2024 - 2025

Boletín de monitorización del discurso de odio en redes sociales:

reacción y eficacia de las plataformas

Este informe ofrece un análisis sobre la notificación y respuesta ante el discurso de odio en redes sociales durante el 2024 y el primer semestre de 2025. Incluyen datos cuantitativos y cualitativos relativos a la identificación de contenidos racistas, xenófobos, antisemitas, antigitanos e islamófobos, así como la respuesta y eficacia de las cinco plataformas (Facebook, Instagram, X, TikTok y YouTube) ante estas notificaciones que se monitorizan desde el Observatorio Español contra el Racismo y la Xenofobia, el OBERAXE, perteneciente al Ministerio de Inclusión, Seguridad Social y Migraciones.

El objetivo es proporcionar una visión de la respuesta de las plataformas y las tendencias en la moderación de estos contenidos, con el fin de apoyar a la toma de decisiones y al diseño de estrategias que contribuyan a un entorno digital más seguro y respetuoso. Uno de los aspectos clave en el ejercicio de la monitorización es evaluar la respuesta de las plataformas.

La moderación de las plataformas se basa en dos pilares fundamentales: la retirada de contenidos ilegales según la legislación nacional de los Estados miembros de la Unión Europea, conforme a lo dispuesto en la Ley de Servicios Digitales (Digital Services Act, DSA); y la eliminación de contenidos que infrinjan las normas de uso internas de cada plataforma, una acción voluntaria que responde a sus propios compromisos, incluida la adhesión al Código de Conducta europeo.

GOBIERNO DE ESPAÑA
MINISTERIO DE INCLUSIÓN, SEGURIDAD SOCIAL Y MIGRACIONES
SECRETARÍA DE ESTADO DE MIGRACIONES
oberaxe Observatorio Español del Racismo y la Xenofobia
LALIGA
FUNDACIÓN LALIGA
Cofinanciado por la Unión Europea

1.1. Contenidos monitorizados y notificados y reacción de las plataformas de redes sociales en 2024.

En 2024**, se notificaron 2.870 contenidos considerados de odio racista, xenófobo, antisemita, antigitano o islamófobo**, que podían ser constitutivos de delito, de infracción administrativa o que violan las normas de conducta de las plataformas digitales, a las cinco redes sociales monitorizadas (Facebook, X, Instagram, TikTok, y YouTube).

En la distribución de comunicaciones realizadas a cada plataforma (gráfico 1) hay un predominio de X con 758 casos (26% del total), seguido de Facebook con 727 casos (25%), Instagram con 538 (19%), TikTok con 478 (17%), y YouTube con 369 (13%). Este desigual volumen de notificación de contenidos obedece, principalmente, al distinto grado de dificultad para su identificación en cada red social.

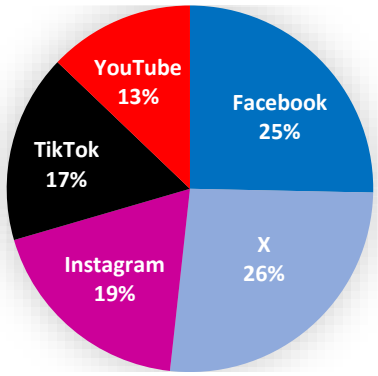

Gráfico 1. Porcentaje de comunicaciones a cada plataforma 2024

Las plataformas retiraron un total de 1.010 contenidos, un 35% de los que les fueron notificados. Del conjunto de contenidos comunicados, **únicamente el 9%** (272) **fueron eliminados cuando la notificación se realizó a través de un perfil de usuario normal**, mientras que el **26% (738) fueron retirados tras ser reportados mediante la figura de *trusted flagger*.** Estos datos evidencian una mayor eficacia en la retirada de contenidos cuando la notificación se realiza a través de canales oficiales o reconocidos como comunicante fiable.

No obstante, **la tasa de retirada del conjunto de las plataformas es mejorable y, en comparación con el año 2023, disminuyó en 14 puntos porcentuales**. Este hecho puede tener una relación causal con los cambios en las políticas de moderación de contenido en las plataformas, en concreto X y META (Facebook e Instagram) que afectan a la retirada de discurso de odio. X ha reducido la eliminación de contenido denunciado, mientras que Meta ha modificado restricciones sobre temas como inmigración y género.

En cuanto a la tasa de retirada con base en las notificaciones realizadas a cada plataforma, **la más eficiente es TikTok que retiró el 69% del total del contenido que se le notificó,** seguida por Instagram (49%), Facebook (29%), YouTube (27%) y X (15%).

1.1.1. Características del contenido retirado a las 24 horas, a las 48 horas y a la semana

La reacción del conjunto de las plataformas en relación con el tiempo de retirada del contenido notificado se muestra en el gráfico 2, donde se observa que la mayoría de los reportes son retirados cuando se notifican como *trusted flagger* (26%). La eficiencia y la rapidez en la retirada de contenidos a las 24h, a las 48h, a la semana, o por la vía de *trusted flagger* se muestra en la tabla 1 y los resultados difieren entre las cinco plataformas, aunque todas ellas son más receptivas a la eliminación de contenido cuando se emplea la vía de *trusted flagger*.

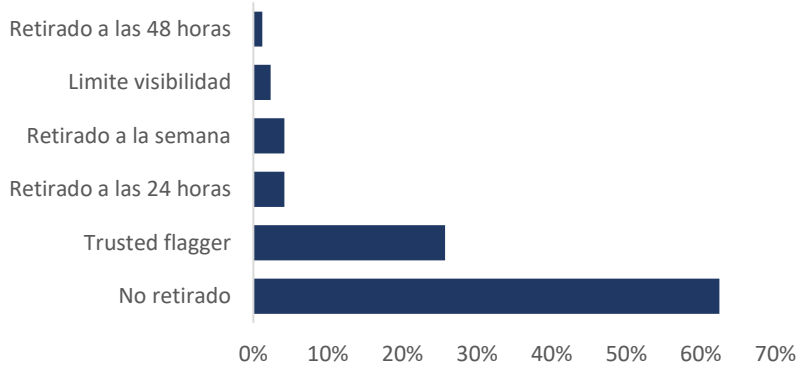

Gráfico 2. Porcentaje de contenidos retirados según el tiempo transcurrido desde la notificación al conjunto de las plataformas monitorizadas, 2024

	Total, contenido retirado	Retirado a las 24 horas	Retirado a las 48 horas	Retirado a la semana	Retirado *trusted flagger*	No retirado
Facebook	29%	0,5%	0,5%	5%	23%	71%
X	14%	4%	0,6%	2,5%	7%	86%
Instagram	49%	3%	2%	7%	37%	51%
TikTok	69%	15%	3%	5%	47%	31%
YouTube	27%	0%	0%	1%	25%	73%
Total	35%	4%	1%	4%	26%	65%

Tabla 1. Porcentaje de contenidos retirados según el tiempo transcurrido desde la notificación y según plataforma, 2024

Los datos evidencian una respuesta poco efectiva por parte de las plataformas ante notificaciones realizadas desde perfiles de usuario normal, especialmente en las primeras 24 y 48 horas, periodos relevantes para minimizar el impacto del discurso de odio. La **baja tasa de retirada inmediata, solo el 4% en 24 horas**, revela un margen considerable de mejora en los sistemas de moderación. Esta reacción por parte de las plataformas facilita que contenidos que deshumanizan, promueven estigmas o incitan a la violencia permanezcan visibles y circulen ampliamente, afectando especialmente a los grupos diana. Estas dinámicas pueden contribuir a la normalización del discurso de odio online, lo que subraya la necesidad de reforzar los mecanismos de moderación y respuesta de las plataformas.

De los contenidos que han sido comunicados como usuario normal, **TikTok es la red social que más contenido retiró en las primeras 24 horas** (15%), seguida de X (4%), Instagram (2%), Facebook (1%), y por último YouTube que solo retiró un 0,3% de sus contenidos en 24 horas.

El 36% de los contenidos retirados a las 24 horas contienen lenguaje que deshumaniza o degrada o que expresa agresividad. El grupo diana mayoritario de los contenidos retirados a las 24 horas ha sido el de las personas originarias del norte de África (43% casos) y predomina la inseguridad ciudadana como episodio prototípico (40%).

El conjunto de plataformas ha retirado el 1% de los contenidos en el plazo de 48 horas desde su notificación. En los contenidos retirados en este plazo, predominan aquellos que deshumanizan o degradan gravemente (53%), seguidos de los que promueven el descredito en base a atributos personales (50%). El episodio prototípico principal sigue siendo la inseguridad ciudadana en un 32% de las notificaciones retiradas, y el grupo diana también es de las personas originarias del norte de África (con un 65% de las notificaciones retiradas).

La plataforma que más contenido ha retirado a las 48 horas ha sido TikTok, que ha eliminado el 3% de los contenidos que se le notificaron, seguida de Instagram (2%), X (1%), Facebook (1%) y YouTube (0,3%).

Respecto al **contenido retirado en el plazo de una semana, el conjunto de las plataformas ha retirado el 4% de las notificaciones, siendo Instagram la plataforma que más contenido ha retirado en este plazo temporal (7%)**, seguida de Facebook (5%), TikTok (5%), X (3%) y YouTube (1%). Cabe resaltar que el 28% de los casos eliminados a la semana no están vinculados a ningún episodio prototípico, y el 59% de las comunicaciones retiradas en ese plazo contienen un lenguaje agresivo explícito.

1.1.2 Contenido retirado como Trusted Flagger

La **vía de *trusted flagger*** continúa consolidándose como el mecanismo más efectivo para la retirada de contenido de discurso de odio por parte de las plataformas. De las 2.870

notificaciones, **un 26% fueron eliminadas tras ser comunicadas mediante esta vía**, en contraste con el 9% de efectividad observada cuando la notificación se realizó desde un perfil de usuario normal. La diferencia en las tasas de retirada según la vía es especialmente significativa, evidenciando que las plataformas dan prioridad a los comunicantes fiables.

Al desagregar los datos por plataforma se aprecian diferencias significativas en el nivel de eficacia en la retirada de contenidos. TikTok es la plataforma más eficiente ante el comunicante fiable con un 47% de contenidos retirados por esta vía. Le siguen Instagram, con un 37%; YouTube, con un 25%; Facebook, con un 23%; y, en último lugar, X, que tiene una tasa de retirada del 7% a través de esta vía.

1.1.3. Características del contenido no retirado en 2024

El porcentaje de contenido no retirado ha sido del 65% (1.860 casos). Este porcentaje incluye un 2% de notificaciones para las que se ha reducido la visibilidad por parte de la red social X, la cual estableció este mecanismo en 2023 como acción positiva para disminuir el efecto del contenido de odio, que, aunque continúa circulando por la red, es menos visible para las personas usuarias.

A pesar de las normas y mecanismos que establecen las plataformas en el marco del Código de Conducta y la normativa establecida por la DSA, la retirada de contenido de discurso de odio sigue siendo insuficiente teniendo en cuenta que en **un 96% de las comunicaciones se infringen las propias normas establecidas por cada una de las plataformas**. El **análisis cualitativo de los 1.860 contenidos no retirados** muestra lo siguiente:

- En un 38% de los casos se promueve el descrédito hacia atributos personales.
- El 36% deshumaniza o degrada gravemente al grupo diana.
- Un 28% incita a la violencia mediante amenazas directas o indirectas.
- Un 17% incita a la expulsión de personas de origen extranjero.
- Predomina la narrativa de la vinculación de la inseguridad ciudadana con los grupos diana.

1.2. Contenidos monitorizados y notificados y reacción de las plataformas de redes sociales en el primer semestre de 2025.

En el periodo entre el 1 de enero y el 25 de julio de 2025, el **monitor FARO ha detectado 531.727 contenidos** de discurso de odio reportables de los cuales se han realizado 1.809 notificaciones a las plataformas[1], y estas **han retirado el 33% de los contenidos reportados**.

En la distribución de comunicaciones realizadas a cada plataforma (gráfico 3) hay un predominio de Instagram con 456 casos (25% del total), seguido de X con 415 casos (23%), Facebook con 322 (18%), TikTok con 315 (17%) y YouTube con 301 (17%). Este desigual volumen de notificación de contenidos obedece, principalmente, al distinto grado de dificultad para su identificación en cada red social.

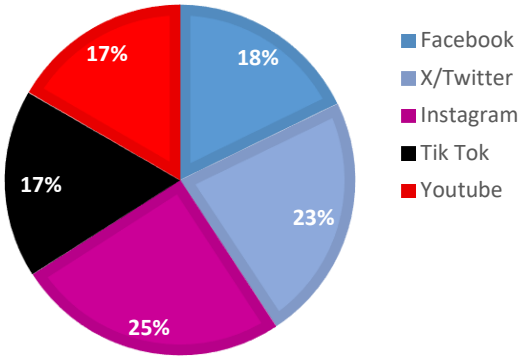

Gráfico 3. Porcentaje de comunicaciones a cada plataforma 2025

Las plataformas en su conjunto han retirado 595 contenidos del total de las 1.809 notificaciones realizadas por el OBERAXE, es decir, el 33%.

Por un lado, se han retirado 186 notificaciones como **usuario normal (10%).** Las plataformas han retirado tan solo un 2% de las notificaciones en un plazo inferior a las 24 horas, un 2% en las 48 horas y un 6% el transcurso de una semana.

A su vez, mediante la **vía *trusted flagger***, empleada este trimestre en 1.623 ocasiones, las plataformas han retirado 409 publicaciones más, lo que supone un **23% adicional**. A continuación, se muestra el gráfico de la reacción del conjunto de las plataformas ante la retirada de contenido:

[1] Al respecto, conviene precisar que los procesos de notificaciones no se realizan de manera automatizada, sino que requieren de una intervención, valoración y tratamiento humanos.

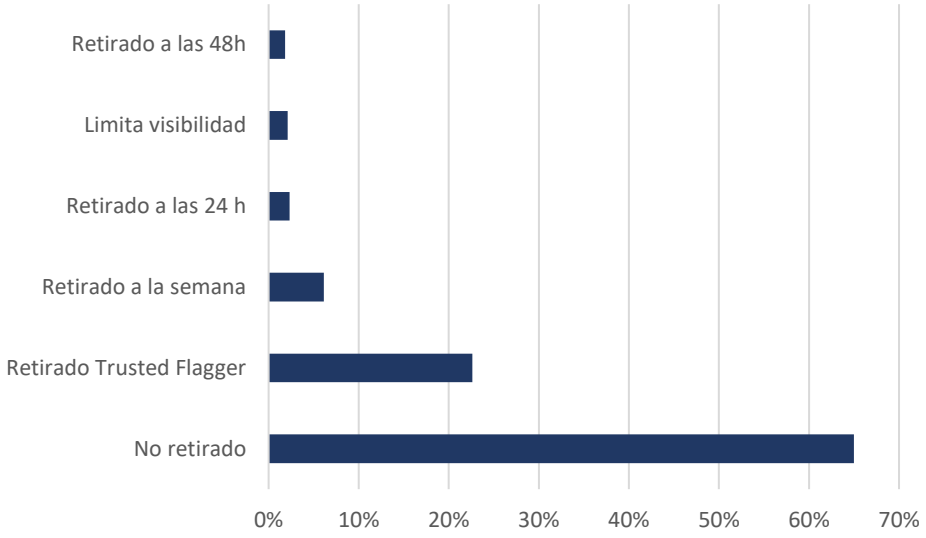

Gráfico 4. Porcentaje de reacción del conjunto de las plataformas ante la retirada de contenido en 2025

Al analizar los contenidos retirados de acuerdo con las comunicaciones realizadas a cada plataforma, se observan diferencias significativas en la respuesta de cada una de ellas. A continuación, se presenta el gráfico 5, que refleja el número de contenidos retirados por cada plataforma durante el periodo analizado. Destaca **TikTok**, que **ha retirado casi la totalidad de los contenidos notificados (89%)**, siendo la plataforma más eficaz ante la retirada de contenido. Facebook e Instagram ocupan posiciones intermedias, y tienen una tasa de retirada del 36% y 32% respectivamente, mientras que X y YouTube reflejan una baja tasa de retirada (del 9% y 5% respectivamente). Estos porcentajes se calculan sobre el total de contenidos notificados a cada plataforma, reflejando así la eficacia relativa de retirada en cada caso.

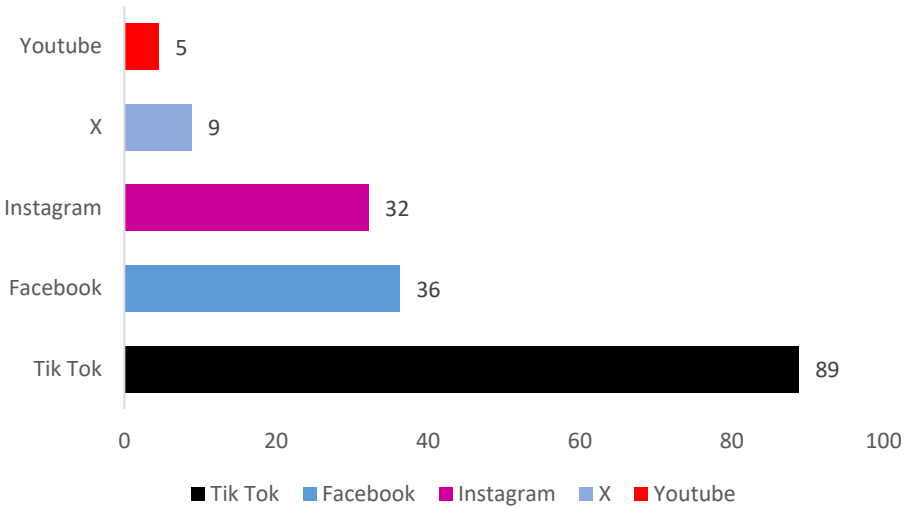

Gráfico 5. Porcentaje de retirada de contenido de cada plataforma 2025

Por otro lado, al analizar el plazo de retirada de contenido notificado a las plataformas, se observan patrones diferenciados en función de la vía de retirada (usuario normal o *trusted flagger*) y los tiempos de retirada. **A pesar de que X tiene una tasa de retirada muy baja (9%), ha retirado el 22% del contenido en las 24 horas desde su notificación** y un 54% en el plazo de una semana, lo que sugiere una gestión basada en criterios de inmediatez. En el caso de **YouTube, la práctica totalidad de los contenidos retirados (86%) se produce exclusivamente a través de la vía *trusted flagger*,** evidenciando que solo es efectiva cuando se emplea el comunicante fiable. Facebook muestra un comportamiento similar, con un 84% de retiradas vía *trusted flagger* y mínimos porcentajes en las primeras 24 y 48 horas (3% y 1%, respectivamente). Por su parte, TikTok muestra una combinación de efectividad y rapidez: un 27% de los contenidos son retirados en plazos inferiores a una semana, mientras que el 73 % restante se elimina a través de la vía *trusted flagger*. Por último, Instagram presenta una distribución más equilibrada, con una retirada del 12 % en las primeras 24 horas, 29% en el plazo de una semana y un 59% adicional mediante la vía *trusted flagger*.

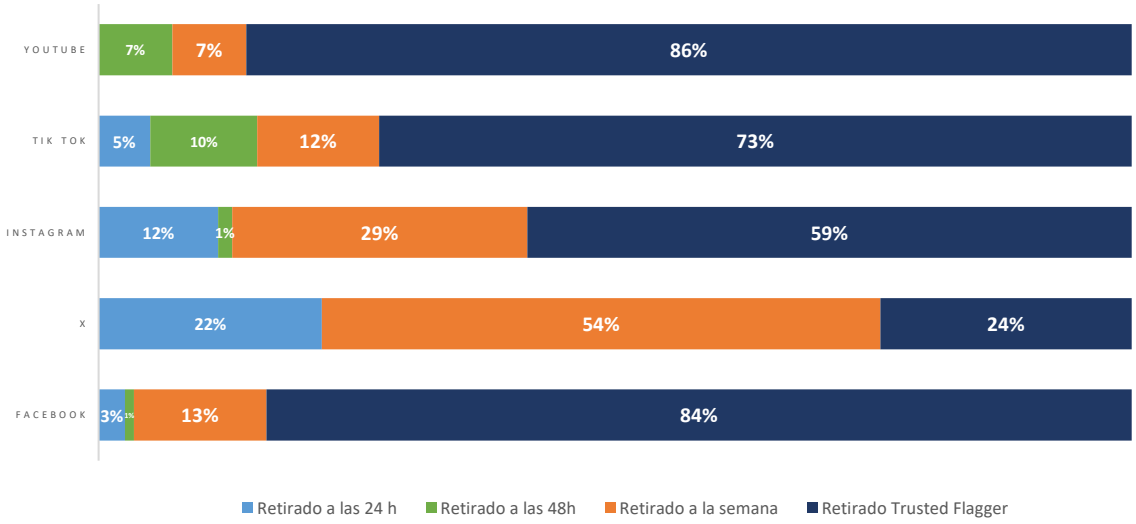

Gráfico 6. Plazo de retirada de contenido de cada plataforma 2025

1.2.1. Características del contenido no retirado 2025

El análisis del contenido que no ha sido retirado por el conjunto de las plataformas (67%) en 2025 revela una alta carga de hostilidad y estigmatización hacía las personas de origen extranjero. El 53% deshumaniza o degrada gravemente a personas migrantes, mientras que el 36% promueve el descrédito en base a atributos personales y el 26% incita directamente a la expulsión de los grupos diana. Además, el 26% presenta a los diferentes grupos diana como una amenaza para la ciudadanía. Por otro lado, el 43% de los casos no retirados están vinculados a narrativas de inseguridad ciudadana, y el 40% contiene un lenguaje agresivo explícito.

1.3. Conclusiones

- **Volumen y persistencia del discurso de odio**. En 2024 se notificaron cerca de 3.000 contenidos de discurso de odio racista, xenófobo, antisemita, antigitano o islamófobo, al conjunto de las plataformas, muchos de ellos con potencial de constituir delito o infracción. Asimismo, **en el primer semestre de 2025 se ha detectado por el monitor FARO cerca de medio millón de contenidos**. Estos datos evidencian la persistencia y gravedad del discurso de odio en las principales redes sociales.

- **Baja tasa de retirada inmediata**: **la moderación efectiva y rápida es clave para limitar el alcance del discurso de odio**. Sin embargo, actualmente la mayoría de las plataformas cuentan con un amplio margen de mejora para la retirada, especialmente en las primeras 24 y 48 horas tras la notificación. **Solo el 2% de los contenidos notificados por el OBERAXE en 2025 han sido retirados en las primeras 24 horas, un período crucial para minimizar su impacto**. Esta circunstancia facilita que mensajes que deshumanizan, estigmatizan o incitan a la violencia permanezcan visibles, afectando especialmente a los diferentes grupos diana como son las personas originarias del norte de África, y contribuyendo así a la normalización del discurso de odio online.

- **Diferencia en la eficacia según la vía de notificación**: **las plataformas son más efectivas cuándo se emplea la vía de comunicante fiable (*trusted flagger*)**. En el primer semestre de 2025, se ha retirado a través de esta vía el 23% del contenido notificado, y tan solo el 10% ha sido eliminado tras la notificación como usuario normal. Se evidencia la necesidad de fortalecer la denuncia a través de usuario normal y promover su uso por parte de la ciudadanía.

- **Variabilidad en la eficacia según plataforma y políticas internas**. TikTok destaca con una tasa de retirada del 69% en 2024 y casi el 90% en 2025, mientras que otras plataformas han reducido su eficacia tras cambios en sus políticas de moderación, presentando un desafío para garantizar un entorno digital seguro y respetuoso.

- **Compromiso de mejorar los mecanismos de moderación**. Aunque existen marcos normativos vigentes y la adhesión al Código de Conducta de la UE, **el 67% del contenido denunciado sigue visible**. Esta cifra pone de manifiesto la necesidad urgente de **fortalecer la colaboración entre autoridades y plataformas para optimizar los sistemas de detección y moderación**. Para avanzar hacia un entorno digital más seguro es imprescindible intensificar la cooperación entre todos los actores involucrados (plataformas, autoridades y sociedad civil), mejorar las tecnologías de detección y moderación, y adaptar tanto los marcos normativos como los procedimientos internos a la evolución constante del discurso de odio en línea, garantizando así una respuesta más efectiva y protectora para las personas migrantes.