Informe de monitoreo
Social Media 4 Peace Colombia

versión 2.0 - 10 de abril de 2023

1. Introducción

En el marco del programa 'Social Media 4 Peace' promovido por Unesco, desarrollamos un monitoreo de contenido potencialmente dañino en Twitter (en adelante, usaremos el término completo o la sigla CPD). El objetivo del ejercicio es observar cómo ciertas comunidades o grupos poblaciones en Colombia están expuestos a publicaciones que riñen con las reglas de las plataformas y que eventualmente pueden llegar a ser consideradas como incitación al odio, incitación a la violencia o acoso.

Como criterio base para definir el CPD se tomaron, además de las de Twitter, las normas comunitarias de Meta y YouTube, todas las cuales establecen sus propias reglas de interacción y contenido a la vez que ofrecen parámetros y ejemplos para identificar publicaciones problemáticas. A partir de cinco políticas relacionadas con material potencialmente dañino en estas plataformas –amenazas, comportamiento abusivo, incitación al odio, bullying y acoso, y violencia e incitación– organizamos un esquema que nos permitiera clasificar el contenido en tres categorías:

- Incitación al odio. Comentarios basados en características protegidas y que incluyen calumnias epítetos, tropos racistas o sexistas y expresiones para degradar a otra persona; comentarios deshumanizantes; llamados a la exclusión o discriminación; o vinculación con actividades delictivas, grupos criminales o terroristas.

- Acoso. Insultos o términos despectivos para atacar basado en actividades sexuales; burlarse o negar una tragedia; insultos y lenguaje obsceno en general; comentarios sexualizados graves; o ataques a una persona por ser víctima de violencia.

- Incitación a la violencia. Declaraciones a favor de la violencia; amenazas directas o deseo de daños.

Para detectar esta clase de contenido, monitoreamos la conversación en Twitter alrededor de siete eventos ocurridos entre 2022 y 2023 y extrajimos una muestra relevante para el análisis. Se trata de siete hechos que tocan asuntos sociales, migratorios, sobre paz y justicia transicional, género y discriminación racial, en Colombia. A continuación enumeramos los hechos y subrayamos el grupo vulnerable o minoritario en el que enfocamos la observación sobre el posible contenido problemático del que fueron destinatarios:

- El enfrentamiento entre miembros de la comunidad emberá y la Policía Nacional ocurrido en Bogotá en noviembre de 2022.

- La resolución de conclusiones emitida por la Jurisdicción Especial para la Paz (JEP) relativa a la "toma de rehenes, graves privaciones de la libertad y otros crímenes cometidos por las FARC-EP", y los excombatientes firmantes del acuerdo.

- La propuesta del gobierno nacional de liberar a jóvenes manifestantes de la primera línea, detenidos durante las movilizaciones sociales de años anteriores, para nombrarlos como gestores de paz.

- La captura y el procesamiento judicial de una banda delincuencial conformada por ciudadanos venezolanos en Bogotá.

- La reanudación de las mesas de diálogo con el Ejército de Liberación Nacional (ELN), los negociadores del grupo ilegal y firmantes de acuerdos pasados.

- El feminicidio de la DJ Valentina Trespalacios y las víctimas de violencias basadas en género.

- La conversación alrededor del esquema de seguridad de la vicepresidenta Francia Márquez y las comunidades afrocolombianas.

Este trabajo incluyó dos elementos adicionales sobre las normas comunitarias y su aplicación: por un lado, el seguimiento a los cambios de estas reglas de las plataformas que resultaran relevantes para este análisis. Esta observación permite evaluar su evolución e idoneidad para gestionar el CPD en sus espacios. Este punto será desarrollado en el apartado sobre las tensiones con las reglas de las plataformas.

Por otro lado, quisimos identificar en Twitter episodios de fallas en la moderación de contenido que afectaran la participación en línea de los miembros de las comunidades afectadas –ya fuera en esa red social o en otra que desde allí se denuncia–. Sin embargo, no hallamos episodios de esta naturaleza a través de la metodología desarrollada. En el aparte de conclusiones ofrecemos una explicación sobre la ausencia de denuncias de casos.

Finalmente, hacemos dos aclaraciones fundamentales sobre el alcance y naturaleza de este monitoreo:

- Como se explica en la metodología, no tuvimos en cuenta como CPD los tuits que respondían o comentaban denuncias de agresiones o daños a bienes

públicos (por ejemplo, cuando circula un video en el que un gestor de convivencia es atacado por un manifestante). Las normas comunitarias ofrecen un mayor nivel de protección a las expresiones de indignación o rechazo que se dan en estos contextos.

- Este ejercicio no busca equiparar el CPD identificado con categorías jurídicas de odio o incitación a la violencia. Tampoco pretende presentar este monitoreo como un diagnóstico exhaustivo de la conversación en Twitter sobre estos temas. El propósito, en últimas, es observar estos contenidos y sus interacciones y tener mayores elementos de juicio para analizar las tensiones entre la participación de los usuarios, las normas comunitarias y el ejercicio de derechos como la libertad de expresión.

2. Metodología

El monitoreo fue desarrollado en tres fases:

- **Exploración.** Con la selección de los casos se elaboraron diccionarios para cada uno de los eventos: palabras clave, sinónimos y expresiones relacionadas con contenido potencialmente dañino. Estos siete diccionarios fueron utilizados como llave de búsqueda en Meltwater Explorer, una herramienta de escucha social (ver Anexo: diccionario de búsquedas).

- **Captura.** A partir de estos diccionarios se capturaron las siete conversaciones alrededor de los eventos. Los periodos seleccionados para la captura corresponden a picos en la conversación. De cada captura se seleccionaron las diez publicaciones con mayor interacción, es decir, las que habían generado mayor actividad mediante réplicas, citas y respuestas.

- **Análisis.** A partir de las publicaciones con mayor interacción caracterizamos el contenido que podría ser considerado potencialmente dañino y lo clasificamos de acuerdo a nuestro esquema, según se tratara de incitación a la violencia, incitación al odio o acoso. En uno de los casos en que la muestra arrojó muy pocos resultados a partir del diccionario de expresiones de CPD, ampliamos la observación a la conversación general del tema.

Como se menciona en la introducción, para la fase de análisis no se tuvieron en cuenta como CPD aquellas expresiones problemáticas que se hubieran dado como forma de indignación o rechazo a un evento violento o delictivo. La tolerancia a este tipo de manifestaciones riesgosas en estos contextos no solo hace parte del amparo de la libertad de expresión, sino que en ocasiones se encuentra contemplada en las políticas de las plataformas.

Reconociendo la indignación que puede desencadenar un episodio de estos, la norma de comportamiento abusivo de Twitter permite desear daños graves contra una persona acusada de violencia grave. Medidas de este tipo también han sido adoptadas por Meta para situaciones específicas, como cuando permitió desear la muerte de Vladimir Putin en el contexto de la invasión a Ucrania, o la del ayatola Jamenei durante las protestas en Irán.

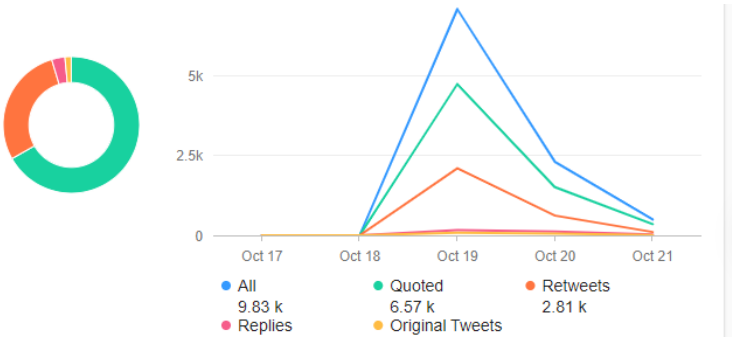
Por último, es importante resaltar que no toda la muestra capturada constituye CPD. Al filtrar los contenidos con los diccionarios se obtiene una muestra de publicaciones que incluyen estos términos, pero es necesaria la evaluación del contexto y el sentido de cada tuit para determinar si en efecto lo es.

3. Hallazgos por episodio

a. Comunidad Emberá

- El 19 de octubre de 2022 se presentaron enfrentamientos entre la policía nacional y miembros de la comunidad indígena Emberá, quienes habían organizado una protesta en el centro de Bogotá para reclamar por las condiciones de los albergues en los que fueron reubicados luego de haber vivido durante meses en campamentos en el Parque Nacional, a donde llegaron tras haber sido víctimas de violencia en sus regiones. Los enfrentamientos fueron noticia nacional; imágenes de las agresiones, tanto a miembros de la comunidad como a la policía y gestores de convivencia del Distrito, fueron divulgadas por redes sociales al mismo tiempo que se desarrollaban los hechos.

- La conversación fue capturada en el periodo del 17 al 21 de octubre, en el que se detectaron cerca de 9.830 tuits, con un pico el día del evento.



- Dado que en efecto ese día circularon imágenes de agresiones a la policía, gestores de convivencia y a bienes públicos, ciertas expresiones de indignación o rechazo que podían incluir términos agresivos o condenatorios no fueron tenidos en cuenta como CPD.

- Al evaluar los diez principales tuits por interacciones se encuentra que dos de ellos constituyen CPD. En el primer caso, un usuario se refiere a los indígenas como "indios farianos", asociándolos directamente a la guerrilla de las FARC. En el segundo, una usuaria los llama "plaga". Ambos comentarios podrían ser considerados contenido de incitación al odio. En el primero, por la asociación con grupos delincuenciales o terroristas y, en el segundo, como una comparación deshumanizante.

- En las respuestas a los top tuits vemos que comentarios de esta naturaleza atraviesan la conversación. Sin hacer ninguna distinción, se ataca de manera general a la población indígena llamándola "narcoindígena", "terroristas indígenas" o "asesinos y narcos", además de otras referencias a la guerrilla de las FARC o del ELN.

- A su vez, el top tuit que se refiere a los indígenas como "plaga", da lugar a que en las respuestas prevalezcan comentarios deshumanizantes en los que se le llama a la población "parásitos de la sociedad", "plaga delincuencial" o "animales de monte". Expresiones semejantes se encuentran igualmente en las respuestas a otros top tuits, en los que las referencias a la población indígena parecen plantearse en la dicotomía civilización-barbarie, donde los emberá son llamados "salvajes" o "incapaces de vivir en la civilización".

- Por esta misma línea se encuentran comentarios de incitación al odio bajo la modalidad de exclusión o discriminación, en los que se llama a que la comunidad indígena sea regresada a sus territorios.

- Cuatro de los diez principales tuits por interacción incluyen videos en los que aparecen miembros de la comunidad indígena agrediendo a agentes de la policía o a gestores de convivencia del distrito. Las respuestas a estas publicaciones

dan lugar a expresiones de incitación a la violencia que trascienden la indignación o el rechazo, y que podrían constituir declaraciones a favor de la violencia. Es el caso de los llamados a que se les dispare a los miembros de la comunidad indígena (a,b,c).

- La muestra arrojó también contenidos que podrían ser considerados como amenazas, donde se afirma que esta población es "un problema que hay que acabar de raíz" o se plantea la idea de acudir a prácticas de exterminio del pasado.

- Por último, la conversación permite ver publicaciones que podrían ser consideradas CPD a través de acoso, en el sentido de que se emplean insultos y lenguaje obsceno contra la población indígena (a,b,c,d).

b. Decisión de la JEP en el caso de secuestros

- El 25 de noviembre de 2022 la Sala de Reconocimiento de Verdad de la Justicia Especial para la Paz (JEP) –tribunal de justicia transicional creado tras el acuerdo de paz con la guerrilla de las FARC– emitió una resolución de conclusiones para el caso 01: "toma de rehenes, graves privaciones de la libertad y otros crímenes", lo que en la opinión pública se reconoce como secuestros. El documento, que no constituye una decisión final, recomendó penas restaurativas para exmiembros del antiguo secretariado de las FARC.

- La conversación fue capturada en el periodo del 22 al 28 de noviembre de 2022, en la que se detectaron cerca de 5.570 tuits, con un pico el 25, día en que se anunció la decisión.

- Dado el contexto de la conversación, que involucra a personas que en efecto participaron en diferentes delitos durante el conflicto, muchas de las expresiones de indignación que implican acusaciones contra excombatientes parten de un supuesto fáctico y no constituyen CPD, como lo permiten el derecho a la libertad de expresión y las normas comunitarias que orientaron este monitoreo. Por esta razón, al revisar el contenido de los top tuits, disminuye el número de expresiones que podrían considerarse como incitación al odio; el CPD se concentra en declaraciones a favor de la violencia: llamados a la "justicia por mano propia" o a atacar a los excombatientes.

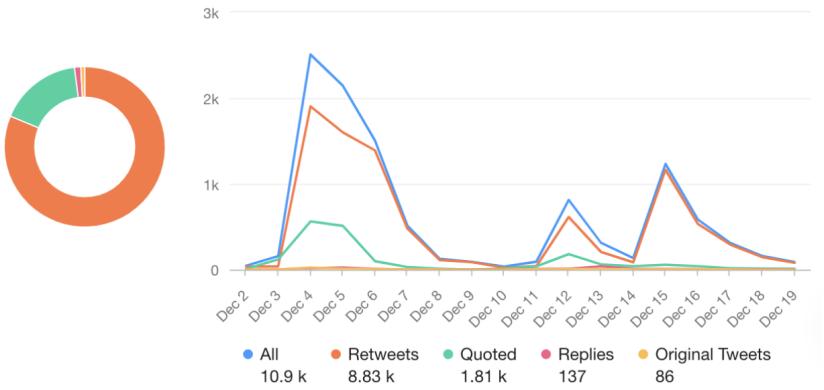
- Ninguno de los principales tuits por interacción pueden ser considerados CPD. Al revisar las respuestas, solo se encuentra una publicación que podría ser considerada incitación a la violencia: en respuesta a un tuit de una reconocida periodista, un usuario llama a la ciudadanía a armarse y atacar a los excombatientes en las calles.

- Al depurar y revisar la base total de la conversación, se encontraron otros cuatro ejemplos que de igual forma podrían constituir incitación a la violencia, en los que se llama a ejercer justicia por mano propia y a penas crueles e inhumanas.

c. Liberación de manifestantes encarcelados

- El 3 de diciembre de 2022, el gobierno nacional anunció su intención de liberar a jóvenes manifestantes que habían sido detenidos durante las protestas sociales de los últimos años, incluidos miembros del grupo Primera Línea, para convertirlos en gestores de paz.

- La conversación fue capturada en el periodo del 2 al 19 de diciembre, en el que se detectaron cerca de 10.900 tuits. Los picos de la conversación, visibles en la siguiente gráfica, corresponden al día del anuncio y a una supuesta exigencia del ELN para liberar a los miembros de la Primera Línea, lo que reactivó la conversación.



- Al igual que en otros casos, la conversación está vinculada con eventos del pasado que pudieron haber involucrado actos de violencia, por lo cual no todas las expresiones de rechazo o indignación, aunque incluyan términos agresivos, pueden ser considerados CPD.

- Ninguno de los principales tuits por interacción constituye en sí mismo CPD. Sin embargo, la revisión de las respuestas permite ver que prevalecen ciertas publicaciones en las que los manifestantes o miembros de la Primera Línea son vinculados con el ELN, el Bloque Occidental de esta guerrilla o se les llama "narcoterroristas" o miembros de "milicias urbanas". De acuerdo con nuestra caracterización, esta clase de vinculaciones podrían considerarse una forma de incitación al odio.

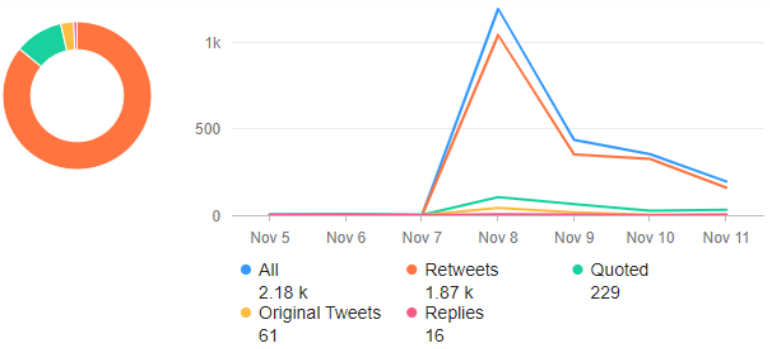
- La conversación incluye en menor medida contenidos que podrían ser considerados incitación a la violencia, como llamados a ejercer justicia por mano propia o declaraciones que de manera menos directa se refieren a la muerte de los involucrados, como ocurre con un usuario que pide para ellos "cárcel o bolsas negras".

- Esta conversación permitió encontrar CPD dirigido no solo contra los involucrados, sino también contra los autores de los top tuits, quienes en algunos casos eran periodistas:
    - En las respuestas a uno de los principales tuits por interacción, en el que se cita una publicación de una periodista, se encuentra contenido que sexualiza a esta última llamándola "prepago", no siempre con la connotación de ser una periodista que trabaja a favor de otros intereses.
    - A una periodista que también aparece en los principales tuits por interacción con una publicación en la que pide diferenciar falsos positivos judiciales con delincuentes de Primera Línea que sí cometieron delitos, se le ataca en las respuestas llamándole "guerrillera".

d. Captura y procesamiento de ciudadanos venezolanos

- El 8 de noviembre se conoció la noticia de que había sido capturada una banda de 13 personas de nacionalidad venezolana dedicada a asaltar con armas de fuego a pasajeros en Transmilenio, el sistema de transporte masivo de Bogotá. Ese mismo día, una juez de control de garantías decidió no dictar medida de aseguramiento, una figura legal que en Colombia permite recluir a personas en establecimientos carcelarios mientras se dicta sentencia. Uno de los argumentos de la juez para dejar en libertad a los implicados fue el costo que tiene para el Estado mantener a un ciudadano privado de su libertad. La decisión aumentó el interés alrededor de la noticia e involucró a la alcaldesa Claudia López, quien criticó que la liberación de presuntos criminales se diera por razones económicas.

- La búsqueda tuvo como referencia el periodo de 5 a 11 de noviembre de 2022, con una captura de cerca de 2.180 tuits. Para este caso, la llave de búsqueda que incluye expresiones de CPD arrojó muy pocos resultados, por lo cual se hizo la búsqueda en la conversación general. Con este enfoque ampliado, la mitad de los top tuits son publicaciones de medios de comunicación o líderes de noticias.



- Al tratarse de un evento que incluye un hecho delictivo, ciertas expresiones de rechazo o indignación, como se explicó, son legítimas y no se consideran CPD. La conversación, sin embargo, permite ver reacciones que trascienden la noticia y se dirigen contra la población migrante en general.

- Al igual que en el caso de la JEP, este episodio tiene como una de sus aristas la insatisfacción frente a las decisiones del sistema judicial. Cinco de los diez principales tuits por interacción mencionan que la banda quedó libre, y en cuatro de ellos se cuestiona la situación. Las respuestas a estas publicaciones incluyeron distintos comentarios que podrían ser considerados como incitación a la violencia: llamados a "fusilar" a los involucrados, darles de baja, o declaraciones a favor de una "limpieza social".

- La decisión de dejar a los involucrados en libertad por razones económicas dio lugar a respuestas que podrían ser consideradas CPD y que reaccionan directamente a los argumentos de la juez. Es el caso de usuarios que sugieren que si tener a una persona en la cárcel es muy costoso para el Estado, lo mejor es matarlas, o de quienes dicen que las balas o el cementerio son más baratos.

- La conversación también permite observar contenido que podría considerarse incitación al odio mediante comentarios deshumanizantes, pues se encuentran referencias a la población migrante venezolana en Colombia como una "plaga" (a,b,c).

- Se observa contenido de incitación al odio en forma de expresiones que llaman a la discriminación o a la exclusión de esta población en general. Es el caso de un usuario que llama a no ayudar a los venezolanos en Colombia, ni a darles dinero o arrendarles propiedades, o quienes piden revisar el estatus migratorio de esta población o críticas a los vuelos desde este país.

- En algunos tuits se asegura de manera generalizada que la población migrante venezolana llegó a Colombia para delinquir. En ocasiones, algunos usuarios intentan hacer la salvedad de que no todos son malos, "pero sí la mayoría",

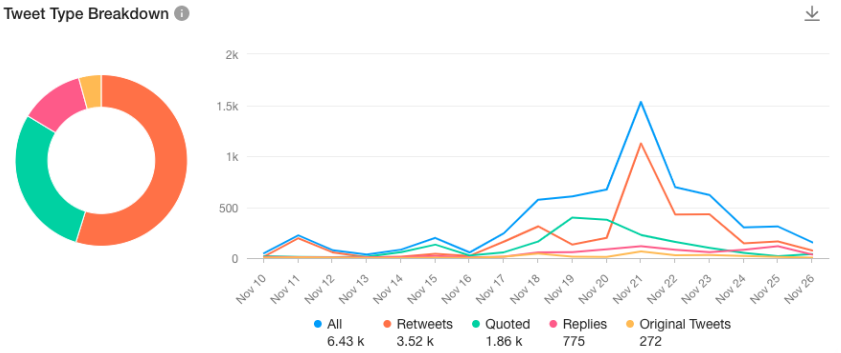
mientras que otros afirman que la mayor cantidad de delincuentes en el país son de esta nacionalidad.

- La conversación permite observar también casos de acoso, en el que la población migrante venezolana es atacada con insultos y lenguaje obsceno en general (a, b, c, d, e).

e. Reanudación de los diálogos de paz con el ELN

- El 21 de noviembre el gobierno nacional reanudó las conversaciones para llegar a un acuerdo de paz con la guerrilla del ELN. El proceso había sido interrumpido por el gobierno anterior en 2019, luego de un atentado terrorista de este grupo contra la Escuela de Cadetes de Policía General Santander en el que murieron 22 personas.

- La conversación fue capturada en el periodo de 10 a 26 de noviembre de 2022, y se encontraron alrededor de 6.430 tuits, con un pico el día del anuncio de la reanudación.



- Al igual que en los otros casos, la conversación incluye muchas publicaciones agresivas que responden al contexto político del país. A los negociadores de las mesas de diálogo y a los miembros del ELN en general se les acusa de toda clase de crímenes –violaciones, atentados, secuestros, asesinatos, delitos contra el medio ambiente, entre otros– que parten de la base fáctica del conflicto a lo largo de décadas. Por estas razones, algunos insultos y comentarios despectivos que en otro caso podrían ser considerados como acoso o incitación al odio, no se tienen como CPD.

- El análisis de top tuits y respuestas solo arroja un contenido que podría ser considerado potencialmente dañino, en la medida en que hace declaraciones a favor de la violencia. En respuesta a la publicación de un periodista que critica la reanudación de los diálogos, un usuario asegura que con delincuentes no se negocia, sino que se les "neutraliza".

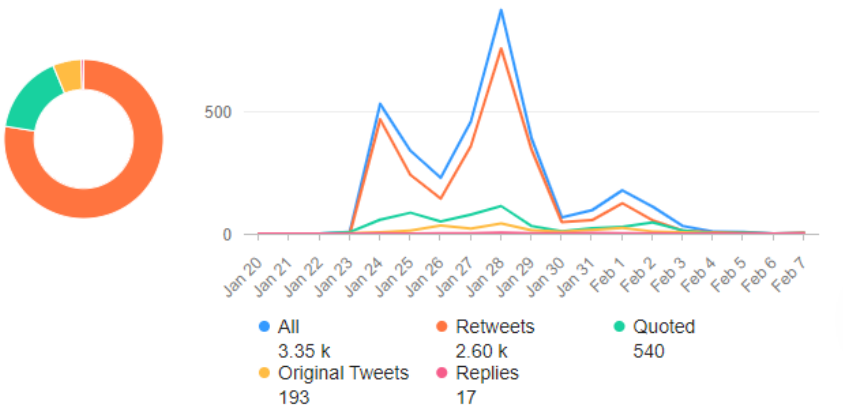
- La revisión de la base de datos de la conversación permite observar otras publicaciones que podrían ser consideradas CPD y que igualmente constituyen declaraciones a favor de la violencia. Es el caso de dos usuarios que explícitamente desean la muerte de los involucrados o celebran una "limpieza social" ejecutada por otros bandos políticos (a,b,c).

- El posible CPD también se refiere a firmantes de acuerdos de paz anteriores. Es el caso de publicaciones que contienen declaraciones a favor de la violencia dirigidas contra miembros del partido Comunes –fundado por excombatientes de las Farc tras el acuerdo de paz de 2016– o contra personas que están en el Congreso como resultado de ese mismo acuerdo.

f. Feminicidio de Valentina Trespalacios

- El 22 de enero de 2023 el cuerpo de la DJ Valentina Trespalacios fue encontrado con señales de tortura en un contenedor de basura en el occidente de Bogotá. Desde el momento del hallazgo, se tuvo como principal sospechoso a su pareja, el ciudadano estadounidense John Poulos. Durante los días y las semanas siguientes los medios de comunicación cubrieron amplia y exhaustivamente el caso, informando sobre la captura de Poulos, la relación entre la víctima y el presunto victimario y detalles de la vida personal de cada uno que fueron divulgados y comentados a través de redes sociales.

- La conversación observada se limitó al periodo entre 20 de enero y 7 de febrero y se capturaron alrededor de 3.350 tuits. El primer pico señalado en la gráfica corresponde al momento en que se da la noticia por primera vez, mientras que el segundo corresponde a una ola de rechazo a comentarios que pretendían justificar en crimen basados en datos sobre la vida privada de la víctima.



- Ocho de los diez principales tuits por interacción rechazan las posturas de quienes intentan justificar el feminicidio de Valentina Trespalacios a raíz de información o rumores sobre su vida privada y su relación con John Poulos,

mientras que los otros dos reproducen rumores sobre una supuesta infidelidad como móvil del crimen.

- A diferencia de otras conversaciones, la mayoría de las cuentas que publicaron los principales tuits por interacción no corresponde a tribunas de medios de comunicación, periodistas, políticos o figuras públicas, sino a cuentas personales de usuarios sin una gran audiencia. Esto, a excepción de un usuario que asegura informar "lo que los medios callan", quien reprodujo rumores sobre una supuesta infidelidad acompañados de imágenes del levantamiento del cadáver de la víctima, que aparece con un intersticial de advertencia de contenido potencialmente delicado.

- La conversación permitió observar distintas muestras de acoso. Es el caso de los contenidos que atacan a Valentina Trespalacios como víctima de violencia, cuestionando su inteligencia por, supuestamente, no haber atendido las advertencias de su familia o por sus supuestos intereses en la relación con Poulos.

- Dos de las usuarias que rechazaron los intentos por justificar el delito fueron a su vez víctimas de acoso por parte de usuarios que las atacaron con insultos basados en actividades sexuales (a,b).

- La conversación dio lugar también a un comentario sexualizado grave que alude a la necrofilia, así como de publicaciones que se burlan de la víctima llamándola "angelita".

- Un único comentario podría constituir una declaración a favor de la violencia, agradeciendo a Poulos por haber cometido el feminicidio. El tuit fue publicado por una cuenta anónima y con pocos seguidores que, de acuerdo con su nombre de usuario, está a favor del patriarcado.

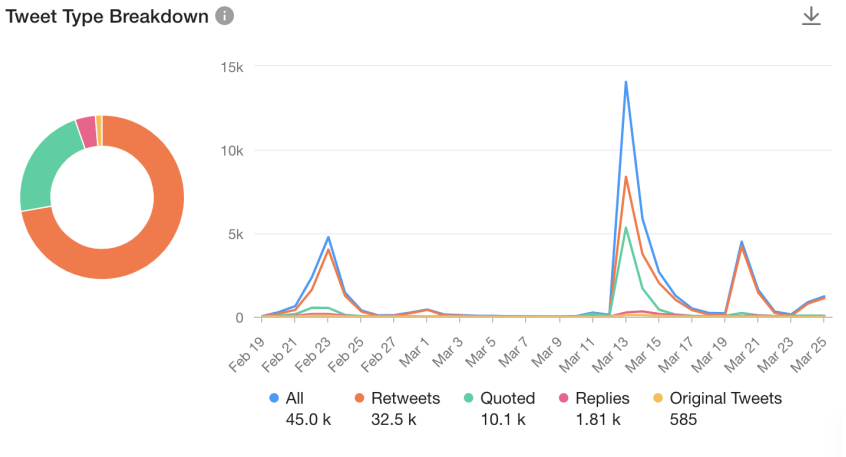
- La muestra permitió observar distintas publicaciones que adjudican la culpa a la propia víctima (a,b,c,d). Ninguna de las normas comunitarias que orientaron la clasificación del CPD sanciona este tipo de afirmaciones, lo cual podría constituir un área gris en el ámbito de protección de estas normas, pues tanto en el caso de Valentina Trespalacios como en otros de violencias basadas en género, contenido de esta naturaleza podría ser considerado revictimizante.

g. Discusión sobre el esquema de seguridad de Francia Márquez

- En febrero de 2023 la senadora María Fernanda Cabal criticó que la vicepresidenta Francia Márquez se transportara en helicóptero para llegar a su casa en Dapa, una localidad cercana a Cali. Márquez, quien el mes anterior había

sido objeto de un frustrado atentado con explosivos, respondió que la aeronave hacía parte de su esquema de seguridad. En los siguientes días la controversia dio lugar a una conversación alrededor de los supuestos lujos de la Vicepresidenta tras su llegada al poder y de defensores suyos que aseguraban que las críticas tenían un trasfondo racista. Unas semanas después la conversación se reactivó, luego de que, en una entrevista con la revista Semana, Márquez expresara que seguiría utilizando un helicóptero para movilizarse. "Pueden gritar, pueden llorar, pueden hacer todo lo que quieran, me pueden ir a demandar si quieren, y que sea un juez el que defina si estoy haciendo algo ilegal", dijo en esa oportunidad.

- La conversación observada se limitó al periodo entre 20 de febrero y 24 de marzo y se capturaron cerca de 45.000 tuits. El primer pico de la conversación corresponde al inicio de la discusión entre Francia Márquez y la senadora Cabal, mientras que el segundo está relacionado con su entrevista en Semana.



- La conversación permitió encontrar distintas publicaciones que podrían ser consideradas incitación al odio al tratarse de comparaciones deshumanizantes o actos de discriminación, en los que se utilizan imágenes, videos, GIFs o emoticones para representar a la vicepresidenta como un simio (a, b, c, d, e, f, g, h). En ocasiones, la vicepresidenta es llamada directamente "simio" o "gorila".

- Al hablarse de los supuestos lujos del sector en el que se ubica la casa de la vicepresidenta, sale a flote una publicación que igualmente podría ser considerada incitación al odio por ser un acto de exclusión, en el que un usuario asegura que Márquez debería vivir en una "choza en el Chocó".

- El uso de epítetos racistas también tiene presencia en la conversación. Para el caso, es relevante destacar que el uso de la palabra "negra" no implica automáticamente una acepción discriminatoria. Sin embargo, cuando está acompañado de insultos o en un contexto que indica un sentido peyorativo podría ser considerado como CPD (a, b, c, d). En otro caso, se asegura que la vicepresidenta "detesta a los blancos".

- Ninguna de las plataformas de redes sociales cuyas normas comunitarias sirvieron como criterio orientador para establecer la clasificación de CPD utilizada en esta investigación incluye la clase social de una persona como una categoría protegida por las políticas de incitación al odio. Sin embargo, dado el contexto de la conversación y el origen humilde de Francia Márquez, encontramos distintas publicaciones en las que tropos, insultos y expresiones peyorativas están basadas en esta forma de discriminación. Es el caso del uso de las palabras "zarrapastrosa", "levantada" o "manteca", como de los usuarios que atacan a la Vicepresidenta por considerarla una resentida social (a,b). A su vez, otras publicaciones que podrían constituir ataques, bien con un sentido de discriminación racial o social, se fundan en la forma en que Márquez pronuncia ciertas palabras (a,b).

- En el material se encuentran publicaciones que incluyen ataques o burlas fundamentados en la condición de víctima de la Vicepresidenta, bien por motivos racistas o por ser objeto de intentos de atentados (a, b). Como en el punto anterior, las normas comunitarias de las plataformas que orientaron la clasificación del CPD para esta investigación no tienen reglas establecidas para esta clase de publicaciones. En ocasiones, las redes sociales prohíben publicaciones de acoso o incitación al odio contra víctimas de acoso y abuso sexual o de eventos de violencia grave. Aunque no es posible afirmar que en el sentido de esta conversación las publicaciones detectadas puedan constituir CPD, el fenómeno expone una tensión entre las políticas de las plataformas y situaciones en las que se dirigen ataques o burlas contra una persona que en efecto ha sido víctima en distintos sentidos.

- Por último, algunas publicaciones encontradas en este monitoreo podrían constituir formas de acoso basadas en el peso corporal de Francia Márquez. Las formas de acoso relacionadas con el aspecto de una figura pública son un asunto en el que no concuerdan las normas comunitarias de las plataformas. Mientras que YouTube, por ejemplo, tiene este tipo de publicaciones como una excepción a sus políticas de acoso, Meta las sanciona. El asunto pone de presente una nueva tensión entre la moderación de contenidos y la libertad de expresión, teniendo en cuenta que las figuras públicas cuentan con un menor umbral de protección frente a las críticas, como lo ha reconocido la jurisprudencia interamericana. No obstante, vale la pena resaltar que en el contexto de esta conversación convergen asuntos de género y raza que complejizan la evaluación de este tipo de contenidos. En ocasiones, los comentarios sobre el aspecto físico coinciden con comparaciones deshumanizantes.

4. Análisis y conclusiones

a. Tensión con reglas de las plataformas

Este trabajo tuvo como insumo para el análisis un seguimiento de las normas de las plataformas relevantes, de manera que pudiera evaluarse su evolución en el último año y su idoneidad para prevenir contenido potencialmente dañino.

Las políticas de contenido de las plataformas de Meta, Twitter y YouTube cubren, desde sus propias aproximaciones, conductas que podrían considerarse contenido potencialmente dañino. La política de amenazas violentas de Twitter, por ejemplo, incluye las amenazas de muerte, la agresión sexual y los actos violentos; Meta, por su parte, prohíbe en la política de lenguaje que incita al odio la incitación a la violencia y las comparaciones deshumanizantes.

A pesar de las prohibiciones de las normas, estos tipos de contenido abundan en las redes sociales debido al volumen con el que se generan y la dificultad de abordarlos a partir de un contexto suficiente. El carácter dañino de este material no siempre es evidente, o se enmarca dentro de situaciones sociales y políticas complejas que pueden superar la capacidad de las plataformas para hacer una evaluación adecuada. Por esta razón, plataformas como Twitter matizan o alivianan las sanciones relacionadas con CPD en que los intercambios o reacciones se producen al calor de un evento donde se evidencian agresiones o violencia de sujetos que después son receptores de ese contenido problemático.

En el último año, Meta ha modificado sus políticas de violencia e incitación para prohibir el contenido que exprese la intención de llevar armas a lugares en las que hay señales de riesgo elevado de violencia, lo que según la propia plataforma incluye contextos de protestas o brotes de violencia locales. Si bien en las conversaciones revisadas no encontramos contenidos de esta naturaleza, queda claro que de manera creciente las plataformas orientan el diseño de sus políticas para prevenir daños en la vida real durante situaciones concretas. Estas situaciones, a su vez, pueden ser detonantes de CPD, como se evidencia en el caso de los enfrentamientos de la comunidad Emberá y la policía, o en el de la propuesta de liberación de los manifestantes de la primera línea.

Meta ha modificado sus normas para dar mayor claridad sobre algunos comentarios que podrían ser considerados lenguaje que incita al odio. En noviembre de 2022, cambió esta política para ampliar la prohibición de comentarios que ataquen a una categoría protegida con expresiones de “infrahumanidad” –llamándolos salvajes o primitivos–. Esta clase de inclusiones amplían el alcance de la norma y ofrecen criterios

para restringir expresiones de CPD que, en el caso de Twitter, pudimos observar en la conversación sobre la comunidad Emberá.

El caso del feminicidio de la DJ Valentina Trespalacios expone una nueva tensión en relación con las normas comunitarias y el CPD. Si bien algunas plataformas protegen a las víctimas de algunos delitos sexuales o violentos de ser objeto de comportamientos abusivos, burlas, o comentarios que nieguen o cuestionen que dichos eventos ocurrieron, no hay ninguna regla que prevenga otro tipo de contenido que podría ser revictimazante: el que adjudica la culpa de una tragedia a la persona que la padeció. Como señalamos arriba, comentarios de esta naturaleza están presentes en la conversación, y aunque no incumplen ninguna norma comunitaria, podrían llegar a afectar a víctimas de violencias basadas en género.

A su vez, la conversación alrededor de la vicepresidenta Francia Márquez pone de presente una forma de discriminación que no es tenida en cuanta por las normas comunitarias de las plataformas: la que se da por clase u origen social. Como se expuso en su momento, el monitoreo encontró distintas expresiones peyorativas relacionadas con el origen y trayectoria de la Vicepresidenta, quien proviene de una zona históricamente excluida y en el pasado trabajó en el servicio doméstico. En el caso de Colombia, la relación entre discrminación racial y pobreza ha sido estudiada en distintas oportunidades, bien de manera amplia como en el contexto específico de la campaña electoral que antecedieron la llegada de Márquez a la vicepresidencia. Si bien es necesario adelantar un análisis independiente para esta clase de contenidos, es posible que en esta clase de conversaciones nos encontremos ante una suerte de discriminación interseccional en la que comentarios ofensivos, basados en clase, tienen un vínculo con características protegidas por las normas comunitarias.

Para este mismo caso, vale exponer una última tensión presente entre las normas comunitarias y ciertas publicaciones que se refieren negativamente al peso corporal de Francia Márquez. En este tipo de contenidos, la libertad de expresión, protegida de manera más amplia por estar relacionada con una figura pública que ocupa un cargo de poder, puede reñir con ciertas formas de acoso, en las que de nuevo se involucran asuntos de raza y género.

b. Conclusiones

- La complejidad de las conversaciones analizadas y de los asuntos políticos y sociales del país ponen de presente la relevancia del contexto al momento de evaluar un contenido y señalarlo como CPD. Esta relevancia, presente en todos los casos analizados, es más marcada en conversaciones como la que suscitó la reanudación de los diálogos de paz con el ELN o la decisión de la JEP sobre el caso de secuestros. Allí, buena parte de la conversación incluía expresiones de

indignación y de rechazo que podían darse en términos agresivos o despectivos, pero que deben ser entendidas en el marco de la historia del conflicto armado colombiano.

- En cuanto a la conversación, el análisis de los principales tuits por interacción permite ver que el contenido potencialmente dañino no se encuentra por lo general en las publicaciones que tienen amplificación relevante, sino en las respuestas de otros usuarios a estas publicaciones. Y entre éstas, hay varios indicios de actividad inauténtica.

- Solo dos de los principales tuits por interacción analizados podrían constituir en sí mismos contenido potencialmente dañino, ambos en la conversación sobre la población Emberá (a,b). En las dos circunstancias las cuentas corresponden a tribunas de nicho, en comparación con los demás autores de los principales tuits. En el caso de la usuaria que llamó a la población a la población “plaga”, muchos usuarios responden con esa misma o expresión o con otros comentarios deshumanizantes.

- En este mismo hecho, seis de los principales tuits por interacción incluyen videos en los que aparecen miembros de la comunidad agrediendo a policías o gestores de paz. La mayoría de las respuestas que podrían ser consideradas CPD son reacciones a estas publicaciones, con declaraciones a favor de la violencia y con expresiones discriminatorias que se dirigen contra la población indígena en general.

- En el caso de los ciudadanos venezolanos, cinco de los diez principales tuits por interacción mencionan la decisión de la juez de dejar en libertad a los implicados, y cuatro de ellos la critican. La insatisfacción con la administración de justicia da lugar a expresiones que pueden ser consideradas CPD, en especial declaraciones a favor de la violencia. Una situación similar se presenta en la conversación alrededor de la resolución de conclusiones emitida por la JEP contra los exmiembros del antiguo secretariado de las Farc en el caso de secuestros.

- Mientras que en las conversaciones sobre la decisión de la JEP, la reanudación de diálogos con el ELN y la propuesta para liberar a manifestantes, prevalecen los contenidos que podrían considerarse incitación a la violencia, en las conversaciones que involucran poblaciones con características protegidas –como etnia y nacionalidad– salen a flote, además, otros tipos de contenido potencialmente dañino, como el que podría considerarse acoso, exclusión o discriminación y comentarios deshumanizantes.

- El hecho de que el contenido que puede considerarse potencialmente dañino no esté en los principales tuits por interacción sino en las respuestas, pone sobre la mesa la pregunta sobre el impacto que este material tiene en la discusión y el daño que puede ocasionar. Salvo los casos ya mencionados en la conversación que involucra a la comunidad Emberá, el CPD proviene de cuentas que no constituyen nodos en las conversaciones ni tienen amplificación relevante.

- Un monitoreo distinto de CPD alrededor de eventos que no involucren posibles delitos o que no incluyan personas con un pasado cuestionado (excombatientes, por ejemplo), arrojaría un resultado distinto. Dicho de otra forma, cualquier conclusión que se haga sobre CPD en una red social como Twitter depende de variables como el hecho, los actores y el momento en que se desarrolla la conversación en ese espacio.

- Como se explicó antes en este documento, este ejercicio de monitoreo incluyó la posible detección de casos de moderación de contenidos alrededor de estos hechos. Sin embargo, no se encontró ninguno. Esto concuerda de alguna forma con lo que hemos observado desde Circuito: a menos que se trate de usuarios con tribuna o influencia, la gran mayoría de ciudadanos que enfrentan problemas de moderación de contenidos no saben bien cómo navegarlos ni suelen exponerlos públicamente.

- Durante esta observación detectamos algunas cuentas con indicios de automatización o de hacer parte de un esquema de call-center o bodega (a, b, c, d): nombres de usuarios con muchos números, ausencia de foto de perfil, pocos seguidores, contenido repetido (1,2) y retuits constantes. Si bien este trabajo no buscaba hacer un análisis sobre autenticidad, resulta preciso subrayar la relación que suele existir entre CPD y actividad coordinada.

- Las conversaciones sobre los casos de Valentina Trespalacios y Francia Márquez evidencian algunas zonas grises en la protección de las normas comunitarias frente a ciertas formas de acoso y discriminación, bien como revictimización o como discriminación por características que las políticas de contenido no consideran protegidas, como es el caso de la discriminación social.

- El presente ejercicio constituye un insumo relevante para aproximarse a la pregunta sobre la regulación de las plataformas en relación con el contenido que publican sus usuarios:

    - Primero, pone de presente la necesidad de entender los matices y variables del contenido problemático en estos espacios: la relación con episodios offline, el interés público subyacente y la libertad de expresión. La búsqueda de categorizaciones y diagnósticos cuantitativos pueden

interferir en el propósito de tener una adecuada caracterización del problema.

- Segundo, plantea el interrogante sobre el verdadero impacto de este material, teniendo en cuenta su amplificación limitada y el tipo de actores que lo producen. Por esa vía, también es necesario incluir en el análisis los 'ofrecimientos' de las plataformas para que el usuario gestione este contenido problemático –ocultar respuesta, silenciar o bloquear, entre otros– como alternativa a las soluciones prescriptivas.

- Tercero, se relaciona con los incentivos que tienen las plataformas –y que tendrían a la luz de un cambio legislativo– para expandir y profundizar su rol como árbitros del debate en sus espacios. Una regulación en la materia no puede enfocarse simplemente en las conductas y los contenidos, sino que debe tomar en consideración los problemas de escala, consistencia y proporcionalidad que tienen hoy en día estos actores a la hora de diseñar y aplicar sus normas comunitarias.

* * *

Anexo: diccionario de búsquedas

- Caso de la comunidad emberá:
  - (enfrentamient* OR disturbios OR agresi* OR agred* OR golp*) AND (ind?genas OR indios OR ember?) AND (polic?a* OR esmad OR smad) AND (pepazo OR borrach* OR trago OR coca OR cocalero* OR droga OR salvajes OR v?ndalos OR bandidos OR elimin* OR narcos OR delincuen* OR crimen OR criminal* OR terrorist* OR di?idencias OR asesin* OR guerrill* OR maton* OR par?sit* OR besti* OR fuera OR largo OR tiro OR indio* OR hps OR hijueput* or plaga) NOT (Mexico OR Jilotepec OR Bazin OR Vietnam OR Santo Domingo)
  - Periodo: 17 de octubre - 21 de octubre de 2022.

- Caso de la decisión de la JEP sobre secuestros:
  - (JEP OR "justicia especial" OR magistrad* OR jurisdicc* OR tribunal OR sala) AND (secuestr* OR retenc* OR rehe* OR decisi?n OR resoluci?n) AND (FARC or secretariad* OR fari* OR guerrill* OR desmovil* OR firmant* OR exjef* OR congres* OR senad* OR exsecretariad*) AND (comunist* OR asesin* OR genocid* OR sicari* OR mercenar* OR tortur* OR muert* OR mate* OR mata* OR muer* OR sufr* OR plomo OR bala* OR tiro* OR pepaz* OR merece OR disid* OR dicid* OR faria* OR colga* OR plaga OR paria* OR par?sit* OR narco* OR marquetalia)
  - Periodo: 22 de noviembre - 28 de noviembre de 2022.

- Caso de la propuesta de liberación de jóvenes manifestantes:
  - (aminist?a OR indulto OR "perd?n judicial" OR beneficios OR judicial* OR "sacar c?rcel" OR excarcelar OR libera* OR custodia) AND ("paro nacional" OR protesta* OR paro) AND (j?ven* OR "primera l?nea" OR manifestante* OR "gestores de paz" OR "j?venes detenidos") AND (v?ndalo OR b?ndalo OR "toma guerrillera" OR guerrill* OR criminal OR asesino* OR mercenario* OR bandido* OR terrorista* OR cobarde* OR delincuencia OR terrorismo OR encapuchado* OR gamine* OR desadaptado* OR mamerto* OR vicioso* OR tortur* OR matar* OR bajen OR mueran OR deber?an OR sangre OR espero OR ojal? OR pistola) NOT (country: pe)
  - Periodo: 2 de diciembre - 19 de diciembre de 2022.

- Caso de la captura y proceso de ciudadanos venezolanos:
  - (Venezolanos OR extran?er* OR Venezuela OR veneco* OR veneca*) AND (band* OR grup* OR pandill* OR 10 OR 13 OR diez OR trece) AND (delincuen* OR criminal* OR atrac* OR roba* OR robo* OR ladr?n* OR hurt* OR hamp*) AND (transmilenio OR bus OR buses OR estaci?n* OR transporte) NOT (Chile OR Per? OR peruan* OR EE.UU OR M?xic*)

    - Periodo: 5 de noviembre - 11 de noviembre de 2022.

- Caso de la reanudación de diálogos de paz con el ELN:
    - ("proceso de paz" OR "acuerdo de paz" OR acuerdo OR agenda OR mesa OR di?logo* OR "mesa de conversacion" OR "mesa de conversación" OR conversaci?n* OR negociaci?n) AND (guerrill* OR eleno* OR heleno* OR exguerrillero* OR desmovilizado* OR FARC OR ELN) AND (bandido* OR genocida* OR sicario* OR mercenario* OR matar OR torturen OR asesino* OR mueran OR est?pid* OR mugrient* OR enfermo* OR pervertid* OR sucio* OR apestos* OR depredador* OR vomit* OR golpe OR golpes OR bajen OR pistola OR merecer* OR apuñal* OR callen OR calla* OR criminal* OR sangre OR colgar OR plomo OR tiro OR fusil* OR cementerio* OR sepultur*)
    - Periodo: 10 de noviembre - 26 de noviembre de 2022.

- Caso del feminicidio de Valentina Trespalacios
    - ("Valentina Trespalacios" OR (DJ near Valentina) OR (DJ near Trespalacios)) AND (feminicidio OR asesin* OR homicidio OR muert? OR crimen OR caso) AND (perra OR zorra OR prostitu* OR puta OR cachos OR infiel OR prepago OR infidelidad OR cuernos OR amante OR mozo OR moso OR vagabunda OR bandida OR buscona OR aprovechada OR vividora OR ingenua OR mantenida OR marrano OR interesada OR provocadora OR provocativa OR boba OR tonta OR bruta OR est?pida OR "se lo busc?" OR "por andar buscando" OR "algo habr? hecho" OR "eso le pas?" OR "plata f?cil" OR "dinero f?cil" OR "qui?n la manda")
    - Periodo: 20 de enero - 7 de febrero de 2023

- Caso del esquema de seguridad de Francia Márquez
    - (helic?ptero OR elic?ptero OR elicotero OR "black hawk" OR "fuerzas armadas" OR viaje OR viajes OR traslados OR dapa OR residencia OR reside OR casa OR lujosa OR vivienda OR seguridad OR familia OR “plata de los colombianos” OR “de malas” OR demalas OR indignante) AND ("Francia Márquez" OR "Francia Marquez" OR vicepresidenta OR vice) AND (negra OR negro* OR simio* OR animal OR malparida OR piroba OR malnacida OR ladrona OR corrupta OR hijueputa OR viva OR atraca* OR roba* OR tiro OR plomo OR bala OR muert* OR muera* OR sacar OR saquen OR merece*) AND metaData.source.socialOriginType:“twitter”
    - Periodo: 20 de febrero - 24 de marzo de 2023