NOTA METODOLÓGICA

SISTEMA FARO

Filtrado y Análisis de Odio en las Redes Sociales

Monitorización del discurso de odio racista y/o xenófobo, islamófobo, antisemita y antigitano en redes sociales

Versión 1, actualizada a fecha de 11 de marzo de 2025.

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

ÍNDICE

1. Introducción

La monitorización del discurso de odio realizada por el **Observatorio Español del Racismo y la Xenofobia (OBERAXE)** desde el año 2020 consiste en la identificación, análisis y notificación a las plataformas de contenidos de discurso de odio[1] con motivación racista, xenófoba, islamófoba, antisemita y antigitana, publicados en cinco plataformas de redes sociales (Facebook, Instagram, TikTok, YouTube y X); y que puedan ser constitutivos de delito, de infracción administrativa, o que infrinjan las normas de uso de las propias plataformas de prestación de servicios digitales.

Los objetivos principales de la monitorización son dos. Por una parte, obtener una fotografía de la situación del racismo y la xenofobia en España y su evolución que contribuya a inspirar el desarrollo de políticas públicas en favor de la inclusión y de los derechos humanos de las personas de origen migrante y de los diferentes colectivos relacionados[2]. Por otra, la notificación de los contenidos de discurso de odio a las plataformas de redes sociales solicitando su retirada, con el propósito de evaluar la respuesta de las plataformas ante los contenidos de discurso de odio.

La base de la metodología de monitorización parte inicialmente del modelo establecido en los ejercicios de evaluación del cumplimiento del *Código de Conducta para la lucha contra la incitación ilegal al odio en Internet,* firmado en 2016 por la Comisión Europea junto con las plataformas de prestación de servicios digitales; y renovado en 2025 a través del lanzamiento del *Código de Conducta +*[3], que refuerza el compromiso adquirido por las plataformas para hacer frente al discurso de odio en internet complementando la aplicación de la *Ley de Servicios Digitales (DSA)*.[4]

Partiendo del precedente del Código de Conducta (2016), el OBERAXE llevó a cabo en 2020 una monitorización específica impulsada por la necesidad de analizar el discurso de odio contra las personas asiáticas que se estaba produciendo en el contexto de la irrupción de la pandemia de COVID-19. Tras esta experiencia, se decidió ampliar el análisis al discurso de odio con motivación racista, xenófoba, islamófoba, antisemita y antigitana; y continuar con una monitorización diaria de las redes sociales.

En un primer proceso de desarrollo de la metodología y de los recursos disponibles para la recogida de datos, en 2022 se empezó a hacer uso de ALERTODIO, aplicación de desarrollo propio destinada al registro de los contenidos identificados en las redes. Desde ese momento se abrió un periodo de mejora y consolidación metodológica que se ha visto reflejada en la elaboración de boletines e informes a través de los cuales se presentan los datos obtenidos, con periodicidad bimestral (hasta 2024 incluido), mensual (desde la segunda mitad de 2024), trimestral (desde 2025) y anual (primera edición publicada en 2024 con los datos obtenidos a lo largo de 2023).

El 24 de octubre de 2024 el **Ministerio de Inclusión, Seguridad Social y Migraciones**, y **LALIGA**, firmaron un convenio de colaboración [5] para reforzar su compromiso en la lucha contra el discurso de odio en el deporte. Uno de los puntos centrales del convenio es la cesión al Ministerio, por parte de LALIGA, de la herramienta MOOD (Monitor para la Observación del Odio en el Deporte)[6]. Tal y como se detalla a continuación, esta herramienta ha sido adaptada al objeto de estudio del OBERAXE e incorpora el uso de la inteligencia artificial para la identificación y análisis a tiempo real de contenidos de discurso de odio en las redes sociales; lo que permite trabajar con un mayor volumen de contenidos contribuyendo a alcanzar un análisis más preciso y detallado de la realidad del discurso de odio en España.

---

[1] Se toma de referencia la definición de discurso de odio establecida por la Comisión Europea contra el racismo y la Intolerancia del Consejo de Europa (ECRI) en la *Recomendación General nº 15 relativa a la lucha contra el discurso de odio*.

[2] Consultar apartado 3.6 (Grupos diana).

[3] El *Código de Conducta para la lucha contra la incitación al odio ilegal en línea +*, lanzado el 25 de enero de 2025, fue firmado por la Comisión Europea y las plataformas (Meta, X, Google, TikTok, LinkedIn, Twitch, Viber, Dailymotion, jeuxvideo.com, Snap Inc. y Microsoft).

[4] La *Ley de Servicios Digitales* (DSA), de aplicación desde el 17 de febrero de 2024, es un reglamento europeo que integra un conjunto de normas y obligaciones aplicables en la UE a los proveedores de servicios digitales, con el objetivo de garantizar una mayor protección de los derechos fundamentales de todos los usuarios.

[5] El Convenio refuerza el compromiso de ambas entidades en la lucha contra el discurso de odio y permite desarrollar acciones específicas para monitorizar, sensibilizar y prevenir el discurso de odio y el racismo tanto en redes sociales como en el ámbito deportivo.

[6] El Monitor para la Observación del Odio en el Deporte (MOOD) es un monitor semanal que audita el nivel de odio y racismo que se produce en las redes sociales en torno a LALIGA.

2. Consideraciones generales

2.1. Sistema FARO

El **Sistema FARO (Filtrado y Análisis de Odio en las Redes Sociales)** es la metodología que emplea el OBERAXE, desde su lanzamiento en marzo de 2025, para la identificación y análisis a tiempo real de los contenidos de discurso de odio con motivación racista, xenófoba, islamófoba, antisemita y antigitana, publicados en cinco plataformas de redes sociales (Facebook, Instagram, TikTok, YouTube y X[7]).

El desarrollo del Sistema FARO nace de la adaptación específica de la herramienta MOOD al objeto de estudio de la monitorización del OBERAXE, incorporando el uso de la tecnología de la inteligencia artificial al conocimiento y la experiencia que atesora el OBERAXE en la monitorización del discurso de odio en las redes sociales.

El Sistema FARO es la conjunción del uso de dos herramientas (Monitor FARO y ALERTODIO) aunadas en una nueva metodología de trabajo que atañe tanto a la identificación de contenidos como al análisis y presentación de los resultados a través de un monitor de visualización.

2.1.1. Monitor FARO desarrollado por Séntisis Intelligence

Por una parte, se cuenta con el **monitor FARO**, que constituye la adaptación específica de la herramienta MOOD al objeto de estudio del OBERAXE. El monitor FARO permite, de manera automatizada y a través de la incorporación de inteligencia artificial, la captación de contenidos de discurso de odio en las redes sociales y su categorización en tiempo real en base a las variables empleadas en el análisis. El desarrollo técnico del monitor ha sido realizado a través de Séntisis Intelligence.

2.1.2. ALERTODIO

Por otra, en lo referente al reporte de los contenidos a las plataformas y a la evaluación de las respuestas de estas, se emplea la herramienta **ALERTODIO** que permite registrar los contenidos notificados y llevar un seguimiento de las acciones tomadas por las plataformas ante las notificaciones remitidas (limita la visibilidad, contenido retirado a las 24 horas, a las 48 horas, a la semana, retirado a través de la vía cualificada de los alertadores de seguimiento; y contenido no retirado).

2.2. Fases de monitorización a través del Sistema FARO

Las fases de monitorización a través del Sistema FARO son: captación y filtrado de los contenidos; análisis de los contenidos; reporte de los contenidos y seguimiento de las plataformas; y presentación de los datos.

2.2.1. Captación y filtrado de los contenidos

El monitor FARO se conecta a las API oficiales de las distintas plataformas de forma directa o a través de proveedores autorizados para detectar en tiempo real el discurso de odio. La búsqueda de la información se realiza a través de *queries*[8] o búsquedas específicas que dependen de cada plataforma. Estas búsquedas incorporan palabras clave, combinaciones de términos (*queries* booleanas[9]), expresiones complejas; y canales y perfiles específicos.

Séntisis Intelligence cuenta con una tecnología propia de análisis del lenguaje especializada en el idioma español, que combina algoritmos avanzados y técnicas de inteligencia artificial con más de 100.000 reglas semánticas para poder filtrar la información y clasificarla según los criterios establecidos por el OBERAXE.

[7] La red social Twitter cambió de nombre y pasó a llamarse X el 24 de julio de 2023.

[8] Términos o conceptos que se escriben en los buscadores al realizar una búsqueda por palabra clave.

[9] La búsqueda booleana se realiza utilizando una combinación de palabras clave y los operadores booleanos principales (AND, OR y NOT).

La captación y filtrado de contenidos de discurso de odio en las plataformas de redes sociales de forma automatizada actualmente se ciñe a datos públicos y está basada fundamentalmente en el análisis textual, aunque se está trabajando para ampliar el proceso de identificación y filtrado automatizado a las publicaciones basadas íntegramente, o de manera principal, en imagen, video y audio, de forma paralela al avance de los medios tecnológicos que permitirán este desarrollo.

Conviene aclarar algunos conceptos que se emplean tanto en la nota metodológica como en la visualización de los datos:

- **Mensajes detectados**

Dentro del total de mensajes captados cada día, la categoría de "mensajes detectados" hace referencia a la suma de los contenidos filtrados a través de dos capas: por una parte, figuran los contenidos de discurso de odio, entendiendo el concepto de una manera estricta (como se detalla a continuación) y; por otra, se añade una capa más amplia que integra los contenidos de discurso odioso y las narrativas discriminatorias.

- **Contenidos de discurso de odio**

En el concepto de discurso de odio en la metodología aplicada en la monitorización se refiere a los mensajes y contenidos publicados en las redes sociales que son potencialmente ilegales (ya que pueden incurrir en delito o infracción administrativa) y/o que infringen las normas comunitarias de las plataformas de redes sociales. En base al concepto teórico, en el filtrado de estos mensajes es una condición fundamental que el contenido haga referencia o se dirija de manera directa a las personas que constituyen un grupo diana (p.ej: personas migrantes, personas del norte de África, personas musulmanas) y, a su vez, que exista una motivación de odio; es decir, que las personas del grupo identificado sean atacadas en base a sus condiciones personales o estados (origen étnico, procedencia nacional, religión, etc.) y no por cualquier otro motivo.

De esta capa de filtrado, la más reducida del análisis, se selecciona la muestra de casos sobre la que se realiza el reporte a las plataformas de redes sociales.

- **Contenidos de discurso odioso y narrativas discriminatorias**

Los contenidos de discurso odioso hacen referencia a los mensajes y publicaciones que, si bien resultan discriminatorias e intolerantes, no son consideradas ilegales y pueden quedar amparadas bajo el paraguas de la libertad de expresión.

Esto puede ocurrir en contenidos en los que no se identifica al grupo diana de manera inequívoca; en los que, debido al uso u omisión de determinadas palabras o expresiones, ya sea de forma deliberada o involuntaria, la motivación de odio pueda resultar ambigua a pesar de contener elementos violentos, de deshumanización o intolerancia contra las personas; o en los que la incitación al odio, si bien contribuye a alimentar el clima de odio y de hostilidad, no cuenta con la gravedad o intensidad suficiente para ser considerado ilegal o ilícito en virtud de las políticas de uso de las plataformas.

Asimismo, dentro de esta categoría se incluyen los contenidos en los que, si bien no se identifica un ataque directo a las personas, sí se ataca o se hace referencia de manera explícita, a través de elementos de incitación a la violencia o discriminación, a instituciones asociadas a los grupos victimizados, a las religiones, a lugares de culto u a otros elementos o conceptos vinculados a la realidad de los diferentes colectivos.

Además, se tienen en cuenta contenidos que promueven y/o apoyan políticas o acciones antiinmigración que atentan contra los derechos humanos; así como mensajes de enaltecimiento del nazismo o que incluyan referencias supremacistas; y publicaciones que difunden teorías o informaciones, vinculadas con la desinformación, que estigmatizan a las personas de origen migrante y de los diferentes grupos diana, y fomentan la discriminación y la negación de los derechos humanos de las personas que conforman el grupo.

2.2.2. Análisis de los contenidos

El análisis de los contenidos consta de dos fases consecutivas: el análisis automatizado a través del monitor FARO y la revisión humana por parte del equipo de monitorización.

- **Análisis automatizado**

Esta primera fase del análisis se realiza de manera automatizada a través del monitor FARO partiendo del total de mensajes captados.

Los miles de mensajes captados cada día se analizan en tiempo real utilizando un modelo de clasificación con más de 100 categorías desarrolladas específicamente para el sistema FARO. Estas categorías se basan en técnicas de **inteligencia artificial** y reglas semánticas para determinar qué temáticas, emociones e intenciones hay detrás de cada mensaje.

El resultado es un modelo multi-categoría en el que se identifican para cada mensaje distintos atributos, tales como: grupo diana, episodio que suscita el discurso de odio, tipo de contenido del mensaje, expresión del lenguaje y género.

- **Revisión humana**

La primera fase de análisis, ejecutada a través de procesos automatizados, va sucedida de un análisis manual por parte del equipo de monitorización de OBERAXE que posibilita la revisión de los mensajes identificados y clasificados a través del monitor FARO con dos objetivos principales.

El primer objetivo es trabajar sobre una muestra mensual de los contenidos de discurso de odio detectados por el monitor FARO y entrenar a la herramienta con el fin de ir alcanzando cada vez un mayor nivel de precisión en el análisis a tiempo real realizado a través de la incorporación de la IA.

El segundo objetivo es efectuar un análisis detallado y pormenorizado de cada uno de los contenidos de la muestra. Cada uno de los contenidos de la muestra es analizado por una persona integrante del equipo de monitorización y revisada por uno de los responsables del mismo. De esta manera, la revisión humana permite, al mismo tiempo, entrenar a la IA y poner el foco en el análisis de los contenidos de una manera detallada, haciendo uso del sistema jerarquizado de categorías y subcategorías establecidas en las diferentes variables del análisis.

La **muestra** se compone de un conjunto de contenidos captados por el monitor FARO y clasificados por la propia herramienta en la categoría de discurso de odio. La muestra contará con un mismo tamaño en cada uno de los periodos mensuales, aunque puede sufrir cambios en función de las adaptaciones técnicas y metodológicas, así como de los recursos disponibles para la captación, análisis, reporte de los contenidos y seguimiento de los casos notificados.

Para 2025, primer año de vigencia del Sistema FARO, se prevé una muestra anual de 3.500 contenidos, dividida en partes iguales (por periodos mensuales), sobre las que se realizarán las labores de revisión, análisis y reporte a las plataformas. El número total de contenidos de discurso de odio notificados por el OBERAXE a lo largo del año 2024 fue de 2.871, por lo que el tamaño de la muestra anual inicial supondría un aumento de más del 20% sobre la cifra de los contenidos reportados en el año anterior. No obstante, se contempla la posibilidad de realizar un aumento en el tamaño muestral a medida que se realicen las adaptaciones metodológicas necesarias.

Al margen de la revisión y análisis de los contenidos de la muestra, el **factor humano** interviene en el trabajo de identificación y análisis del discurso de odio de manera permanente, principalmente a través de dos vías.

La primera de estas vías son los ajustes y mejoras continuas en el monitor FARO para perfilar los procesos de captación y filtrado de los mensajes, ya sea a través de la introducción de nuevas palabras y expresiones clave, nuevos perfiles para ampliar la escucha, o modificaciones en las reglas internas para precisar la categorización de los contenidos.

La segunda vía es la observación diaria de las redes sociales realizada por las personas del equipo de monitorización, una labor que permite de primera mano la identificación de las tendencias actuales asociadas

al discurso de odio en internet, como la creación de nuevas expresiones y *dog whistle*[10] vinculadas a este tipo de discurso, el surgimiento de nuevos perfiles relevantes en la difusión de discurso de odio; y los nuevos usos, significados y temas predominantes en las redes, una cuestión esencial en un escenario de innovación y transformación constante.

La observación diaria de los contenidos en redes sociales permite, a su vez, entrenar el monitor FARO y alimentarla con nuevos términos, enfoques, significados y particularidades asociadas al discurso de odio a medida que van surgiendo.

2.2.3. Reporte de los contenidos y seguimiento de las plataformas

El **reporte a las redes sociales** consiste en la notificación de contenidos de discurso de odio a las plataformas con el propósito de solicitar su retirada y, al mismo tiempo, evaluar cómo es su respuesta ante dicho reporte y qué acciones despliega para abordar el contenido notificado.

Los contenidos reportados a las plataformas se obtienen de la muestra seleccionada tras haber sido captados y filtrados a través del monitor FARO. Se reportan los casos obtenidos de la muestra que cumplen efectivamente las condiciones para ser reportados, por lo que deben ser contenidos de discurso de odio constitutivos de delito, de infracción administrativa, o que infrinjan las normas de uso de las propias plataformas de prestación de servicios digitales.

El procedimiento de **seguimiento de la respuesta de las plataformas** dentro de la metodología del Sistema FARO se inspira en los ejercicios anuales de monitorización realizados en el marco del Código de Conducta (2016) y continúa con las variables ya establecidas o introducidas a lo largo de la monitorización realizada por el OBERAXE desde 2020.

Se establece una diferenciación entre los contenidos notificados como usuario normal de la red social y los notificados a través de los procedimientos de reporte prioritario habilitados por las plataformas de manera exclusiva para instituciones y entidades que atesoran conocimiento y experiencia en la notificación de contenidos de discurso de odio. Esta vía cualificada de reporte, conocida como *Trusted Flagger*, fue introducida en el marco de los ejercicios de evaluación del Código de Conducta y ha sido reformulada con la entrada en vigor de la *Ley de Servicios Digitales* y el lanzamiento del nuevo Código de Conducta+ (2015)[11].

- **Categorías del seguimiento de los contenidos reportados**

Contenido retirado a las 24 horas: Hace referencia a los contenidos reportados como usuario normal que la plataforma ha retirado, o que el propio usuario ha eliminado, en un periodo inferior a las 24 horas desde que fueron notificados.

Contenido retirado a las 48 horas: Contenidos que han sido retirados en un periodo inferior a las 48 horas desde que fueron notificados como usuario normal.

Contenido retirado a la semana: Contenidos que han sido retirados en un periodo superior a las 48 horas e inferior a la semana desde que fueron notificados como usuario normal.

Contenido retirado tras su notificación vía alertador de seguimiento (*Trusted Flagger*): Contenidos que han sido retirados por la plataforma tras haber sido notificados a través de los procedimientos habilitados para las instituciones cualificadas.

[10] Códigos lingüísticos que, mediante términos y palabras clave que funcionan en forma de eufemismos y dobles sentidos, promueven mensajes de odio en determinados círculos con el propósito de no ser detectados por el público general.
[11] En este nuevo escenario, esta vía de prioritaria de notificación actualmente no cuenta con una nomenclatura oficial o unificada. En el marco de los ejercicios del Código de Conducta + se denomina *Monitoring Reporter*, mientras que en las plataformas posee diferentes nombres, aunque estos mecanismos cumplen con funciones análogas y procedimientos similares. Ej: *Priority Flagger* (YouTube), o *Approved Partner* (TikTok).

Esta vía de notificación se activa una vez haya transcurrido una semana desde que el contenido fuera notificado como usuario normal y aún siga publicado en la red social; ya hayamos recibido una respuesta negativa por parte de la plataforma al reporte enviado o la notificación aún siga sin revisar.

Limita la visibilidad: Contenidos reportados que, tras haber sido revisados por la plataforma, esta ha decidido no retirar; aunque sí ha llevado a cabo acciones para limitar su visibilidad en la red como dejar de sugerir la publicación o restringir su visualización al acceso mediante la url, al considerarlas publicaciones sensibles.

Contenido no retirado: Contenidos que permanecen publicados en la plataforma tras haber sido reportados. Una vez que hayan transcurrido los plazos de reporte como usuario normal, se haya efectuado la notificación vía *Trusted Flagger*, y el contenido siga publicado, bien porque se ha recibido respuesta negativa a la notificación, o porque esta aún siga sin ser revisada por parte de la plataforma, el contenido figurará como no retirado de manera definitiva.

Todo lo relativo a la información del seguimiento de las respuestas de las plataformas ante los contenidos reportados se obtiene a partir de la anotación de los mismos en la aplicación ALERTODIO.

Ante la identificación de contenidos que potencialmente puedan constituir un delito de odio, en 2023 se comenzó a hacer uso de la vía judicial a través de la remisión de denuncias a la Fiscalía de Delitos de Odio de la Fiscalía General del Estado.

2.2.4. Presentación de los datos

Los datos sobre discurso de odio en las redes sociales obtenidos a través del Sistema FARO se presentan principalmente a través de dos formatos: el visualizador en la web del OBERAXE y los boletines de monitorización.

- **Visualizador de datos del Sistema FARO**

La presentación de los datos de los contenidos de discurso de odio monitorizados en tiempo real se realiza a través de dos versiones de un visualizador de datos, insertado en la página de discurso de odio del OBERAXE, con datos actualizados en tiempo real que se obtienen de la integración de la información obtenida del monitor FARO y de ALERTODIO.

En el **visualizador de datos principal** se muestran gráficos con datos cuyo periodo de referencia son los últimos 30 días.

En el **visualizador ampliado**, al que se accede a través del botón "consultar todos los datos" se muestran los datos referentes al acumulado en el año (desde el 1 de enero de 2025). Además de los gráficos que se muestran en el cuadro principal (mensajes detectados, mensajes reportados, mensajes retirados, evolución de los mensajes detectados, evolución de los mensajes retirados, grupo diana, episodio que suscita el discurso de odio), se añaden los siguientes: tipo de contenido, expresión del lenguaje y género.

- **Boletines de Monitorización**

El análisis de los datos de los contenidos de discurso de odio obtenidos a través del Sistema FARO se difunden a través de los boletines de monitorización, que cuentan con periodicidad **mensual** y **trimestral**. A estos boletines se suma el **Informe Anual de Monitorización**.

3. Variables de monitorización del Sistema FARO

Las variables que se disponen a continuación corresponden a los gráficos incluidos en el visualizador de datos del Sistema FARO, incluyendo los gráficos de la versión ampliada.

Estas variables son fundamentales para comprender las dinámicas y la evolución del fenómeno en cuestión. Se emplean métricas cuantitativas que permiten identificar patrones, frecuencias e intensidad en los discursos de odio. Asimismo, se incorporan factores socio-demográficos y contextuales que facilitan el análisis de los segmentos poblacionales más afectados. A través de estos gráficos, se busca ofrecer una visión clara y precisa de los flujos discursivos y su impacto social, lo cual es esencial para la implementación de políticas públicas eficaces y la promoción de un entorno digital más inclusivo.

Los datos tienen una periodicidad anual (1 de enero al 31 de diciembre). En la consulta del dato a tiempo real se toma de referencia los últimos 30 días y el acumulado en el año (desde el 1 de enero de 2025).

La publicación del análisis del dato se realiza con periodicidad mensual, trimestral y anual a través de los boletines de monitorización, que se pueden consultar a través del siguiente enlace: boletines y visualizador de datos.

3.1. Mensajes detectados

Descripción	Hace referencia a los contenidos identificados en base a dos capas de filtrado: por una parte, se incluyen los contenidos de discurso de odio y por otra, se añade una capa más amplia con contenidos de discurso odioso y narrativas discriminatorias.
Origen del dato	Monitor FARO.
Observaciones	El porcentaje de mensajes detectados en cada red social no es uniforme. La búsqueda de información se realiza a través de *queries* o búsquedas específicas que dependen de las diferentes funcionalidades de las API de cada plataforma. Además, influyen las diferentes características de uso y del discurso de cada red en relación con las posibilidades de captación y filtrado que ofrece la tecnología disponible.

3.2. Mensajes reportados

Descripción	Casos obtenidos de la muestra (configurada con contenidos de discurso de odio de las diferentes plataformas monitorizadas) que se reportan a las plataformas.
Variables de clasificación	Son contenidos de discurso de odio constitutivos de delito, de infracción administrativa, o que infrinjan las normas de uso de las propias plataformas de prestación de servicios digitales.
Origen del dato	Monitor FARO y ALERTODIO.
Observaciones	El volumen de los contenidos reportados mensualmente puede sufrir alteraciones debido a varios factores como el ajuste de la muestra a los contenidos considerados de discurso de odio en un sentido estricto de la definición (cumpliendo, por tanto, las condiciones para ser reportados al ser potencialmente ilegales o infringir las políticas de uso de la plataforma); así como a la capacidad de recursos técnicos y de personal para poder remitir las notificaciones y efectuar el seguimiento diario de los contenidos reportados.

3.3. Mensajes retirados

Descripción	Suma de los contenidos de discurso de odio (potencialmente constitutivos de delito, de infracción administrativa, o que infringen las normas de uso de las propias plataformas de prestación de servicios digitales) que han sido eliminados por las plataformas tras haber sido reportados por el OBERAXE.
Variables de clasificación	24 horas, 48 horas, semana, limita visibilidad y *trusted flagger*
Origen del dato	ALERTODIO.

3.4. Evolución de los mensajes reportados

Descripción	Muestra la evolución de los contenidos de discurso de odio reportados
Origen del dato	Monitor FARO y ALERTODIO.

3.5. Evolución de los mensajes retirados

Descripción	Muestra la evolución del número de contenidos que han sido retirados por las plataformas tras haber sido reportados por el OBERAXE. Esta variable refleja la respuesta de las plataformas ante los contenidos notificados; y se compone del sumatorio de las dos vías de retirada: usuario normal y *trusted flagger*.
Variables de clasificación	24 horas, 48 horas, semana, y *trusted flagger*
Origen del dato	ALERTODIO.

3.6. Grupo Diana

Descripción	Grupos de población que son objeto del discurso de odio atendiendo a la motivación racista, xenófoba, islamófoba, antisemita o antigitana del mensaje detectado. Esta variable permite identificar a los colectivos que son más frecuentemente atacados en los contenidos de discurso de odio identificados, facilitando el análisis de las dinámicas de discriminación y proporcionando información clave para el diseño de estrategias y políticas públicas de intervención y prevención del racismo y la xenofobia.
Variables de clasificación	Personas africanas y afrodescendientes; personas asiáticas; personas europeas; personas inmigrantes; personas latinoamericanas; personas musulmanas; personas con origen en el norte de África; personas refugiadas; comunidad gitana; comunidad judía; niños, niñas y adolescentes no acompañados; y otros grupos.
Origen del dato	Monitor FARO.

Observaciones	- En el gráfico solo se muestran los seis grupos diana con mayor número de contenidos monitorizados en el periodo de referencia. Dadas las limitaciones de caracteres en la visualización del cuadro se presenta una versión reducida de la nomenclatura de cada variable de clasificación. - Para la definición de las diferentes categorías se toma de referencia la distribución geográfica el geoesquema de las Naciones Unidas. - En la categoría "personas con origen en el norte de África" se incluyen los contenidos dirigidos a personas procedentes u originarias de Marruecos, Argelia, Libia, Egipto o Túnez. - Aunque en gran parte de las expresiones de discurso de odio hacia las personas del norte de África y las personas musulmanas no se observa una distinción entre origen étnico / nacional y religión por parte del usuario que difunde el mensaje, en las variables contempladas sí se mantiene esta diferenciación considerando la categoría "personas musulmanas" para los contenidos en los que la parte central del mensaje hace referencia a elementos relacionados con el islam.

3.7. Episodio que suscita el discurso de odio

Descripción	Se centra en la identificación de los eventos, acontecimientos o situaciones que comúnmente desencadenan discursos de odio en las redes sociales. Este concepto permite analizar los contextos sociales, políticos o culturales específicos que sirven de catalizador para la expresión de actitudes hostiles, excluyentes o discriminatorias contra los grupos diana. El análisis de los episodios más representativos que dan lugar a estos discursos permite identificar patrones recurrentes que revelan las tensiones sociales o los conflictos que subyacen tras este tipo de discurso. La agenda política y mediática, así como las tendencias temáticas predominantes en las redes; junto con otros factores como la desinformación, influyen en la configuración y predominancia de los episodios prototípicos.
Variables de clasificación	Agresión sexual; conflicto armado; ámbito del deporte; ámbito económico; elecciones; inseguridad ciudadana; llegada de embarcaciones a las costas; llegada de personas inmigrantes a través de otras vías; okupaciones; políticas públicas o acción de un gobierno; ámbito religioso; salto de valla; ámbito de la salud; terrorismo; ningún acontecimiento; y otro acontecimiento.
Origen del dato	Monitor FARO

Observaciones	- En el gráfico solo se muestran los episodios con mayor número de contenidos monitorizados en el periodo de referencia. Dadas las limitaciones de caracteres en la visualización del cuadro se presenta una versión reducida de la nomenclatura de cada variable de clasificación. - El episodio "inseguridad ciudadana" incluye los mensajes que hacen referencia a actos e incidentes violentos, como robos y agresiones, que generalmente se atribuyen a personas de origen inmigrante o de los diferentes grupos target, ya sea de forma verídica o falsa. - "Conflicto armado" abarca los mensajes que hacen referencia a guerras y conflictos armados, como el conflicto en Oriente Medio y la guerra en Ucrania, entre otros. Se considera el concepto de conflicto armado desde una perspectiva amplia, no limitándose solo a los hechos bélicos e incluyendo también otros acontecimientos vinculados al conflicto, como las reacciones de políticos y gobiernos; y los desplazamientos forzados ocasionados, así como otras consecuencias humanitarias. - En "salto de valla" se incluyen los acontecimientos vinculados a la llegada de personas inmigrantes a través de la frontera terrestre de Ceuta y Melilla. - En "llegada de personas inmigrantes a través de otras vías" se incluyen mensajes suscitados a partir de movimientos migratorios al margen de los contemplados en las categorías de "llegada de embarcaciones" y "salto de valla"; como, por ejemplo, los desplazamientos en avión o el cruce de fronteras a pie o a través de vehículos. - La categoría "terrorismo" abarca los mensajes que hacen referencia a actos e incidentes terroristas, ya sean producidos en España o en otros países, o a otros acontecimientos relacionados como la detención de personas radicalizadas ante delitos de captación, adoctrinamiento o propagación de material propagandístico de organizaciones terroristas, entre otros.

3.8. Tipología del contenido

Descripción	Se recogen los diferentes tipos de manifestaciones de discurso de odio según la tipología del contenido del mensaje a través del cual los usuarios atacan al grupo diana. La clasificación de la tipología del contenido incluye una gran variedad de términos, manifestaciones y conductas, agrupadas en 6 categorías principales, sobre las que se asienta el núcleo del mensaje de discurso de odio.
Variables de clasificación	Presenta al grupo como una amenaza; alaba o apoya a quien atenta contra el grupo diana; deshumaniza o degrada; incita a la expulsión del colectivo; incita a la violencia; y promueve el descrédito (en base a atributos personales del grupo o sin dar ningún argumento más que la pertenencia al grupo).
Origen del dato	Monitor FARO.
Observaciones	Los términos grupo diana, grupo target y colectivo se emplean de forma equivalente.

3.9. Expresión del lenguaje

Descripción	Se refiere a las características que adopta el lenguaje empleado por los usuarios para transmitir los mensajes de discurso de odio. Esta variable analiza la forma en la que se construyen los mensajes atendiendo a diferentes elementos como el tono y los recursos lingüísticos empleados, a través de los cuales los mensajes se difunden, ya sea de manera directa (explícita) o encubierta (implícita). Su estudio es fundamental ya que el lenguaje denota el uso de estrategias de comunicación para difundir mensajes de discurso de odio en determinados círculos y contextos de manera eficaz; y, en ocasiones, con el propósito principal de evitar los mecanismos de moderación por parte de las plataformas.
Variables de clasificación	Agresivo explícito; discriminatorio no agresivo; e ironía o sarcasmo.
Origen del dato	Monitor FARO.
Observaciones	- El lenguaje "agresivo explícito" se caracteriza por emplear un tono hostil, con insultos, ataques directos u otras expresiones agresivas. - El "discriminatorio no agresivo" se refiere a contenidos que promueven discurso de odio a través de mensajes discriminatorios que no incluyen insultos u otras expresiones agresivas. - La categoría "ironía o sarcasmo" incluye mensajes que emplean la ironía, el sarcasmo o un cierto tono humorístico como vehículo para difundir mensajes de odio. En gran parte de las ocasiones esta categoría de expresión del lenguaje se usa para enmascarar el discurso y presentarlo de una forma velada.

3.10. Género

Descripción	Se refiere a la identificación y clasificación de los mensajes de discurso de odio según el género de las personas a los que van dirigidos, en base a las variables gramaticales de género. Permite analizar cómo el discurso de odio adopta diferentes formas dependiendo de si se dirige a hombres o a mujeres. Además, se utiliza para calcular el porcentaje de contenidos destinados a cada género, lo que facilita el estudio de las dinámicas de género en la expresión del odio y explorar cómo estas se intersecan con otras formas de discriminación.
Variables de clasificación	hombre; mujer.
Origen del dato	Monitor FARO.