Hatemedia (PID2020-114584GB-I00)

EL ODIO QUE PERMANECE

Moderación digital y la pervivencia del discurso hostil en los perfiles sociales de medios informativos.

Delete

Financiado por

MINISTERIO DE CIENCIA E INNOVACIÓN

AGENCIA ESTATAL DE INVESTIGACIÓN

Hatemedia

Proyecto PID2020-114584GB-I00
https://www.hatemedia.es/

Editores
Elias Said-Hung
Julio Montero-Díaz
Jacobo Herrero Izquierdo

Diseño y maquetación:
Jacobo Herrero Izquierdo

Este informe contó con la colaboración del grupo de investigación SIMI: Inclusión socioeducativa e intercultural, Sociedad y Medios de UNIR.
https://gruposinvestigacion.unir.net/simi/

Como citar este informe:
Said Hung, E., Montero-Díaz, J., & Herrero Izquierdo, J. (2025). El odio que permanece. Informe sobre la moderación digital y la pervivencia del discurso hostil en los perfiles sociales de medios informativos. Hatemedia Project. https://doi.org/10.5281/zenodo.15595651

Financiado por

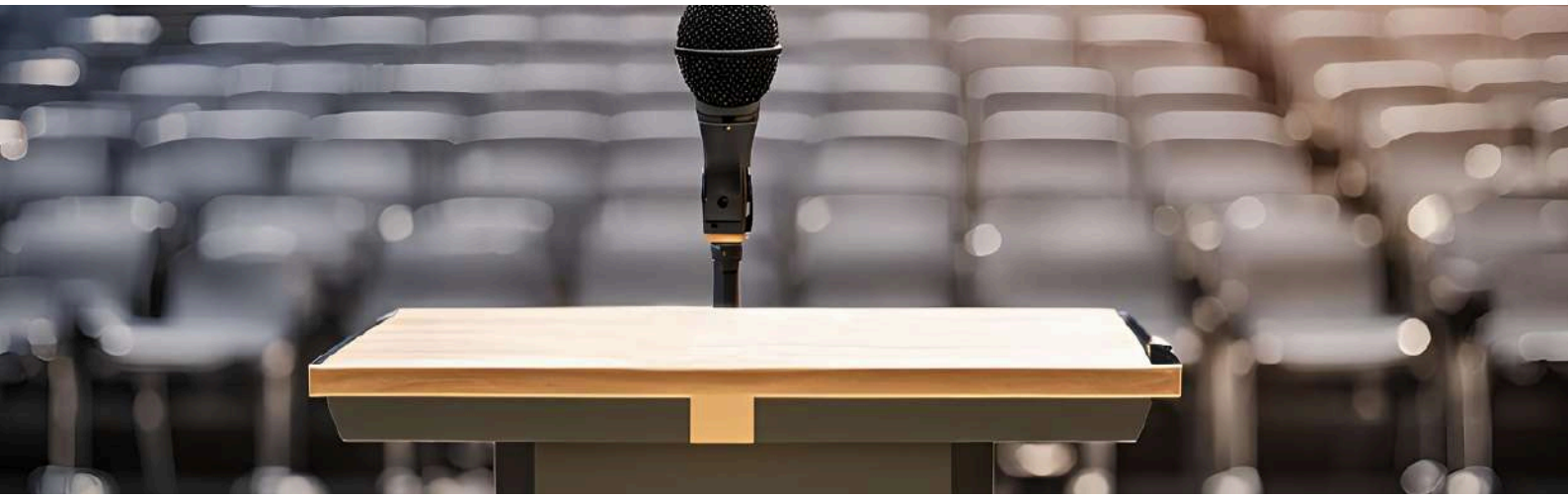
Ministerio de Ciencia e Innovación
Agencia Estatal de Investigación

Entidades ejecutoras:

unir LA UNIVERSIDAD EN INTERNET

Entidades participantes:

Universidad de Murcia
Universidad de Vigo
Universidad Camilo José Cela
Universidad de Burgos
Universidad de Navarra
UCLM Universidad de Castilla-La Mancha



1. INTRODUCCIÓN

El discurso de odio forma parte habitual del entorno digital. No es un exceso puntual ni una disfunción aislada, sino un elemento persistente en la conversación pública en redes sociales. Puede expresarse de forma explícita o velada, pero siempre contribuye a degradar el debate, erosionar la convivencia y reforzar dinámicas de exclusión.

Durante los últimos años, las plataformas han defendido la eficacia de sus sistemas de moderación. Sin embargo, los contenidos hostiles siguen circulando con normalidad, incluso cuando vulneran sus propias normas. Lejos de desaparecer, muchos de estos mensajes se mantienen activos con el paso del tiempo, sin intervención aparente. Esta permanencia plantea interrogantes de fondo: ¿qué mensajes se eliminan y cuáles no? ¿De qué depende que el odio desaparezca o se mantenga? ¿Qué margen real tienen las plataformas, los medios y los usuarios para frenar su propagación?

Este informe se sitúa en ese terreno. Examina cómo se gestiona —o se deja sin gestionar— el discurso de odio cuando ya no está bajo los focos. Y lo hace con una intención clara: ofrecer datos, detectar patrones y reflexionar sobre las consecuencias de esa permanencia. Porque si el odio no se borra, no es solo por fallo técnico. A veces, como veremos, también es por decisión.

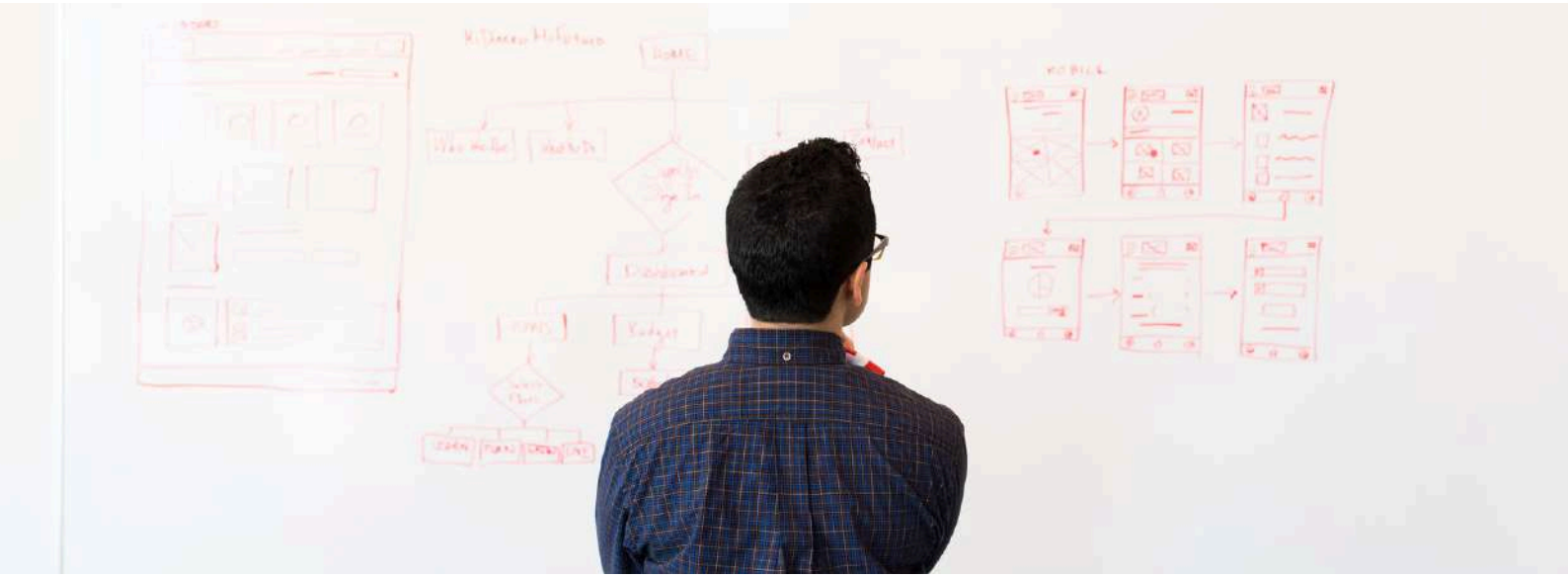
2. LA INVESTIGACIÓN

2.1 El proyecto HATEMEDIA

HATEMEDIA (Exp. PID2020-114584GB-I00) es un proyecto financiado por el Ministerio de Ciencia e Innovación del Gobierno de España. Su propósito es analizar cómo se manifiestan las expresiones de odio en espacios digitales vinculados a medios informativos y qué mecanismos —tecnológicos, editoriales, sociales— determinan su proliferación, su impacto y su permanencia. Desarrollado por la Universidad Internacional de La Rioja (UNIR) en colaboración con otras entidades académicas, el proyecto aborda el fenómeno desde una perspectiva multidisciplinar, combinando el análisis computacional con la reflexión comunicativa y el estudio de la gobernanza digital.

2.2 Objetivos del estudio

La investigación recogida en este informe parte de una pregunta concreta: ¿Qué mensajes de odio desaparecen en redes sociales y cuáles se quedan? El objetivo ha sido identificar los factores que explican la supervivencia o eliminación de mensajes hostiles en una plataforma como X, especialmente cuando estos se generan en torno a contenidos informativos. Este análisis pretende aportar datos verificables que ayuden a evaluar la eficacia de las políticas actuales de moderación, detectar posibles sesgos en la gestión del odio digital y generar herramientas interpretativas útiles tanto para investigadores como para profesionales de la comunicación.

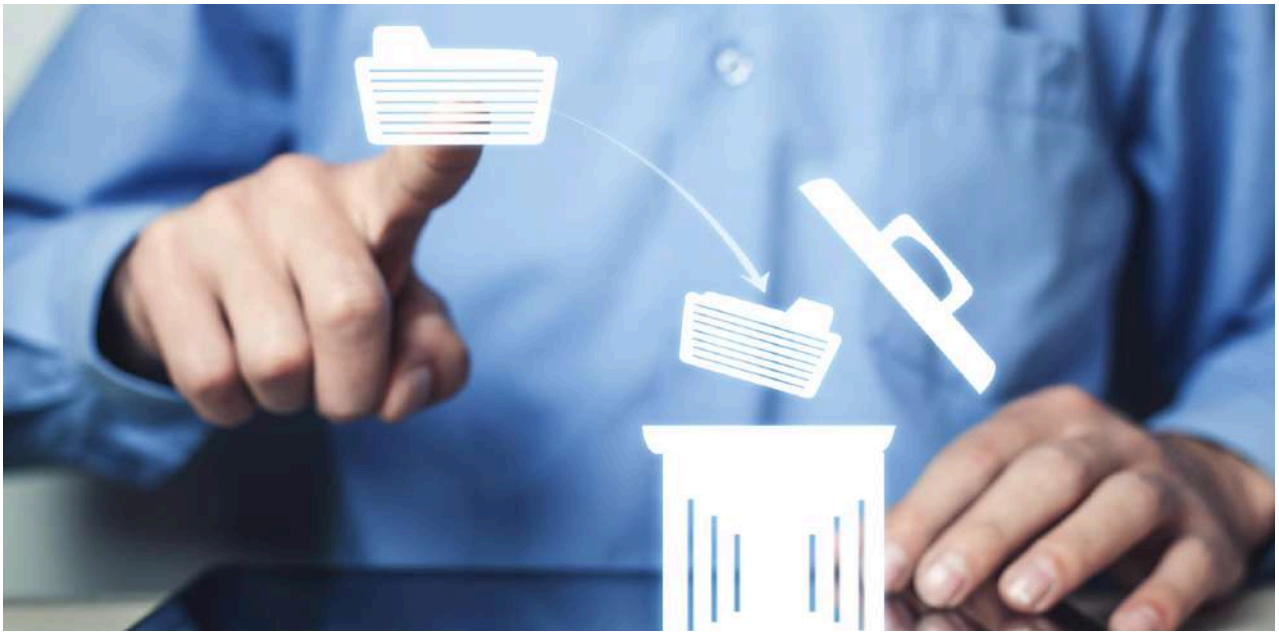


2.3 METODOLOGÍA

Para llevar a cabo este trabajo, se partió de un universo de **2.157.920 mensajes** publicados en la plataforma X (antes Twitter) entre enero de 2021 y mayo de 2022. Todos los mensajes fueron generados por terceros (usuarios ajenos a los medios) en torno a contenidos difundidos por los perfiles oficiales de cinco medios informativos digitales españoles: "El País", "El Mundo", "ABC", "La Vanguardia" y "20 minutos". La recolección se realizó mediante la API académica de Twitter (versión 2.0), priorizando aquellos tuits que respondían, comentaban o citaban publicaciones de estos medios. Para automatizar el proceso de recogida, se emplearon DAGs programados en Apeche Airflow, lo que permitió una extracción sistemática, diaria y escalable de los mensajes.
Una vez completada la recolección, los mensajes fueron sometidos a un proceso de depuración y normalización lingüística, eliminación de duplicados, URLs, emojis, caracteres especiales, transformación a minúsculas, lematizacion y tokenizacion. Posteriormente, se procedió a la clasificación automática de los mensajes mediante una arquitectura secuencial de tres algoritmos entrenados para el contexto español:

- **Detección de odio** (odio/no odio)
- **Clasificación por tipo de odio:** político, misógino, xenófobo, sexual o general
- **Clasificación por nivel de intensidad:** desde tono incívico hasta amenaza explícita (niveles 1 a 4)

Gracias a este sistema de etiquetado (validado con técnicas de revisión cruzada) se pudo obtener un corpus estructurado y confiable. Los modelos empleados alcanzaron niveles de precision y F1-score adecuados para cada dimensión, aunque con algunas limitaciones en los niveles de mayor intensidad debido a su baja frecuencia relativa. A partir de este corpus, se diseñó una **muestra representativa** con un objetivo específico: comprobar qué mensajes con expresiones de odio seguían activos tres años después de su publicación.



Para ello, se seleccionaron **9.894 tuits** a través de un muestreo estratificado proporcional, que tuvo en cuenta el tipo de odio, su intensidad y la distribución mensual del volumen detectado. Esta estrategia permitió evitar distorsiones y mantener la diversidad del fenómeno. La revisión de la existencia o eliminación de estos mensajes se llevó a cabo de forma manual en los meses de noviembre y diciembre de 2024, lo que permitió incorporar una dimensión temporal al análisis y evaluar la persistencia real de los mensajes tras un periodo prolongado.

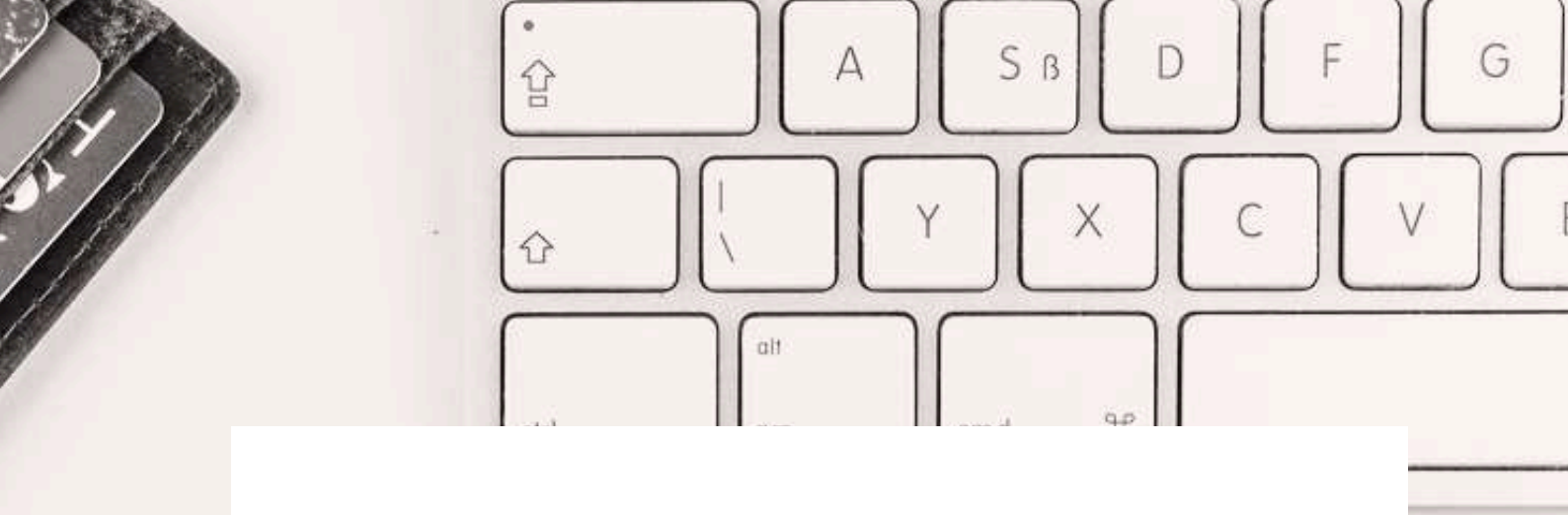
En esta muestra verificada se analizaron las siguientes variables como posibles factores explicativos de la eliminación de mensajes:

- Tipo de odio
- Nivel de intensidad
- Medio citado
- Nivel de engagement (respuestas, retuits, favoritos, marcadores)
- Carga emocional del mensaje (basada en el léxico NRC)

El análisis combinó técnicas descriptivas (frecuencias, medias, varianzas) con modelos estadísticos más complejos. Para identificar relaciones significativas se aplicó un **análisis de varianza (ANOVA)** y para estimar la capacidad predictiva de las variables, se utilizó un modelo de **Random Forest,** que permitió jerarquizar los factores según su influencia en la permanencia del mensaje.

Resumen metodológico

Recolección	Clasificación	Muestra	Verificación	Análisis
2.157.920 tuis recopilados (2021-2022)	Identificación automática de odio (tipo e intensidad)	Selección estratificada: 9.894 mensajes	Revisión manual en el año 2024: activo/eliminado	Cruce variables y modelos predictivos



3. RESULTADOS

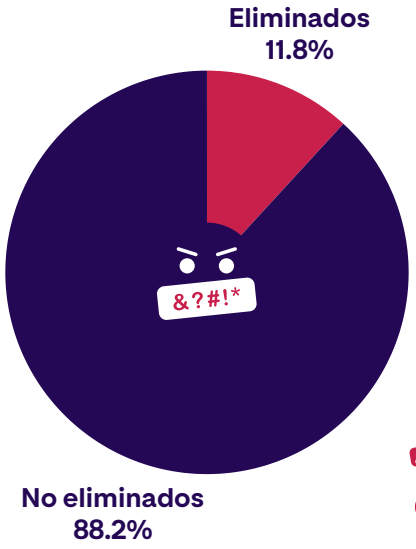
Panorama general: lo que se queda, lo que desaparece

El análisis de los mensajes verificados revela una conclusión central: **el discurso de odio rara vez desaparece por moderación activa.** Más de dos años después de su publicación, la gran mayoría de los mensajes identificados como hostiles seguían visibles en la plataforma. Este dato desafía la idea de que las plataformas sociales eliminan automáticamente los contenidos más ofensivos y pone en cuestión la efectividad de los sistemas de control. En lugar de seguir una lógica proporcional —en la que los mensajes más agresivos desaparecen antes o con más frecuencia—, los datos muestran una moderación irregular, opaca y poco relacionada con la gravedad del contenido. En este contexto, conviene observar qué factores sí parecen influir en la permanencia del odio digital y cuáles quedan neutralizados por la lógica del sistema.

3.1. **Baja tasa** de eliminación

Como adelantábamos, del total de tuits analizados solo el 12 % fue eliminado frente a un 88 % que seguía accesible al cierre del trabajo de verificación. Esta proporción se mantuvo estable incluso en mensajes publicados tres años antes. No se encontraron indicios sólidos de que estos tuits hubieran sido eliminados como resultado de una política sistemática de moderación. En la mayoría de los casos, no constaba denuncia visible ni razones explícitas de retirada. El resultado es coherente con estudios previos que señalan una caída en la vigilancia moderadora tras la retirada de programas de verificación en plataformas como X.

Eliminados vs. No eliminados (%)

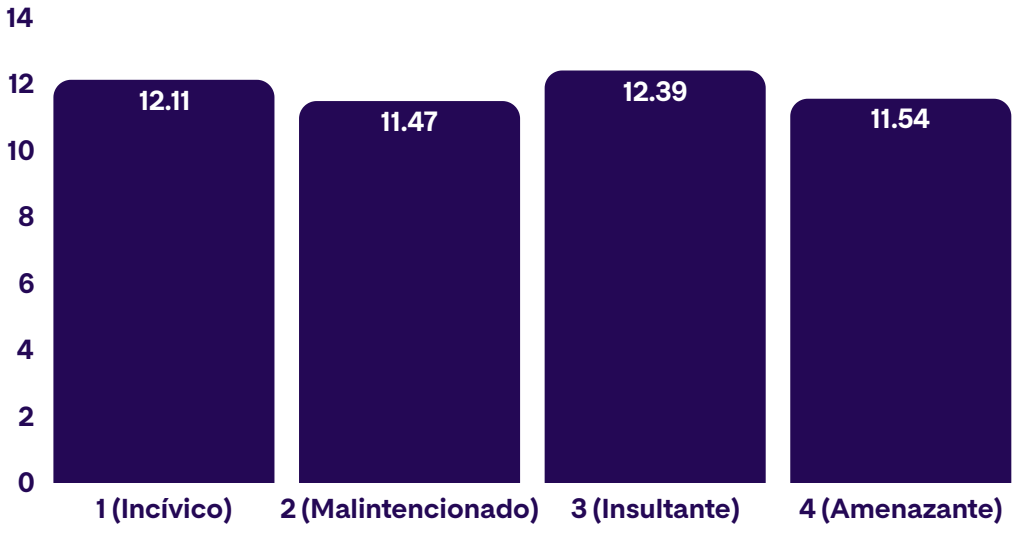


En otras palabras, más de 8.700 mensajes de odio (de un total de 9.894) seguían activos varios años después de haber sido posteados. La persistencia de estos tuis implica que las plataformas no están realizando una eliminación proactiva o no cuentan con los sitemas adecuados para eliminar eficazmente el odio digital.

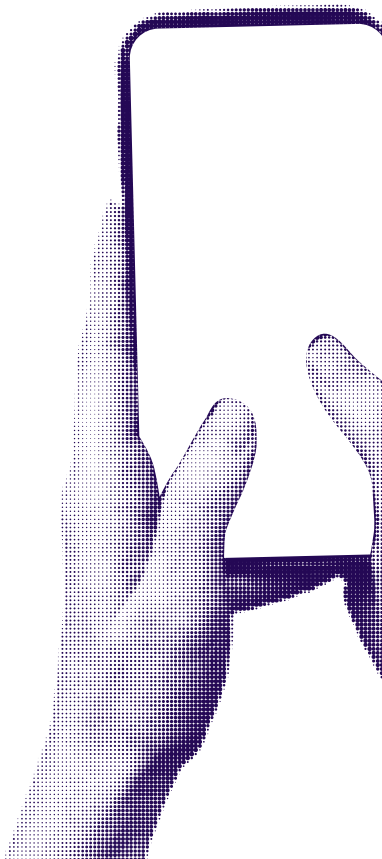
Casi el 90% de los mensajes siguen activos dos años después de publicarse.

3. 2 La **intensidad** de odio no determina su eliminación

Los mensajes más agresivos —insultos, amenazas, expresiones violentas— deben ser los primeros en desaparecer. Ese es uno de los supuestos más extendidos en las estrategias de moderación. Sin embargo, los datos muestran que esta lógica no se cumple en la práctica. Los tuits clasificados con los niveles más altos de intensidad (niveles 3 y 4, correspondientes a insultos graves y amenazas) no presentaron tasas de eliminación significativamente superiores a los de niveles más bajos. De hecho, la distribución de los mensajes eliminados se reparte de forma casi uniforme entre los distintos niveles de gravedad. Este resultado cuestiona directamente modelos como la **"pirámide del odio"**, que parten de una progresión lineal entre intensidad del discurso y riesgo de daño. En el contexto analizado, la severidad del contenido no se tradujo en una mayor intervención. Todo indica que los sistemas de moderación aplicados —tanto humanos como automatizados— no priorizan la eliminación en función de la agresividad del mensaje.



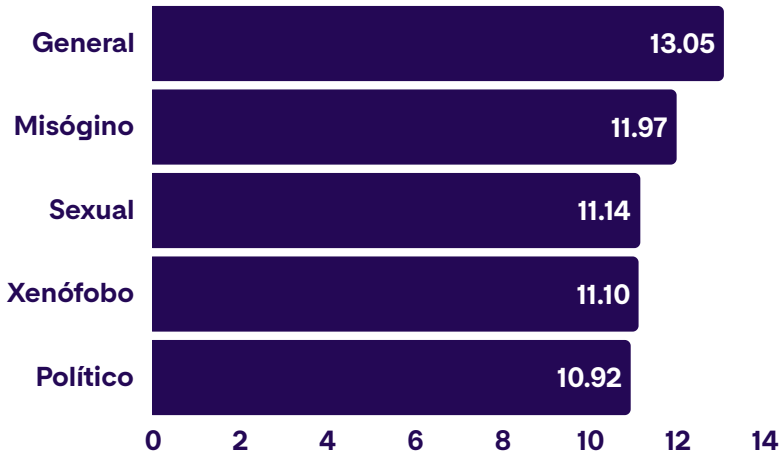
Tasa de eliminación según la intensidad de odio (%)



3.3 El **tipo** de odio sí influye

A diferencia de la intensidad, el tipo de odio expresado en el mensaje sí muestra una relación significativa con su permanencia o eliminación. Según el análisis estadístico, esta fue la única variable categórica que presentó una diferencia significativa en el modelo ANOVA aplicado. Entre los distintos tipos, el odio político fue el menos eliminado, seguido del sexual y el xenófobo. En cambio, los mensajes clasificados como "odio general" —aquellos que no señalan un colectivo concreto o presentan un rechazo difuso— registraron una mayor tasa de eliminación.

Tasa de eliminación por tipo de odio (%)



Esta diferencia no puede explicarse por una mayor agresividad de unos mensajes frente a otros, ni por su carga emocional. Como veremos más adelante, lo que sí sugiere es que el contenido ideológico se tolera con más frecuencia, incluso cuando reproduce patrones claramente hostiles. El sistema —algorítmico o institucional— parece más permisivo con aquello que se inscribe en la disputa política que con los discursos de odio más genéricos o simbólicamente neutros.

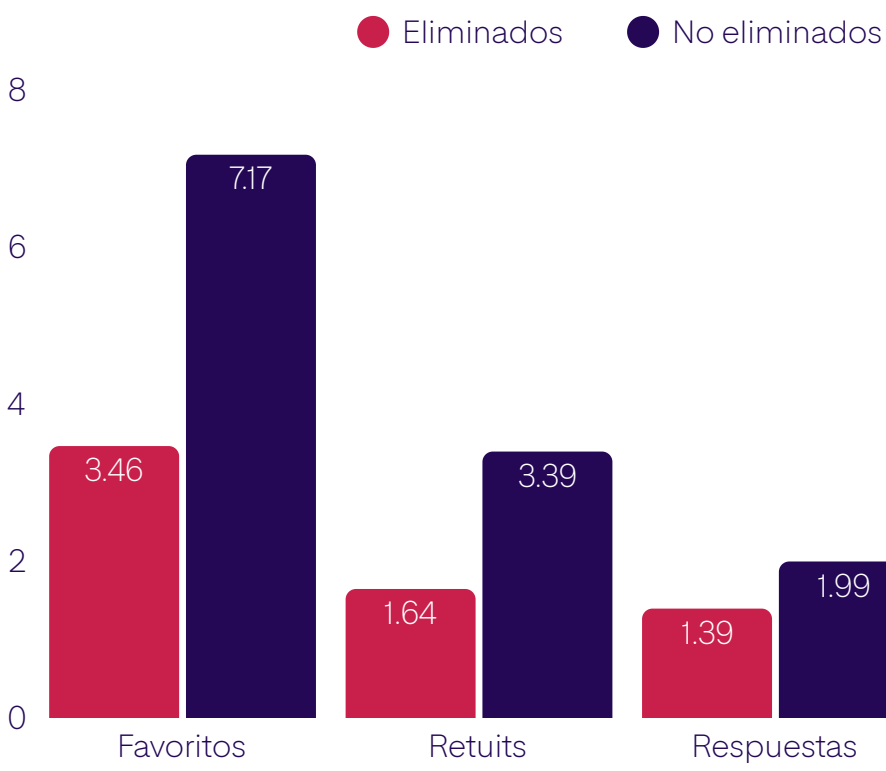
3.4 Y el **medio** también importa

El medio al que se dirige el mensaje influye: no todos los entornos moderan del mismo modo.

Además, la eliminación del discurso de odio no depende solo del contenido del mensaje. También del entorno mediático en el que circula. En el análisis se observaron diferencias claras según el medio implicado. De este modo, los mensajes vinculados a "20 Minutos" presentaron la menor tasa de eliminación, mientras que "ABC" y "La Vanguardia" registraron los porcentajes más altos. Estas variaciones se asientan en factores como el perfil del medio, su volumen de interacción o posibles criterios internos de moderación. Es cierto que resulta difícil establecer una relación causal directa, pero el dato es claro: el medio al que se dirige el mensaje condiciona, en parte, su permanencia o eliminación.

3.5 El ***engagement*** protege al mensaje

Interacción media en tuits eliminados y no eliminados



En cuanto a los mensajes que acumulan más interacciones —respuestas, retuits, favoritos o marcadores—, se ha comprobado que estos tienen más probabilidades de permanecer activos. Los datos muestran que el nivel de engagement suele ser un factor que actúa como barrera frente a la eliminación. Este patrón se repite de forma consistente en todos los tipos de interacción analizados. Cuanto mayor es la visibilidad pública del mensaje, menor es la intervención posterior. Dicho hallazgo refuerza la idea de que la lógica de la viralidad puede entrar en conflicto con la de la moderación. Lo que más circula no es lo más vigilado, pero sí, en muchos casos, lo más tolerado.

3.6. La **carga emocional** no es un criterio de moderación

Que un mensaje sea más visible lo protege. Sin embargo, que su tono sea más violento no ejerce como factor determinante. Tras analizar miles de mensajes en función de su carga emocional —especialmente aquellas emociones asociadas al odio, como la ira o el miedo—, los resultados fueron contundentes: la emocionalidad del contenido no influye en su eliminación.

Comentarios con un alto grado de agresividad emocional permanecen activos igual que aquellos con expresiones más neutras. No se observa una intervención proporcional al mensaje, ni mayor vigilancia sobre los contenidos más cargados de hostilidad. Esto apunta a un fallo estructural, ya que ni los algoritmos ni la moderación manual parecen estar calibrados para interpretar el componente afectivo del discurso. La emoción no activa la alerta. Y cuando no se detecta el tono, el sistema deja pasar el fondo.

El odio puede gritar, pero si no genera clics, pasa desapercibido.

4. **ANÁLISIS** E INTERPRETACIÓN

4.1 Una **intervención** sin prioridades claras

Estos resultados muestran una realidad difícil de esquivar: la eliminación del discurso de odio no responde a criterios consistentes. Ni la intensidad del mensaje, ni la manera en que se comunica actúan como factores determinantes para su desaparición. Por el contrario, elementos secundarios como el tipo de odio o el nivel de engagement parecen influir más en su permanencia. Esto revela un problema importante en las estrategias de moderación, pues no se está actuando sobre los contenidos más dañinos. Lo que se borra, en muchos casos, es lo que no molesta demasiado, y lo que se queda, aunque sea más hostil, es aquello que circula. En este sentido, el sistema no filtra el odio, ya que lo ordena según su capacidad de adaptación al entorno digital.

4.2 Una **lógica** que prioriza lo que más impacto genera

Lo hemos visto. La lógica de las plataformas está pensada para maximizar la propagación de contenidos. Cuanto más potencial de interacción tiene un mensaje —por conflictivo, provocador o polarizante—, más se amplifica. Esto convierte la viralidad en un criterio práctico de relevancia, aunque lo que se diga sea dañino. En el terreno de las redes sociales, no se distingue entre lo que aporta y lo que daña, por lo que los mensajes más visibles acaban teniendo más valor que los más justos o necesarios.

El problema es que esta dinámica desplaza todo sentido de protección. No hay mecanismos que retiren los discursos de odio de forma automática, ni una política clara que limite su alcance. Esta forma de funcionamiento hace que los comentarios y opiniones más extremos circulen con normalidad y queden integrados en la conversación.

4.3 Tolerancia estructural al discurso de odio

Asimismo, cabe insistir en una idea clave. Algunos tipos de odio son más propensos a permanecer. El odio con connotaciones ideológicas (particularmente el de carácter político) encuentra mayor margen para permanecer en los entornos digitales. Esta tolerancia no parece casual ni puramente técnica y se sostiene sobre una ambigüedad funcional que protege a estos discursos como si fuesen parte del debate legítimo. Las expresiones de este tipo se benefician de un marco interpretativo más indulgente, donde el límite entre confrontación política y discurso de odio se difumina.

Así, la moderación tiende a ser más reacia a intervenir cuando el contenido se inscribe en disputas ideológicas. Seguramente ocurra por miedo a incurrir en censura o cortar el tipo de conversación que mantiene activa la red. El resultado es una forma de impunidad discursiva. Si el odio ideológico permanece, lo hace no porque sea menos agresivo, sino porque se percibe como más defendible, más opinable o más propio del terreno de la disputa argumentativa. Sin embargo, esa permisividad acaba naturalizando expresiones de hostilidad que no se permitirían en otros contextos.

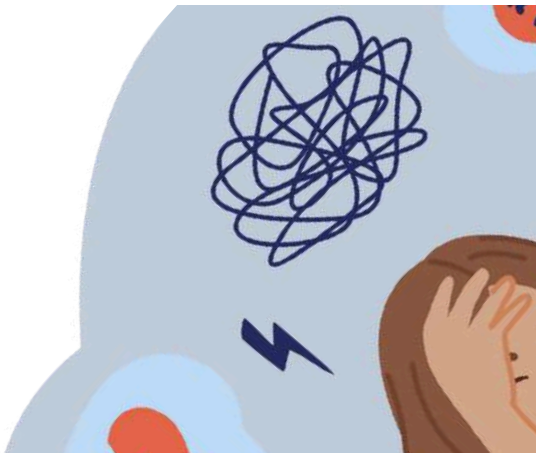
4.4. Una **gestión desigual** que fragmenta la respuesta

Por otra parte, el hecho de que la visibilidad del odio dependa del medio al que se dirige plantea un problema de fondo. No estamos ante decisiones conscientes ni estrategias editoriales, sino ante una gestión condicionada por el desorden y la falta de herramientas. Los medios operan dentro de plataformas que imponen sus propias reglas, pero no ofrecen soluciones claras ni margen de actuación suficiente. Por eso, aunque la exposición al discurso de odio es constante, la capacidad de intervenir no siempre está en manos de quienes lo reciben. Además, la ausencia de criterios compartidos entre los propios medios debilita la eficacia de la moderación, generando un paisaje desigual en el que la agresividad se concentra allí donde encuentra menos fricción institucional o comunitaria.

Los medios aplican criterios desiguales frente a contenidos hostiles y cuando esto ocurre, el odio aprende por dónde "colarse"

4.5. La **emoción** no activa ninguna alerta

También es importante recordar que la forma en que se expresa el odio —más airada, más cargada de tensión emocional— no incrementa su probabilidad de ser eliminado. Incluso las expresiones marcadas por emociones intensas como ira o miedo permanecen activas con la misma frecuencia que aquellas más neutras o frías. Esto indica que la moderación digital no incorpora criterios sensibles al tono emocional del lenguaje. Se actúa sobre lo que se dice, pero no sobre cómo se dice. En esa capa de análisis, una parte significativa del odio más perturbador se queda fuera del radar. Esta limitación tiene consecuencias importantes, dado que ignorar el registro emocional de las expresiones hostiles contribuye a normalizarlas.



5. CONCLUSIONES

A la vista de todo lo anterior, este informe demuestra con datos fidedignos que la permanencia del discurso de odio en entornos digitales no responde a una lógica coherente ni proporcional al daño potencial que pueda causar. El análisis realizado prueba que la mayoría de las expresiones hostiles permanece en la plataforma con el paso del tiempo, con independencia de su gravedad o función desestabilizadora. En lugar de formar parte de una estrategia ordenada de contención, **la gestión de contenidos hostiles opera de forma dispersa, sin mecanismos de verificación estables ni umbrales compartidos entre actores**. Este patrón refleja una lógica ineficaz, así como una forma de permisividad estructural. A modo de recapitulación:

- La mayoría de los mensajes de odio analizados siguen accesibles años después de su publicación.
- La agresividad o el tono emocional no incrementan las probabilidades de eliminación.
- El tipo de odio y el medio al que se dirige influyen más que el contenido en sí.
- Cuanto mayor es el engagement, más probable es que el mensaje se mantenga.
- No existe una política común de intervención entre plataformas, medios y comunidades.
- Esta fragmentación favorece la normalización del odio y debilita la respuesta institucional.

6. RECOMENDACIONES

Las conclusiones expuestas nos confirman que la permanencia del discurso de odio en redes sociales es alta, la intervención es irregular y los criterios para actuar son, en el mejor de los casos, poco consistentes. Ante esta situación, la pregunta que toca hacerse es clara: **¿qué se puede hacer para corregirlo?** En los siguientes apartados recogemos una serie de propuestas realistas, aplicables y directamente conectadas con los problemas identificados en el análisis. No han de verse como soluciones definitivas, si bien pueden funcionar como líneas de actuación que son capaces de contribuir a mejorar la gestión del odio digital y reducir su impacto en el entorno público.

Nuestras metas y objetivos

Desde el proyecto HATEMEDIA, entendemos que investigar el odio digital no implica solo medirlo, describirlo o clasificarlo. También supone **comprometerse con su reducción** y con el desarrollo de herramientas prácticas para abordarlo. Por eso, las siguientes recomendaciones buscan ir más allá del análisis académico y pretenden convertirse en un punto de partida para la acción, tanto en el plano tecnológico como en el institucional, mediático o educativo.

6.1 Mejorar los sistemas de detección

Uno de los problemas más evidentes es que muchos mensajes hostiles no se identifican a tiempo o directamente pasan desapercibidos. Los filtros actuales no están preparados para detectar todas las formas que puede adoptar el odio digital. Funcionan bien con insultos directos, pero fallan cuando el lenguaje es más sutil, irónico o encubierto. Además, los sistemas de detección suelen centrarse solo en palabras clave o combinaciones concretas. Eso los hace fáciles de esquivar, ya que no interpretan el contexto o no consideran la intencionalidad indirecta que muchas veces acompaña a estos mensajes. Por consiguiente, es necesario:

1

INCORPORAR tecnologías capaces de reconocer modos encubiertos de agresión: insinuaciones, sarcasmos hostiles, desprecio indirecto.

2

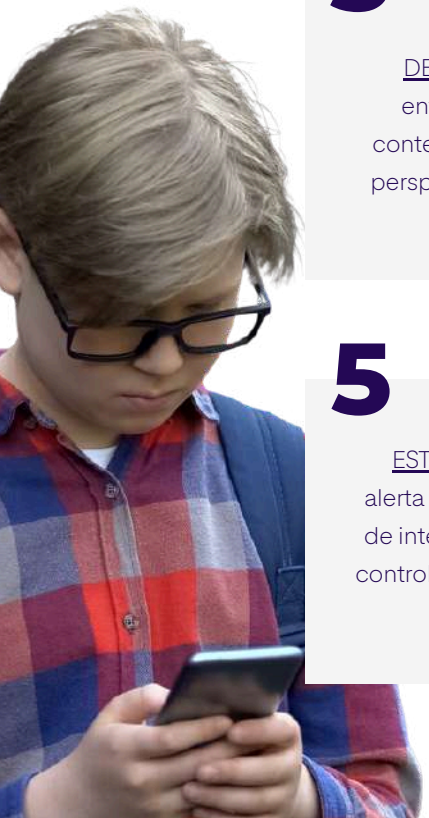
INVERTIR en herramientas que integren análisis semántico y no solo listas de palabras para entender el significado completo de las frases

3

DESARROLLAR modelos entrenados en diferentes contextos socioculturales, con perspectiva de género, racial y política.

4

COMBINAR la automatización con revisión humana especializada, especialmente en los casos ambiguos.



5

ESTABLECER umbrales de alerta adaptativos según el nivel de interacción, para no dejar sin control los mensajes más visibles.

6

PROMOVER la colaboración entre investigadores, tecnólogos y profesionales de la comunicación, para mejorar la fiabilidad de los sistemas.

6.2 **Identificar** el odio disfrazado de opinión

Como se ha visto, hay formas de odio que resultan más difíciles de detectar no porque sean menos dañinas, sino porque se expresan de forma más aceptada, más disimulada o bajo la apariencia de debate legítimo. Es lo que ocurre con muchos mensajes de desprecio ideológico o con ciertas formas de misoginia encubierta. El análisis del proyecto HATEMEDIA muestra que este tipo de mensajes tiende a mantenerse con más facilidad en las plataformas.

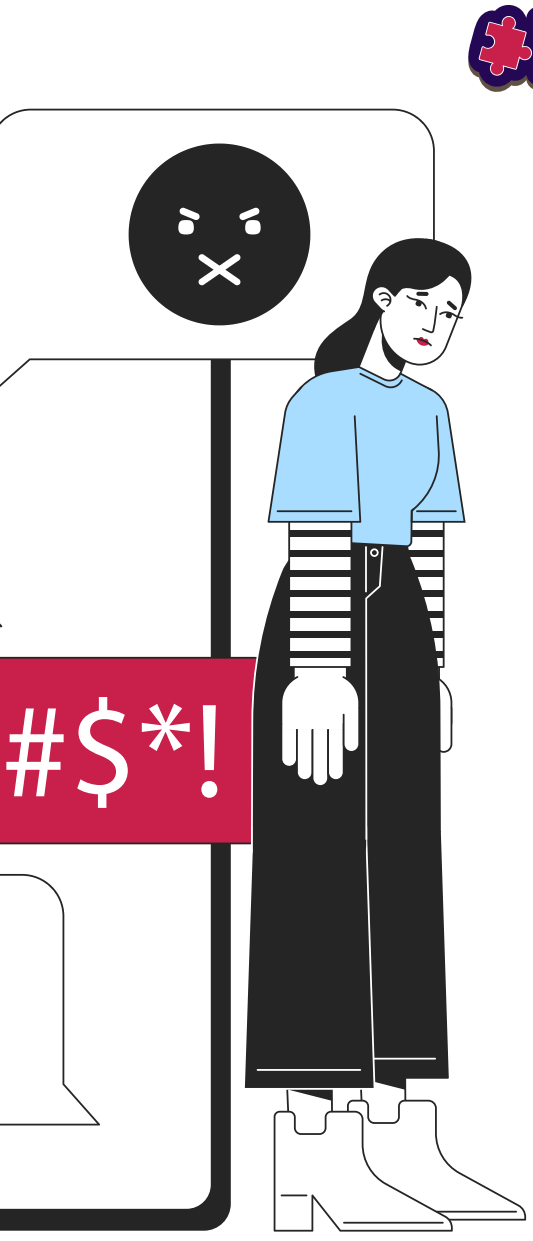
- **Ejemplo práctico:** cuando alguien desacredita a un colectivo diciendo que "no están preparados para la vida pública" o que "deberían callarse más", sin insultos directos, está aplicando una lógica de exclusión disfrazada de opinión. Eso también puede bordear los límites e incluso ser discurso de odio.

¿Cómo solucionarlo?

Es importante que el tipo de odio NO se trate como un dato secundario. Debe formar parte del núcleo de análisis y respuesta. Hay que prestar **especial atención a los casos donde la agresión se esconde bajo un tono razonable o se presenta como una simple opinión**. Para ello, es clave que quienes moderan sepan reconocer estos patrones, que se revise con más cuidado el contenido ideológico que normaliza el desprecio y que se usen herramientas capaces de detectar la carga simbólica que muchas veces se esconde tras palabras aparentemente inofensivas. Y sobre todo, hay que evitar que la libertad de expresión se utilice como excusa para dejar pasar lo que claramente busca excluir, señalar o degradar.

6. 3 **Protocolos** claros para actuar con coherencia

Hoy cada medio y cada plataforma responde a su manera. No hay un criterio común, y eso provoca respuestas muy distintas ante situaciones parecidas. A veces se elimina un mensaje de inmediato, pero otras no se hace nada, provocando una sensación de descontrol y desigualdad que debilita cualquier intento de intervención coherente. Explicado fácilmente: un medio puede tener filtros en su página web, pero no en sus redes sociales; o puede bloquear automáticamente ciertas palabras, aunque eso deje pasar otros mensajes más dañinos que no usan insultos explícitos. Para ello, lo más útil es contar con una serie de medidas concretas que ayuden a ordenar la respuesta y a reducir la disparidad entre entornos. Entre ellas:



Establecer protocolos claros y públicos que definan cómo actuar ante los discursos de odio.

Acordar criterios mínimos comunes entre medios, redes y profesionales que gestionan contenidos.

Diseñar guías prácticas y realistas, adaptadas al trabajo diario de quienes moderan.

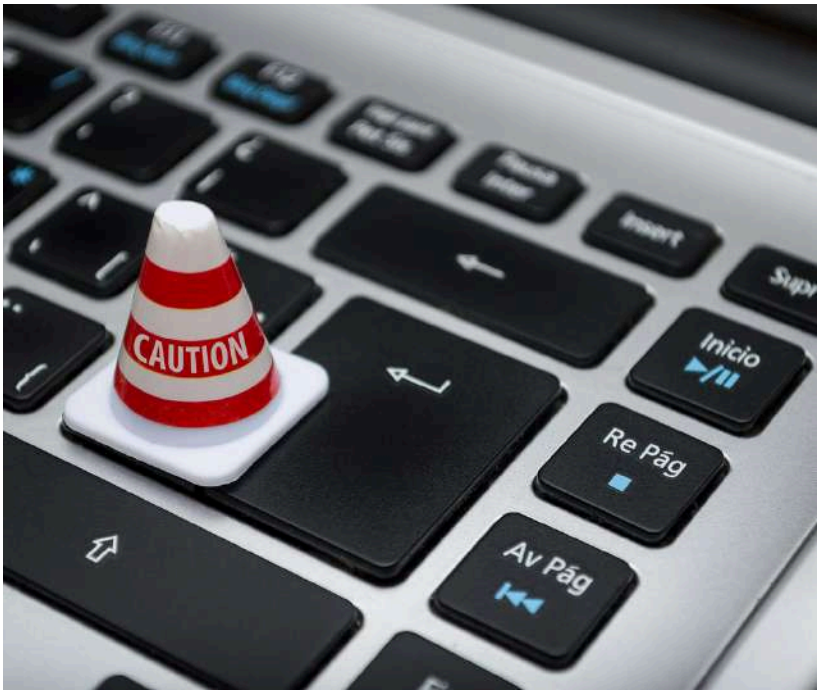
Crear canales de coordinación entre medios y plataformas para compartir decisiones y experiencias.

Evitar la desconexión entre espacios propios (web, redes, secciones de opinión o comentarios).

6.4 Vigilar lo que más circula

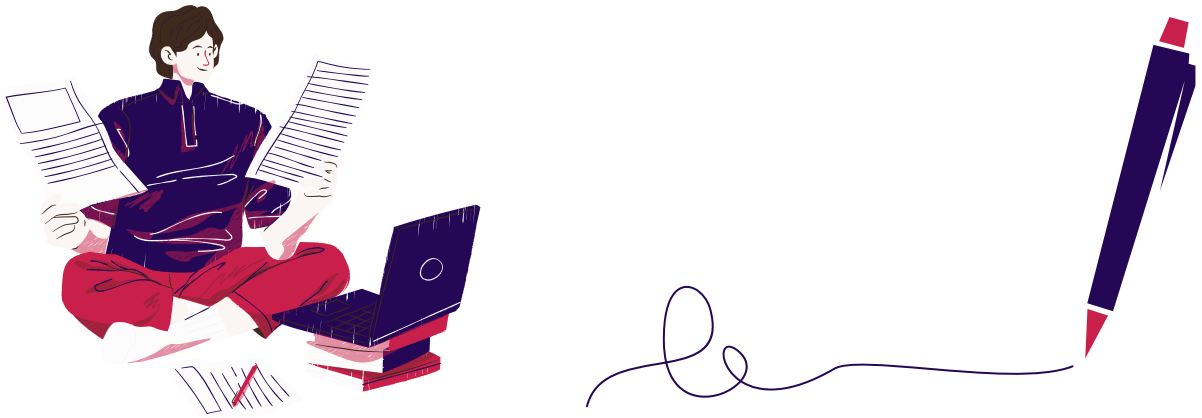
Muchos de los mensajes que generan más odio son también los que más atención reciben. No es algo que ocurra por casualidad, ya que hemos comprobado cómo la interacción y el engagement actúan como una forma de blindaje, de manera que si algo funciona, se mantiene en el espacio digital. Esto crea un problema evidente, que ha de ser combatido con una vigilancia más activa sobre aquello que más se comparte y más visibilidad alcanza. Se requiere:

- Revisar con mayor atención los mensajes que reciben más interacción.
- No tratar la viralidad como un criterio de valor, sino como un factor de riesgo.
- Aplicar filtros que se activen no solo por lo que se dice, sino por cuánto circula.
- Establecer sistemas que marquen contenidos populares para revisión prioritaria.
- Evitar que la popularidad de un mensaje desactive los mecanismos de control.



6.5 Cuidar el enfoque informativo

La manera en que se cuenta una historia influye directamente en cómo se recibe. Algunos tratamientos poco cuidados, ciertos titulares llamativos o coberturas que insisten en estereotipos pueden reforzar sin querer los mismos discursos que alimentan el odio. No basta con publicar los hechos: también importa el tono, el encuadre, las palabras que se eligen y las que se dejan fuera. Una imagen, una cita destacada o el orden en que se presenta una información pueden condicionar la interpretación del público y, en algunos casos, abrir la puerta a comentarios hostiles, desinformación o justificaciones encubiertas. Por eso, es clave revisar con atención cómo se presentan los temas sensibles, sobre todo cuando afectan a colectivos vulnerables.



Por ejemplo, una noticia sobre violencia de género que reproduce de forma acrítica la versión del agresor o que desvía el foco hacia su situación personal puede terminar generando confusión e incluso burlas. En esos casos, la falta de cuidado narrativo no solo desinforma, dado también contribuye a consolidar una mirada dañina **Cuidar el enfoque informativo implica responsabilidad**. Significa evitar el morbo, explicar los contextos, priorizar voces informadas y, sobre todo, tener claro que la forma también comunica.

6.6 **Apostar** por la intervención temprana

Cuando un mensaje de odio se mantiene en circulación durante mucho tiempo gana visibilidad. Pero no solo eso: se normaliza, se replica y, muchas veces, se convierte en la base de nuevas agresiones. Por eso es tan importante detectar los casos problemáticos lo antes posible. La denuncia manual suele ser lenta y los filtros automáticos no siempre reaccionan con agilidad. Mientras tanto, los contenidos hostiles se mueven con libertad y generan interacción.

Muchas veces, cuando se revisa una publicación problemática, ya ha tenido un impacto considerable. Ha sido vista, compartida y comentada por miles de personas. En ese punto, la intervención llega tarde. Por eso es importante actuar antes, sin que eso signifique censurar a la primera ni eliminar contenido sin criterio. Intervenir a tiempo implica tener herramientas que detecten patrones que se repiten, situaciones delicadas o señales que indiquen que puede haber un problema. Y, a partir de ahí, activar protocolos de forma rápida y ajustada al caso.

ACCIONES CONCRETAS

- Detectar patrones de hostilidad que se repiten
- Activar alertas ante picos inusuales de reacción
- Identificar temas sensibles antes de que escalen
- Responder rápido sin depender solo de denuncias
- Revisar lo que empieza a viralizarse

7. PARA TERMINAR

El odio digital no desaparece solo. No se corrige con el paso del tiempo ni se elimina sin intervención. Si sigue presente, es porque algo lo permite. Y si se convierte en parte del paisaje, es porque no se está haciendo lo suficiente para evitarlo. Este informe no trae soluciones mágicas, pero deja claro que hay margen para actuar. Lo que se tolera se extiende y lo que se enfrenta con criterio se puede reducir. Desde el proyecto HATEMEDIA seguimos investigando cómo circula el odio en entornos digitales y qué herramientas pueden ayudar a limitarlo. Nuestra apuesta combina análisis riguroso, propuestas concretas y compromiso con una conversación más sana. Puedes conocer más sobre nuestro trabajo en:

SCAN HERE

PUBLICACIONES Y RECURSOS

Artículos en revistas

ANTONA JIMENO, T., MAYAGOITIA-SORIA, A., & ĐORĐEVIĆ, J. (2024). Investigación sobre el discurso de odio. Una propuesta de análisis bibliométrico en España y LATAM entre 2021 y 2022. Revista ICONO14, 22(1), e2128. https://doi.org/10.7195/ri14.v22i1.2128

ARCE-GARCÍA, S., SAID-HUNG, E., & MONTERO-DÍAZ, J. (2024). Unmasking coordinated hate: Analysing hate speech on Spanish digital news media. New Media & Society. https://doi.org/10.1177/14614448241259715

BRÄNDLE, G., CÁCERES-ZAPATERO, M. D., & PAZ-REBOLLO, M. A. (en prensa). Sentir el odio: análisis de la gravedad percibida de los discursos de odio en la población española. Revista Española de Sociología.

BRÄNDLE, G., ZAPATERO, M. D., & PAZ-REBOLLO, M. A. (2024). Sentir el odio: análisis de la gravedad percibida de los discursos de odio en la población española. Revista Española de Sociología, 33(2), a219. https://doi.org/10.22325/fes/res.2024.219

CÁCERES-ZAPATERO, M. D., BRÄNDLE, G., & PAZ-REBOLLO, M. A. (2023). Stances on hate speech: Population opinions and attitudes. Profesional de la Información, 32(4). https://doi.org/10.3145/epi.2023.jul.10

GONZÁLEZ-AGUILAR, J., SEGADO-BOJ, F., & MAKHORTYKH, M. (2023). Populist right parties on TikTok: spectacularization, personalization, and hate speech. Media and Communication, 11(2). https://doi.org/10.17645/mac.v11i2.6358

MARTÍNEZ VALERIO, L. (2021). Mensajes de odio hacia la comunidad LGTBIQ+: análisis de los perfiles de Instagram de la prensa española durante la "Semana del Orgullo". Revista Latina de Comunicación Social, 80, 363–388. https://doi.org/10.4185/RLCS-2022-1749

MATARÍN, E., GÓMEZ, T., & RODRÍGUEZ-PERAL, D. (2025). Propagation of hate speech on social network X: Trends and approaches. Social Inclusion, 13. https://doi.org/10.17645/si.9317

MORENO-LÓPEZ, R., & ARGÜELLO-GUTIÉRREZ, C. (2025). Violence, hate speech, and discrimination in video games: A systematic review. Social Inclusion, 13. https://doi.org/10.17645/si.9401

RÖMER-PIERETTI, M., SAID-HUNG, E., & MONTERO-DÍAZ, J. (2025). Semiotic analysis of hate discourse in Spanish digital news media: Biden's inauguration case study. Social Inclusion, 13, 1–26. https://doi.org/10.17645/si.9295

SAID-HUNG, E., MORENO-LÓPEZ, R., & MOTTAREALE-CALVANESE, D. (2023). Promotion of hate speech by Spanish political actors on Twitter. Policy & Internet. https://doi.org/10.1002/poi3.353

SÁNCHEZ, M., VÁZQUEZ DIÉGUEZ, I., & MERINO, D. (2024). La semiótica del odio en los bulos sobre inmigrantes detectados por plataformas de fact-checking en España, Grecia e Italia. Revista ICONO14, 22(2). https://doi.org/10.7195/ri14.v22i2.2083

TELLO-DÍAZ, L., & MARTÍNEZ-VALERIO, L. (2025). Hate speech directed at Spanish female actors: Penélope Cruz —A case study. Social Inclusion, 13. https://doi.org/10.17645/si.9250

Monografías y capítulos de libro

CUBILLAS, J. J., DE GREGORIO VICENTE, O., & RODRÍGUEZ-CABALLERO, C. V. (2023). Approximation of hate detection processes in Spanish and other non-Anglo-Saxon languages. En E. SAID-HUNG & J. MONTERO (Eds.), News Media and Hate Speech Promotion in Mediterranean Countries (pp. 65–80). https://doi.org/10.4018/978-1-6684-8427-2.ch005

DE GREGORIO, O., & CUBILLAS, J. J. (2024). Aproximación de procesos de detección de odio en castellano en España. En A. MORENO, E. SAID-HUNG, & M. RÖMER (Eds.), Expresiones de odio en entornos digitales españoles. Tirant Lo Blanch.

MARTÍNEZ, L., & TELLO, L. (2024). Discursos de odio dirigido a la industria cinematográfica española. En A. MORENO, E. SAID-HUNG, & M. RÖMER (Eds.), Expresiones de odio en entornos digitales españoles. Tirant Lo Blanch.

MORENO, A., SAID-HUNG, E., & RÖMER-PIERETTI, M. (2024). Expresiones de odio en entornos digitales españoles. Tirant Lo Blanch.

MORENO-DELGADO, A. (2024). La investigación en discursos del odio en Comunicación: evolución, temática y metodología. En A. MORENO, E. SAID-HUNG, & M. RÖMER (Eds.), Expresiones de odio en entornos digitales españoles. Tirant Lo Blanch.

PIERETTI, M. R., MONTERO-DÍAZ, J., & SAID-HUNG, E. S. (2023). The semiotics of xenophobia and misogyny on digital media. En E. SAID-HUNG & J. MONTERO (Eds.), Advances in Media, Entertainment and the Arts (pp. 111–135). https://doi.org/10.4018/978-1-6684-8427-2.ch007

RÖMER, M., MONTERO-DÍAZ, J., & SAID-HUNG, E. (2024). La semiótica de la xenofobia y misoginia en los medios digitales: un estudio de caso en España. En A. MORENO, E. SAID-HUNG, & M. RÖMER (Eds.), Expresiones de odio en entornos digitales españoles. Tirant Lo Blanch.

SAID-HUNG, E., ARCE-GARCÍA, S., & MONTERO-DÍAZ, J. (2025). Patterns of dissemination of expressions of hate and polarization in Ibero-America. En A. CASERO & P. LÓPEZ (Eds.), The Routledge Handbook of Political Communication in Ibero-America. https://doi.org/10.6084/m9.figshare.26042818

SAID-HUNG, E., & MONTERO-DÍAZ, J. (Eds.). (2023). News media and hate speech promotion in Mediterranean countries. Advances in Media, Entertainment and the Arts (AMEA). https://doi.org/10.4018/978-1-6684-8427-2

SÁNCHEZ, M., DIÉGUEZ, I., & ARRIBAS, M. (2023). Mapping stigmatizing hoaxes towards immigrants on Twitter and digital media: Case study in Spain, Greece, and Italy. En E. SAID-HUNG & J. MONTERO (Eds.), Advances in Media, Entertainment and the Arts (pp. 136–161). https://doi.org/10.4018/978-1-6684-8427-2.ch008

ZAMORA-MARTÍNEZ, P., & ANTONA, T. (2024). El juicio de la monarquía en Twitter. En A. MORENO, E. SAID-HUNG, & M. RÖMER (Eds.), Expresiones de odio en entornos digitales españoles. Tirant Lo Blanch.

Congresos y Seminarios

ANTONA JIMENO, T., SAID-HUNG, E., & MONTERO, J. (2022, octubre). Hate speech's taxonomy in Spanish professional news media. Ponencia presentada en ECREA Pre-conference 'News Media, Disinformation, and Hate Speech's Promotion', Lisboa, Portugal.

ARCE-GARCÍA, S., & MONTERO-DÍAZ, J. (2024, octubre). The Mapping of Patterns Project: Dissemination of political hate in Spanish digital media. 2024 IEEE Digital Platforms and Societal Harms, Nueva York, EE. UU.

GONZÁLEZ-AGUILAR, J., VICENT-IBÁÑEZ, M., & PAZ-REBOLLO, M. A. (2024, abril). El humor y la metáfora visual en la polarización política: Los memes sobre Puigdemont. X Congreso Comunicación Política y Sociedades Polarizadas, Universidad Complutense de Madrid.

MARTÍNEZ VALERIO, L. (2021, diciembre). Mensajes de odio hacia la comunidad LGTBIQ+ en los perfiles de Instagram de la prensa española. XIII Congreso Latina de Comunicación Social, Universidad de La Laguna, Tenerife.

MONTERO-DÍAZ, J. (2024, diciembre). Un indicador global sobre el odio en los escenarios digitales. I Congreso Internacional Expresiones de Odio, Medios Digitales y Procesos de Detección (CIOMD), Logroño, España.

MONTERO-DÍAZ, J., & SAID-HUNG, E. (2024, octubre). The Hatemedia Project. Analyzing and monitoring hate expressions in digital environments in Spain. 2024 IEEE Digital Platforms and Societal Harms, Nueva York, EE. UU.

SAID-HUNG, E. (2022, mayo). Sobre discursos de odio: el proyecto HATEMEDIA. Seminario para profesores, Pontificia Università della Santa Croce, Roma, Italia.

SAID-HUNG, E., RÖMER, M., & MONTERO-DÍAZ, J. (2021, noviembre). Discurso de odio y medios digitales en España: Reto de su detección en la web. XIV Congreso Internacional de Ciberperiodismo, Universidad del País Vasco.

SÁNCHEZ, M., & MERINO, A. (2024, mayo). El papel del fact-checking en la detección de bulos estigmatizantes contra inmigrantes en España, Grecia e Italia. XXIX Congreso Internacional de la Sociedad Española de Periodística, Universidad de Valladolid.

VÁZQUEZ DIÉGUEZ, I., & BRÄNDLE, G. (2024, octubre). Opiniones y actitudes de la población española ante los discursos de odio. Seminario online "Los discursos de odio y sus efectos en el Estado de Bienestar", Universidad Complutense de Madrid.

Manuales técnicos y Datasets

DE GREGORIO VICENTE, O., BLANCO, X., RUIZ-INIESTA, A., PÉREZ-PALAU, D., CUBILLAS, J., SAID-HUNG, E., et al. (2024). Informe técnico: desarrollo de algoritmo de clasificación de odio/no odio en medios informativos digitales españoles en X (Twitter), Facebook y portales web. Hatemedia Project. https://doi.org/10.6084/m9.figshare.26085688.v1

RUIZ-INIESTA, A., BLANCO-VALENCIA, X., PÉREZ, D., DE GREGORIO-VICENTE, O., CUBILLAS-MERCADO, J., MONTERO-DÍAZ, J., & SAID-HUNG, E. (2024). Informe final sobre el scrapeo de datos brutos obtenidos en medios informativos digitales españoles en X. Proyecto HATEMEDIA. https://doi.org/10.6084/m9.figshare.25187591.v2

SAID-HUNG, E., MONTERO-DÍAZ, J., BLANCO, X., RUIZ-INIESTA, A., PÉREZ-PALAU, D., DE GREGORIO VICENTE, O., & CUBILLAS, J. (2024). Datos brutos de mensajes publicados en usuarios vinculados a medios informativos digitales españoles en X, Facebook y Web [Dataset]. Hatemedia Project. https://doi.org/10.6084/m9.figshare.25222118.v3

SAID-HUNG, E., MONTERO-DÍAZ, J., RUIZ-INIESTA, A., BLANCO-VALENCIA, X., PÉREZ, D., DE GREGORIO-VICENTE, O., & CUBILLAS-MERCADO, J. (2024). El odio en los medios informativos digitales en España, 2023. Hatemedia Project. https://doi.org/10.5281/zenodo.13964530

SAID-HUNG, E., MONTERO-DÍAZ, J., RUIZ-INIESTA, A., BLANCO-VALENCIA, X., PÉREZ, D., DE GREGORIO-VICENTE, O., & CUBILLAS-MERCADO, J. (2024). Informe de librerías de odio por intensidad y tipos en medios informativos digitales en España. figshare. https://doi.org/10.6084/m9.figshare.26197934.v2

SAID-HUNG, E., MONTERO-DÍAZ, J., RUIZ-INIESTA, A., BLANCO-VALENCIA, X., DE GREGORIO-VICENTE, O., & CUBILLAS-MERCADO, J. (2024). Datos limpios asociados a mensajes recabados en medios informativos españoles en X (Twitter), Facebook y Web, entre 2021 y 2022 [Dataset]. Proyecto HATEMEDIA. https://doi.org/10.6084/m9.figshare.25091603.v1

SAID-HUNG, E., MONTERO-DÍAZ, J., BLANCO, X., RUIZ-INIESTA, A., PÉREZ PALAU, D., DE GREGORIO VICENTE, O., et al. (2024). Dataset usado para entrenamientos de modelos de algoritmos de clasificación de odio, por tipos e intensidades. figshare. https://doi.org/10.6084/m9.figshare.26085700.v1

Otras aportaciones

COORDINACIÓN EDITORIAL

SAID-HUNG, E., MONTERO-DÍAZ, J., & SÁNCHEZ, M. (Coords.). (2024). Coordinación del número especial Expresiones de Odio, Medios Digitales y Procesos de Detección en ICONO14.

RÖMER-PIERETTI, M., & ESTEBAN-RAMIRO, B. (Coords.). (2024). Coordinación del número especial Violence, Hate Speech, and Gender Bias en Social Inclusion, 13.

SAID-HUNG, E., MONTERO-DÍAZ, J., & SÁNCHEZ, M. (Coords.). (2024). Coordinación del número especial Hate Speech and Journalism Practice en Journalism Practice, 28(2).

PODCASTS Y DIVULGACIÓN CIENTÍFICA

SAID-HUNG, E., MONTERO-DÍAZ, J., et al. (2023–2025). Serie de podcasts En pocas palabras, Vicerrectorado de Transferencia de UNIR. Episodios sobre discurso de odio en medios, videojuegos, redes sociales y casos mediáticos.

SAID-HUNG, E., IZQUIERDO, D., & RÖMER-PIERETTI, M. (2022–2024). Entrevistas en medios como RTVE, La Sexta, Cadena SER, Newtral, COPE, The Conversation y Deutsche Welle, entre otros.

EVENTOS ORGANIZADOS

SAID-HUNG, E., MONTERO-DÍAZ, J., & JERÓNIMO, P. (2022). ECREA Pre-conference: News Media, Disinformation and Hate Speech's Promotion. Lisboa.

SAID-HUNG, E., & PAZ-REBOLLO, M. A. (2023). Congreso Internacional Expresiones de Odio, Medios Digitales y Procesos de Detección (CIOMD). Logroño.

MONTERO-DÍAZ, J., ANTONA, T., & SAID-HUNG, E. (2024). Panel MPS – The Role of Local News Media in Promoting Hate Speech and Disinformation.

FORMACIÓN IMPARTIDA

SAID-HUNG, E. (2024). De la discriminación a la convivencia positiva en el aula. Taller de formación para docentes, organizado por CEPAIM.

SAID-HUNG, E. (2023). Redes sociales: desafíos y oportunidades para la juventud. Ponencia en IV Jornadas de la Juventud, Cadena SER Rioja.

Hatemedia

Proyecto PID2020-114584GB-I00

https://www.hatemedia.es/

Financiado por:

MINISTERIO DE CIENCIA E INNOVACIÓN

AGENCIA ESTATAL DE INVESTIGACIÓN