierno o grupos de personas están confrontando estas narrativas? ¿Cómo lo hacen? ¿Qué argumentos utilizan para refutarlas o desacreditarlas?	

13 Entman, R. (1993). Framing: Towards Clarification of a Fractured Paradigm. Journal of Communication, 43 (4), 51-58.

- **Historias:** Las historias son lo que se dice sobre un tema. *"Tienen comienzo, medio y final, tienen héroes y villanos, lecciones y moralejas implícitas".*[14] Identificar el contenido de los mensajes que difunden los promotores de odio es identificar las historias que se repiten en él y la manera en que se articulan para despertar emociones generalmente asociadas al miedo y a la rabia.

Preguntas clave:

¿Qué historias y metáforas utiliza la narrativa violenta o incendiaria?	
¿Con qué valores se asocian las historias?	
¿Cómo se presenta el grupo objetivo en el relato? ¿Quiénes son los héroes y villanos de la historia?	

- **Tono:** el tono es el estilo empleado para comunicar el mensaje en un contexto específico. Este determina la intención del hablante e invita a hacer una lectura específica de lo que se dice. El tono es útil para indagar en el objetivo del mensaje.

Preguntas clave:

¿Cuál es el tono predominante en los mensajes?	
¿Qué tipo de conducta promueven los mensajes?	
¿Qué emociones buscan despertar en la audiencia?	

14 ORS Impact (2021). Op. Cit.

El contexto

Ningún discurso adversarial o ataque puede entenderse por fuera del contexto social, histórico y cultural en el que se desarrolla. En efecto, la severidad de un discurso dependerá no solo de su contenido, sino también de la forma, el tiempo y el espacio en el que se expresa. En el análisis de las expresiones de odio es importante recopilar la mayor cantidad de elementos de contexto disponibles que sirvan para diseñar la narrativa propia.

Preguntas clave:

¿Qué eventos del pasado son relevantes para comprender el discurso adversarial actual? ¿Existe un historial de violencia entre los promotores de la narrativa violenta y el grupo objetivo?

¿Qué situaciones o actores del contexto actual generan preocupación y por qué? ¿Existen condiciones y estructuras dadas para que la violencia contra el grupo objetivo emerja (por ejemplo, apoyo de una élite poderosa, presencia de milicias y grupos armados ilegales, entre otros)?

¿Qué esfuerzos se están haciendo para contrarrestar la narrativa, quiénes los están liderando, qué estrategias utilizan y qué tan efectivas han demostrado ser?

La audiencia

Para que una narrativa adversarial prospere, es necesario que haya una audiencia receptiva. La exposición reiterada a mensajes que difunden miedo, el exceso de información, las crisis sociales y económicas, el recuerdo de hechos sociales traumáticos, entre otros, contribuyen a la generación de unas condiciones favorables para la manipulación y el discurso extremista. *"Ni el mensaje más incendiario puede inspirar violencia si la audiencia no es susceptible a tales nociones"*[15].

Esto es aún más notorio en el entorno digital, gracias a la existencia de un ecosistema de información mucho más difuso que en el pasado, en el que una multitud de actores con intereses propios se disputan la atención limitada de las personas, por lo que la efectividad de este tipo de narrativas reside en la explotación de agravios y emociones presentes en la audiencia y las percepciones de esta frente a otros grupos e instituciones.

Preguntas clave:

¿A qué público le hablan los promotores de la narrativa adversarial o violenta? ¿Cómo está reaccionando este público? ¿Lo repite? ¿De qué manera lo hace?

¿Cuál es el grupo objetivo de esta narrativa? ¿Por qué son atacados?

¿Qué creencias y valores son utilizados para justificar estas narrativas en contra del grupo objetivo?

15 Dangerous Speech Project (2021). Op. Cit.

El espacio

El ecosistema de información cuenta con multiplicidad de medios (redes sociales, foros, medios de comunicación online, correo, plataformas de streaming, etc.) que hacen que la tarea de identificar y desagregar las narrativas que circulan sea difícil. El medio o espacio a través del cual se difunde una narrativa tiene efectos tanto en el alcance como en la comprensión del mensaje, por lo que la identificación de los medios que están siendo utilizados para difundir las expresiones extremas o violentas, es un criterio relevante en la evaluación de su peligrosidad.

Este elemento será útil también al momento de definir el objetivo de nuestra intervención ya que la estrategia será distinta si se trata de una narrativa ampliamente difundida en medios tradicionales, —en cuyo caso la contranarrativa deberá confrontarla— o si se trata de una narrativa adversarial que permanece en discusiones de nicho, —en cuyo caso la estrategia será evitar que conquiste nuevos medios y llegue a audiencias más amplias—.

Preguntas clave:

¿Cuál ha sido el alcance geográfico del discurso violento?

¿Qué tipo de medios han sido utilizados para difundir la narrativa violenta y cuál es su nivel de influencia?

En caso de que la haya tenido, ¿cuál ha sido la cobertura de los medios mainstream alrededor de esta narrativa?

¿Con qué frecuencia se ha transmitido el mensaje?

El vocero

Los voceros o promotores del discurso de odio son personas con capacidad de influencia sobre un determinado grupo o comunidad. Esta influencia puede derivar del carisma, el estatus social, la ocupación, la autoridad u otros factores —desde personas respetadas en su círculo familiar hasta líderes políticos, religiosos o comunitarios—. Algunas organizaciones y gobiernos también contribuyen ampliamente a la diseminación de discursos peligrosos.

En muchos casos, resulta difícil identificar la fuente, o fuentes, que promueven las expresiones de odio, especialmente en el ecosistema digital, en donde la identidad de los promotores del discurso puede resultar difusa. Para ello es recomendable hacer un monitoreo sostenido del ecosistema que se busca impactar, a fin de obtener información sobre los atacantes.

Por último, es importante aclarar que los voceros no son solo quienes crean mensajes incendiarios, también quienes distorsionan, tergiversan y difunden mensajes de otros ampliando su alcance a nuevas audiencias.

Preguntas clave:

¿Quién o quiénes son los promotores del discurso incendiario?

¿Cuál es el nivel de influencia de los promotores? ¿Qué tan exitosos son en la diseminación del discurso de odio?

ii. Diseñar la contranarrativa

En el diseño de contranarrativas es fundamental evitar reproducir relaciones de poder asimétricas[16], esto significa generar una actitud reflexiva acerca del lenguaje que utilizamos para referirnos a un determinado tema o grupo de personas, ya que en muchas ocasiones el uso de expresiones incorporadas de manera inconsciente, incluso en los trabajadores del sector social, puede llevar a una acción con daño. Para evitarlo, es recomendable involucrar a personas afectadas directamente por la narrativa adversarial en el diseño de la contranarrativa o adquirir un conocimiento amplio sobre estas personas y el grupo social al que pertenecen.

Por su parte, cuando hablamos de diseñar la contranarrativa hablamos de la posibilidad de crear una estrategia de reacción y contención a una narrativa adversa, pero también de la posibilidad de reencuadrar algunos aspectos de la narrativa dominante para avanzar hacia una narrativa más positiva, ofreciendo una lógica distinta para la interpretación de mismo un asunto o problema, o incluso de buscar otro enmarcado que cambie el foco de atención de la conversación. A continuación se detallan

16 Movimiento frente al discurso de odio (2018). ¡Sí podemos! Actuar contra el discurso de odio a través de las contranarrativas y narrativas alternas. Disponible en: https://rm.coe.int/spanish-we-can-si-podemos-2019-ax/16809819fe

algunos factores a tener en cuenta en el diseño de contranarrativas:

- **Usar un lenguaje nuevo.** En su libro *No pienses como un elefante*, George Lakoff explica la importancia de no usar ninguno de los términos, imágenes o etiquetas del encuadre original: *"Debido a que el lenguaje activa marcos, se requiere un nuevo lenguaje para activar nuevos marcos. Pensar de manera diferente requiere hablar de manera diferente"*[17]. Es decir que debemos evitar responder al ataque negándolo, pues estaríamos usando el mismo marco del agresor.

- **Evitar chivos expiatorios.** En las contranarrativas es importante excluir cualquier forma de odio y discriminación. Expresiones como *"no odies a los pobres, odia al sistema"* son igualmente dañinas puesto que utilizan la culpa para trasladar y justificar el odio contra otras personas o grupos.

- **Ir más allá de los datos.** La evidencia fáctica, aunque necesaria, es poco efectiva para hacer frente al discurso violento. Los ataques generalmente no buscan respuestas argumentadas, lógicas o con apelaciones a la razón. El registro de enunciación suele ser moral o emocional, por lo que la contranarrativa deberá apelar a valores y emociones de la audiencia.

- **Desarrollar mensajes inclusivos.** Las narrativas violentas y las expresiones de odio se alimentan de la división entre "ellos" y "nosotros". Esta división se utiliza para inducir al miedo y representar a un grupo externo como una amenaza. Las contranarrativas deberán evitar esta división y por el contrario, contener mensajes inclusivos basados en los derechos humanos.

- **Trabajar en red y mantener un enfoque local.** Trabajar con socios locales permite tener una mejor lectura del contexto en el que se inscriben las violencias y diseñar respuestas efectivas. Es clave buscar aliados que apoyen el trabajo contranarrativo.

Pasos para el diseño de la contranarrativa. **Más allá de una campaña**

1 Definir el objetivo

El objetivo funciona a la vez como punto de partida y de llegada: es un elemento transversal a todo el trabajo narrativo, guía la toma de decisiones y permite evaluar el impacto de la estrategia. Aunque reducir el impacto de la narrativa adversarial es un objetivo en sí mismo, algunas consideraciones acerca del porqué buscamos hacerlo, cómo vamos a lograrlo, qué necesitamos para ello —conocimiento, alianzas, recursos, herramientas, etc.— y cómo sabremos si estamos siendo efectivos en nuestro propósito, son importantes.

Al momento de definir el objetivo, es necesario promover el pensamiento crítico en el equipo e indagar profundamente en las motivaciones de la organización. Omitir este paso —o subestimarlo— es quizás el error más frecuente en el diseño de intervenciones y estrategias y el más costoso en tiempo y dinero.

Por último, la definición del objetivo también pasa por tener un conocimiento profundo de nuestras capacidades y limitaciones como organización, con el fin de hacer una buena lectura del contexto en el que buscamos intervenir y diseñar una estrategia que encuentre un balance entre lo que es factible, lo que es viable y lo que es deseable.

17 Lakoff, George (2014). The all new. Don't think of an elephant!: know your values and frame the debate : the essential guide for progressives. White River Junction, Vt. :Chelsea Green Pub. Co.

2 Segmentar la audiencia

Una audiencia es, sencillamente, un conjunto de usuarios a los que hablamos o nos dirigimos. Suelen compartir algunos rasgos que hacen más fácil su identificación y posterior segmentación. La definición de la audiencia es clave para determinar los mensajes, los canales y los mensajeros.

En este punto, el ejercicio de comprensión de las dinámicas de la narrativa violenta ya ha proporcionado información valiosa acerca de las audiencias a las que estarían dirigidos los esfuerzos contranarrativos. Sin embargo, la audiencia no es homogénea, por lo que es fundamental hacer un análisis detallado de los distintos grupos que la conforman.

Entender cómo se relaciona cada grupo con la narrativa adversa nos ayudará a definir objetivos específicos para cada audiencia. Por su parte, segmentar la audiencia en grupos facilita la creación de perfiles representativos de cada uno. Esto es, la creación de personajes ficticios que reúnen las características, necesidades y comportamientos de un público más amplio.

Rachel Brown[18] propone una segmentación de audiencias basada en los roles que estas desempeñan –o pueden llegar a desempeñar– tanto en la dispersión del discurso violento como en su contención. Las categorías son:

Individuos influyentes. Tienen un alto nivel de influencia en su comunidad. Pueden estar relacionados de distintas maneras con el discurso extremista o no tener relación alguna.

Miembros reticentes de la audiencia. Presencian o escuchan expresiones de odio y discursos violentos de manera renuente.

Miembros comprometidos de la audiencia. Son receptivos al discurso violento pero no se ubican en el extremo.

Personas que probablemente participen en daños al grupo objetivo (de manera renuente). Participan —o podrían participar en el futuro— en el daño a otro grupo, pero tienen reservas. No están completamente convencidos de hacerlo y pueden estar manipulados por ciertos individuos influyentes.

Promotores del discurso de odio. Diseminan discurso violento y alientan a la acción.

Distribuidores de información. Tienen un papel importante en la diseminación de información. Son fuente principal de información en sus comunidades. Al igual que los individuos influyentes, pueden estar relacionados de distintas maneras con el discurso extremista o no tener relación alguna.

Personas que probablemente participen en daños al grupo objetivo (de manera voluntaria). Participan o están dispuestos a participar voluntariamente y sin reservas en daños a otro grupo. Este grupo puede ser difícil de alcanzar y el objetivo frente a él dependerá de una evaluación del contexto.

Personas de la audiencia que difunden el discurso violento. Repiten las expresiones de odio difundidas por otros y contribuyen a ampliar su alcance.

Personas de la audiencia que contrarrestan el discurso violento. Difunden mensajes que de manera directa o indirecta contrarrestan estas narrativas.

Pese a que estas categorías son valiosas para establecer las relaciones existentes y probables de las audiencias con las narrativas adversariales, no son exclusivas y pueden variar según el contexto en el cual se esté trabajando y el objetivo que haya sido definido por la organización.

18 Brown, R. (2016). Op. Cit.

3 Seleccionar los medios

Los medios determinan la manera en que vamos a hacer llegar nuestro mensaje a las audiencias para influir en ellas. Hay dos elementos esenciales a tener en cuenta: no todas nuestras audiencias están en el mismo lugar y no todos los medios cumplen la misma función. El análisis y la segmentación de las audiencias proporcionará información necesaria para seleccionar los medios en relación con su alcance, la relevancia para la audiencia y el modo y tiempo de consumo. Una estrategia de medios integrada deberá ofrecer claridad sobre a quién le hablamos, en qué circunstancias y con qué frecuencia.

4 Construir los mensajes

Los mensajes deben estar basados en una comprensión amplia de la audiencia. En el diseño de los mensajes es esencial reconocer las motivaciones, valores y las barreras que pueden llegar a limitar el accionar de nuestra audiencia en contra de la narrativa violenta. También se deben reconocer los riesgos que puede tener el contenido de la contranarrativa, es decir, el daño involuntario a las personas objeto del discurso violento debido a consecuencias no deseadas del mensaje; riesgos para la seguridad e integridad de quienes difunden el mensaje; riesgo de apatía por parte de la audiencia frente al mensaje, entre muchos otros, y diseñar planes de mitigación.

Tenga en cuenta que las personas tienden a rechazar información que se opone a su sistema de creencias[19] por lo que trabajar con base en los conceptos de encuadres, narrativas y valores nos ayudará a crear y a posicionar mensajes a partir de la activación de ciertos esquemas de interpretación en la audiencia, el uso de un lenguaje y tono diferente al utilizado por los promotores del odio y la apelación a valores compartidos.

Construir mensajes efectivos es una labor difícil, por eso es recomendable dedicarle tiempo a este paso y testear distintos tipos de contenidos entre la audiencia de manera que se pueda recopilar información útil para mejorar y corregir aquellos que no tengan el impacto esperado.

19 Brown, R. Select and Design Mediums, Speakers & Message Content. Workbook 3.
Disponible en: https://www.ushmm.org/genocide-prevention/reports-and-resources/defusing-hate-a-guide-to-counteract-dangerous-speech

5 Pensar en los encuadres

Los encuadres son estructuras mentales que le dan forma y sentido al mundo. Estas estructuras se ubican en nuestro inconsciente y son activadas a través del lenguaje. Por ejemplo, cuando se aborda el tema de drogas, un encuadre puede ser el libre desarrollo de la personalidad, mientras que otro puede estar relacionado con el derecho a la salud. Ambos se refieren a la misma problemática, pero el cambio de encuadre activa ciertos esquemas mentales y no otros. De la misma manera, el trabajo por el medio ambiente se puede encuadrar como un trabajo de regulación ambiental o de protección ambiental[20]. La variación en los términos tendrá una variación en la interpretación que se hace del asunto.

El objetivo de los encuadres es que resuenen en la audiencia objetivo; que las personas, al pensar en alguna situación, acudan al encuadre que difundimos. La resonancia de un encuadre depende de su credibilidad y protagonismo. La credibilidad se refiere a (1) la correspondencia del encuadre con la realidad o los hechos empíricos (debe ser verosímil); (2) la consistencia entre las creencias, peticiones y acciones que activa y, (3) la credibilidad de los mensajeros en la audiencia objetivo. El protagonismo, por su parte, hace referencia a que (1) los valores y creencias que activa el encuadre sean importantes para la audiencia; (2) que se relacionen con la experiencia cotidiana de las personas y; (3) que haya correspondencia o cercanía entre el encuadre que diseñamos y otros encuadres e ideologías que ya existen en la sociedad.[21]

Para diseñar encuadres efectivos debemos tener en cuenta tres prerrequisitos y analizar el encuadre en tres secciones:

Que el encuadre esté disponible: las personas tienen que haber encontrado el encuadre alguna vez para que esté "disponible" para ser activado.

Que el encuadre sea accesible: se refiere a la probabilidad de que el encuadre sea recordado y activado cuando el receptor vea el mensaje. Una manera de lograr este efecto es por medio de la repetición[22] [23].

Que el encuadre sea aplicable: el receptor debe percibir que el mensaje es relevante y aplicable en su propia vida. En otras palabras, debe lograr que el individuo sienta que puede actuar al respecto[24].

Por su parte, los encuadres pueden componerse de un diagnóstico, que presenta la situación o problema a la que se le hace frente; un pronóstico, que articula las soluciones; y una motivación, que provee caminos para llevar a cabo acciones colectivas de mejoramiento de las condiciones actuales[25].

20 Narrative Initiative (s.f.). Op. cit.
21 Benford, R., & Snow, D. (2000). Framing Processes and Social Movements: An Overview and Assessment. Annual Review of Sociology, 26, 611-639. Disponible en: http://www.jstor.org/stable/223459
22 Entman, R. M., Matthes, J., & Pellicano, L. (2008). Op. cit.
23 Lakoff, G. (2018, enero 13). A Modest Proposal #Protect The Truth. Disponible en: https://medium.com/@GeorgeLakoff/a-modest-proposal-protectthetruth-c7c32713f827
24 Entman, R. M., Matthes, J., & Pellicano, L. (2008). Op. cit.
25 Benford, R., & Snow, D. (2000). Op. cit.

Matriz para el diseño de mensajes

Componente	Descripción
Valor	*¿Cuáles son los valores que sostienen mi contranarrativa?*
Problema	*¿Cuáles son los valores que sostienen la narrativa violenta y por qué deben ser cambiados?*
Aspiración	*¿Cómo debería lucir el mensaje basado en los valores de mi narrativa? ¿Qué mensaje deseo posicionar a partir de la contranarrativa?*
Solución	*¿Qué propone nuestra contranarrativa, cuál es el llamado?*

6 Identificar a los mensajeros

Los mensajeros son personas que tienen la confianza de la audiencia a la que buscamos alcanzar y pueden ayudarnos a difundir y amplificar nuestros mensajes de una manera efectiva. Su elección no es intuitiva, es decir que no necesariamente se trata de activistas o de personas que ostentan un determinado cargo, sino de aquellas que logran conectar con la audiencia bien sea por su tipo de liderazgo, porque comparten una historia en común o por su capacidad para transmitir confianza y credibilidad entre quienes lo escuchan.

Así como los medios no son los mismos para todas las audiencias, los mensajeros tampoco. Cada segmento de audiencia tiene unos mensajeros en los que ya confían o en los que es muy probable que puedan llegar a confiar. Un mensajero influyente aumentará las probabilidades de que el mensaje sea bien recibido.

iii. Evaluar los resultados

La pregunta por los resultados no es totalizante, es más bien reflexiva: permite estructurar un marco de monitoreo y evaluación para tomar acciones correctivas en la medida en que se avanza en la estrategia, y una vez concluida la intervención, permite considerar en qué medida se logró la contribución esperada, cuáles fueron sus efectos y qué elementos podrían mejorarse en futuras intervenciones.

Tener conciencia de nuestro lugar en la intervención es fundamental para evaluar nuestros aportes. Tenga en cuenta que el discurso de odio no se combate con una campaña; es un asunto estructural que requiere del trabajo de muchas personas y organizaciones en distintos momentos y contextos. Tampoco es un asunto estático, pues los promotores de odio constantemente sofistican sus estrategias para alcanzar a más personas, lo que exige que quienes diseñen contranarrativas estén en capacidad de adaptar también sus intervenciones a nuevos escenarios.

Para evaluar los efectos e impactos de la contranarrativa es necesario establecer indicadores y variables aplicables durante todo el proceso. Utilice la siguiente tabla como guía para comparar sus resultados esperados con sus resultados reales:

Antes de la acción	Después de la acción
¿Cuáles son nuestros resultados previstos?	¿Cuáles fueron nuestros resultados previstos?
¿Cómo deberá lucir una acción exitosa?	¿Cuáles fueron nuestros resultados reales?
¿Qué desafíos podemos encontrar?	¿Qué causó nuestros resultados?
¿Qué hemos aprendido de situaciones similares?	¿Qué mantendremos o mejoraremos?
¿Qué nos hará exitosos esta vez?	¿Cuál es nuestra próxima oportunidad para probar lo que aprendimos?
¿Cuándo haremos nuestra próxima revisión después de la acción?	¿Cuándo haremos nuestra próxima revisión antes de la acción?

Fuente: Measuring Narrative Change: Tracking and Responding to Changes in Context

8. Consideraciones finales

El debate digital está en constante movimiento. Aunque quisiéramos tomar una foto, la realidad es que el entorno es dinámico: una multiplicidad de actores se disputa la atención en temas que se activan y desactivan constantemente de acuerdo a coyunturas políticas o agendas particulares. Al mismo tiempo, la moderación de contenidos y la pugna entre las plataformas y ciertos gobiernos por regular estos espacios, generan un ambiente dinámico e impredecible.

Por eso, al igual que una estación meteorológica hace mediciones todos los días sobre el clima para tener puntos de comparación y medir los eventos extremos, la volatilidad del ecosistema digital también exige un trabajo de monitoreo constante para detectar cambios en una conversación o el surgimiento de nuevos ataques. Es recomendable hacer un barrido sencillo de las cuentas, las etiquetas y los enlaces que se están compartiendo en las redes sobre algún tema en particular. Seguir palabras claves puede dar una idea de cómo se está desarrollando alguna discusión en particular y en qué punto puede volverse crítica.

Por su parte, es vital ahorrar energía en la respuesta a los ataques. Las cuentas falsas son fáciles de identificar, pero solo merecen un poco de atención cuando se trata de una operación conjunta automatizada. No vale la pena concentrar esfuerzos en una cacería de brujas que, en cualquier caso, se transforma continuamente.

Finalmente, es importante pensar los ataques desde la estrategia, para lo cual compartimos once maneras en las que se puede mitigar el impacto de los ataques y controlar los riesgos:

Evitar el pánico. El pánico puede llevar a una pérdida de proporción sobre las dimensiones del ataque y sus posibles impactos, dando lugar a respuestas apresuradas que no contribuyen a minimizar sus riesgos.

Buscar aliados. Usar el poder y la voz que se tiene en redes para buscar aliados que apoyen y justifiquen el trabajo que ha sido atacado.

Diseñar un plan de acción. En el caso de las organizaciones, es importante diseñar un plan de acción para momentos de crisis que facilite la toma de decisiones. Es clave definir con antelación qué se va a considerar como riesgo, así como construir un historial de métricas y etiquetas que permita definir con antelación a partir de qué volumen de actividad y bajo qué parámetros cierta conversación digital amerita prender las alarmas de la organización.

Monitorear. Monitorear. Monitorear. El monitoreo no debe hacerse solo cuando hay una amenaza inminente. Con el monitoreo podemos identificar trolls, medir ataques o seguir las opiniones que circulan sobre algún tema. También podemos emprender acciones frente a los ataques que pueden aparecer cuando se introduce una nueva conversación digital.

Contemplar el silencio como respuesta. Hay olas de menciones y notificaciones que al estar asociadas a una burbuja específica de conversación, no ameritan respuesta. Es más, responder puede llegar a amplificar aquellas narrativas que hasta entonces tenían un alcance limitado. Esto se puede identificar gracias al monitoreo regular de ciertas conversaciones.

Prender las alertas cuando los medios mainstream amplifican. Se debe evaluar qué tanto está involucrada la organización y determinar si es necesaria una respuesta.

Promover la coordinación interna. Es importante delegar las situaciones de riesgo a un miembro del equipo o a un comité de crisis que tenga una guía de acción para enfrentar situaciones riesgosas.

Promover la coordinación externa. Debemos identificar también oportunidades para forjar relaciones y campañas con nuestros aliados, estar atentos a las conversaciones con propósitos similares a los de la organización y buscar una manera de coordinarse durante coyunturas clave.

Responder de manera prudente. No debemos responder en caliente o sin seguir el protocolo de gestión de crisis. En todo caso, es importante tener en cuenta el factor tiempo: mientras hay ataques que pueden esperar, hay otros que requieren de una respuesta inmediata.

Tomar medidas para fortalecer el ecosistema digital. Esto se logra si ciertas conversaciones se llevan a espacios como cuentas dedicadas a tratar temas sensibles para la organización y manejadas siempre por voceros previamente definidos. En estos espacios se puede controlar mejor a los contradictores para impedir hasta cierto punto que tengan amplificación o herramientas suficientes para desplegar o agrandar un ataque.

Leer el contexto offline. En la evaluación de la respuesta al ataque es útil tener un termómetro de nuestras comunidades tanto en espacios en línea como *offline*. Ninguna acción de respuesta en línea debe estar desligada del entorno *offline* de la organización. Por el contrario, es importante aprovechar este espacio para diseñar respuestas coordinadas y conocer mejor a las audiencias.

Referencias

Alizadeh, M. S. (2020). Content-based features predict social media influence operations. *Science Advances*.Benesch, S. (2014). *Countering dangerous speech: new ideas for genocide prevention*. United States Holocaust Memorial Museum.

Benford, R., & Snow, D. (s.f.). Framing Processes and Social Movements: An Overview and Assessment. *Annual Review of Sociology*, 2000.

Blackmore, E., & Sanderson, B. (2017). *Framing equality toolkit*. ILGA-Europe & PIRC

Bradshaw, S., Bailey, H., & & Howard, P. N. (2020). *Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation*. Obtenido de Oxford Internet Institute: https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/127/2021/01/CyberTroop-Report20-FINALv.3.pdf

Brown, R. H. (2016). *Defunsing Hate: A Strategic Communication Guide to Counteract*. United States Holocaust Memorial Museum.

Coombes, T., & Krizna, G. (2019). *Be the Narrative. How changing the narrative could revolutionize what it means to do human rights*. JustLabs.

Crabtree-Condor, I. (2020). *Narrative Power and Collective Action: Conversations with people working to change narratives for social good - Part 1*. Oxfam On Think Tanks.

Crabtree-Condor, I. (2020). *Narrative Power and Collective Action: Conversations with people working to change narratives for social good – Part 2*. Oxfam On Think Tanks.

Dangerous Speech: A Practical Guide. (2018). *Dangerous Speech Project*. Obtenido de dangerousspeech.org

Decker, B. (2019). *Adversarial Narratives: A New Model for Disinformation*. Obtenido de Global Disinformation Index: https://www.disinformationindex.org/blog/2019-8-1-adversarial-narratives-are-the-new-model-for-disinformation/

Entman, R. (1993). Framing: Towards Clarification of a Fractured Paradigm. . *Journal of Communication,*, 43 (4), 51-58.

Entman, R., Matthes, J., & Pellicano, L. (2009). Nature, sources, and effects of news framing. *The handbook of journalism studies*.

FrameWorks. (2020). *Mindset Shifts: What Are They? Why Do They Matter? How Do They Happen?* . Robert Wood Johnson Foundation.

Gleicher, N. (2018). Coordinated Inauthentic Behavior Explained. *Facebook Newsroom*. Obtenido de https://about.fb.com/news/2018/12/inside-feed-coordinated-inauthentic-behavior/

HeartMob. (s.f.). *Social Media Safety Guides*. Obtenido de https://iheartmob.org/resources/safety_guides

Kalra, N., Farfan, C. B., Robles, L., & Stachowiak, a. S. (2021). *Measuring Narrative Change Understanding Progress and Navigating Complexity*. ORSImpact.

Lakoff, G. (2014). *Don't think of an elephant!* Chelsea Green Publishing.

Lakoff, G. (s.f.). *A Modest Proposal: #ProtectTheTruth*. Obtenido de https://georgelakoff.wordpress.com/2018/01/13/a-modest-proposal-protectthetruth/

Latour, A. d., Perger, N., Salaj, R., Tocchi, C., & Otero, P. V. (2019). *Sí Podemos Actuar Contra el Discurso de Odio Mediante Contranarrativas y Narrativas Alternas*. Ediciones Conapred.

Luntz, F. (2008). *Words at Work: It's Not What You Say, It's What People Hear* . Boston: Hachette Books.

Meedan. (2020). *Content Moderation Toolkit*. Obtenido de https://meedan.com/reports/content-moderation-toolkit/

Pacheco, D., Hui, P.-M., Torres-Lugo, C., Truong, B. T., Flammini, A., & Menczer, F. (2021). Uncovering Coordinated Networks on Social Media: Methods and Case Studies. Proceedings of the International AAAI Conference on Web and Social Media, 15(1), 455-466.

Pen America. (2021). *Online Harassment Field Manual*. Obtenido de https://onlineharassmentfieldmanual.pen.org/es/

Pew Research Center. (2021). *The State of Online Harassment*. Obtenido de https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/

Phillips, W. (2018). *The Oxygen of Amplification. Better Practices for Reporting on Extremists, Antagonists, and Manipulators Online*. Obtenido de Data & Society: https://datasociety.net/wp-content/uploads/2018/05/FULLREPORT_Oxygen_of_Amplification_DS.pdf

Rhodes Artificial Intelligence Lab (RAIL). (2018). *The challenge of identifying subtle forms of toxicity online*. Obtenido de https://medium.com/jigsaw/the-challenge-of-identifying-subtle-forms-of-toxicity-online-465505b6c4c9

Smith, R., & Dotto, C. (2019). *The not-so-simple science of social media bots*. Obtenido de First Draft: https://firstdraftnews.org/articles/the-not-so-simple-science-of-social-media-bots/

The Narrative Initiative. (2017). *"Toward New Gravity: Charting a New Course for the Narrative Initiative"* . Obtenido de https://narrativeinitiative.org/wp-content/uploads/2019/08/TowardNewGravity-June2017.pdf.

Tim Holmes, E. B. (2012). *The Common Cause Handbook: A Guide to Values and Frames for Campaigners, Community Organisers, Civil Servants, Fundraisers, Educators, Social Entrepreneurs, Activists, Funders, Politicians, and everyone in between*. Obtenido de http://www.commoncause.com.au/uploads/1/2/9/4/12943361/common_cause_handbook.pdf

Linterna Verde