BOLETÍN DE MONITORIZACIÓN DEL DISCURSO DE ODIO EN REDES SOCIALES

2024

1 SEPTIEMBRE 31 OCTUBRE

El Observatorio Español del Racismo y la Xenofobia (OBERAXE) colabora desde 2017 con la Comisión Europea en los ejercicios de monitorización que se realizan en cumplimiento del *Código de Conducta para la lucha contra la incitación ilegal al odio en Internet*, que fue firmado con las empresas prestadoras de servicios de alojamiento de datos con mayor presencia en la Unión Europea.

Desde mayo de 2020, el OBERAXE monitoriza a diario el discurso de odio en España en cinco redes sociales (X, Facebook, YouTube, Instagram y TikTok) y notifica aquellos contenidos considerados de odio racista y/o xenófobo, antisemita, antigitano o islamófobo que pueden ser constitutivos de delito, infracción administrativa o que violan las normas de conducta de las plataformas.

En el periodo entre el 1 de septiembre y el 31 octubre de 2024 se han realizado 336 notificaciones a las plataformas, las cuales han retirado el 25% de los contenidos. El 38% de los contenidos de discurso de odio notificados no está vinculado a un acontecimiento concreto; aunque la inseguridad ciudadana sigue suscitando una parte importante de los mensajes de odio (30%), dirigiéndose principalmente a las personas del norte de África y a las personas inmigrantes en general. Asimismo, se han detectado contenidos que promueven actitudes y narrativas antiinmigratorias.

CONTENIDOS ANALIZADOS

338 contenidos de discurso de odio notificados

El número total de contenidos de discurso de odio notificados es de 338. En la Figura 1 se observa cierto predominio en el número de contenidos notificados a X con 122 casos comunicados, seguido de Instagram (66), Facebook (61), YouTube (47), y TikTok (42). El desigual volumen de contenidos notificados obedece a los diferentes usos y características propias de cada red social.

El conjunto de las plataformas ha retirado un cuarto de los contenidos notificados

Las plataformas en su conjunto han retirado 85 contenidos del total de las 338 notificaciones realizadas por el OBERAXE, es decir, el 25%; con un descenso de doce puntos porcentuales respecto al bimestre anterior.

Por un lado, se han retirado 41 contenidos notificados como usuario normal, lo que supone solo un 12% sobre el total. A su vez, mediante la vía *trusted flagger*, empleada este bimestre en 297 ocasiones, las plataformas han retirado 44 publicaciones más, lo que supone un 13% adicional.

Facebook es la plataforma con mayor tasa de retirada de contenidos de discurso de odio (51%)

Al analizar los contenidos retirados por cada plataforma en relación con el total de casos notificados a cada una de ellas (Figura 2), Facebook ha sido, en este periodo, la red social con mayor tasa de retirada (51%), incrementándose en cuarenta y dos puntos porcentuales respecto al periodo anterior (9%).

La siguen Instagram (35%), que empeora su tasa en treinta puntos porcentuales, TikTok (33%) y X (13%). Por el contrario, YouTube en este bimestre solo ha retirado el 2% de las comunicaciones realizadas a su plataforma.

*Conforme a la RECOMENDACIÓN (UE) 2018/334 DE LA COMISIÓN de 1 de marzo de 2018 sobre medidas para combatir eficazmente los contenidos ilícitos en línea se entiende como "comunicante fiable" (*trusted flagger*): aquella persona física o jurídica que un prestador de servicios de alojamiento de datos considere que posee competencias y responsabilidades particulares a efectos de la lucha contra los contenidos ilícitos en línea.

FIGURA 1.
CONTENIDOS COMUNICADOS

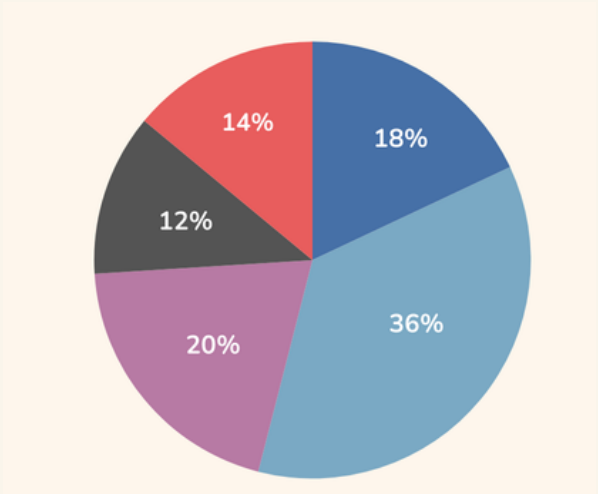
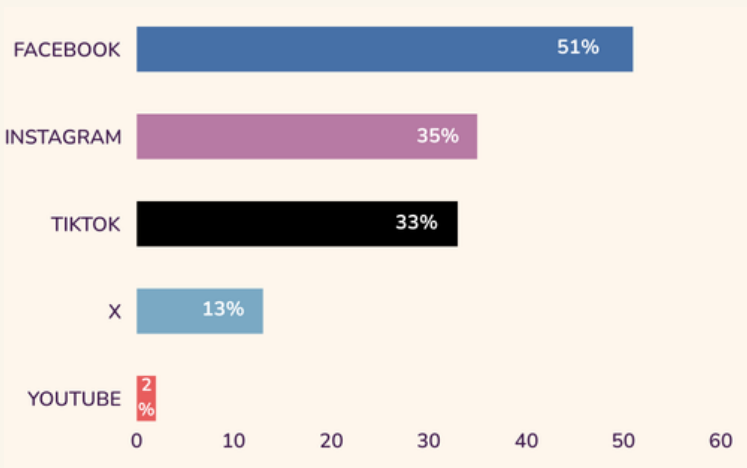


FIGURA 2.
% CONTENIDOS RETIRADOS



INSTAGRAM FACEBOOK X TIKTOK YOUTUBE

EVOLUCIÓN EN LA RETIRADA DE CONTENIDOS

La Figura 3 presenta los plazos en los que las plataformas retiran el contenido que se les ha notificado*: 24 horas, 48 horas o una semana. Asimismo, incorpora el número de contenidos retirados tras la comunicación como *trusted flagger*.

TikTok ha retirado el 86% de las comunicaciones en las primeras 24 horas

Del total de contenidos comunicados al conjunto de las plataformas, se ha retirado el 8% de las notificaciones a las 24 horas, el 1% a las 48 horas, y el 3% a la semana.

Al analizar los contenidos retirados en base a las notificaciones realizadas a cada plataforma, TikTok ha eliminado el 86% de sus casos en un plazo inferior a las 24 horas; X, el 38%; Instagram, el 30%; y Facebook, el 6%.

Por el contrario, YouTube, aunque solo ha retirado un contenido de los cuarenta y siete que les han sido notificados, lo ha eliminado en un periodo inferior a las 24 horas.

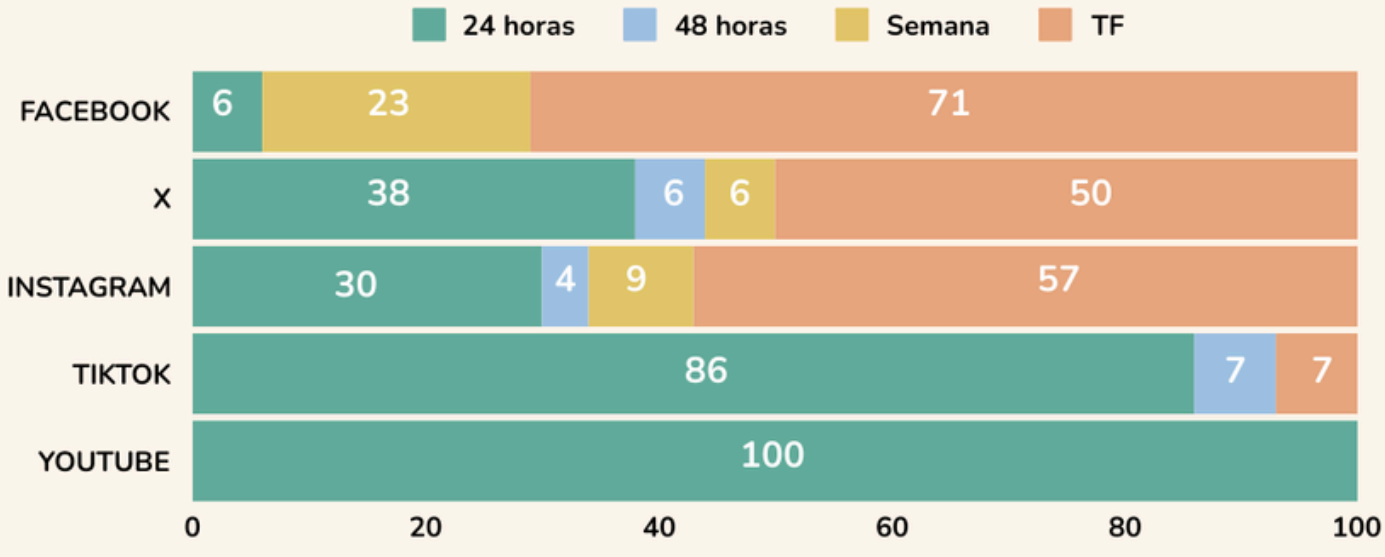
El 71% de los contenidos retirados en Facebook responde a notificaciones realizadas vía *trusted flagger*

Facebook ha retirado el 71% del total de los contenidos que ha eliminado tras su comunicación como *trusted flagger*. En la tasa de retirada a través de esta vía le siguen X (50%), Instagram (57%) y TikTok (7%).

La proporción de casos retirados vía *trusted flagger* en comparación con los retirados a través de la notificación como usuario normal, muestra la importancia de la figura de los comunicantes fiables en la identificación y acción sobre los contenidos de discurso; así como la necesaria mejora por parte de las plataformas en la moderación de los contenidos reportados por los usuarios comunes a través de los mecanismos habilitados en las propias redes.

*Si transcurrida una semana, la plataforma no retira el contenido denunciado como usuario normal, se procede a utilizar el procedimiento de comunicante fiable o *trusted flagger.*

FIGURA 3.
% TIEMPO DE RETIRADA DE CADA PLATAFORMA



El 23% de los contenidos incita a la violencia mediante amenazas

Las tipologías de contenido de discurso de odio más predominantes en este bimestre son el descrédito en base a atributos personales o del grupo (38%); y, en segundo lugar, la deshumanización o degradación (37%). Asimismo, un 26% de los mensajes presentan al grupo como una amenaza, fomentando el alarmismo y afectando de manera negativa a la cohesión social.

Por otra parte, el 23% de los contenidos analizados incitan a la violencia a través de amenazas directas o indirectas, de forma que contribuyen a generar un ambiente de hostilidad que puede desembocar en la comisión de actos violentos. Además, en un 11% de los mensajes se promueve la expulsión de las personas de origen extranjero.

El 48% de los mensajes contienen un lenguaje agresivo explicito

El lenguaje agresivo explícito, con el uso de insultos y otras expresiones agresivas, es el más común en los mensajes de discurso de odio, representando el 48% del total de contenidos notificados.

Por otro lado, el lenguaje discriminatorio no agresivo también es un componente relevante en el discurso (44%), con mensajes no necesariamente violentos pero sí de carácter discriminatorio (Figura 4).

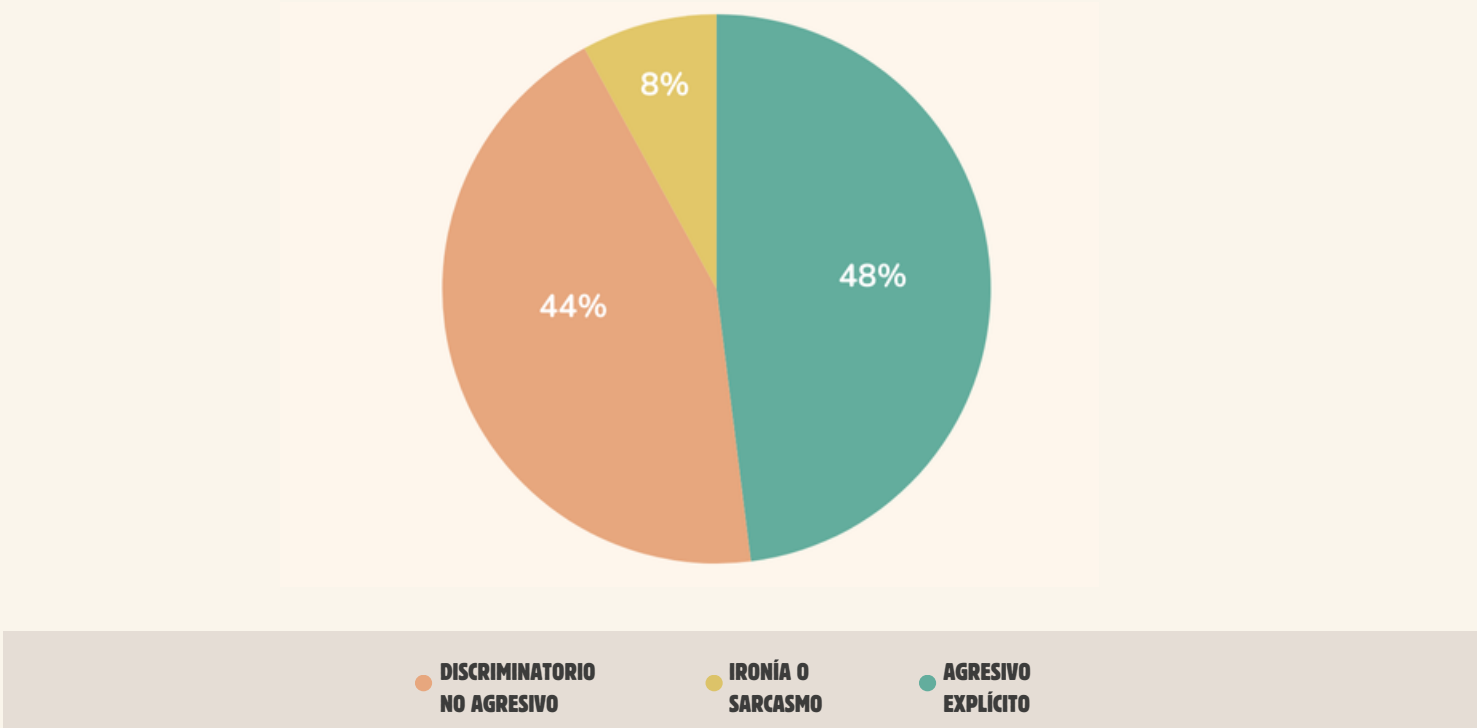
A su vez, el 8% de los mensajes presentan un tono irónico o sarcástico, empleado como recurso para difundir de forma velada contenidos de carácter racista y/o xenófobo, dificultando su detección por parte de los sistemas de moderación de las plataformas al no incluir de manera directa expresiones explícitas o no permitidas por las plataformas.

El uso de emojis refleja una estrategia de adaptación del lenguaje agresivo

Aunque la gran mayoría de los mensajes se manifiestan exclusivamente en formato de texto (79%), un porcentaje considerable (15%) integra imágenes, videos o memes, promoviendo directamente discurso de odio mediante medios visuales, lo que contribuye a obtener un mayor alcance en las redes.

Además, la inclusión de emojis como elementos necesarios en el significado del mensaje (6%) refleja una estrategia de adaptación para difundir un lenguaje agresivo de forma velada.

FIGURA 4.
% TIPO DE LENGUAJE EN EL DISCURSO DE ODIO



El discurso de odio se dirige en primer término hacia las personas del norte de África

Según el análisis de los datos sobre los grupos diana (Figura 5) a los que se dirige el discurso de odio, el grupo que mayor porcentaje de discurso de odio recibe es el conformado por las personas con origen en el norte de África (35%), seguido de las personas inmigrantes en general (26%), las personas africanas y afrodescendientes (22%); y las personas musulmanas (19%).

El 38% de los contenidos no se vinculan a un acontecimiento concreto

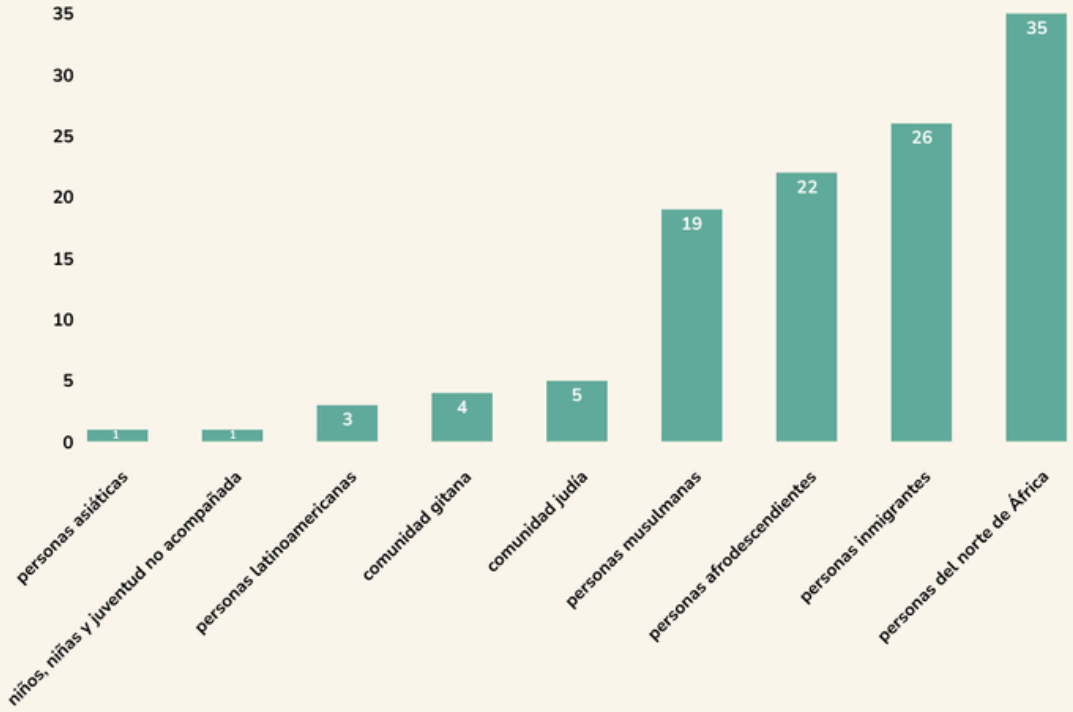
Durante este periodo no se ha identificado un tema específico que se configure como el principal detonante del discurso de odio. En este sentido, el 38% de los contenidos identificados promueven mensajes de discurso de odio sin estar vinculados a ningún acontecimiento o ámbito temático.

La inseguridad ciudadana sigue siendo el episodio prototípico principal (30%)

La inseguridad ciudadana es el episodio prototípico asociado al discurso de odio con mayor frecuencia (30%), a través de mensajes que atribuyen la comisión de incidentes como robos y agresiones a inmigrantes y personas procedentes del norte de África, principalmente, a las que se criminaliza.

El 47% de los contenidos clasificados vinculados al episodio prototípico de inseguridad ciudadana promueven la percepción de amenaza pero no se refieren a hechos verídicos, actuales y producidos en España; y están descontextualizados o basados en información falsa. Otros de los episodios que han suscitado el discurso de odio son la llegada de embarcaciones y las políticas públicas. Además, se han identificado contenidos que difunden narrativas antiinmigración; y que perciben la llegada de inmigrantes como una amenaza para la sociedad española y europea.

FIGURA 5.
% COLECTIVOS VICTIMIZADOS EN EL DISCURSO DE ODIO



Edita y distribuye: Observatorio Español del Racismo y la Xenofobia.

NIPO: 121-23-006-3