EMONITOR+

INFORME ESPECIAL

VIOLENCIA DIGITAL CONTRA PERIODISTAS Y MEDIOS DE COMUNICACIÓN

ANÁLISIS DEL
01.12.2022 – 31.03.2023

Autores:
Álvaro Beltrán y Alisson Ramírez.

Edición a cargo de:
Salvador Candia, Sally Jabiel, Daniella Toce y David Hidalgo.

ÍNDICE DEL DOCUMENTO

1. Resumen ejecutivo

La violencia digital contra periodistas y medios de comunicación, ejercida principalmente a través de lo que en el ámbito internacional se ha definido como “comunicación tóxica” y “discursos de odio”, tiene consecuencias profundas en la vida de quienes ejercen esta profesión y su derecho a la libertad de expresión.

Las violencias digitales tienen consecuencias más allá de las pantallas. Diversos estudios demuestran que las víctimas no solo experimentan dificultades psicológicas, como mayor miedo, desconfianza o ansiedad; sino que también sufren consecuencias fisiológicas, incluidos mayores índices de presión arterial y estrés. Aún más preocupante, existe una correlación entre la violencia digital que se ejerce selectivamente contra grupos poblacionales y el incremento de ataques físicos contra personas de la identidad afectada. Esto se debe a que la comunicación tóxica y los discursos de odio sirven como vectores que profundizan la polarización y la radicalización en la sociedad.

El Perú no está exento de esta dualidad. En tan solo los últimos meses, se ha visto cómo periodistas han sido forzadas a dejar su labor debido al sostenido acoso digital que sufrían. Asimismo, las plataformas digitales se han usado como medios de organización para violencias físicas, incluidas acciones de acoso a viviendas de periodistas.

Para enfrentar este problema hay que entenderlo. Este informe busca contribuir a cerrar la brecha de evidencia sobre las formas y objetivos de este tipo de ataques contra periodistas y medios de comunicación. A diferencia de otros informes de análisis de redes sociales, el equipo de investigación de eMonitor+ utilizó inteligencia artificial para procesar mayores cantidades de datos. Esto permitió analizar horizontes más amplios de la conversación digital, reduciendo los posibles sesgos políticos por parte del equipo humano.

1.1. Conceptos clave

Para los propósitos de este informe, se utilizaron los conceptos planteados en la Estrategia y Plan de Acción de las Naciones Unidas para la Lucha contra los Discursos de Odio (2019). Esta sugiere que:

Es un discurso de odio cualquier acto de comunicación, textual, verbal o visual, que ataque a una persona por características de su identidad que son difíciles o imposibles de cambiar, como puede ser su ascendencia racial o étnica; su género; su condición de migración; su afiliación ideológica o religiosa; entre otras.

Es comunicación tóxica cualquier discurso que usa formas violentas del lenguaje (insultos, amenazas, contenido profano, entre otros) para excluir, deslegitimar o separar de la conversación a individuos o grupos específicos.

Se entiende de estas definiciones que, para la argumentación de este informe, todo discurso de odio es en simultáneo un caso de comunicación tóxica, pero no toda comunicación tóxica es necesariamente un discurso de odio.

1.2. Muestra y metodología de análisis

EMONITOR+

eMonitor+ es un sistema digital que combina herramientas de captura automatizada, inteligencias artificiales y un equipo humano especializado en análisis del discurso para monitorear el potencial uso de comunicación tóxica y discursos de odio en la conversación política digital.

Este análisis se realiza a partir de 113,852 publicaciones difundidas por 1,942 cuentas pertenecientes a 1,166 personas o instituciones. Esta es una muestra no aleatoria que responde al modelo teórico de “centralidad de la influencia”. Éste evalúa las capacidades y relaciones que poseen ciertos actores en una determinada red social para afectar la percepción y conducta de otros actores. Para este informe, se definieron criterios de centralidad en relación a la posibilidad de afectar el tono y los focos de la conversación política digital. Alrededor del 62% de estas cuentas ha sido categorizado de “alcance nacional”, mientras que el 38% es de “alcance regional”. El presente estudio considera publicaciones realizadas

por las cuentas monitoreadas entre el 1 de diciembre de 2022 y el 31 de marzo de 2023 en Facebook, Twitter, Instagram y YouTube.

El desafío de trabajar con grandes cantidades de datos, como las mencionadas, está en el tiempo requerido para hacer viable un análisis certero. Es ahí donde eMonitor+ se vuelve una pieza diferencial en este trabajo. Esta tecnología inicia su proceso con "robots de captura" que realizan copias digitales de las publicaciones de las cuentas monitoreadas. Acto seguido, cuatro inteligencias artificiales realizan predicciones en rangos del 0 al 1 (negativo - positivo) sobre las posibilidades de que un texto contenga comunicación tóxica o discursos de odio. Las publicaciones con mayor porcentaje de inferencia positiva son revisadas por un equipo humano que no solo evalúa la predicción automatizada sobre la presencia de violencia digital, sino que también añade información cualitativa adicional para entender la publicación en contexto político más amplio.

Luego de este procesamiento de la muestra, 9,938 publicaciones (9% del total capturado) fueron identificadas como de alta preocupación por eMonitor+ y confirmadas por el equipo humano.

1.3. Principales hallazgos

1

Del 1 de diciembre de 2022 al 31 de marzo de 2023, 1,041 de los 9,938 casos de comunicación tóxica o discurso de odio detectados tuvieron como víctimas a periodistas o medios de comunicación. Más del 80% de agresiones contra este grupo se realizó a partir de febrero de 2023, lo que sugiere un escenario de animadversión creciente entre las cuentas de mayor influencia en la conversación política digital hacia quienes se dedican a la práctica periodística.

2

Los mayores picos de violencia digital contra periodistas reflejan un uso reactivo de este tipo de discursos. La relación entre las concentraciones de los datos con eventos disruptivos o enfrentamientos offline sugiere que la violencia es usada como respuesta a opiniones discordantes a la perspectiva del agente de la publicación. Esto reafirma conclusiones de informes anteriores de eMonitor+, que demuestran que un uso común de los discursos de odio y la violencia basada en género son las falacias ad hominem. Es decir, la búsqueda de deslegitimación a través del ataque al emisor del mensaje antes que a su contenido.

3

El 55.4% de las publicaciones violentas contra periodistas contenía discursos de odio, es decir, atacaba directamente la identidad de las personas afectadas. Esta cifra es 17 puntos porcentuales mayor que la presencia media de discursos de odio en la muestra completa de 9,939 publicaciones capturadas en el periodo analizado.

4

Las formas más extremas de lenguaje, en la forma de justificación o llamamientos a la violencia contra periodistas y medios de comunicación, se usaron en 24 ocasiones. Aunque esto representa un porcentaje menor de las publicaciones capturadas, no debemos subestimar el impacto que este tipo de publicaciones puede tener en sus víctimas.

5

Las cuentas que más han contribuido a la violencia digital contra periodistas y medios de comunicación son aquellas que eMonitor+ ha definido como "cuentas de influencia digital". Éstas no poseen atributos de influencia en espacios offline, sino que han consolidado audiencias relevantes en plataformas digitales. Tres de cada cuatro casos de violencia digital contra periodistas vino de este grupo, a pesar de representar solo el 15% de las cuentas monitoreadas. Los niveles crecientes de anonimidad real o funcional podrían empoderar a este tipo de usuarios en el uso de discursos tóxicos y de odio.

6

No es correcto pensar que estas cuentas están desvinculadas de otros grupos monitoreados. En muchas ocasiones, las formas más frontales de violencia son ejercidas por cuentas que gozan de mayores niveles de anonimato, pero son compartidas por cuentas políticas o de líderes de opinión. En estos casos, el segundo grupo sirve como plataforma de amplificación de estos discursos más radicales.

7

El análisis sugiere que los canales informales de información (77.8%); líderes de opinión con trayectoria política (59.7%) y las cuentas individuales con influencia digital (56.6%) son más propensos a usar formas más intensas del lenguaje contra periodistas y medios de comunicación. Es de particular interés de

este informe el caso de los canales informales de información, ya que a pesar de no auto-describirse como espacios periodísticos ni someterse a la ética de la profesión, han logrado consolidar audiencias importantes, ante las cuales ejercen un rol importante a nivel informativo. Esta información, sin embargo, llega mediada por un discurso altamente polarizador.

8

Se identificaron 100 casos de violencia basada en género (VBG) contra periodistas o medios de comunicación, es decir que aproximadamente uno de cada diez casos de violencia digital usó discursos que profundizaron la desigualdad de género.

9

Los datos confirman la evidencia de diversos estudios acerca de que periodistas y medios de comunicación son particularmente propensos a experimentar violencia basada en género. La VBG contra periodistas representa uno de cada 5 de los 474 casos de VBG identificados en la muestra total. Asimismo, este tipo de discurso fue usado de manera más recurrente contra la prensa que contra otros grupos de la muestra completa.

10

Se detectaron 71 casos de violencia basada en género contra mujeres, 13 contra hombres y 16 contra instituciones. Sin embargo, en todos los casos, la VBG ejercida contra periodistas profundizaba estigmas y estereotipos que ponían lo femenino en un espacio inferior. De ahí que, mientras que las mujeres recibieron ataques por su género usando estereotipos o juzgando su apariencia, los hombres recibieron ataques que cuestionan su orientación sexual o pretendían hacer burla de una supuesta masculinidad ejercida "de manera incorrecta".

11

Las principales formas de VBG empleadas durante el periodo de análisis fueron los insultos (30.6%), los estereotipos (23.7%), los comentarios sobre la apariencia física (12.7%) y el acoso político (11.2%). Lamentablemente, se confirman conclusiones realizadas en informes anteriores de eMonitor+, según las cuales la mayor parte de ataques de VBG busca excluir o deslegitimar a las víctimas.

2. Escenario de desarrollo y conceptos clave

El discurso académico ha cambiado alrededor del potencial democratizador de las plataformas digitales. Mientras que el "tecno-optimismo" (McGee et al. 2018:25) de la década pasada hablaba de estas herramientas como un 'espacio público' de mayor horizontalidad, hoy la evidencia demuestra que transferir el debate político y social a espacios digitales ha tenido impactos negativos en las democracias y los derechos humanos. Siguiendo las ideas de Roberts (2017), es posible leer esta dicotomía en el abordaje académico como una "tecno-centralidad" en búsqueda de soluciones técnicas y apolíticas para desafíos inherentemente políticos, como la desigualdad. Franks (2021:427) recoge ese sentimiento cuando sugiere que la analogía de la 'plaza pública', tantas veces adjudicada a las redes sociales, estaba fallada desde el inicio, ya que en ningún momento de la historia la plaza pública ha sido un "mercado libre de ideas", si no que sistémicamente ha sido un espacio de refuerzo del status quo y de violencias que han excluido y silenciado a diversos grupos.

Las violencias digitales sirven propósitos similares: Reforzar, excluir, silenciar. No obstante, las consecuencias de estos ataques no se mantienen dentro de las pantallas. Una amplia revisión de literatura por Stevens, Nurse & Arief (2020), sugiere un consenso académico de que la violencia digital produce niveles elevados de depresión, ansiedad, ideación suicida y ataques de pánico. Estudios como el de LaBarron et al. (2017) demuestran que esto también impacta la salud física, a través de niveles elevados de presión arterial y estrés. Franks (2021:444-5) lamenta que, como medida de protección, muchas de las víctimas de este tipo de violencia optan por la "autocensura". Sin embargo, tomando en consideración que las personas más expuestas a estas prácticas son quienes ya experimentaban vulnerabilidades en los espacios offline (Alorainy et al. 2018), esto sugiere un espacio digital profundamente excluyente.

Las personas que ejercen el periodismo viven las consecuencias de este escenario a diario. Por un lado, representan un grupo históricamente en tensión con otros grupos de poder, vulnerable a las violencias comunicacionales y físicas por su labor. Por el otro, la continuidad de su práctica depende de la exposición de sus voces, lo que hace que mecanismos de autocensura sean menos viables. Waisbord (2020:1037) teoriza que el incremento de violencia contra periodistas en línea responde a un fenómeno de "censura de masas", en el que la violencia digital se usa por "ciudadanos comunes" para forzar a la prensa a actuar según sus intereses. Sugiere, además, que este proceso ha sido profundizado en los últimos años por tres dinámicas: 1) El acceso fácil a periodistas a través de su exposición

en línea; 2) La consolidación de grupos digitales radicalizados, principalmente de extrema derecha; y 3) la demonización de la prensa por parte de fuerzas políticas populistas. En un estudio posterior realizado por Henrichsen & Shelton (2022), se sugiere que, aunque los y las periodistas pueden identificar, a grandes rasgos, el sesgo político de sus atacantes, existen serias brechas para entender el fenómeno a nivel social, en particular las motivaciones emocionales y objetivos pragmáticos de sus atacantes. Obermaier (2023) sugiere que el principal predictor de violencia contra periodistas es su postura política o línea editorial, y que aquellas personas que tienen rasgos vinculables a identidades marginalizadas (e.g.- raza, género, etc.) experimentan también discriminación basada en esas características.

Harlow, Wallace & Chueca-Chacón (2022:6) sugieren que, en Latinoamérica, la respuesta a este escenario "tiende a ser visto como un problema personal". Esto significa que son periodistas, en su condición de individuos, quienes deben identificar formas para enfrentar las violencias. Por ello, comentan los autores, recurren a soluciones enfocadas en las emociones que sienten, como la supresión de pensamientos negativos o la búsqueda de soporte en colegas; o a soluciones enfocadas en el problema, que suelen vincularse a la autocensura, la reducción de interacciones con su audiencia, o incluso la renuncia a la práctica periodística en su totalidad. En ambos casos, lo que demuestran los autores es que la violencia digital contra periodistas en la región tiene un impacto directo en el libre ejercicio de su derecho de expresión y pone en riesgo la libertad de prensa.

Para enfrentar este problema, el paso inicial es entenderlo. Es por ello que, en años recientes, la cantidad de estudios que buscan entender las violencias discursivas en redes sociales se han multiplicado. Dos términos son particularmente comunes en este cuerpo académico, la 'comunicación tóxica' y los 'discursos de odio'. Sin embargo, a pesar de su popularidad, los límites y especificidades de ambos conceptos no poseen consenso académico (Alkomah & Ma 2022). Gagliardone & Pohjonen (2022) comentan que esto presenta tres grandes retos para el estudio de la violencia digital: 1) A nivel conceptual, representa un campo atrapado en divergencias académicas, legales y sociales; 2) A nivel ético, calificar de 'tóxico' o 'de odio' el discurso de un determinado grupo social puede profundizar su exclusión y/o radicalización; 3) A nivel metodológico, diversas herramientas tecnológicas y modelos teóricos permitirán siempre una exploración sólo parcial de la realidad.

No obstante, la inercia no es una alternativa. Particularmente porque diversos estudios empiezan a demostrar una correlación entre la proliferación de comunicación tóxica y discursos de odio con el incremento de violencia física contra poblaciones vulnerables. Gallacher y Bright (2021) sugieren que esto sucede porque los discursos tóxicos y de odio sirven como vectores de radicalización a través de dos fenómenos: 1) La consolidación de sentimientos de pertenencia grupal, lo que incluye la definiciones de 'enemigos' comunes; 2) la deshumanización y

reducción de la significancia de la violencia ejercida contra quienes se consideran 'los otros'. Los autores llaman a este proceso "contagio del odio" (ibid:5), el cual es acelerado durante momentos de disrupción offline (Burnap et al. 2014).

El presente informe busca cerrar una brecha de información sobre las violencias digitales, en la forma de comunicación tóxica y discursos de odio, ejercida contra periodistas en el Perú. Para ello, usa las definiciones de la Estrategia y Plan de Acción de las Naciones Unidas para la Lucha contra los Discursos de Odio (2019), la cual sugiere que:

Es un discurso de odio cualquier acto de comunicación, textual, verbal o visual, que ataque a una persona por características de su identidad que son difíciles o imposibles de cambiar, como lo puede ser su ascendencia racial o étnica; su género; su condición de migración; su afiliación ideológica o religiosa; entre otras.

Es comunicación tóxica cualquier discurso que usa formas violentas del lenguaje (insultos, amenazas, contenido profano, entre otros) para excluir, deslegitimar o separar de la conversación a individuos o grupos específicos.

Se entiende de estas definiciones que, para la argumentación de este informe, todo discurso de odio es en simultáneo un caso de comunicación tóxica, pero no toda comunicación tóxica es necesariamente un discurso de odio.

Para abordar los desafíos metodológicos planteados por Gagliardone y Pohjonen (2022), este informe utiliza por primera vez en Perú un método tecnológico basado en inteligencia artificial: "eMonitor+". Este sistema está diseñado para potenciar la capacidad analítica de un equipo humano especializado en análisis del discurso y entrenado específicamente en el enfoque de Naciones Unidas sobre esta problemática.

Con estos nuevos datos, esperamos contribuir al mejor diseño de respuestas políticas e institucionales para proteger la libertad de prensa y reducir la polarización en el país.

3. Metodología

3.1. Revisión de la tecnología: eMonitor+

eMonitor+ es un sistema digital diseñado por el Programa de las Naciones Unidas para el Desarrollo (PNUD) que aprovecha modelos de inteligencia artificial (IA), y concretamente grandes modelos lingüísticos (LLMs por sus siglas en inglés), para realizar predicciones binarias y clasificación automatizada sobre la posibilidad de que un determinado texto evaluado contenga comunicación tóxica o discursos de odio. El modelo pone especial énfasis en clasificar las formas de estas expresiones violentas del lenguaje, para contribuir al diseño de estrategias especializadas a sus diversas manifestaciones.

Los LLMs son modelos computacionales diseñados para el "procesamiento natural del lenguaje". Es decir, estas tecnologías son capaces de interpretar textos a pesar de los usos no-ortodoxos del lenguaje, lo que incluye variaciones conscientes (expresiones regionales, jerga y otros) e inconscientes (errores de tipeo y otros). Esto se debe a que un LLM no opera a través de relaciones entre significantes y significados, sino más bien en patrones comunes entre las palabras usadas con ciertas intenciones. Para arribar a este tipo de análisis, los LLMs son entrenados con enormes cantidades de datos que han sido previamente categorizados por el equipo de desarrollo. El sonado término 'machine learning' se refiere al proceso guiado y autónomo a través del cual los LLMs identifican el proceso más eficiente para que los nuevos datos que reciban puedan ser categorizados según las instrucciones originales de sus desarrolladores, en el mayor número de casos posible.

El PNUD describe eMonitor+ como un 'sistema' o 'portafolio de herramientas digitales' porque utiliza diversos aplicativos; robots para el rastreo y extracción de datos; interfaces interactivas para facilitar la interacción con datos; y no solo uno, sino cuatro LLMs para el procesamiento de lenguaje. Tres de estos últimos fueron desarrollados por agentes externos, mientras que un cuarto LLM fue específicamente desarrollado para responder al concepto de 'discurso de odio' según la interpretación de la ONU. El conjunto de estos aplicativos le permite evaluar cuatro grupos de características:

Cuadro 1: Características y categorías evaluadas por los LLMs de eMonitor+	
Característica del texto	Categorías de evaluación
Sentimiento del texto	1) Positivo; 2) negativo
Formas de toxicidad	1) Toxicidad; 2) Toxicidad severa; 3) Ataques a la identidad; 4) Profanidad; 5) Insultos; 6) Amenazas
Intensidad de la toxicidad	1) Muy tóxico; 2) Tóxico; 3) Nada tóxico; 4) Difícil de estimar
Discurso de odio	1) Normal; 2) Ofensivo; 3) Abusivo; 4) Con formas de odio

Los modelos operativos de cada uno de estos LLM pueden ser consultados en el ANEXO 1.

eMonitor+ ha sido usado con éxito en países que experimentan altos grados de polarización, como Libia, Líbano y Túnez. En particular, ha contribuido a que instituciones públicas puedan mapear la extensión con la que actores políticos, medios de comunicación y otros actores con alta influencia han usado la comunicación tóxica y los discursos de odio. Hasta marzo de 2023, esta tecnología ha sido utilizada para analizar 2 millones de publicaciones. Cuando ha sido usada por organismos electorales, eMonitor+ ha servido como una herramienta de alerta temprana para detener campañas políticas basadas en la desinformación y la polarización. Perú es el primer país en Latinoamérica en usar eMonitor+.

3.2. Criterios de selección de la muestra

eMonitor+ identifica, captura y analiza datos que provienen de una muestra no aleatoria de cuentas en redes sociales, las cuales pertenecen a una serie de actores públicos y privados altamente influyentes. Uno de los principios clave en el diseño de esta muestra fue el de "centralidad de la influencia", el cual refiere a las capacidades y relaciones que tienen ciertas cuentas para influir en el discurso de otras en plataformas digitales. Al 31 de marzo, la muestra estaba compuesta por 1,942 cuentas que pertenecen a 1,166 personas o instituciones. 62% de éstas han sido categorizadas como de "alcance nacional", mientras que 38% tienen "alcance regional". El presente estudio considera 113,852 publicaciones compartidas en Facebook, Twitter, Instagram y Youtube por las cuentas de la muestra.

El estudio de la centralidad de la influencia data de la década de 1940 (Freeman 1978). A lo largo de esta larga historia, tres conceptos resultan claves para esta aproximación teórica (Hanneman & Riddle 2019):

Grados	Cercanía	Intermediación
La capacidad objetiva de un actor de proveer un recurso a otros.	Capacidad de un actor de llegar a otro de la forma más rápida posible.	Posición de un actor como conexión o barrera entre dos o más actores.

Dada la gran cantidad de datos producidos a través del uso de plataformas digitales, muchas investigaciones han usado este enfoque para entender cómo grados de centralidad, cercanía e intermediación pueden explicar la influencia en la comunicación online y los comportamientos offline. Nouh & Nurse (2015), por ejemplo, usaron esta aproximación para identificar las voces líderes en grupos de activismo digital. Los autores también usaron análisis de sentimiento para inferir si estos usuarios estaban contribuyendo a establecer ideas más constructivas o destructivas dentro de sus redes de influencia. Modelos similares han sido usados durante discusiones de políticas específicas (Pirim et al. 2022; Struweg 2020), lo que incluye apuntes interesantes sobre cómo cuentas falsas y bots son usadas para manipular e influenciar (Weng & Lin 2022). En una revisión de literatura extensa, Riquelme & González-Cantergiani (2016) demostraron que los predictores de influencia digital más efectivos eran los retweets, mientras que las interacciones y el número de seguidores también eran relevantes.

eMonitor+ es particularmente efectivo en la lectura de este tipo de métricas. Esto lo vuelve una herramienta idónea para el análisis de plataformas digitales desde un enfoque de centralidad de la influencia. Es por ello que el equipo de eMonitor+ realizó su muestreo considerando esta aproximación teórica. Esto no solo permite aprovechar la gran cantidad de datos disponibles (aunque no consolidados), sino también enfocar el análisis en cómo el discurso político es afectado por actores centrales en las redes evaluadas, antes que actores menores que pueden experimentar momentos de gran exposición, pero poco impacto a largo plazo.

Para los propósitos de su operación en Perú, eMonitor+ estableció grados de centralidad, cercanía e intermediación que pueden ser consultados en el ANEXO 2. Con ello, en septiembre de 2022, se diseñó una primera muestra no aleatoria de aproximadamente 500 cuentas públicas, construida bajo la supervisión del equipo país del PNUD y la red de medios Ama Llulla. Las cuentas fueron distribuidas en cinco categorías: 1) Cuentas políticas; 2) Medios de comunicación y periodistas; 3) Líderes de opinión; 4) Espacios informales de información; 5) Cuentas influyentes en la conversación digital. Entre septiembre y noviembre de 2022, esta muestra fue actualizada de manera bisemanal para reflejar las interacciones entre cuentas monitoreadas y no monitoreadas, y así establecer relaciones de gradualidad, centralidad e intermediación. La muestra se estabilizó hacia finales de 2022, arrojando un total de 1,942 cuentas que pertenecen a 1,166 personas o instituciones. El Cuadro 2 resume su distribución.

Cuadro 2: Resumen de la muestra de eMonitor+					
Categoría	Total[1]		Subtipo de cuenta	Subtotal[2]	
	Personas	Cuentas		Per	Cue
Cuentas políticas	487	857	Personas en cargos de representación	183	381
			Personas en altos cargos de confianza	17	23
			Candidatos/as o ex-candidatos/as	156	217
			Líderes de partidos políticos	14	38
			Cuentas de partidos políticos	41	90
			Cuentas de movimientos políticos	76	108
Medios de comunicación y periodistas	318	635	Televisión	51	119
			Radio	48	107
			Web/Digital	123	198
			Prensa escrita	40	121
			Periodista	56	90
			Otro	0	0
Líderes de opinión	75	132	Sector académico	17	34
			Sector político	44	77
			Sector artístico-cultural	6	10
			Sector profesional	8	11
Espacios informales	74	79	Espacios informales de información	74	79
Influencia	212	318	Cuentas individuales con influencia	149	165
			Cuentas de memes con influencia	63	74
Total	1,166	1,942			

[1] El total de las cuentas monitoreadas
[2] El subtotal por personas (per) y cuentas (cue) monitoreadas

3.3. Metodología de análisis

eMonitor+ utiliza robots de rastreo y extracción para capturar las publicaciones realizadas por las cuentas monitoreadas. Cada LLM realiza una predicción en el rango de 0 a 1 (negativo a positivo) sobre las posibilidades de que un texto contenga alguna de las categorías de toxicidad y odio que han sido entrenadas para evaluar. Cada categoría recibe una evaluación individual. Las publicaciones son categorizadas como de alta o baja preocupación, siguiendo límites permisibles de inferencia que se explican en el ANEXO 4. Estos límites hacen referencia a la posibilidad de contener formas más violentas de lenguaje. Todas las preocupaciones de alta preocupación son evaluadas por un equipo de monitores que no solo confirma el análisis automatizado a través de un modelo validado por la ONU, si no que también añade información cualitativa adicional para leer la publicación en relación al escenario político más amplio. En este periodo, 9,939 de alta preocupación fueron identificadas por eMonitor+ y validadas por el equipo humano. Esto representa 9% de la muestra total.

El trabajo del equipo humano tiene tres fases. Primero, se filtran todas las publicaciones que usan lenguaje violento pero no hacen referencia a un tema político (e.g.- deportes, entretenimiento). Segundo, se confirma que la inferencia automatizada responde a los conceptos de comunicación tóxica o discursos de odio acordados; 3) Proveer información cualitativa adicional para profundizar el análisis e identificar patrones. Esta revisión de datos se realiza de manera diaria en reuniones editoriales que reúnen al equipo de PNUD y Ama Llulla. Esta submuestra es el insumo principal para el análisis de eMonitor+, incluido este informe. Históricamente, menos del 10% de las inferencias de eMonitor+ son rechazadas como falsos positivos por el equipo humano.

Del total de 113,852 publicaciones capturadas como muestra total entre el 1ro de diciembre de 2022 y el 31 de marzo de 2023, 9,938 publicaciones fueron identificadas como de alta preocupación y validadas por el equipo humano. Esto representa el 9% de la muestra total.

Para los fines de este informe, el equipo de eMonitor+ trabajó también una sub-muestra referida a cuentas de medios de comunicación y periodistas sujetos de violencia digital. Esto significó un trabajo de revisión de cada uno de los objetos de violencia identificados en la muestra de 9,938 publicaciones de alta preocupación. Como resultado, se identificaron 105 sujetos de violencia digital asociados al trabajo de prensa. Como grupo de control, se trabajaron también sub-muestras referidas a las personas manifestantes y a miembros del Poder Ejecutivo.

3.4. Limitaciones

Límites conceptuales

Aunque la definición de 'discurso de odio' provista por la ONU brinda un marco amplio de trabajo, quienes lean este informe deben ser conscientes de las complejidades sociales y legales del uso de este término en diversos foros. En Perú, no existe una figura legal que defina los 'discursos de odio' per se. Por el contrario, existen una serie de leyes y políticas públicas que constituyen el marco nacional para identificar ciertas formas de violencia basadas en la identidad, como por ejemplo en relación a raza, género o religión. Otras identidades no gonzan los mismos estándares de protección, como es el caso de la orientación sexual o las identidades transgénero. eMonitor+ reconoce estos debates abiertos y busca contribuir a la discusión de política como un mapeo general del panorama, antes que como una herramienta para la persecución legal o el desprestigio político.

Límites de la muestra

eMonitor+ se enfoca en cuentas con alto potencial de influencia. Esto significa que, inevitablemente, ciertas conversaciones e incluso liderazgos pueden haber sido dejados fuera de la muestra. Para reducir la posibilidad de sesgos, la muestra inicial de eMonitor+ se construyó incluyendo al entonces Presidente Castillo y su consejo de Ministros; cada uno de los congresistas elegidos para el periodo 2021-2026; todos los partidos políticos con representación congresal; una lista de los medios nacionales y regionales más importantes, provista por Ama Llulla; cada candidato/a postulando a Gobiernos Regionales en diez regiones; y una lista inicial de influenciadores digitales construida en colaboración de PNUD y Ama Llulla. Sin embargo, a pesar de que la muestra actual es cuatro veces más grande que la inicial, tiene dos limitaciones percibibles:

1) En una mayor parte, produce conversaciones de alcance nacional antes que regional.

2) No registra lo suficiente partidos y movimientos políticos menos visibles en la conversación política.

Por ello, el análisis de eMonitor+ puede ser usado de manera más precisa como una fotografía parcial de la realidad política nacional. Futuros análisis pueden enfocarse en realidades regionales, locales o enfocadas en discusiones de política específicas, y eMonitor+ estará disponible para este trabajo, dada la eficiencia y costo-efectividad que ha demostrado hasta el momento.

Límites de captura

eMonitor+ usa robots de rastreo y extracción de datos que dependen de interfaces provistas por terceras partes. Aunque la aspiración es capturar el 100% de publicaciones de la muestra, estas interfaces fueron afectadas en diversos momentos del recojo de datos, lo que puede haber impactado en la muestra final.

Límites metodológicos

Aunque el equipo humano de eMonitor+ es un componente clave para minimizar la posibilidad de falsos positivos, las grandes cantidades de datos capturados en la muestra completa hace más difícil evaluar falsos negativos. Esto significa que, aunque ya hay un protocolo para asegurar que solo publicaciones avaladas por el equipo de monitores ingresen a la muestra de análisis, aún no se ha desarrollado una metodología sistemática para evitar que se queden fuera de la muestra las publicaciones equivocadamente designadas como de baja preocupación por las LLM. Por ello, aunque eMonitor+ reduce horas de trabajo manual de rastreo, extracción y análisis, no solo depende de publicaciones capturadas por robots. Por el contrario, tiene mecanismos que permiten la incorporación manual de publicaciones adicionales que el equipo considera no han sido representadas de manera correcta. Esto se discute en reuniones editoriales realizadas de manera diaria por el equipo humano.

4. Discusión

Entre el 01 de diciembre de 2022 y el 31 de marzo de 2023, eMonitor+ capturó 113,852 publicaciones de 1,942 cuentas que pertenecen a 1,166 personas o instituciones. Luego de ser analizadas por las inteligencias artificiales de eMonitor+, recibir verificación por parte del equipo de monitoreo humano y aplicar los filtros para extraer de la muestra cualquier publicación vinculada a temas no políticos, se identificaron 9,938 publicaciones con contenido tóxico o discursos de odio, aproximadamente 9% de la muestra total. De ellas, 1,041 usaban estos tipos de violencia digital contra cuentas de periodistas o medios de comunicación.

Cuadro 3: Publicaciones capturadas entre DIC/22 y MAR/23		
Total de publicaciones	**Publicaciones con presencia de discursos tóxicos o de odio**	**Casos de discurso tóxico o de odio contra periodistas/ medios**
113,852	**9,938**	**1,041**

4.1. Formas del discurso y evolución en el tiempo

Las palabras tóxicas o cargadas de odio utilizadas con mayor frecuencia vincularon a la actividad periodística con acciones ilícitas, en particular recibir dinero por parte de actores públicos y privados para defender intereses específicos. Este es el caso de términos como “prensa basura”, “mermeleros” o “sicarios mediáticos”. También se usó la afiliación ideológica, particularmente posturas de centro o centro-izquierda tildadas de “caviares”, como un argumento de deslegitimación. Sin embargo, como se explicará más adelante, las formas utilizadas en la violencia digital contra periodistas han presentado diversas tendencias durante los cuatro meses analizados.

Gráfico 1: Palabras tóxicas o de odio más utilizadas contra periodistas

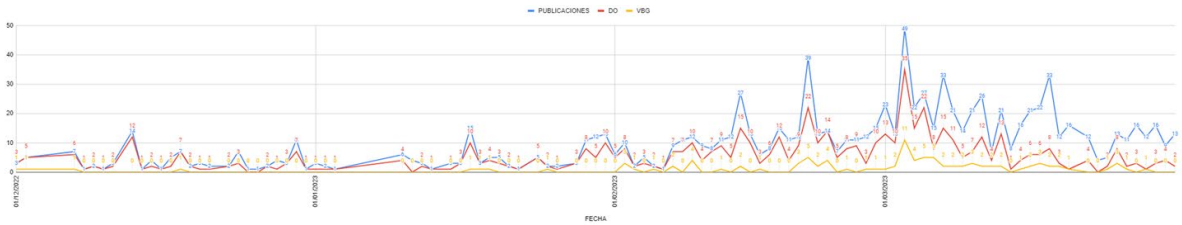
Elaboración: Propia sobre data de eMonitor+

Una mirada longitudinal a los datos arrojados (Gráfico 2) refleja dos momentos distintos en la forma en la que la muestra se expresaba en relación a periodistas y medios de comunicación:

1. Entre el 1 de diciembre y el 31 de enero, se detectaron 198 publicaciones contra periodistas y medios de comunicación. Esto representa el 19% del total de ataques contra este grupo detectado por eMonitor+, a un ratio de 3.2 publicaciones por día. La mayoría de las publicaciones capturadas usa el discurso tóxico o de odio para criticar un supuesto alineamiento entre el Gobierno de Dina Boluarte y los medios de comunicación, expresado a través del silencio ante hechos de violencia. De ahí que las principales expresiones utilizadas incluían menciones a "prensa basura", "mermelera" o "golpista". Antes que ataques contra medios o periodistas en concreto, la mayoría de estas publicaciones (81%) en estos dos meses atacan la práctica periodística en su conjunto.

2. Entre el 1 de febrero y el 31 de marzo, se detectaron 843 publicaciones contra periodistas o medios de comunicación, el 81% de la muestra total. Es de particular preocupación el rápido ascenso en el ratio de publicaciones por día, escalando de 3.2 a 14.2. En este periodo cambian también las formas del discurso en dos maneras: 1) Se hace mayor énfasis en la afiliación ideológica de periodistas como un elemento de deslegitimación o intento de exclusión de la conversación. Esto se refleja en el incremento de ataques que incluyen las palabras "caviar", "terrorista" y, hacia el final del periodo analizado, las alusiones a una "mafia caviar" asociada al expresidente Alejandro Toledo; 2) Se dejan las referencias abstractas a medios de comunicación para concentrar el discurso en ataques personalizados hacia periodistas, principalmente como respuesta a comentarios realizados por ellos o ellas.

Gráfico 2: Número de ataques digitales contra periodistas o medios por día



Elaboración: Propia sobre data de eMonitor+

Para entender el porqué de este quiebre entre los dos periodos, sirve recordar conclusiones presentadas en algunos informes anteriores de eMonitor+. En particular, el hecho de que la muestra monitoreada presenta dinámicas de interacción muy reactivas. En una gran mayoría de casos, los discursos tóxicos y de odio son utilizados como falacias ad hominem, es decir, intentos de deslegitimar o excluir de la conversación política a cualquier voz crítica a los intereses de quien emite el mensaje. Esto es más visible, como expresamos en la sección 2 de este informe, durante periodos disruptivos offline que incrementan la tensión online, y hace que la conversación digital sea más propensa al uso de formas más extremas de lenguaje.

Para probar este argumento, resulta esclarecedor revisar los tres picos de mayor concentración de violencia digital contra periodistas en el periodo analizado. Estos sucedieron el 13 de febrero, el 3 de marzo y el 17 de marzo. En las tres ocasiones, como se resume en el Cuadro 4, un hecho offline desencadena un momento de tensión, los periodistas objeto de violencia realizan una opinión que impacta a actores influyentes en la conversación digital y, subsecuentemente, se incrementa el uso de discursos tóxicos o de odio contra periodistas, como una forma de deslegitimar sus opiniones.

Cuadro 4: Dinámicas online/offline en picos de violencia contra periodistas

Fecha	Hecho disruptivo offline	Número de ataques[1]			Principales objetos de la violencia	Palabras más utilizadas
		TOT[2]	DO[3]	VBG[4]		
13/FEB	Fallecimiento de 7 policías en el VRAEM. Críticas emergen luego de comentarios sobre el uso laxo del término "terrorista" por grupos políticos de derecha.	50	29	03	Rosa María Palacios; Gustavo Gorriti; IDL Reporteros	Terroristas; terrorismo; rojos; terror; Sendero
13/MAR	Enfrentamiento durante entrevista en vivo entre la periodista Juliana Oxenford y la congresista Tania Ramírez.	84	50	17	Juliana Oxenford; Rosa María Palacios; César Hildebrandt	Caviares; sicaria mediática
17/MAR	La entrevista de René Gastelumendi al Alcalde de Lima, Rafael López Aliaga, hace que surjan críticas sobre la efectividad de su gestión.	76	20	07	Juliana Oxenford; Rosa María Palacios; Gustavo Gorriti; Josefina Townsend; América TV	Mercenaria; prostituta; ridícula; basura; sicaria

Elaboración: Propia sobre data de eMonitor+.

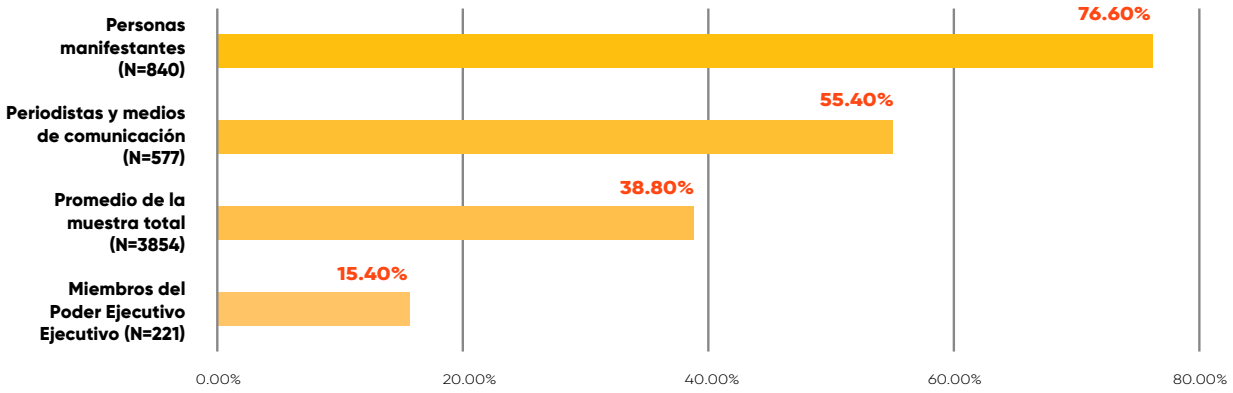
[1] El número de ataques representa el número de ataques en el día del hecho offline y los dos días consecutivos
[2] TOT = Total de publicaciones con discurso tóxico o de odio
[3] DO = Publicaciones con discurso de odio
[4] VBG = Publicaciones con violencia basada en género

Lo que los datos sugieren es que la relación entre la muestra de alta influencia en la conversación digital y la prensa exhibe una dinámica de animadversión ascendente. Aunque este escenario está altamente influenciado por hechos disruptivos offline, el discurso se mantiene latente y puede ser usado como falacia ad hominem cuando un nuevo caso enfrenta a cuentas influyentes y periodistas, dentro o fuera de plataformas digitales. Futuros estudios pueden evaluar la relación de esta dinámica con una profundización de la desconfianza en los medios de comunicación, particularmente en quienes tienen proximidad a las cuentas influyentes que lideran los ataques digitales.

4.2. Intensidad del discurso y contraste con otros grupos poblacionales

En el periodo analizado, se detectaron 577 mensajes con discurso de odio contra periodistas y medios de comunicación. Esto significa que el 55.4% de las publicaciones violentas contra personas en este grupo atacaban directamente su identidad. De manera comparativa, esta cifra es 17 puntos porcentuales mayor que la proporción media de discursos de odio encontrados en la muestra completa de 9,939 publicaciones del periodo analizado. Los datos también demuestran que la intensidad del discurso contra periodistas y medios es inferior al ejercido contra personas manifestantes, pero bastante superior a los usados contra miembros del Poder Ejecutivo. Esta comparación se visibiliza en el Gráfico 3.

Gráfico 3: Proporción de discursos de odio en relación con la muestra total



Elaboración: Propia sobre data de eMonitor+.
N = Número total de discursos de odio capturados en contra de cada grupo poblacional

Como se sugiere en la sección anterior, las principales palabras utilizadas en los discurso de odio incluyen la vinculación con hechos delictivos, particularmente a la recepción de dinero para orientar las líneas editoriales; la vinculación con el terrorismo; la descalificación según afiliación ideológica; y, como exploraremos más adelante, también violencia basada en género.

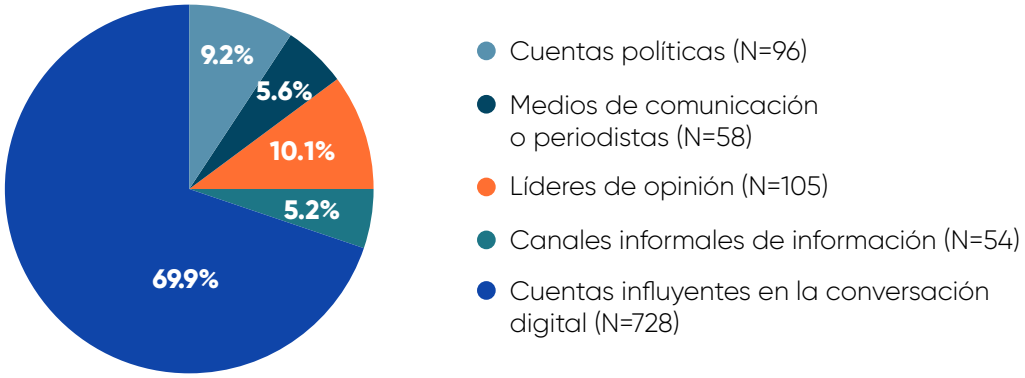
Las formas más extremas de lenguaje, en la forma de justificación o llamamientos a la violencia contra periodistas y medios de comunicación, se usaron en 24 ocasiones. Aunque esto representa un porcentaje menor de las publicaciones capturadas, no debemos subestimar el impacto que este tipo de publicaciones puede tener en sus víctimas. También cabe precisar que, debido a la visibilidad que tienen las cuentas de nuestra muestra, es posible que exista mayor aprehensión de usar lenguaje extremo. Otras herramientas digitales de análisis de redes sociales podrían complementar esta información para medir la virulencia de cuentas que, aunque menos influyentes, podrían reaccionar con mayor violencia dadas condiciones de mayor anonimato, real o funcional.

4.3. Agentes y objetos de la violencia digital

Las cuentas que más han contribuido a la violencia digital contra periodistas y medios de comunicación son aquellas que eMonitor+ ha definido como “cuentas de influencia digital”. Como se explica en la sección 3.2., éstas no poseen atributos de influencia en espacios offline, pero han logrado consolidar audiencias relevantes en plataformas digitales. Este tipo de cuentas produjeron aproximadamente tres de cada cuatro casos de violencia digital contra periodistas, a pesar de representar solo el 15% de las cuentas monitoreadas. Niveles mayores de anonimidad real o funcional podrían empoderar a este tipo de usuarios en el uso de discursos tóxicos y de odio.

El Gráfico 4 muestra el porcentaje de contribución a la muestra total de violencia digital, según las cinco categorías de cuentas definidas en la sección 3.2.

Gráfico 4: Contribución a la violencia digital por categoría de cuenta



Elaboración: Propia sobre data de eMonitor+.
N = Número total de publicaciones con discurso tóxico o de odio ejercido por cada categoría de cuenta

Es importante mencionar que, como se ha descrito en informes anteriores de eMonitor+, es incorrecto pensar en estas categorías de manera disociada. En muchas ocasiones, una publicación con niveles altos de violencia digital será compartida por cuentas políticas o por líderes de opinión, con la típica excusa de "Retweet no es endoso". Nuestro análisis anterior sugiere que esta práctica se realiza como una forma de separarse de las formas más extremas de discurso y evitar las posibles repercusiones. Así, mientras las cuentas más anónimas juegan un rol frontal en el ataque, las cuentas menos anónimas sirven un rol de plataforma o mediación con audiencias más grandes.

Como se describe en la sección 3.2. y en el ANEXO 2, la muestra de este informe está dividida en cinco categorías y 19 subtipos. Esto permite analizar también qué subtipos de cuentas usan de manera más recurrente formas de lenguaje más extremo, contribuyendo a una mayor polarización en la conversación digital. Para evaluar este rasgo, eMonitor+ contrastó el número de publicaciones que contienen discurso de odio con el total de publicaciones emitidas por cada subtipo de cuentas. La data sugiere que los medios informales de información (77.8%); los líderes de opinión por trayectoria política (59.7%) y las cuentas individuales con influencia digital (56.6%) fueron los grupos que usaron los discursos de odio en una mayor proporción a la media de la muestra total. Los ex candidatos y candidatas a gobiernos regionales y los movimientos políticos no registrados también muestran porcentajes de uso de discurso de odio superiores a la media, pero el número de publicaciones capturadas para estos subtipos es menor, por lo que la data no es lo suficientemente sólida. El desagregado de todos los subtipos se presenta en el Cuadro 5.

Cuadro 5: Intensidad de la violencia digital por subtipo de cuenta				
Categoría	Subtipo	TOT[1]	DO[2]	%[3]
Cuentas políticas	Personas en cargos de representación	45	14	31,82
	Personas en altos cargos de confianza	0	0	NA
	Candidatos/as o ex-candidatos/as	11	7	63,64
	Líderes de partidos políticos	26	13	52,00
	Cuentas de partidos políticos	0	0	NA
	Cuentas de movimientos políticos	16	15	93,75
Medios y periodistas	Televisión	2	0	0,00
	Radio	0	0	NA
	Web/Digital	16	6	54,55
	Prensa escrita	8	2	33,33
	Periodista	52	17	42,50
	Otro	0	0	NA
Líderes de opinión	Sector académico	0	0	NA
	Sector político	77	46	59,74
	Sector artístico-cultural	9	4	50,00
	Sector profesional	19	4	20,00
Informal	Canales informales de comunicación	66	42	77,8
Influencia digital	Cuentas individuales (real o parodia) con influencia digital	641	391	56,58
	Cuentas de memes	37	16	43,24
Total	Agregado total de la muestra	1025	577	55.4

Elaboración: Propia sobre data de eMonitor+.

[1] TOT = Total de publicaciones con discurso tóxico o de odio publicadas por un subtipo de cuenta

[2] DO = Total de publicaciones con discurso de odio publicadas por un subtipo de cuenta

[3] %= Porcentaje de publicaciones de un subtipo de cuenta que contiene discurso de odio

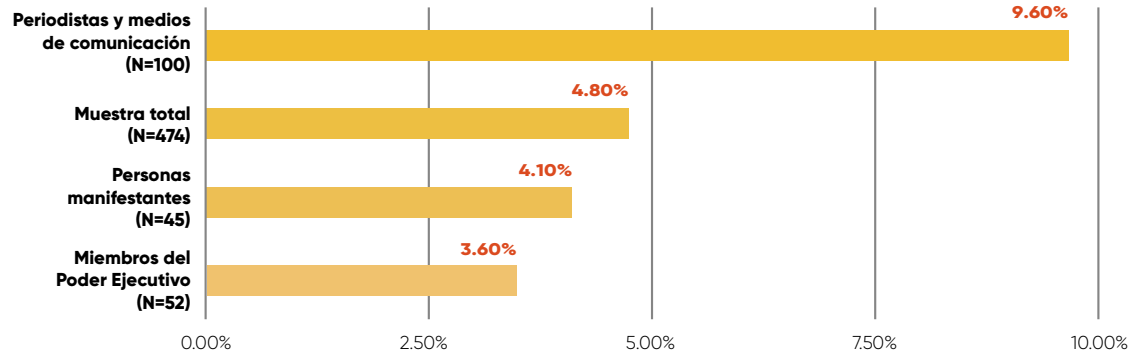
Requiere particular atención el caso de los canales informales de comunicación. Como se describe en la sección 3.2., se incluyen dentro de esta categoría espacios digitales cuyo contenido es principalmente informativo y han logrado establecer audiencias de al menos 5,000 seguidores en alguna red social. Sin embargo, a diferencia de las cuentas de medios digitales mapeadas en otra categoría, estas cuentas no se autodefinen como periodísticas ni se rigen por los estándares de la ética periodística. Sin embargo, la forma de su contenido y la relevancia para sus audiencias representan un alto potencial de polarización, ya que la información que comparten es mediada por discursos altos en comunicación tóxica y de odio.

4.4. Violencia basada en género contra periodistas

Diversos estudios prueban que las personas que ejercen el periodismo, en particular mujeres y personas LGBTI+, son particularmente vulnerables a la violencia basada en género (VBG) en redes sociales. La data capturada por eMonitor+ confirma esta afirmación. Se identificaron 100 casos de VBG contra periodistas o medios de comunicación en el periodo analizado. Esto quiere decir que aproximadamente uno de cada diez ataques digitales contra este grupo poblacional usó la desigualdad de género como un instrumento de violencia.

De hecho, uno de cada cinco casos de violencia basada en género encontrados en la muestra total tuvo como víctima a una persona que ejerce el periodismo. Asimismo, mientras que la proporción media de casos de VBG sobre la muestra total fue de 4.8%, el número de ataques cargados de violencia basada en género contra periodistas se elevó a 9.6%. La proporción de ataques con VBG sobre el total de publicaciones contra periodistas también fue proporcionalmente superior a la ejercida contra otros grupos poblacionales, como miembros del Poder Ejecutivo o personas manifestantes. El Gráfico 5 muestra el contraste entre los diferentes grupos analizados.

Gráfico 5: Proporción del uso de VBG sobre el total de publicaciones



Elaboración: Propia sobre data de eMonitor+.
N = Número total de casos de VBG capturados en contra de cada grupo

Se detectaron 71 casos de violencia basada en género contra mujeres, 13 contra hombres y 16 contra instituciones. En todos los casos, este tipo de mensajes profundizaron la estigmatización y subyugación de lo femenino, en favor de lo masculino. Esto se refleja en las palabras utilizadas en cada uno de estos ataques. La VBG contra mujeres usaba alusiones a la apariencia física, la supuesta vinculación sentimental con políticos, y en líneas generales un tinte de desprecio a la opinión ejercida. Mientras tanto, la VBG contra hombres hacía referencias

a la homosexualidad o la incapacidad de ejercer una masculinidad hegemónica. El Gráfico 6 compara ambas series de palabras.

Gráfico 6: Palabras usadas en VBG contra mujeres y hombres periodistas

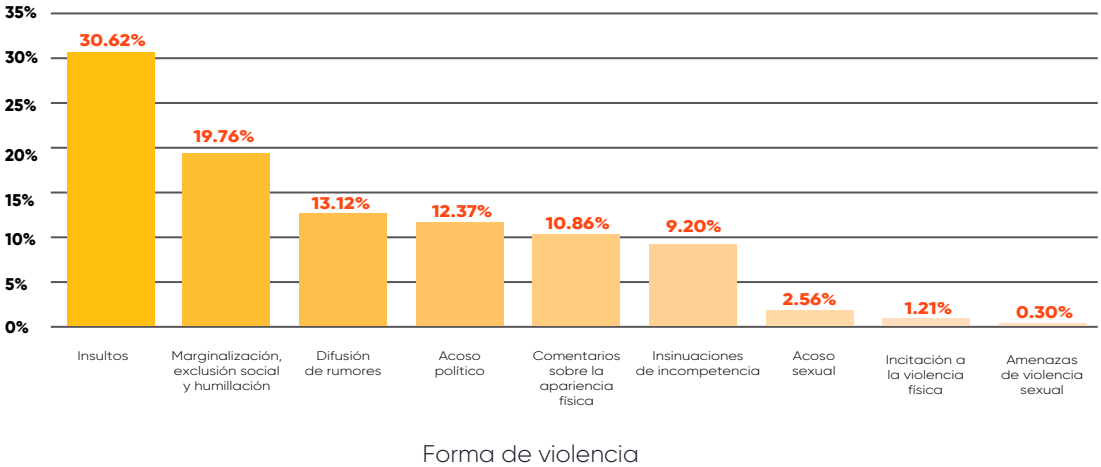
6.1. Palabras contra mujeres

6.2. Palabras contra hombres

Elaboración: Propia sobre data de eMonitor+.

Informes anteriores de eMonitor+ sugieren que las principales razones para el ejercicio de la violencia basada en género se relacionan a los intentos de excluir la opinión de las personas afectadas de la conversación política. Diversas formas de VBG son usadas para lograrlo, siendo los insultos (30.6%); los intentos de marginalización o estereotipificación (23.7%); los comentarios sobre la apariencia física (12.7%) y el acoso político (11.2%) las más comunes en el periodo analizado. El contraste entre esta y otras formas se presenta en el Gráfico 7.

Gráfico 7: Formas de VBG utilizadas durante el periodo analizado



Elaboración: Propia sobre data de eMonitor+.

ANEXO 1:
MODELOS OPERATIVOS DE LOS GRANDES MODELOS LINGÜÍSTICOS (LLMS) USADOS EN EMONITOR+

Para el procesamiento de la información, eMonitor+ utiliza tres modelos de lenguaje desarrollados por terceras partes y uno desarrollado directamente por el PNUD. Los tres primeros son:

BERT procesado a través de la plataforma de analítica KNIME:
BERT es un LLM diseñado por Google en 2018, uno de los primeros en el mundo en aplicar el 'modelo de transformador' aplicado al procesamiento natural de lenguaje. La contribución clave del modelo de transformador es su 'mecanismo de auto-atención', un modelo matemático que contribuye a una 'traducción' ágil de relaciones semánticas a pesos numéricos. Este proceso de 'codificación' de cada insumo incorporado al LLM se enfoca en la proximidad, similaridad y relación de sintaxis entre cada fragmento de dato (e.g.- cada palabra de una oración) en contraste con todos los otros fragmentos de datos presentes en el mismo insumo. Al resultado de este procesamiento numérico se le llama una 'representación', la cual no es un solo valor matemático, sino una matriz de valores interconectados. Las LLM basadas en el modelo de transformador realizan contrastes entre un nuevo insumo y los patrones de matrices que fueron codificados durante su proceso de aprendizaje. eMonitor+ accede a las capacidades de BERT a través de un software abierto llamado KNIME. Su contribución al sistema de eMonitor+ es la evaluación de si un texto tiene sentimiento positivo o negativo.

Perspective API
Esta tecnología fue desarrollada por Jigsaw y el Equipo de Google para la Prevención del Abuso en 2017. Aunque originalmente utilizó otro modelo grande lingüístico, ahora utiliza BERT como su sistema operacional central, tanto para la codificación/decodificación de insumos como también la traducción en tiempo real entre 18 idiomas. La contribución principal de este modelo a eMonitor+ es su capacidad de predecir seis atributos de la comunicación tóxica. Estos son: 1) Toxicidad; 2) Toxicidad severa; 3) Ataques a la identidad; 4) Insultos; 5) Profanidad; y 6) Amenazas. La herramienta fue entrenada a través de millones de comentarios realizados en plataformas abiertas, como Wikipedia, y cada uno de ellos recibió categorización por parte de 3 a 10 evaluadores humanos. Este grupo crowdsource definió los parámetros iniciales del proceso de aprendizaje automatizado. Dos auditorías adicionales se realizaron para reducir los sesgos de la categorización inicial. Esto fue particularmente relevante en relación a términos de identidad, los cuales inicialmente tenían mayor probabilidad de ser detectados como tóxicos, a pesar de ser usados en contextos no tóxicos.

Detoxify

Esta herramienta, desarrollada por Unitary en 2018, usa un modelo de transformador no-basado en BERT. A diferencia de los dos LLMs descritos arriba, Detoxify fue entrenado sobre la base de bases de datos abiertas disponibles en Kaggle, una comunidad abierta de desarrollo. Aunque su data es más limitada que BERT, Detoxify provee una perspectiva distinta que se basa en un proceso de inferencia distinto. He ahí su principal contribución a eMonitor+, sirviendo como una confirmación secundaria sobre si una publicación es 1) Muy tóxica; 2) Tóxica; 3) No tóxica; o 4) Difícil de estimar. En 2018, 2019 y 2020, Detoxify tuvo uno de los mejores desempeños en concursos organizados por Jigsaw para desarrollar rutas alternativas para optimizar LLMs para la identificación de comunicación tóxica y discursos de odio.

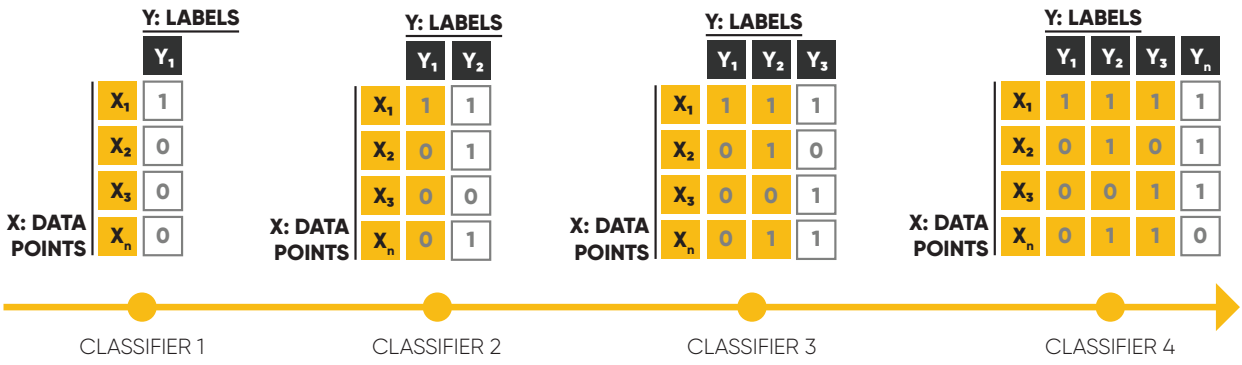
Cuadro 6: LLMs usados en eMonitor+ que fueron desarrollados por terceras partes

Nombre del LLM	Desarrollo	Procesador	Contribución a eMonitor+
BERT operado a través de KNIME	Google; KNIME	Modelo de transformador BERT	Análisis de sentimiento
Perspective API	Google; Jigsaw	Modelo de transformador BERT	Seis atributos: Toxicidad; toxicidad severa; ataque a la identidad; insulto; profanidad; amenaza
Detoxify	Unitary	Modelo de transformador no-BERT	Confirmación secundaria de toxicidad

Elaboración: Propia

La pieza final y más importante de eMonitor+ es un LLM desarrollado por el PNUD llamado “Analizador de Discurso de Odio”. Como se define en la sección 2, las divergencias legales, sociales y políticas que rodean al concepto de ‘discurso de odio’ han hecho que muchos equipos de desarrollo tecnológico lo eviten como un concepto operacional en su diseño. Por ello, las herramientas construidas para mantener las comunidades seguras no suelen abordar las complejidades del discurso de odio y prefiere abordar la toxicidad en términos más amplios. No obstante, la Estrategia y Plan de Acción de la ONU para la Lucha contra los Discursos de Odio provee una definición que permite operativizar este concepto en el análisis de discurso. Por ello, PNUD es la primera agencia del Sistema ONU que lo usa como concepto central para el procesamiento natural de lenguaje.

Graphic 8: Modelo de clasificación en cadena usado en el "Analizador de Discursos de Odio"



Elaboración: Propia. Fuente: Selmi (np).

A diferencia de los tres LLMs desarrollados por terceras partes descritos en párrafos anteriores, el modelo de PNUD utiliza un método de clasificación en cadena (CC) para su proceso de aprendizaje. Las CCs son usadas para hacer la clasificación multi-categoría un proceso más ágil, deconstruyendo problemas complejos en una serie de inferencias individuales, lo que produce resultados binarios (positivo/negativo). Para producir una estimación final, los clasificadores de cadena no se detienen en la inferencia fragmentada, si no que contrastan cada nivel de toma de decisión en contraste con los otros. Esto ayuda a evitar la pérdida de información en ninguno de los procesos de análisis. Luego de este proceso, se generan inferencias finales. El modelo se visualiza en el Gráfico 8.

El Analizador de Discurso de Odio del PNUD utiliza cuatro categorías para clasificar los insumos ingresados. Estos son 1) Normal; 2) Ofensivo; 3) Abusivo y 4) Con contenido de odio. El modelo fue entrenado usando una base de datos de 75,000 entradas individuales y recibió supervisión de equipos de la ONU para asegurar una codificación, decodificación y categorización adecuadas. Al igual que con los modelos de transformador, cualquier nuevo insumo añadido a la plataforma para análisis es contrastado con los patrones aprendidos por el LLM para su correcta clasificación.

ANEXO 2:
ATRIBUTOS DE CENTRALIDAD USADOS EN LA CONSTRUCCIÓN DE LA MUESTRA DE EMONITOR+

Al inicio de la operación de eMonitor+ en Perú, no existían bases de datos disponibles lo suficientemente amplias como para realizar un análisis certero de centralidad de la influencia. Es por ello que los primeros seis meses del proyecto (desde Septiembre 2022) han estado enfocados en la recolección de esta información. Siguiendo el enfoque teórico de centralidad de la influencia, el equipo estableció características para identificar qué cuentas deberían ser sujeto de análisis. La lógica macro fue la de predecir la posibilidad de un usuario de influir la conversación política digital. Cada atributo se puede revisar en el cuadro 7:

Cuadro 7: Atributos relevantes de centralidad, online y offline, de la muestra de eMonitor+

Categoría	Offline	Online
Grado	- Capacidad legal para definir y rechazar políticas públicas. - Capacidad legal para proveer, mejorar o detener servicios públicos.	- Provee actualizaciones o estados relevantes y de primera mano sobre asuntos relacionados a políticas públicas.
Proximidad	- Provee asesoría relevante o influencia en el diseño de políticas. - Tiene una relación relevante y formal con grupos amplios de personas. Por ejemplo, en la forma de membresías, militancias, alianzas y otros tipos.	- Su contenido digital es compartido por actores que han sido previamente establecidos como centrales en el sistema. - Tienen al menos 5,000 seguidores en una red social (número no agregativo entre diferentes sitios).
Intermediación	- Tiene control editorial sobre redes de telecomunicación masiva, lo que le brinda capacidad de brindar o no acceso a audiencias amplias.	- Ha establecido una audiencia cautiva. Esta relación de fidelidad se expresa a través de niveles altos de interacción por parte de quienes consumen su contenido.

Siguiendo los atributos online y offline descritos arriba, el equipo de eMonitor+ propuso hipotéticamente cinco categorías que serían sujeto de niveles elevados de centralidad, las cuales fueron usadas para desarrollar el mapeo inicial de cuentas para la muestra. Estas son:

- **Cuentas políticas**
  Cuentan con grados de centralidad ya que tienen el mandato legal de desarrollar políticas y distribuir servicios públicos. Experimentan relaciones de proximidad con otros tomadores de decisión y acceso a bases políticas a través de mecanismos de partido. En línea, pueden proveer actualizaciones de primera mano sobre asuntos vinculados a política. Estas cuentas son relevantes incluso cuando su presencia digital no está del todo desarrollada. En este grupo se incluyen personas elegidas para cargos públicos; partidos y movimientos políticos; así como los líderes de éstos.

- **Medios de comunicación y periodistas**
  Los medios y sus colaboradores pueden influenciar la toma de decisión a través del comentario y la veeduría. De manera clave, los medios masivos tienen un inmenso poder de decisión sobre las voces que accederán a sus plataformas y alcanzarán audiencias masivas. En línea y fuera de ella, pueden proveer actualizaciones de primera mano producto de investigaciones periodísticas. Su modelo se basa en incrementar niveles de alcance e interacción, por lo que experimentan niveles altos de proximidad e intermediación con el público general. Esta categoría incluye todos los medios de comunicación y quienes trabajan en ellos, incluidos los medios digitales, siempre y cuando se autorreconozcan como espacio periodístico.

- **Líderes de opinión**
  Las cuentas en esta categoría gozan de niveles adicionales de influencia en el diseño de política debido a su trayectoria académica, profesional, artística o política. Esto les da un nivel de proximidad clave offline. eMonitor+ mapea a líderes de opinión que además han sido capaces de generar grados, proximidad e intermediación de centralidad online.

- **Canales informales de información**
  Las cuentas en esta categoría no poseen atributos de centralidad offline. Sin embargo, sirven como una fuente de información sobre temas vinculados a política para una audiencia considerable de al menos 5,000 seguidores, los cuales producen altos niveles de interacción. Las cuentas más importantes de esta categoría, además, han conseguido que su contenido sea compartido o apreciado por personas o instituciones que poseen atributos de centralidad offline.

- **Cuentas con influencia digital**
  En esta categoría se incluyen cuentas que no poseen atributos de centralidad offline. Estas son cuentas no-informativas, ya que la forma de su contenido no es la de actualizaciones de noticias o similares. Sin embargo, sí realizan comentarios sobre temas vinculados a políticas y, en líneas generales, sobre el escenario político. También han logrado consolidar una audiencia de al menos 5,000 seguidores, quienes producen niveles altos de interacción. Se incluyen en esta categoría cuentas de memes y shitposting (enfocadas en comentarios provocadores), así como cuentas individualizadas que interactúan como si fuesen personas, sean reales o no.

Una muestra no aleatoria de 500 cuentas fue desarrollada siguiendo estas categorías y atributos. El proceso fue liderado por el equipo país del PNUD y la red de verificación de noticias "Ama Llulla". Sin embargo, este acercamiento no científico sólo tenía la intención de servir como un punto de partida para la captura de datos. Entre septiembre de 2022 y noviembre de 2022, la muestra fue actualizada de manera bisemanal para reflejar las interacciones entre las cuentas de la muestra y las cuentas externas. Cada nueva interacción era evaluada por sus atributos de centralidad online y offline. Las cuentas fueron incluidas en la muestra si:

1) Tenían cualquiera de los atributos de centralidad offline.

2) Poseían el grado de centralidad online.

3) Presentaban todos los atributos de proximidad e intermediación online.

ANEXO 3:

MUESTRA DETALLADA DE EMONITOR+

Tipo de cuenta	Personas	Cuentas	Descripción	Subtipo	Personas												Cuentas											
					NAC	MML	LIM	CJA	AQP	LOR	AMZ	CLL	CUZ	MOQ	PAS	PIU	NAC	MML	LIM	CJA	AQP	LOR	AMZ	CLL	CUZ	MOQ	PAS	PIU
Cuentas políticas	487	857	Personas que ejercen actualmente cargos públicos, en particular quienes fueron elegidos para cargos de representación.	Personas en cargos de representación	132	7	1	11	5	9	4	0	12	0	1	1	319	7	1	12	7	9	9	0	15	0	1	1
				Personas en altos cargos de confianza	11	0	0	0	0	4	0	0	0	2	0	0	15	0	0	0	0	6	0	0	0	2	0	0
				Candidatos/as o ex-candidatos/as	0	77	1	31	24	17	2	2	2	0	0	0	0	97	1	32	28	38	6	7	8	0	0	0
				Líderes de partidos políticos	9	0	0	0	1	0	0	0	0	0	2	2	32	0	0	0	1	1	0	0	0	0	2	2
				Cuentas de partidos políticos	14	0	1	1	5	10	2	2	1	2	1	2	36	0	1	1	6	25	5	7	3	3	1	2
				Cuentas de movimientos políticos	71	0	0	4	0	0	0	0	0	0	1	0	101	0	0	5	0	1	0	0	0	0	1	0
				Subtotal	237	84	3	47	35	40	8	4	15	4	5	5	503	104	3	50	42	80	20	14	26	5	5	5
Medios de comunicación	318	635	Espacios que se autorreconocen como espacios periodísticos y, en consecuencia, dicen ceñirse a la ética de la práctica	Televisión	13	0	1	12	8	4	6	1	3	0	2	1	38	0	1	36	14	4	10	5	8	0	2	1
				Radio	5	0	3	10	5	8	8	2	1	3	3	0	14	0	3	35	10	16	13	5	2	6	3	0
				Web/Digital	26	0	10	10	13	14	12	5	6	9	9	9	68	0	10	14	26	15	19	9	8	11	9	9
				Prensa escrita	20	0	0	1	3	3	2	2	4	1	2	2	67	0	0	2	11	8	7	7	14	1	2	2
				Periodista	56	0	0	0	0	0	0	0	0	0	0	0	90	0	0	0	0	0	0	0	0	0	0	0
				Otro	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
				Subtotal	120	0	14	33	29	29	28	10	14	13	16	12	277	0	14	87	61	43	49	26	32	18	16	12
Líderes de opinión	75	132	Personas que, debido a su trayectoria académica, profesional o política, son reconocidos como especialistas en su materia.	Sector académico	16	0	1	0	0	0	0	0	0	0	0	0	33	0	1	0	0	0	0	0	0	0	0	0
				Sector político	44	0	0	0	0	0	0	0	0	0	0	0	77	0	0	0	0	0	0	0	0	0	0	0
				Sector artístico-cultural	5	0	1	0	0	0	0	0	0	0	0	0	9	0	1	0	0	0	0	0	0	0	0	0
				Sector profesional	8	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0
				Subtotal	73	0	2	0	0	0	0	0	0	0	0	0	130	0	2	0	0	0	0	0	0	0	0	0
Canales informales de comunicación	74	79	Cuentas que, aunque no se reconocen como espacios periodísticos, utilizan y comentan información de coyuntura con un fin informativo	No aplica subtipo	73	0	0	0	0	0	0	0	0	0	1	0	78	0	0	0	0	0	0	0	0	0	1	0
Cuentas influyentes en la conversación digital	212	239	Cuentas que tienen más de 5000 seguidores y cubren, principalmente, temas políticos. Se incluyen cuentas de memes y parodias.	Cuentas individuales con influencia	147	0	2	0	0	0	0	0	0	0	0	0	163	0	2	0	0	0	0	0	0	0	0	0
				Cuentas de memes con influencia	63	0	0	0	0	0	0	0	0	0	0	0	74	0	0	0	0	0	0	0	0	0	0	0
				Subtotal	210	0	2	0	0	0	0	0	0	0	0	0	237	0	2	0	0	0	0	0	0	0	0	0
Subtotal personas / cuentas	1166	1942			713	84	21	80	64	69	36	14	29	17	22	17	1225	104	21	137	103	123	69	40	58	23	22	17

Leyenda: Alcance nacional (NAC); Municipalidad Metropolitana de Lima (MML); Lima Región (LIM); Cajamarca (CJA); Loreto (LOR); Amazonas (AMZ); Callao (CLL); Cuzco (CUZ); Moquegua (MOQ); Pasco (PAS); Piura (PIU). La columna "Cuentas" hace referencia al total de cuentas individuales monitoreadas, mientras que la columna "Personas" hace referencia al número de personas o instituciones que son dueñas de las cuentas. El número difiere debido a que una sola persona o institución puede usar más de una cuenta en distintas o la misma red social.

ANEXO 4:
LÍMITES PARA LA IDENTIFICACIÓN DE PUBLICACIONES DE ALTA O BAJA PREOCUPACIÓN EN EMONITOR+

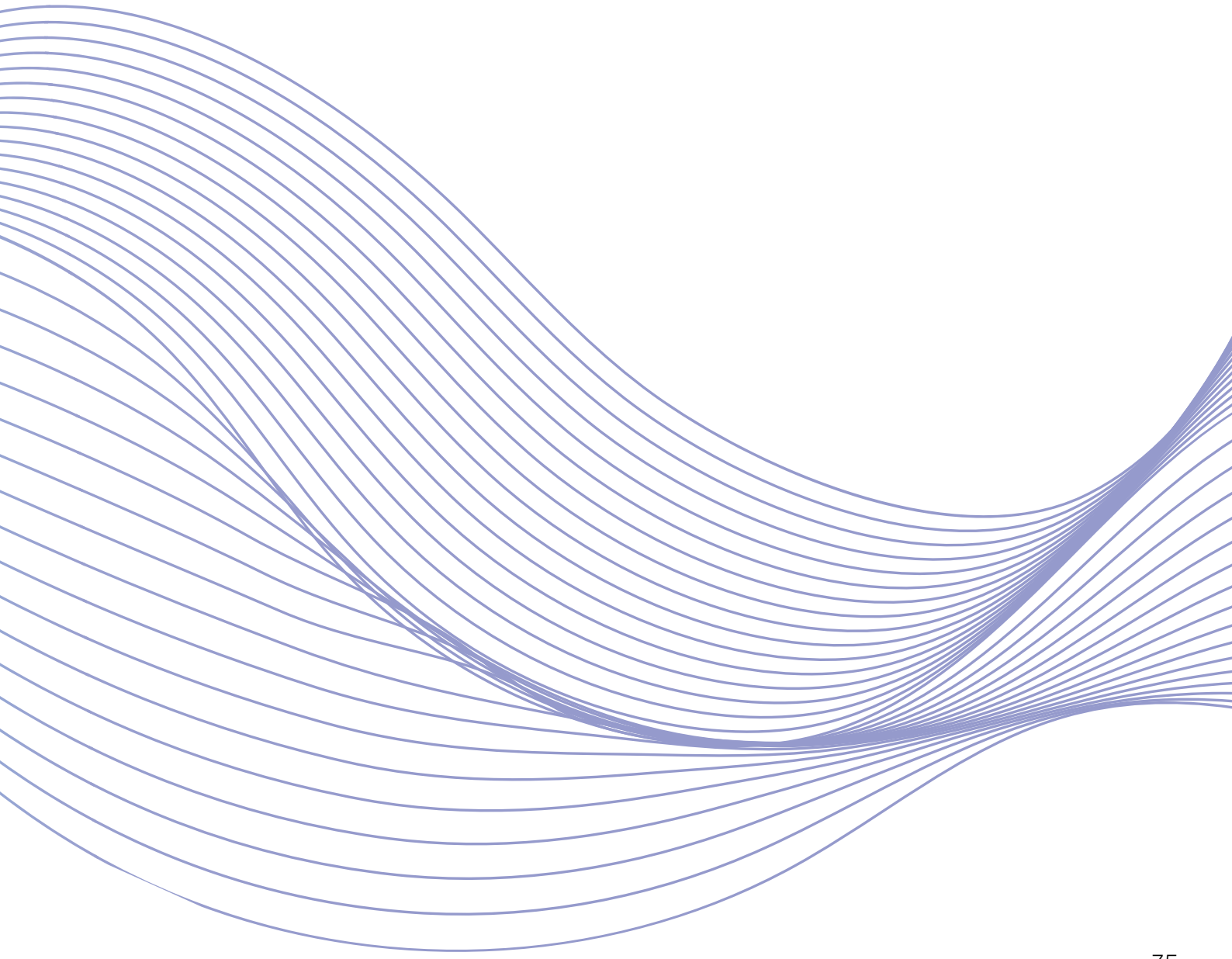
Para procesar la información, los insumos recibidos por eMonitor+ deben ser provistos en formatos simplificados de texto. Es por ello que eMonitor+ utiliza robots de rastreo y extracción de información, los cuales capturan, transforman y guardan las publicaciones realizadas en una base de datos centralizada. En ésta se registra todo el contenido escrito de una publicación, así como su metadata (información del usuario y de interacción), en valores simplificados de texto separados por comas (CSV por sus siglas en inglés). Este proceso se limita a la metadata que ya estaba disponible al momento de captura (por ejemplo, la descripción de una imagen en la metadata de una publicación). Actualmente, eMonitor+ no posee capacidades para la interpretación de imagen-a-texto o de audio-a-texto.

Posteriormente, cada publicación es analizada por los LLMs. La inferencia de cada modelo resulta en una predicción del 0 al 1. Un valor de 0 significa una inferencia absolutamente negativa, mientras que una inferencia de 1 significa una inferencia absolutamente positiva. El cuadro 8 muestra los límites que usa eMonitor+ para definir cuándo una publicación pasa de ser de baja preocupación a alta preocupación, lo que sugiere una mayor posibilidad de contener violencia digital.

Cuadro 8: Límites para la inferencia de violencia digital establecidos para eMonitor+

Modelo	Categoría evaluada	Límite establecido
BERT + KNIME	Sentimiento positivo	No considerado
	Sentimiento negativo	0.97
Perspective API	Toxicidad	0.4
	Toxicidad severa	0.1
	Ataque a la identidad	0.1
	Insulto	0.1
	Profanidad	0.1
	Amenaza	0.1

Detoxify	Toxicidad muy alta	0.1
	Toxicidad	0.4
	Difícil de estimar	No considerado
	Ausencia de toxicidad	No considerado
Hate Speech Analyzer	Normal	No considerado
	Ofensivo	0.4
	Abusivo	0.1
	Hateful	0.1



ANEXO 5:
MODELO DE EVALUACIÓN CUALITATIVA

El equipo humano de eMonitor+ sigue una metodología de análisis cualitativo que fue diseñada por el Buró del PNUD para los Países árabes, durante la implementación inicial de esta tecnología en Túnez y Líbano. El método utiliza tres cuestionarios compuestos por preguntas, en su mayoría de respuesta cerrada, que permiten entender las formas y objetivos de la comunicación tóxica y discursos de odio que han sido identificados por las LLMs y confirmados por el equipo humano.

Estos cuestionarios fueron desarrollados a través de tres grandes procesos:

- **Una revisión de literatura** en los temas claves, utilizando insumos de las Naciones Unidas y también investigaciones externas.
- **Consultas con equipos especializados del PNUD**, como lo fueron el equipo global de gobernanza, de prevención de conflictos electorales, y de igualdad de género.
- **Mesas de trabajo, grupos focales y encuestas** a través de la cual se recibió la opinión de más de 100 miembros de grupos de intereses, entre los que se incluyó funcionarios públicos, candidatos y candidatas a cargos públicos, organizaciones de la sociedad civil y otros organismos internacionales.

Cuadro 9: Cuestionarios para el recojo de datos cualitativos

Cuestionario	Pregunta	Respuestas admitidas
Información general	Tipo de publicación	Texto; imagen; audio; video; otro.
	Tipo de página o medio	Medio digital de comunicación; medio tradicional de comunicación; Facebook; Twitter; Youtube; Instagram; TikTok; Whatsapp u otro aplicativo de mensajería; otro.
	Origen de la publicación	Candidatos/as; partido político; autoridad elegida para un cargo de representación; funcionario/a público; líder de opinión; independiente; otro.
	Tema de la publicación	Electoral; política; económica; social; ciencia y salud; cultura, artes y entretenimiento; otro.
	¿La publicación promueve a un actor político?	Sí; No (Si la respuesta es positiva, la plataforma solicitará registrar el nombre).

Información general	¿La publicación promueve un partido político?	Sí; No (Si la respuesta es positiva, la plataforma solicitará registrar el nombre).
	¿A quién ataca la publicación?	Respuesta abierta.
	¿La publicación ataca a un actor político?	Sí; No (Si la respuesta es positiva, la plataforma solicitará registrar el nombre).
	¿La publicación ataca a un partido político?	Sí; No (Si la respuesta es positiva, la plataforma solicitará registrar el nombre).
	¿Contiene publicidad pagada?	Sí; No.
	Audiencia inferida	Respuesta abierta.
Comunicación tóxica y discurso de odio	Tipo de sentimiento	Positivo; negativo; neutral.
	¿Existe odio?	Sí; No.
	¿Cuál es la identidad afectada por el discurso de odio?	Género; ascendencia racial; ascendencia étnica; afiliación religiosa; afiliación ideológica; orientación sexual; capacidad corporal o mental; condición migratoria; clase socioeconómica; forma de trabajo; otra identidad estigmatizada.
	Intensidad del ataque	Nivel 1: Oposición prejuiciosa a la opinión; Nivel 2: Vinculación de una población con actividades negativas o ilegales; Nivel 3: Uso de lenguaje soez relacionado a la identidad; Nivel 4: Demonización o deshumanización de una población o grupo; Nivel 5: Llamamiento a la violencia física contra un individuo o grupo por su identidad; Nivel 6: Llamamiento al asesinato de un individuo o grupo por su identidad.
Violencia basada en género	¿La publicación fue escrita por una mujer o persona LGBTI+?	Sí, por una mujer; Sí, por una personas LGBTI+; No.
	¿La publicación habla de una mujer o una persona LGBTI+?	Sí; No.
	¿La publicación contiene violencia contra mujeres o personas LGBTI+?	Sí; No.
	¿A qué grupo pertenece la mujer o persona LGBTI+ atacada?	Una persona ciudadana o votante; una persona no ciudadana o no votante; alguien que apoya visiblemente una candidatura política; alguien que trabaja para una campaña política; un monitor/a del proceso electoral; miembros de mesa, personeros o personal de organismos electorales en un local de votación; alguien que trabaja para un partido político; alguien que trabaja como servidor público; la pareja de un actor político; alguien que trabaja para el sector privado; otro.

Violencia basada en género	Tipo de violencia basada en género	Psicológica (acoso; bullying; asesinato de carácter; insultos); sexual (acoso sexual, amenazas de violencia sexual, incluida la filtración de fotografías o videos íntimos); física (Amenazas o pruebas de violencia física o asesinato); estructural (estigmatización o refuerzo de estereotipos).
	Fuente de la violencia	Legal; social; económica; doctrinal.
	Técnica de la violencia	Llamamiento a la violencia física; amenazas o intimidación con el objetivo de forzar una retirada de la carrera electoral; acoso sexual; amenaza de violación; insinuación de incompetencia; ataque verbal; difusión de rumores; comentarios sobre la apariencia física; marginalización o estereotipificación; difusión no consensuada de imágenes o videos; otro.
	Herramientas de la violencia	Imágenes editadas; deepfakes (video); deepfakes (audio); ciberataques organizados; otro.
	¿La publicación contiene imágenes que estereotipan a mujeres o personas LGBTI+?	Sí; No.

REFERENCIAS

- Alkomah, F. & Ma, X. (2022) A literature review of textual hate speech detection methods and datasets. Information 2022, 13, 273. https://doi.org/10.3390/info13060273

- Alorainy, W.; Burnap, P.; Liu, H. & Williams, M. (2018). The enemy among us: Detecting hate speech with threats based "othering" language embeddings. ACM Transactions on the Web, 9(4), 1-26. Consultado el 01/05/2023 en http://arxiv.org/abs/1801.07495

- Burnap, P.; Williams, M. L.; Sloan, L.; Rana, O.; Housley, W.; Edwards, A.; Knight, V.; Morgan, J.; Procter, R. & Voss, A. (2014). Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack. Social Network Analysis and Mining, 4(1), 1-14. https://doi.org/10.1007/s13278-014-0206-4

- Franks, M.A. (2021) Beyond the public square: imagining digital democracy. The Yale Law Journal. Vol 131. pp. 427-453. Consultado el 27/04/2023 en https://www.yalelawjournal.org/forum/beyond-the-public-square-imagining-digital-democracy

- Freeman, L.C. (1978) Centrality in social networks conceptual clarification. Social Networks. Volume 1, Issue 3, 1978-1979. pp. 215-239. https://doi.org/10.1016/0378-8733(78)90021-7

- Gagliardone, I., & Pohjonen, M., Orton-Johnson, K. (Ed.), (2022). How to analyze online hate speech and toxic communication [How-to Guide]. Sage Research Methods: Doing Research Online. https://doi.org/10.4135/9781529609721

- Gallacher, J. & Bright, J. (2021) Hate contagion: Measuring the spread and trajectory of hate on social media. PsyArXiv. 16 February 21. https://doi.org/10.31234/osf.io/b9qhd

- Hanneman, R.A. & Riddle, M (2005) Introduction to social network methods. Riverside, CA: University of California, Riverside (Publicado en http://faculty.ucr.edu/~hanneman/)

- Harlow, S.; Wallace, R. & Cueva Chacón, L. (2022): Digital (In)Security in Latin America: The dimensions of social media violence against the press and journalists' Coping Strategies. Digital Journalism, DOI: 10.1080/21670811.2022.2128390

- Henrichsen, J.R. & Shelton, M. (2022): Expanding the analytical boundaries of mob censorship: How technology and infrastructure enable novel threats to journalists and strategies for mitigation. Digital Journalism, DOI: 10.1080/21670811.2022.2112520

- LaBarron, K.H.; Hoggard, L.S.; Richmond, A.S.; Gray, D.L.; Williams, D.P. & Thayer, J.F. (2017) Examining the association between perceived discrimination and heart rate variability in African Americans. Cultural diversity & ethnic minority psychology. Jan;23(1):5-14. doi: 10.1037/cdp0000076.

- McGee, R. con Edwards, D.; Anderson, C.; Hudson, H. & Feruglio, F. (2018) Appropriating technology for accountability: messages from Making All Voices Count. Making All Voices Count Research Report, Brighton: IDS

- Naciones Unidas – ONU (2019) Estrategia y Plan de Acción de las Naciones Unidas para la Lucha contra el Discurso de Odio. Consultado el 01/02/2023 en https://www.un.org/en/genocideprevention/documents/advising-and-mobilizing/Action_plan_on_hate_speech_ES.pdf

- M. Nouh & J. R. C. Nurse (2015) Identifying Key-Players in Online Activist Groups on the Facebook Social Network. 2015 IEEE International Conference on Data Mining Workshop (ICDMW), Atlantic City, NJ, USA, 2015, pp. 969-978, doi: 10.1109/ICDMW.2015.88.

- Obermaier, M. (2023): Occupational Hazards: Individual and Professional Factors of Why Journalists Become Victims of Online Hate Speech. Journalism Studies, DOI: 10.1080/1461670X.2023.2173955

- Pirim, H.; Nagahi, M.; Larif, O.; Nagahisarchoghaei, M. & Jaradat, R. (2023) Integrated twitter análisis to distinguish systems thinkers at various levels: a case study of COVID-19. Applied Network Science (2023) 8:12 https://doi.org/10.1007/s41109-022-00520-9

- Riquelme, F. & Gonzalez-Cantergiani, P. (2016) Measuring user influence on twitter: A survey, Information Processing and Management, 52 (2016) 949–975.

- Roberts, T. (2017) Digital technology excludes. Making All Voices Count Blog, Brighton: IDS.

- Selmi, G. (No publicado) Hate Speech Internal Analyzer. Documento preparado para explicar funcionamiento del Analizador de Discurso de Odio a audiencias internas del PNUD.

- Stevens, F.; Nurse, J.R.C & Arief, B. (2020) Cyber Stalking, Cyber Harassment and Adult Mental Health: A Systematic Review. Cyberpsychology, Behavior, and Social Networking. ISSN 2152-2715.

- Struweg, I. (2020) A Twitter Social Network Analysis: The South African Health Insurance Bill Case. Responsible Design, Implementation and Use of Information and Communication Technology. 2020 Mar 10; 12067: 120-132. doi: 10.1007/978-3-030-45002-1_11

- Waisbord, S. (2020) *Mob Censorship: Online Harassment of US Journalists in Times of Digital Hate and Populism*. Digital Journalism, 8:8, 1030-1046, DOI: 10.1080/21670811.2020.1818111

- Weng, Z.; Lin, A. (2022) *Public Opinion Manipulation on Social Media: Social Network Analysis of Twitter Bots during the COVID-19 Pandemic*. International Journal of Environmental Research and Public Health 2022, 19, 16376. https://doi.org/10.3390/ijerph192416376

Con el impulso de

AMA LUCLA

TRANSPARENCIA

PNUD