Emojis en contra de las personas negras

2026-001-FB-UA, 2026-002-IG-UA

Resumen

El Consejo asesor de contenido rechazó las decisiones originales de Meta de conservar dos contenidos en los que se usaban emojis para comparar a las personas negras con monos, lo cual constituye una expresión de odio, discriminación y acoso. El Consejo instó a Meta a mejorar sus procesos de moderación manual y automatizada de modo que tengan en cuenta de forma integral el "algospeak", incluidos emojis, a fin de evitar ataques discriminatorios y de odio contra grupos. Esto podría implicar asegurarse de que los datos de entrenamiento para la aplicación automatizada de políticas sean adecuados para el contexto regional y estén actualizados, coordinar iniciativas para interrumpir de forma proactiva campañas de odio y garantizar que sus iniciativas de mitigación incluyan la supervisión activa de contenido con emojis que incitan a la discriminación y la hostilidad durante eventos deportivos importantes, como la Copa Mundial de la FIFA (Federación Internacional de Fútbol Asociación).

Información de los casos

Estos casos se relacionan con dos publicaciones realizadas en mayo de 2025 en las que se usaron emojis de mono en referencia a personas negras.

En el primer caso, un usuario de Brasil publicó un video breve en Facebook con la escena de la película ¿Qué pasó ayer? doblada al portugués, en la que dos personajes se pelean por un mono. En el texto superpuesto en el video, se nombra a los personajes como los clubes de fútbol "Barcelona" y "Real Madrid". Otro texto superpuesto hace referencia a chicos que están cobrando protagonismo en el fútbol brasileño. La descripción consta de un emoji de mono. La publicación se visualizó 22.000 veces y 12 personas la reportaron.

El segundo caso tiene que ver con un comentario publicado en respuesta a un video en una cuenta de Instagram de Irlanda. En el video, el usuario expresa indignación tras ser testigo de un incidente racista en la calle y la descripción insta

a Irlanda a rechazar el racismo. El comentario de otro usuario señala que no respalda el mensaje, sino que quiere que la situación "estalle" y "divertirse a lo grande con todos los [emojis de mono] y en la calle". El comentario además incluye varios emojis de mono, de risa y de plegaria, y enfatiza "días gloriosos por delante". La publicación se visualizó 4.000 veces y 62 personas reportaron el comentario.

Los sistemas automatizados de Meta y, luego de la apelación de los usuarios, los revisores decidieron conservar ambas publicaciones. Los usuarios, entonces, apelaron al Consejo. Luego de que el Consejo seleccionara estos casos para su revisión, Meta determinó como incorrectas sus decisiones originales y eliminó las publicaciones en julio de 2025 por infringir la norma comunitaria sobre conducta que incita al odio de la empresa.

Se puede usar lenguaje cifrado mediante giros idiomáticos o emojis (denominado "algospeak") para transmitir mensajes deshumanizantes o de odio y eludir, al mismo tiempo, los sistemas de moderación de contenido automatizados.

Conclusiones principales

Al Consejo le preocupa la precisión de la aplicación de la política de conducta que incita al odio, en especial en la evaluación de emojis que se emplean como algospeak. Los clasificadores identificaron el contenido, pero no tomaron medida alguna. Meta señala que los revisores deben tener en cuenta todos los aspectos del contenido, como imágenes, descripciones y superposiciones de texto, así como factores que van más allá del contenido inmediato, incluidos la publicación principal y los comentarios. Asimismo, Meta explicó que entrena a sus clasificadores con conjuntos de datos compuestos por ejemplos reportados y etiquetados, incluidos casos en los que se usan emojis de formas potencialmente infractoras. No obstante, las revisiones manuales y automatizadas no lograron evaluar con precisión las publicaciones.

Meta debe auditar de forma periódica sus datos de entrenamiento para mejorar la detección automatizada del uso de emojis infractores. Los procesos de aplicación de políticas siempre deben remitir el contenido a revisores que hablen el idioma en cuestión y cuenten con conocimiento regional.

En respuesta a las preguntas del Consejo, Meta indicó que, tras su comunicado del 7 de enero de 2025, los macromodelos lingüísticos (LLM) ahora se encuentran más ampliamente integrados como un nivel de revisión adicional, incluso para contenido que puede infringir la política de conducta que incita al odio. Según Meta, los LLM no reemplazan los modelos existentes, pero brindan una segunda opinión respecto de las decisiones de aplicación de políticas, ya que se centran en el contenido que se marcó para su eliminación. En estos casos, no participaron LLM en el proceso de revisión.

El Consejo considera que ambas publicaciones infringen la norma sobre conducta que incita al odio, que prohíbe las comparaciones deshumanizantes con animales. Ambas publicaciones usan el emoji de mono para atacar a las personas negras sobre la base de su característica protegida.

Conservar las publicaciones es una medida incoherente con las responsabilidades de Meta con los derechos humanos, ya que los emojis cuyo objeto es deshumanizar a grupos con características protegidas e incitar a la discriminación y la hostilidad contra ellos deberían quedar sujetos a eliminación. Es necesario y proporcionado eliminar ambas publicaciones.

Las dos publicaciones representan formas de algospeak empleado para expresar odio, discriminación y acoso contra grupos con características protegidas específicas. Asimismo, ilustran cómo se pueden usar emojis para incitar a otras personas a realizar acciones discriminatorias y potencialmente hostiles.

La publicación brasileña se realizó en el contexto de hostilidad y racismo sistemático ampliamente documentado en el ámbito del fútbol, en particular contra jugadores negros. El comentario en el caso irlandés se compartió en el contexto de una creciente discriminación racial y un aumento de la afrofobia en Irlanda.

Para coordinar mejor sus iniciativas y proteger a las personas que tal vez no se nombren directamente, pero son el blanco implícito de campañas de odio, Meta debe desarrollar un marco que armonice las medidas que ya impone a fin de interrumpir de forma proactiva campañas de odio, en especial aquellas que implican el uso de emojis. Meta debe garantizar que sus iniciativas de mitigación urgentes, ya sean mediante su centro de operaciones de productos de integridad u otro sistema de mitigación de riesgos, incluyan la supervisión activa de

contenido con emojis que incitan a la discriminación u hostilidad contra un grupo concreto en los días previos a un acontecimiento deportivo importante, durante este y una vez que termina, como es el caso de la Copa Mundial de la FIFA 2026.

Decisión del Consejo asesor de contenido

El Consejo anuló las decisiones originales de Meta de conservar ambos contenidos.

Asimismo, el Consejo le recomendó a Meta:

- Auditar los datos de entrenamiento de los sistemas automatizados que se emplean para aplicar la política de conducta que incita al odio y garantizar que los datos se actualicen de forma periódica a fin de incluir ejemplos de contenido con emojis en todos los idiomas, el uso infractor de emojis y nuevas instancias del uso de emojis con fines de odio.
- Armonizar sus iniciativas actuales para interrumpir de forma proactiva campañas de odio, en especial aquellas que implican el uso de emojis, a fin de proteger mejor a las personas que no se nombran directamente, pero que son el blanco implícito de estas campañas.
- Garantizar que sus iniciativas de mitigación urgentes, ya sea mediante su centro de operaciones de productos de integridad u otro sistema de mitigación de riesgos, incluyan la supervisión activa de contenido con emojis que incitan a la discriminación u hostilidad contra un grupo concreto en los días previos a un acontecimiento deportivo importante, durante este y una vez que termina, como la Copa Mundial de la FIFA.

El Consejo reitera la importancia de la recomendación relevante que hizo previamente y que insta a Meta a:

- Brindar a los usuarios la oportunidad de revertir el error por sus propios medios, comparable a la intervención de fricción en el momento de la publicación que se creó como resultado de la recomendación n.º 6 del caso "Manifestaciones en Rusia a favor de Navalny". Si esta intervención ya no está en vigencia, la empresa debe proporcionar una intervención en el producto similar.

* Los resúmenes de casos ofrecen información general sobre los casos y no sientan precedentes.