



Together Against Online Hate Speech

TAOH

Together Against Online Hate Speech

Deliverable 2.2 Hate Speech Pattern Report

Spanish translation

Sobre TAOH

Together Against Online Hate Speech (TAOHS) es un proyecto internacional impulsado por organizaciones de Alemania, Países Bajos y España para comprender y hacer frente al odio en línea. Analizamos cómo se manifiesta y se propaga el discurso de odio en redes y entornos digitales, y desarrollamos herramientas, formación y recomendaciones prácticas para que organizaciones de la sociedad civil, medios e instituciones puedan prevenirlo y responder de forma segura y eficaz.

Miembros del Consorcio

- SPF - Sharing Perspectives Foundation - Países Bajos | Coordinador
- UvA - University of Amsterdam - Países Bajos
- Build Up - How to Build Up Germany - Alemania
- IBH - ichbinhier e.V. - Alemania
- FundiPau - Fundació par la Pau - España



**Cofinanciado por
la Unión Europea**

Financiado por la Unión Europea. Las opiniones y puntos de vista expresados solo comprometen a su(s) autor(es) y no reflejan necesariamente los de la Unión Europea o los de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser considerados responsables de ellos.



Este trabajo está bajo una [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/)

«Hostilidad en línea: análisis comparativo de Instagram, X y TikTok en España, Alemania y los Países Bajos»

Version 2.2 September 19th, 2026 (translated from English)

Sander van Haperen (s.p.f.vanhaperen@uva.nl)

Abstract

Este documento presenta los resultados de un análisis comparativo de la hostilidad en línea realizado para el proyecto «Together Against Online Hate Speech», llevado a cabo por los siguientes socios: Sharing Perspectives Foundation, BuildUp, ichbinhier e.V., FundiPau y la Universidad de Ámsterdam.

El objetivo de este informe es comparar los patrones de hostilidad en línea en Alemania, los Países Bajos y España en Instagram, X y TikTok, distinguiendo entre categorías, como el racismo y la animadversión basada en el género, así como patrones de comportamiento en términos de focalización y persistencia. Con este fin, se desarrollan modelos de lenguaje para detectar y clasificar distintas formas de hostilidad en línea, que se ponen a disposición como código abierto.

Los hallazgos sirven de base para las estrategias de intervención de organizaciones de la sociedad civil, activistas, cuerpos de seguridad y responsables de la formulación de políticas que abordan la hostilidad en línea. En primer lugar, las categorías impulsadas por acontecimientos (capacitismo, raza, intimidación sexual) requieren respuestas rápidas ante los sucesos, ya que la hostilidad se concentra en pocos objetivos y se propaga principalmente por usuarios transitorios. En segundo lugar, la concentración en determinados objetivos en algunas categorías, como la política, los insultos, la delincuencia y la cultura, exige combinar la monitorización con el apoyo a los posibles objetivos de hostilidad en línea. En tercer lugar, quienes difunden animadversión relacionada con la raza, el lenguaje vulgar y los insultos muestran una persistencia relativamente mayor que quienes difunden otras categorías. Las estrategias de intervención dirigidas a estas categorías podrían centrarse en actores recurrentes que generan hostilidad en línea.

Citación:

van Haperen, Sander (2026). "Online Hostility: Comparative analysis of Instagram, X, and TikTok in Spain, Germany, and the Netherlands". University of Amsterdam. Online resource DOI: 10.21942/uva.33061112.

Índice

Introducción	6
Definición	7
Metodología	8
Recopilación de datos	9
Modelo de lenguaje 1: detección de discurso de odio	11
Modelo de lenguaje 2: clasificador de hostilidad en línea	12
Procedimiento de anotación	12
Construcción de léxicos	14
Rendimiento del modelo	15
Hallazgos	17
Patrones por país	17
Patrones por plataforma	20
Patrones entre objetivos y difusores	22
Patrones de comportamiento	24
Patrones por categoría	25
Descripción de las categorías	25
Delincuencia	27
Cultura	30
Género	32
Insultos y ataques personales	35
Política	37
Raza	39
Religión	42
Intimidación sexual	44
Orientación sexual	47
Lenguaje violento	49
Lenguaje vulgar	51
Conclusiones	53
Estrategias de intervención recomendadas	54
Limitaciones metodológicas	56
Agradecimientos	59
Descargo de responsabilidad legal	60
Declaración de datos y licencia	60
Bibliografía	62

Introducción

Este informe forma parte del proyecto «Together Against Online Hate Speech» (TAOHS). El objetivo general del proyecto es aumentar la capacidad de resiliencia de las organizaciones de la sociedad civil, los medios de comunicación y las autoridades para comprender y abordar el discurso de odio en línea en sus plataformas, así como contribuir a un espacio digital más seguro y plural. Este informe se centra en Alemania, los Países Bajos y España.

El proyecto combina modelos de clasificación lingüística con intervenciones de diálogo dirigidas por personas y acciones de incidencia política, con el fin de comprender y abordar de forma integral la hostilidad en línea. Si bien la moderación de contenidos y los mecanismos de denuncia siguen siendo importantes para abordar contenidos ilegales o que incumplen las políticas de las plataformas, estas estrategias encuentran dificultades ante contenidos que, sin ser ilegales, siguen siendo problemáticos. Asimismo, formas de lenguaje como la intolerancia y la falta de civismo justifican una intervención. Nuestro propósito es apoyar a las personas, organizaciones de la sociedad civil, cuerpos de seguridad y responsables de la formulación de políticas que se enfrentan a este tipo de lenguaje en línea.

El objetivo de este informe es orientar las estrategias de intervención mediante la identificación de patrones de hostilidad en línea. Desde el punto de vista metodológico, este estudio está diseñado para detectar una amplia variedad de formas de hostilidad en línea, dirigidas a una diversidad representativa de objetivos, en distintos espacios de redes sociales, Instagram, X y TikTok, y para comparar estos patrones entre los contextos nacionales de Alemania, los Países Bajos y España. Con este fin, desarrollamos tres modelos de lenguaje para clasificar distintos tipos de hostilidad en línea en alemán, neerlandés y español. Como resultado de una extensa codificación abductiva y de pruebas de rendimiento, estos modelos de lenguaje se ponen a disposición con código abierto, con el fin de que las organizaciones asociadas y otras partes interesadas puedan adaptarlos a análisis específicos.

En este informe empleamos estos modelos para realizar una comparación entre plataformas y países, e identificar patrones relativos a la prevalencia, los objetivos y los difusores de distintos tipos de hostilidad en línea. En general, nuestros hallazgos sugieren que se trata de un problema generalizado, compuesto predominantemente por comentarios hostiles

aislados. No obstante, al diferenciar los comportamientos por categoría, observamos diferencias notables entre los distintos tipos de hostilidad en línea.

De ello se derivan tres conclusiones aplicables. En primer lugar, las categorías que dependen de eventos puntuales (discriminación por discapacidad, racismo, intimidación sexual) pueden ser difíciles de prever, pero pueden beneficiarse de respuestas rápidas a dichos eventos, ya que la hostilidad se concentra en unos pocos objetivos y se propaga principalmente a través de usuarios transitorios. En segundo lugar, las categorías más difusas (política, insultos, delincuencia y cultura) hasta cierto punto pueden llegar a ser más previsibles, al resultar más consistentes con el tiempo, y aunque la atención se desplaza continuamente entre muchos objetivos, estas categorías incluyen a un pequeño número de infractores reincidentes, y relativamente persistentes.

En tercer lugar, quienes difunden hostilidad relacionada con la raza, el lenguaje vulgar y los insultos muestran una persistencia relativamente mayor que quienes difunden otras categorías. Las estrategias de intervención dirigidas a estas categorías podrían centrarse en actores recurrentes que generan hostilidad en línea.

El proyecto «Together Against Online Hate Speech» está compuesto por los siguientes socios: Sharing Perspectives Foundation, Build Up, IchBinHier, FundiPau y la Universidad de Ámsterdam. El proyecto está financiado por la Unión Europea, CERV 2024 CHAR LITI SPEECH. No obstante, los puntos de vista y opiniones expresados corresponden exclusivamente a las personas autoras y no reflejan necesariamente los de la Unión Europea. Ni la Unión Europea ni la autoridad concedente pueden ser consideradas responsables de ellos.

Definición

Al comienzo de nuestro proyecto, formulamos la siguiente definición de «discurso de odio»: «incitar públicamente a la violencia o al odio dirigidos contra un grupo de personas o contra un miembro de dicho grupo, definido por referencia a la raza, el color de piel, la religión, la ascendencia o el origen nacional o étnico». También incluimos referencias al género, la orientación sexual y la capacidad física. Inicialmente, esperamos que este tipo de lenguaje incluya categorías como «deshumanización, sexismo, política, racismo, religión, violencia y amenazas»; adaptación de EOHH, 2026. En la práctica, esta definición nos proporcionó

conceptos sensibilizadores y orientó la calibración entre las personas encargadas de la anotación.

En este informe también utilizamos el término general «hostilidad en línea» para referirnos a un lenguaje que no es estrictamente ilegal ni constituye discurso de odio relacionado con la identidad de las personas, pero que aun así puede perjudicar la calidad de los espacios digitales. Más allá del discurso de odio, otros tipos de hostilidad en línea incluyen el abuso y la falta de civismo. El abuso se refiere al lenguaje destinado a insultar, provocar o causar malestar a una persona concreta (Akhtar y Morrison, 2019; Ward y McLoughlin, 2020). La falta de civismo político se refiere a «una ausencia de respeto por las normas sociales y culturales que regulan las interacciones personales y el funcionamiento de los sistemas democráticos» (Bentivegna y Rega, 2024). Esta categoría abarca discursos que vulneran normas, desde formas relativamente leves de descortesía hasta manifestaciones más abiertas de hostilidad, sin que sea imprescindible que exista intención de causar daño o que se dirijan contra grupos sociales específicos (Rossini, 2019).

Metodología

El objetivo de este estudio es orientar las estrategias de intervención mediante la identificación de patrones de hostilidad en línea. Desde el punto de vista metodológico, este estudio está diseñado para detectar una amplia variedad de formas de hostilidad en línea, dirigidas a una diversidad representativa de objetivos, en distintos espacios de redes sociales, Instagram, X y TikTok, y para comparar estos patrones entre los contextos nacionales de Alemania, los Países Bajos y España.

Con este fin, recopilamos datos de redes sociales relativos a una selección representativa de los objetivos con mayor probabilidad de recibir hostilidad en línea. Analizamos este conjunto de datos mediante dos modelos de lenguaje: un modelo para determinar si una publicación probablemente contiene hostilidad en línea y otro para determinar con qué tema se relaciona la publicación. Esta combinación nos permite identificar y examinar patrones entre distintos tipos de hostilidad en línea. Nos centramos en patrones relativos a la prevalencia, los objetivos y los difusores de hostilidad en línea. Este capítulo examina este procedimiento con mayor detalle.

Recopilación de datos

Recopilamos datos de Instagram, X y TikTok. Estas plataformas fueron seleccionadas por ser espacios en los que probablemente se produce hostilidad en línea y por encontrarse entre las redes sociales más utilizadas en los tres países. La selección también abarca una variedad de bases de usuarios y funcionalidades. En X predominan las opiniones políticas estrechamente vinculadas a acontecimientos de actualidad, Instagram pone el énfasis en los estilos de vida y TikTok presenta una mayor participación de grupos demográficos más jóvenes y un descubrimiento algorítmico abierto.

Las publicaciones y comentarios en X, anteriormente Twitter, incluyen mensajes breves y reacciones que suelen tener 50 palabras o menos, y que pueden contener materiales multimedia, enlaces a URL externas y menciones a otras personas usuarias. Las publicaciones y comentarios en Instagram suelen consistir en textos breves y reacciones de aproximadamente 50 palabras o menos, que pueden contener enlaces a URL externas y menciones a otras personas usuarias. Las publicaciones y comentarios en TikTok suelen consistir en textos breves que acompañan a vídeos, así como reacciones de aproximadamente 50 palabras o menos, que pueden contener enlaces a URL externas y menciones a otras personas usuarias.

El corpus es el resultado del siguiente procedimiento. En primer lugar, identificamos cuentas de personas y organizaciones consideradas como las más propensas a recibir diversas formas de odio, intolerancia y falta de civismo en línea en cada uno de los tres países: España, Alemania y los Países Bajos. La selección de casos en los contextos español, alemán y neerlandés siguió una estrategia de muestreo intencional con el objetivo de reunir un conjunto diverso de cuentas de redes sociales con mayor probabilidad de recibir discurso de odio en línea.

Nuestro objetivo con la recopilación de datos era construir un conjunto de datos que recogiera un amplio espectro de hostilidad en línea, con el fin de apoyar el desarrollo de un modelo de lenguaje sensible tanto a formas explícitas como más sutiles de lenguaje. Para cada uno de los tres países, la selección comenzó con la identificación de personas y organizaciones destacadas que operan en ámbitos caracterizados por una alta visibilidad y participación

públicas. Se seleccionaron cuentas de ámbitos como la política, los medios de comunicación, el entretenimiento y la cultura, el servicio público, el mundo académico, el activismo, el deporte y las organizaciones de la sociedad civil. Dentro de cada ámbito, se garantizó la diversidad interna mediante la incorporación de subgrupos como cargos electos, partidos políticos, periodistas, deportistas y personas influyentes.

Nos basamos en el conocimiento de los contextos nacionales y en fuentes complementarias, como las noticias, para identificar a personas conocidas por recibir un elevado volumen de odio en línea, garantizando que nuestra selección se fundamentara en patrones observables de acoso. En particular, se prestó atención a figuras implicadas en controversias políticas, culturales o mediáticas recientes o en curso, así como a aquellas vinculadas con cuestiones controvertidas, como la migración, los conflictos geopolíticos o los debates nacionales, que se sabe que incrementan la exposición a la hostilidad en línea. Esta estrategia de muestreo se ajustó a nuestra definición de discurso de odio y, al mismo tiempo, permitió recoger diversas manifestaciones de distintos grupos sociales. Identificamos 377 cuentas en España, 217 en Alemania y 360 en los Países Bajos.

Para cada una de estas cuentas iniciales, recopilamos datos de perfiles en X, Instagram y TikTok. Los criterios de selección incluyen comentarios y reacciones publicados en estas cuentas, así como publicaciones que mencionan la cuenta, durante el periodo comprendido entre el 1/1/2024 y el 31/12/2025. En X, esto incluye menciones y republicaciones; en Instagram, menciones y republicaciones; y en TikTok, menciones y respuestas. Se excluyeron las cuentas sin una presencia oficial o identificable en redes sociales. Solo se incluyen publicaciones y comentarios disponibles públicamente. No incluimos contenido que hubiera sido eliminado entre su publicación y la recopilación de datos, realizada entre septiembre y diciembre de 2025. Los datos configurados como privados o eliminados entre el momento de su publicación y nuestro estudio tampoco se incluyen en la recopilación de datos.

Este procedimiento dio lugar a tres conjuntos de datos:

- **Corpus alemán:** 1.452.268 publicaciones, de las cuales 598.865 proceden de Instagram, 730.971 de TikTok y 122.432 de X.
- **Corpus español:** 2.445.553 publicaciones, de las cuales 457.036 proceden de Instagram, 1.953.819 de TikTok y 34.698 de X.

- **Corpus neerlandés:** 5.611.651 publicaciones, de las cuales 1.018.428 proceden de Instagram, 680.296 de TikTok y 3.912.927 de X.

Modelo de lenguaje 1: detección de discurso de odio

Para determinar la probabilidad de que una publicación de nuestro conjunto de datos contenga hostilidad en línea, utilizamos primero clasificadores lingüísticos ya existentes. Tras pruebas preliminares y la consideración de cómo se había entrenado cada modelo, seleccionamos tres modelos independientes de código abierto (Aluru et al., 2002; Pérez et al., 2022; IMSyPP, 2022).

Estos modelos se entrenaron con datos monolingües de redes sociales utilizando el tokenizador BERTje, cada uno con el objetivo de detectar la presencia de discurso de odio (Basile et al., 2019; Pérez et al., 2021; Cinelli et al., 2021; Kralj Novak, 2022; Baider et al., 2023). Aplicamos estos modelos a nuestros conjuntos de datos y obtuvimos los siguientes resultados, recogidos en la Tabla 1.

Tabla 1. Rendimiento del modelo 1, objetivo: detección de discurso de odio

Model	Accuracy	Precision	Recall	F1	AUC_ROC	PR_AUC
Alemán ¹	0.566	0.583	0.315	0.409	0.568	0.405
Español ²	0.616	0.442	0.254	0.322	0.640	0.661
Neerlandés ³	0.650	0.405	0.459	0.430	0.564	0.607

Nota: El modelo alemán aplicó la etiqueta «HATE» con un umbral de 0,94; el modelo español aplicó la etiqueta «hateful» con un umbral de 0,052; y el modelo neerlandés aplicó «Label_2» con un umbral de 0,48. Estos umbrales se incluyen aquí con fines de replicabilidad; sus niveles no son comparables entre los distintos modelos.

Es importante señalar el bajo rendimiento de estos modelos. Fueron seleccionados por presentar el mejor desempeño entre las alternativas disponibles durante el análisis preliminar realizado en la fase de diseño de este estudio. En la práctica, una puntuación AUC de 0,564

¹ <https://huggingface.co/Hate-speech-CNERG/dehatebert-mono-german>

² <https://huggingface.co/pysentimiento/robertuito-hate-speech/tree/main>

³ https://huggingface.co/IMSyPP/hate_speech_nl

ofrece un rendimiento solo ligeramente superior al azar, cuestión que se analiza con mayor detalle en la sección dedicada a las limitaciones metodológicas. Destacamos aquí que desaconsejamos el uso de estos modelos para la toma de decisiones automatizada en la detección de discurso de odio. Para los fines de nuestros análisis, complementamos cada modelo con una clasificación humana de la hostilidad en línea.

Modelo de lenguaje 2: clasificador de hostilidad en línea

Aunque el modelo 1 está diseñado para detectar si una publicación probablemente contiene lenguaje problemático, no permite comprender con qué tema se relaciona. Para identificar distintos tipos de hostilidad, desarrollamos un modelo de lenguaje basado en reglas.

Procedimiento de anotación

Para identificar distintas formas de discurso de odio, adoptamos un enfoque abductivo de codificación (Timmermans y Tavory, 2012). En primer lugar, tres personas expertas en anotación participaron en iteraciones sucesivas de codificación de muestras aleatorias de estos conjuntos de datos, con el objetivo de identificar categorías de hostilidad en línea. Nuestra definición compartida de hostilidad en línea proporcionó conceptos sensibilizadores y orientó los debates sobre la semántica y nuestros objetos de análisis. Para identificar patrones lingüísticos recurrentes en las interacciones de las personas usuarias, interpretamos publicaciones de cuentas pertenecientes a ámbitos similares o que compartían marcadores identitarios, y documentamos palabras clave, frases y expresiones comunes de hostilidad en línea. Este proceso también incluyó la identificación de formas más sutiles de lenguaje dañino, como expresiones codificadas, insultos denigrantes, jerga y alusiones veladas.

En segundo lugar, tras cada ronda de codificación de muestras de 500 publicaciones, nos reunimos para comparar las etiquetas asignadas, evaluar nuestra definición de discurso de odio, dialogar con la literatura previa y ajustar o ampliar nuestro manual de codificación. A partir de nuestra definición inicial, esperábamos encontrar las siguientes categorías: raza, color de piel, religión, ascendencia u origen nacional o étnico, género, orientación sexual y capacidad física. Posteriormente, buscamos complementar estas categorías de manera inductiva, lo que condujo a la incorporación de las categorías «Falsedad», «Desconfianza contra las instituciones», «Teorías conspirativas», «Polarización», «Ucrania», «Rusia»,

«Autocracia», «Oriente Medio», «Clima», «Delincuencia», «Deportes», «Insultos», «Lenguaje vulgar», «Ironía y sarcasmo» e «Intimidación sexual». En esta etapa, aunque seguíamos codificando inductivamente cualquier contenido, mantuvimos las categorías «Informativo», «Crítico» y «Positividad», que posteriormente se eliminaron para centrarnos exclusivamente en la hostilidad en línea.

En tercer lugar, la lista de términos resultante fue revisada y perfeccionada por personas familiarizadas con el contexto digital de cada país, con el fin de garantizar su relevancia cultural y lingüística. Además de identificar distintas categorías de hostilidad en línea y el tema de cada publicación, inicialmente también asignamos una medida de intensidad en una escala de 1 a 10. Esta medida se abandonó finalmente tras una reflexión crítica sobre nuestras distintas interpretaciones de la intensidad. De acuerdo con nuestro enfoque abductivo, alternamos entre la codificación y la revisión de la literatura, lo que nos llevó a distinguir la falta de civismo y la descortesía del discurso de odio (Nithyanand et al., 2017; Theocharis et al., 2020).

Entendemos la falta de civismo como «una ausencia de respeto por las normas sociales y culturales que regulan las interacciones personales y el funcionamiento de los sistemas democráticos» (Bentivegna y Rega, 2024). Esta falta de respeto puede manifestarse como descortesía interpersonal o como discursos que deslegitiman los sistemas democráticos y a sus actores, incluidas formas como la estigmatización, los ataques basados en el rol, la demonización, la descalificación mediante insultos, las amenazas, la incitación a la violencia, la vulgaridad, las acusaciones de ilegalidad y el lenguaje ofensivo (Bentivegna y Rega, 2024). Tras cuatro iteraciones de codificación, consideramos que el manual de codificación era satisfactorio y que las categorías identificadas cubrían adecuadamente los datos encontrados en la cuarta muestra. En ese momento, nuestra codificación y nuestras discusiones ya no generaban nuevas categorías ni etiquetas, lo que interpretamos como un punto de saturación.

En cuarto lugar, aplicamos el manual de codificación para anotar datos no incluidos en la muestra inicial. En este punto, consideramos que el manual era lo suficientemente exhaustivo para nuestro objetivo de identificar hostilidad en línea. Se instruyó a doce personas expertas en anotación, reclutadas a través de los socios del consorcio, para aplicar el manual de codificación a nuevas muestras aleatorias. Este proceso se complementó con ocho sesiones de calibración destinadas a debatir los desacuerdos sobre las etiquetas. Los desacuerdos se resolvieron al alcanzarse un consenso tras la discusión entre las personas anotadoras.

Finalmente, se alcanzó el siguiente nivel de acuerdo entre codificadores:

- Alemán: 7.852 publicaciones de nuestra muestra fueron codificadas por dos personas anotadoras, equivalentes al 0,54 % de las 1.452.268 publicaciones del corpus alemán, y 5.335 fueron codificadas por tres personas anotadoras, equivalentes al 0,37 % del corpus. Se obtuvo un acuerdo aceptable entre codificadores, con un alfa de Krippendorff de 0,73.
- Español: 6.534 publicaciones de nuestra muestra fueron codificadas por dos personas anotadoras, equivalentes al 0,27 % de las 2.445.553 publicaciones del corpus español, y 4.719 fueron codificadas por tres personas anotadoras, equivalentes al 0,19 % del corpus. Se obtuvo un acuerdo aceptable entre codificadores, con un alfa de Krippendorff de 0,67.
- Neerlandés: 12.933 publicaciones de nuestra muestra fueron codificadas por dos personas anotadoras, equivalentes al 0,24 % de las 5.611.651 publicaciones del corpus neerlandés, y 9.091 fueron codificadas por tres personas anotadoras, equivalentes al 0,17 % del corpus. Se obtuvo un buen acuerdo entre codificadores, con un alfa de Krippendorff de 0,83.

Construcción de léxicos

Llevamos a cabo dos rondas de codificación axial con el objetivo de relacionar las categorías con nuestra comprensión del discurso de odio, tal como se desarrolló a lo largo del proceso de codificación (Green y Thorogood, 2005). Para ello, consideramos las relaciones entre las categorías. En última instancia, estas reflexiones nos llevaron a adoptar el término «hostilidad en línea», que es más amplio que «discurso de odio» y, por tanto, representa mejor la variedad de lenguaje encontrada en nuestro conjunto de datos. Utilizamos este término general en todo el análisis, salvo cuando nos referimos explícitamente al odio basado en la identidad y dirigido contra personas, conforme a nuestra definición original de «discurso de odio».

Como resultado de la codificación axial, se fusionaron varias categorías. En el caso de otras, determinamos que no eran suficientemente destacadas o pertinentes a la luz de los objetivos del proyecto. Las categorías centrales finales representadas en estos léxicos son: «Género» [GEN], «Cultura» [CUL], «Raza» [RAC], «Orientación sexual» [SOR], «Religión» [REL],

«Capacitismo» [ABL], «Lenguaje violento» [VIO], «Insultos y ataques personales» [INS], «Lenguaje vulgar» [VUL], «Ironía y sarcasmo» [IRO], «Intimidación sexual» [INT], «Política» [POL], «Falsedad» [UNT] y «Delincuencia» [CRI].

Posteriormente, tres rondas de codificación selectiva sirvieron para identificar, a partir de los datos, palabras clave y secuencias de palabras donde al menos tres personas anotadoras coincidían en la categoría pertinente. Estas se contrastaron con léxicos existentes de discurso de odio: HADES (Tulkens et al., 2016) y HurtLex v1.2 (Bassignana et al., 2018).

Las listas de palabras clave resultantes para cada categoría constituyen los léxicos en los que se basan nuestros modelos de clasificación, disponibles como código abierto (Van Haperen et al., 2026).

Rendimiento del modelo

Aplicamos los modelos de clasificación a nuestros conjuntos de datos y obtenemos los siguientes resultados, presentados en la Tabla 2.

Tabla 2. Rendimiento del modelo 2, objetivo: clasificación

	Alemán					Español					Neerlandés				
	N positivos	Precision	Recall	F1	Prevalence	N positivos	Precision	Recall	F1	Prevalence	N positivos	Precision	Recall	F1	Prevalence
Capacitismo	16	0.05	0.62	0.09	.00	25	0.26	0.32	0.29	.01	62	0.24	0.9	0.38	.01
Crimen	211	0.55	0.91	0.69	.01	28	0.44	1.00	0.62	.04	312	0.37	0.91	0.53	.05
Cultura	162	0.57	0.93	0.70	.07	47	0.71	0.68	0.70	.02	348	0.38	0.95	0.54	.06
Género	140	0.51	0.53	0.52	.03	63	0.54	0.52	0.53	.01	106	0.51	0.79	0.62	.02
Insultos	1453	0.79	0.43	0.56	.29	528	0.77	0.42	0.54	.10	771	0.53	0.67	0.59	.12
Ironía	291	0.07	0.09	0.08	.06	240	0.14	0.09	0.11	.04	418	0.11	0.02	0.04	.07
Política	981	0.63	0.48	0.55	.43	522	0.84	0.46	0.59	.73	3894	0.77	0.44	0.56	.62
Raza	54	0.79	0.89	0.83	.03	19	0.45	0.74	0.56	.00	44	0.39	0.86	0.54	.01
Religión	37	0.56	0.62	0.59	.02	25	0.79	0.44	0.56	.01	360	0.59	0.93	0.72	.06
Intimidación S.	242	0.88	0.77	0.82	.05	24	0.89	0.67	0.76	.01	178	0.79	0.98	0.87	.03
Orientación S.	56	0.44	0.98	0.61	.02	20	0.53	0.90	0.67	.01	50	0.37	0.88	0.52	.01
Falsedad	363	0.24	0.05	0.09	.07	70	0	0	0	.02	194	0.1	0.02	0.03	.03
Violencia	111	0.45	0.57	0.50	.02	53	0.95	0.38	0.54	.02	499	0.85	0.48	0.61	.08
Vulgar	94	0.49	0.55	0.52	.04	330	0.91	0.39	0.55	.06	175	0.79	0.82	0.80	.03

Nota. Cada léxico se elaboró utilizando únicamente datos de entrenamiento. La puntuación F1 aquí indicada se evaluó en un conjunto de prueba independiente, excluido del conjunto de entrenamiento. Para tener en cuenta las grandes diferencias en el número de observaciones y abordar los desequilibrios de clase, se adoptaron umbrales más estrictos para mejorar la recuperación de las categorías más amplias de Política y Insultos en todos los idiomas; Crimen en español y alemán; y Cultura, Raza, Religión, Orientación sexual y Lenguaje vulgar en alemán.

Informamos de la puntuación F1 como medida de la capacidad del modelo para asignar correctamente las publicaciones a las categorías. Observamos que la prevalencia de algunas de nuestras categorías es muy baja. Por ejemplo, en nuestro conjunto de datos alemán, una de cada 1.000 publicaciones recibe la etiqueta «Capacitismo». El rendimiento de los modelos está estrechamente asociado con la prevalencia: las categorías menos frecuentes proporcionan menos datos para entrenar un modelo, lo que dificulta una clasificación fiable. En general, cuanto menor es la puntuación F1, mayor cautela debemos aplicar al examinar la categorización. Consideramos sólidas las puntuaciones F1 superiores a 0,70 e inaceptables las inferiores a 0,50. Esto nos llevó a excluir de los análisis posteriores las categorías «Capacitismo», «Ironía» y «Falsedad».

En los tres idiomas, las categorías que presentan un buen rendimiento son «Intimidación sexual» y «Género», que contienen palabras clave relativamente claras y no ambiguas. La baja prevalencia de las categorías «Capacitismo», «Raza», «Orientación sexual» y «Lenguaje violento» puede explicarse porque probablemente son las categorías sometidas a una moderación más estricta. Las categorías «Política» e «Insultos y ataques personales» constituyen valores atípicos con una alta prevalencia como resultado de decisiones de codificación. Durante la codificación axial, se combinaron muchas subetiquetas bajo la categoría «Política», entre ellas ideología, políticas públicas y figuras políticas, lo que la convirtió en la categoría más amplia y heterogénea. Para abordarlo, se aplicó posteriormente un umbral estricto a esta categoría.

De acuerdo con nuestro enfoque abductivo de codificación, evaluamos críticamente a lo largo de todo el proceso si nuestras categorías seguían reflejando lo que encontrábamos en los datos. Llevamos a cabo una extensa inspección manual de los datos y de las etiquetas asignadas para todas las categorías, plataformas y países, mediante revisión manual y comprobaciones aleatorias puntuales. En particular, examinamos manualmente los falsos positivos y los falsos negativos, y debatimos los casos límite en sesiones de calibración. En la

mayoría de los casos, la inspección manual nos llevó a introducir cambios, principalmente la exclusión de palabras concretas con significados ambiguos.

Un ejemplo es la palabra «Let», que en neerlandés designa a una persona procedente de Letonia y que, por tanto, formaba parte de la categoría «Cultura», pero que aparece con mayor frecuencia en la expresión «Let op», que significa «cuidado» o «presta atención». Del mismo modo, la palabra «Frans», «francés», aparece más habitualmente en neerlandés como nombre propio, por ejemplo, Frans Timmermans, que como referencia a la cultura francesa. Como resultado de ambigüedades similares, del bajo rendimiento de los modelos y de la deliberación a la luz de los objetivos del proyecto, decidimos excluir las categorías «Falsedad» e «Ironía y sarcasmo». La categoría «Capacitismo» surgió durante la codificación inductiva, pero fue demasiado poco frecuente en la anotación posterior como para generar resultados fiables sobre el rendimiento del modelo, por lo que también fue excluida. Estas categorías se mantienen en el léxico y en la metodología, pero se omiten del análisis que sigue en aras de la validez.

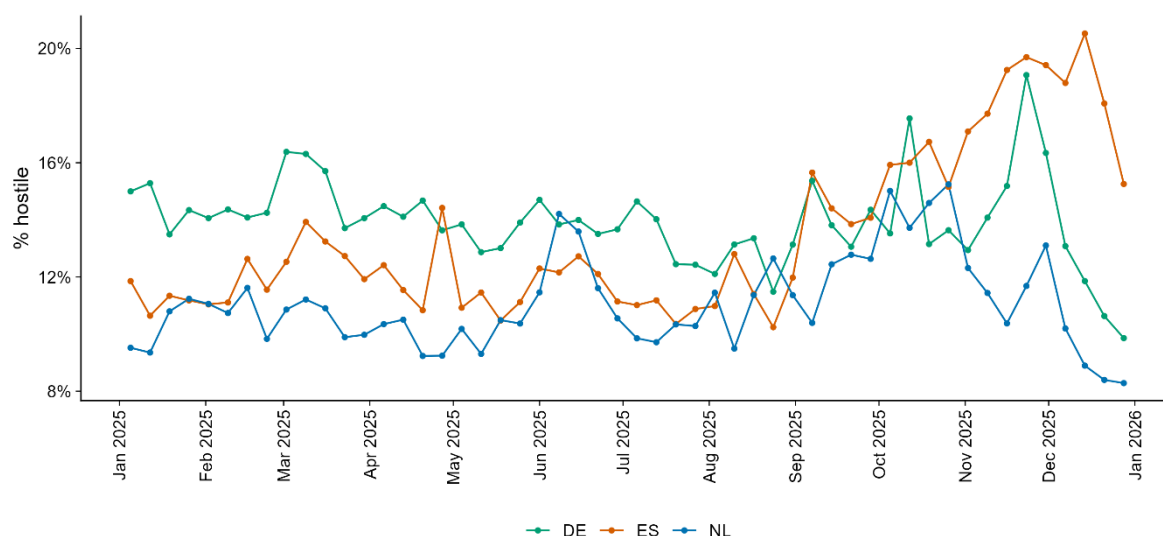
Hallazgos

Esta sección presenta lo siguiente. En primer lugar, se ofrecen descripciones de cada una de las catorce categorías, que aportan una perspectiva cualitativa sobre los distintos tipos de hostilidad en línea identificados y analizados. En segundo lugar, los estadísticos descriptivos generales proporcionan una visión general de las tendencias globales: las proporciones a lo largo del tiempo para cada país, cada plataforma y cada categoría de hostilidad en línea. En tercer lugar, el análisis de la focalización, la persistencia y la agrupación ofrece información sobre distintos tipos de comportamiento de quienes difunden hostilidad. En conjunto, estos análisis proporcionan información aplicable sobre qué tipos de intervenciones son más adecuados para contrarrestar formas específicas de hostilidad en línea en cada contexto.

Patrones por país

La siguiente sección presenta los patrones generales del lenguaje identificado mediante la comparación entre los tres países, véase la Figura 1.

Figura 1. Volumen de hostilidad en línea a lo largo del tiempo en España, Alemania y los Países Bajos



Nota: Porcentaje de posts identificados como hostilidad en línea, agregados por semana y por país.

En nuestro conjunto de datos, identificamos que el 13 % de todas las publicaciones contenían hostilidad en línea, aunque un día típico presenta aproximadamente un 8,8 %. Destacamos que realizamos el muestreo en cuentas con mayor probabilidad de recibir hostilidad, lo que también orientó nuestra anotación. En la práctica, el porcentaje presentado aquí no es representativo de estos países ni de estas plataformas en términos generales y probablemente sobreestima la hostilidad en línea.

La diferencia entre la media y la mediana, así como los picos observados en días concretos y la variabilidad de los datos, con una desviación estándar de 1.543, un mínimo de 1 y un máximo de 21.396, sugieren que la hostilidad en línea está impulsada por incrementos repentinos de atención probablemente relacionados con acontecimientos del mundo real, dinámicas específicas de las plataformas o actividad coordinada. Gran parte del contenido de odio de nuestro conjunto de datos se concentra en días específicos, junto con una actividad de base constante en cada plataforma. Estos incrementos de contenido de odio se agrupan en torno a fechas concretas, en lugar de producirse en días aislados al azar, lo que sugiere que esta actividad está relacionada con acontecimientos informativos específicos que mantienen la atención pública.

Existen diferencias notables entre los países en relación con las distintas categorías de hostilidad en línea, véase la Tabla 3.

Tabla 3. Volumen de hostilidad en línea por país, todas las plataformas

Category	n ES	Mean % ES	Median % ES	SD % ES	n DE	Mean % DE	Median % DE	SD % DE	n NL	Mean % NL	Median % NL	SD % NL
Ableism	593	0.3	0	3.5	668	2.4	0	6.1	58910	3.1	3.1	6.4
Crime	4546	1.8	0	6.9	622	2.7	0	6.4	64428	3.0	3.1	4.0
Culture	611	0.3	0	2.5	261	1.4	0	6.5	72226	4.0	3.8	8.0
Gender	2139	1.9	0	8.5	261	1.1	0	2.4	20354	1.4	1.2	4.5
Insults	5111	3.3	0	10.6	1684	6.4	5.3	7.5	159104	8.4	8.8	10.9
Politics	14847	6.6	0	13.9	1456	5.5	4.2	8.2	253938	10.2	9.2	9.6
Race	1673	0.8	0	4.2	20	0.1	0	0.6	8812	0.6	0.4	3.2
Religion	569	0.2	0	1.4	130	0.6	0	4.1	42986	1.6	1.2	3.5
Sexual intimidation	322	0.3	0	3.5	66	0.2	0	1.2	7475	0.4	0.3	0.5
Sexual orientation	578	0.4	0	4.0	559	2.4	0	5.1	17544	1.0	0.9	1.9
Violent speech	97	0.0	0	0.7	267	1.0	0	2.5	7632	0.4	0.3	0.6
Vulgar language	1418	1.1	0	6.0	504	2.2	0	6.0	12722	0.8	0.6	4.3

Estadísticos descriptivos de los porcentajes por día (número de publicaciones en una categoría como proporción del número total diario de publicaciones identificadas como discursos de odio).

De esta comparación se derivan varias conclusiones. La hostilidad relacionada con la política, los insultos y la delincuencia constituye la categoría más frecuente en cada idioma. En alemán, el lenguaje vulgar y, en menor medida, la orientación sexual son relativamente más frecuentes que en neerlandés o español. En neerlandés, la hostilidad relacionada con la cultura es relativamente más común que en los demás idiomas. La hostilidad relacionada con el género, la raza, la religión, la intimidación sexual, el lenguaje violento y el lenguaje vulgar es relativamente menos destacada en nuestro conjunto de datos para cada uno de los idiomas.

Las medianas difieren de forma notable entre los países, lo que sugiere que la proporción de hostilidad correspondiente a la mayoría de las categorías se produce de forma más constante a lo largo del tiempo en los Países Bajos, mientras que en Alemania y España depende en mayor medida de picos de actividad. Esta conclusión se ve reforzada por las desviaciones estándar relativamente elevadas en categorías como la política en España. Categorías como el lenguaje violento y la intimidación sexual aparecen de forma más constante a lo largo del

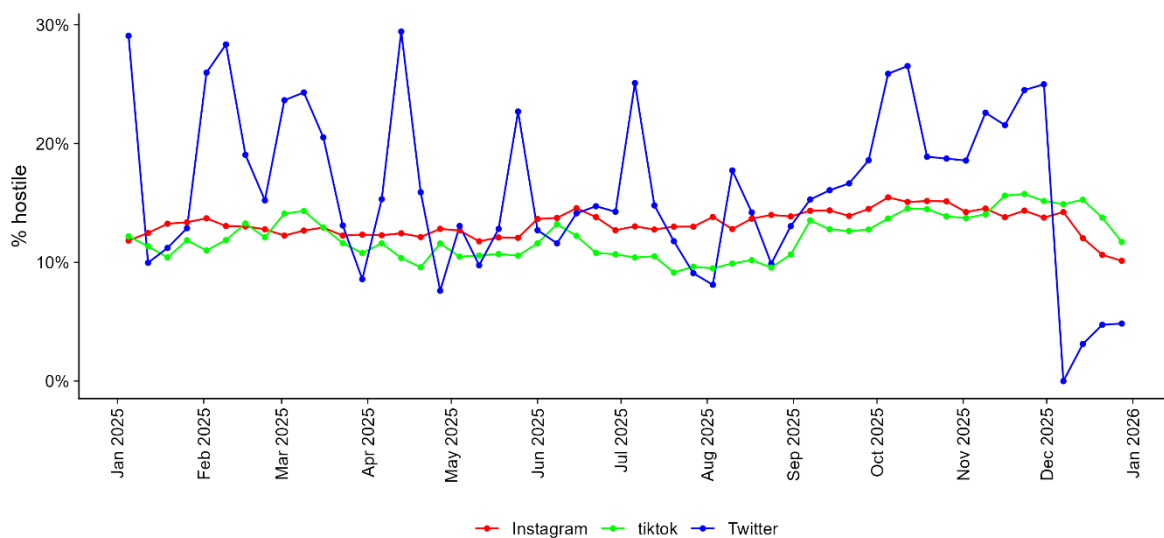
tiempo, mientras que la política y los insultos se concentran más en episodios de actividad, especialmente en España. En conjunto, esto sugiere la existencia de picos de actividad que pueden estar estrechamente vinculados a acontecimientos de actualidad, una dinámica más pronunciada en España.

En la siguiente sección de este informe analizamos con mayor detalle estas diferencias entre países para cada categoría.

Patrones por plataforma

La siguiente sección presenta los patrones generales del lenguaje identificado mediante la comparación entre las tres plataformas, véase la Figura 2.

Figura 2. Volumen de hostilidad en línea a lo largo del tiempo, por plataforma



Nota: Porcentaje de posts identificados como hostilidad en línea, agregados por semana y por plataforma.

Se observan diferencias notables entre las plataformas. En nuestro conjunto de datos, X destaca por presentar la mayor proporción de hostilidad en línea y un patrón volátil a lo largo del tiempo. Instagram presenta proporciones de hostilidad en línea más elevadas que TikTok, pero ambas plataformas parecen relativamente constantes a lo largo del tiempo, sin una tendencia clara al alza ni a la baja. Estas tendencias están impulsadas por incrementos

repentinos de actividad en días concretos, combinados con un nivel de actividad de base en cada plataforma.

Existen diferencias notables entre los países en relación con las distintas categorías de hostilidad en línea, véase la Tabla 4.

Tabla 4. Proporciones por plataforma, todos los países

Categoría	Total n	n Instagram	Mean % Insta	Median % Insta	SD % Insta	n X	Mean % X	Median % X	SD % X	n TikTok	Mean % TikTok	Median % TikTok	SD % TikTok
Crime	69596	4827	2.3	0	7.0	61312	4.7	4.0	7.0	3457	1.5	1.1	2.8
Culture	73098	3314	1.2	0	4.1	68364	4.1	4.1	6.1	1420	1.3	0.3	6.1
Gender	22754	2719	2.1	0	8.1	18705	1.3	1.1	2.7	1330	0.8	0	3.6
Insults	165899	10054	4.9	0	10.6	151369	10.3	11.3	9.4	4476	2.8	1.9	7.3
Politics	270241	9979	6.3	0	13.2	245117	15.0	16.4	12.3	15145	7.0	6.0	8.0
Race	10505	1020	0.6	0	3.5	8346	2.1	0.5	9.9	1139	0.8	0	3.7
Religion	43685	1440	0.4	0	1.9	41604	2.3	1.5	3.8	641	0.5	0	3.6
Sexual intimidation	7863	626	0.4	0	3.4	7047	0.8	0.4	5.7	190	0.1	0	0.3
Sexual orientation	18681	1405	0.7	0	4.1	16421	1.2	0.9	2.7	855	0.6	0	1.2
Violent speech	7996	581	0.2	0	0.8	7188	0.5	0.4	2.2	227	0.1	0	0.6
Vulgar language	14644	1728	1.2	0	5.9	12174	1.9	0.8	7.5	742	0.3	0	0.7

Nota: Estadísticos descriptivos de los porcentajes (número de publicaciones de una categoría como proporción del total diario de publicaciones identificadas como discurso de odio). Téngase en cuenta que las categorías no son mutuamente excluyentes; por ejemplo, una misma publicación puede asignarse tanto a «política» como a «insultos».

En nuestro conjunto de datos, X presenta las mayores proporciones de hostilidad en línea en todas las categorías, especialmente en política e insultos, mientras que TikTok muestra de forma consistente la proporción más baja e Instagram ocupa en general una posición intermedia. La excepción destacada es el contenido relacionado con el género, que es más relevante en Instagram.

En comparación con Instagram y TikTok, X presenta los porcentajes medios más elevados en 11 de las 13 categorías, entre ellas política, 15,0 %; insultos, 10,3 %; delincuencia, 4,7 %; cultura, 4,1 %; religión, 2,3 %; raza, 2,1 %; y lenguaje vulgar, 1,9 %. Instagram ocupa generalmente una posición intermedia. Aunque la mayoría de las categorías son menos pronunciadas que en X, Instagram supera a X en hostilidad relacionada con el género, 2,1 %

frente al 1,3 %, lo que convierte al género en la única categoría en la que Instagram presenta claramente la proporción media más elevada. TikTok presenta de forma consistente la proporción más baja en todas las categorías. Los porcentajes medios son los más bajos para la mayoría de las formas de contenido de odio, particularmente para la religión, 0,5 %; la intimidación sexual, 0,1 %; el lenguaje violento, 0,1 %; y el lenguaje vulgar, 0,3 %. La política sigue siendo la categoría más frecuente en TikTok, pero con una proporción inferior a la mitad de la observada en X.

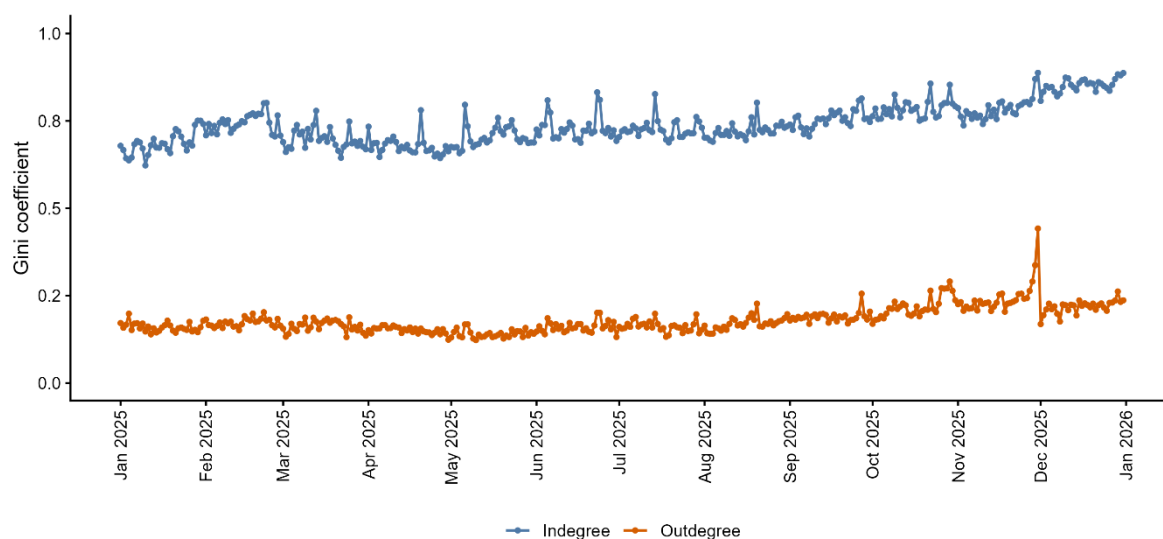
Las medianas sugieren que los incrementos repentinos de actividad generan una gran proporción del contenido hostil en la mayoría de las plataformas. Las medianas de X para política, 16,4 %; insultos, 11,3 %; cultura, 4,1 %; y delincuencia, 4,0 %, sugieren que estos temas mantienen un nivel de base constante a lo largo del tiempo. La variación también es mayor en política e insultos. Las desviaciones estándar son más elevadas para estas categorías en todas las plataformas, especialmente en X, con una desviación estándar de 12,3 para política y de 9,4 para insultos, lo que indica una volatilidad considerable a lo largo del tiempo.

En conjunto, estos hallazgos sugieren diferencias claras entre las plataformas en la composición de la hostilidad en línea. En nuestro conjunto de datos, X se caracteriza tanto por proporciones más elevadas como por una mayor volatilidad entre las categorías, en particular para los insultos y la hostilidad relacionada con la política. Instagram, en cambio, muestra proporciones relativamente más bajas, aunque una mayor proporción de hostilidad relacionada con el género, y una mayor estabilidad a lo largo del tiempo. TikTok presenta las proporciones más bajas para la mayoría de las categorías de contenido hostil que detectamos.

Patrones entre objetivos y difusores

La siguiente sección presenta los patrones generales entre los objetivos y los difusores de hostilidad en línea. En primer lugar, nos centramos en el grado de concentración de la atención hacia objetivos específicos, véase la Figura 3.

Figura 3. Concentración de la hostilidad en línea hacia los objetivos (indegree) y desde los difusores (outdegree)



Esta figura muestra cómo se distribuye la atención entre los objetivos y los difusores de hostilidad en línea. Un coeficiente de Gini más elevado indica una mayor desigualdad entre los objetivos o los difusores: un grupo relativamente reducido de cuentas recibe o envía una proporción desproporcionadamente elevada de publicaciones. La conclusión principal es que existen periodos en los que la atención hostil se concentra más, de modo que un número reducido de objetivos recibe una proporción desproporcionadamente elevada de todo el contenido hostil de nuestro conjunto de datos. Esto indica que determinados acontecimientos generan una atención más intensa hacia objetivos concretos.

La Figura 3 también muestra el grado de concentración de la hostilidad en línea entre los difusores. En este caso, un coeficiente de Gini más elevado indica que una proporción menor de cuentas difusoras es responsable de una parte relativamente elevada del contenido hostil que identificamos en nuestro conjunto de datos. La conclusión principal es el bajo grado general de concentración, lo que indica que los ataques proceden de muchas cuentas diferentes, en lugar de que un pequeño número de cuentas genere la mayor parte del contenido hostil de nuestro conjunto de datos.

Un hallazgo sorprendente es que la concentración, tanto entre los objetivos como entre los difusores, aumenta en momentos similares. En otras palabras, esto ocurre cuando un número reducido de personas usuarias representa una proporción desproporcionada de las publicaciones recibidas o enviadas. También observamos similitudes en el incremento gradual

de la concentración entre difusores y objetivos, aunque esta tendencia es menos pronunciada entre los difusores. Sospechamos que, cuando una persona adquiere notoriedad en la cobertura mediática, los ataques se concentran más en ella y, al mismo tiempo, una proporción relativamente menor de cuentas genera una mayor parte de la hostilidad en línea. Consideramos que esto constituye una indicación adicional de que el contenido hostil está vinculado a acontecimientos de actualidad.

Patrones de comportamiento

Para diferenciar con mayor precisión a quienes difunden hostilidad en línea, examinamos su comportamiento de focalización y persistencia. Al comparar a los difusores de hostilidad en línea según la «focalización», es decir, a cuántos objetivos dirige una persona publicaciones de odio, donde tres objetivos o menos indican una medida de concentración, y la «persistencia», es decir, durante cuántos días una persona dirige publicaciones de odio al mismo objetivo, donde tres días o más a lo largo de un año indican una medida de persistencia, distinguimos cuatro tipos de comportamiento diferentes, véase la Tabla 5.

Tabla 5. Agrupaciones conductuales entre los difusores: focalización y persistencia

Cluster	Tipo	Promedio de focalización	Promedio de persistencia	n
1	Generalistas	11.8	1.05	863
2	Fijadas	1	9.33	3
3	Reincidentes	2.98	2.21	174
4	Oportunistas	1.61	1.00	14669

Un primer grupo está formado por una masa de «oportunistas», personas usuarias de baja intensidad que representan el 94 % de las cuentas que difunden contenido problemático en nuestro conjunto de datos. Se trata de cuentas con una focalización mínima y sin persistencia: publicaciones de odio aisladas dirigidas a uno o dos objetivos. Un segundo grupo es un conjunto de tamaño medio, compuesto por 863 cuentas, de «generalistas»: cuentas agresivas que se dirigen a muchos objetivos distintos, pero con muy poca persistencia. Se trata de ataques aislados dirigidos simultáneamente a numerosas personas diferentes.

Un tercer grupo, el de las personas «reincidentes», es reducido, con 174 cuentas, y presenta una focalización y una persistencia moderadas. Este grupo representa a personas que participan en una variedad de temas durante períodos más prolongados y que lo hacen de forma constante mediante lenguaje problemático. Un último grupo está compuesto por personas «fijadas», casos extremos que se obsesionan con un único objetivo durante más de una semana. Contrariamente a lo esperado, estas cuentas son muy poco frecuentes en nuestro conjunto de datos. Una explicación probable podría ser la eficacia de la moderación de las plataformas: las cuentas más problemáticas pueden ser denunciadas y suspendidas, lo que conllevaría su exclusión de nuestra recopilación de datos.

Patrones por categoría

Descripción de las categorías

Las siguientes categorías son el resultado de varias iteraciones de codificación inductiva. Cabe señalar que estas categorías no son exhaustivas. Aunque cubrieron las muestras extraídas de nuestro conjunto de datos para su anotación, debe asumirse que los datos recopilados en otros países, plataformas o períodos requerirán ampliar esta lista. Además, persiste cierto grado de interpretación y estas categorías pueden incluir lenguaje identificado como odio por nuestras personas anotadoras, aunque otra persona lectora podría no considerarlo como tal. Asimismo, estas categorías no son mutuamente excluyentes: una publicación puede clasificarse simultáneamente como «Lenguaje vulgar» y como «Insulto».

Los volúmenes generales de hostilidad en línea pueden diferenciarse adicionalmente por categoría de la siguiente manera, véase la Tabla 6.

Tabla 6. Estadísticos descriptivos por categoría, todas las plataformas y todos los países

Category	Total n	Mean	Median	Mode	SD
Crime	8318	2.67	2.620	0	1.100
Culture	4793	1.65	1.610	0	0.970
Gender	3539	1.36	1.290	0	0.823
Insults	14467	5.05	5.040	5.38	1.790

Politics	26437	8.320	7.950	8.8	2.830
Race	2142	0.857	0.696	0	0.968
Religion	2124	0.688	0.648	0	0.509
Sexual intimidation	834	0.277	0.244	0	0.291
Sexual orientation	2208	0.845	0.741	0	0.803
Violent speech	818	0.278	0.229	0	0.307
Vulgar language	2252	0.827	0.791	0	0.540

Estadísticos descriptivos de los porcentajes (número de publicaciones en una categoría como proporción del total diario de publicaciones identificadas como discurso de odio).

En primer lugar, varias categorías están impulsadas principalmente por acontecimientos. Estas categorías presentan una elevada volatilidad: «Raza», 0,9 %, con una desviación estándar de 0,968; «Intimidación sexual», 0,3 %, con una desviación estándar de 0,291; y «Lenguaje violento», 0,3 %, con una desviación estándar de 0,307. Este tipo de hostilidad puede ser relativamente difícil de anticipar y podría requerir respuestas rápidas, ya que tiende a concentrarse fuertemente en unos pocos objetivos y a ser difundida mayoritariamente por personas usuarias transitorias.

En segundo lugar, las categorías «Política», «Insultos», «Delincuencia» y «Cultura» combinan un nivel de base constante con volatilidad. Estas categorías representan la mayor proporción de hostilidad en línea en nuestro conjunto de datos. La política y los insultos o ataques personales representan, respectivamente, el 8,3 % y el 5,1 % de toda la hostilidad en línea diaria. La delincuencia, 2,7 %, con una desviación estándar de 1,100, presenta una variabilidad notablemente elevada, lo que sugiere que está impulsada principalmente por acontecimientos. En el caso de la cultura, 1,7 %, con una desviación estándar de 0,970, la variabilidad es relativamente baja, lo que sugiere que este tipo de contenido aparece de forma constante a diario. Dado que la mediana y la moda son similares a la media, tanto la política como los insultos presentan una mayor estabilidad en su nivel diario de base, aunque las desviaciones estándar indican una variabilidad relativamente elevada en los picos. Esto sugiere que el contenido de odio en estas categorías está impulsado tanto por un nivel de actividad constante como por picos episódicos de actividad.

En tercer lugar, las categorías con un nivel de base estable dependen menos de los acontecimientos y aparecen de forma más constante a lo largo del tiempo. Estas son:

«Género», 1,4 %, con una desviación estándar de 0,823; «Religión», 0,7 %, con una desviación estándar de 0,509; «Orientación sexual», 0,8 %, con una desviación estándar de 0,803; y «Lenguaje vulgar», 0,8 %, con una desviación estándar de 0,540. En estos casos, las desviaciones estándar indican una variabilidad relativamente baja, lo que sugiere que este contenido aparece de forma constante a diario en lugar de depender fuertemente de los acontecimientos. Estas categorías constituyen una base estructural diaria de contenido de odio. Podrían beneficiarse de una monitorización continua centrada en objetivos recurrentes, en lugar de intervenciones reactivas o de estrategias enfocadas en personas reincidentes.

Aquí deben destacarse dos salvedades importantes. En primer lugar, como se explica en la metodología, recopilamos datos de redes sociales que aún estaban disponibles públicamente. Para entonces, las publicaciones ya habían sido sometidas a moderación por parte de las plataformas antes de incorporarse a nuestro conjunto de datos. Es plausible que las categorías del primer grupo sean las más moderadas y que, por ello, aparezcan como relativamente menos frecuentes y más volátiles en nuestro análisis. En segundo lugar, subrayamos que la prevalencia no equivale al impacto. Las publicaciones de estas categorías contienen las formas más virulentas de hostilidad en línea de nuestro conjunto de datos y cabe suponer razonablemente que tienen un impacto desproporcionado, incluso si son menos frecuentes.

En conjunto, la política y los insultos personales representan la mayor proporción de publicaciones marcadas como problemáticas por nuestros modelos. Esto sugiere que la hostilidad en línea en los distintos países y plataformas no es aleatoria, sino que presenta características estructurales: algunas formas dependen en gran medida de acontecimientos, mientras que otras tienen un nivel de base más constante. Esto puede ayudar a decidir dónde invertir recursos para intervenciones contra la hostilidad en línea.

A continuación, examinamos los objetivos y el comportamiento de los difusores en cada una de las categorías.

Delincuencia

Esta categoría se refiere al lenguaje sobre comportamiento delictivo, actividades ilegales y la policía.

- **Ejemplo en español:** «quien venga a delinquir fuera».
- **Ejemplo en alemán:** «Das sind kriminell die wo andere Mitmenschen gefährden die müssen alle in Therapie und weggesperrt werden !!!».
- **Ejemplo en neerlandés:** «Die gore crimineel zou harder gestraft moeten worden».

El contenido relacionado con la delincuencia es moderadamente frecuente en nuestro conjunto de datos. En Instagram se observa un aumento gradual, mientras que en TikTok los niveles son más variables y presentan un ligero descenso general tras los picos iniciales. Esta categoría es volátil en X, sin una tendencia sostenida. Entre países, se trata de un tipo de contenido relativamente destacado en Alemania, moderadamente destacado en España y uno de los tipos de contenido menos destacados en los Países Bajos, en comparación con otras categorías.

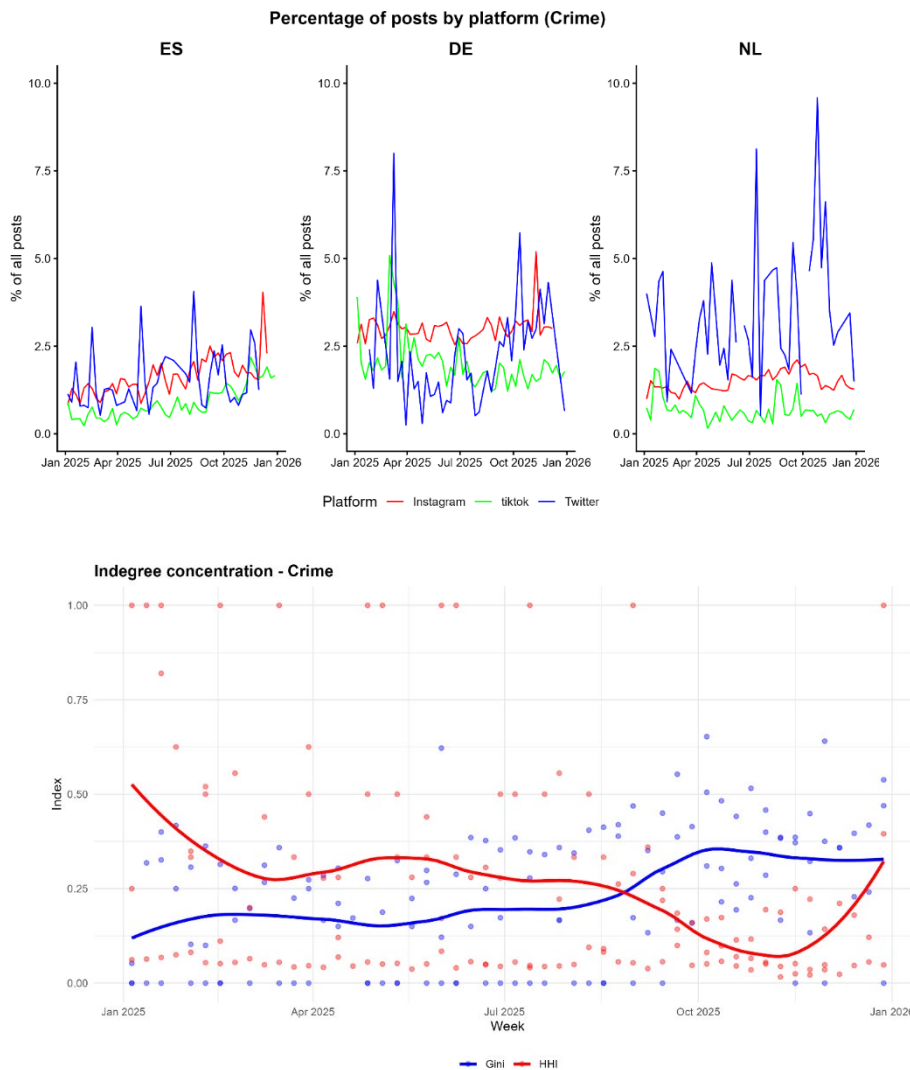
En cuanto a los objetivos, observamos una concentración moderada combinada con un conjunto de objetivos relativamente estable, con un coeficiente de Gini medio de 0,23 y un HHI de 0,25. La concentración se mantiene constante a lo largo del tiempo, con escasas tendencias sistemáticas de aumento o disminución. Se observa una disminución del HHI paralela a un aumento de la actividad hacia el final del año, lo que indica una atención fluctuante a lo largo del tiempo. No obstante, los objetivos de esta categoría se mantienen relativamente concentrados de forma constante. Esto sugiere un patrón de desplazamiento de la atención entre un conjunto cambiante de objetivos, en lugar de una atención sostenida sobre cuentas específicas.

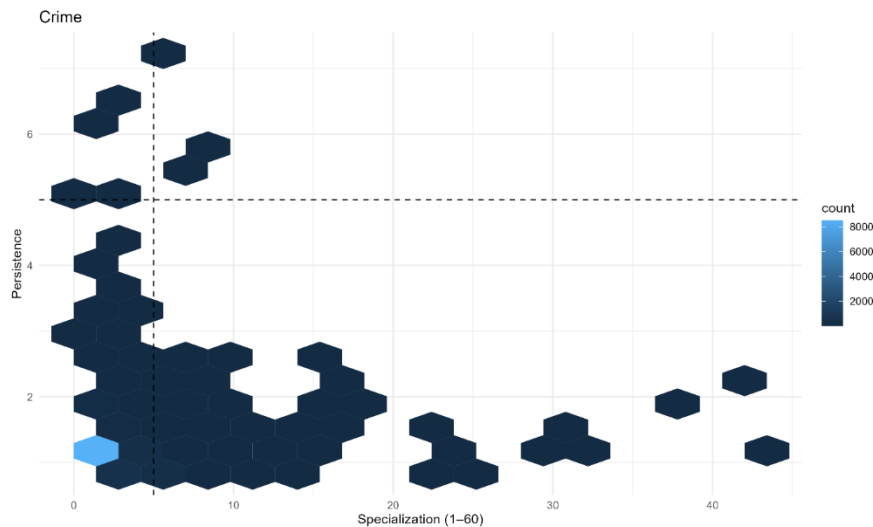
Los difusores de contenido relacionado con la delincuencia están, en comparación con otras categorías de nuestro conjunto de datos, igualmente concentrados: el 95 % presenta una focalización concentrada y baja persistencia. Sin embargo, esta categoría presenta una de las proporciones más elevadas de focalización diversificada, 4,9 % con focalización difusa y baja persistencia, y una proporción insignificante del 0,1 % con focalización concentrada y alta persistencia. Esto sugiere que la hostilidad en línea relacionada con la delincuencia rara vez se mantiene en el tiempo y se combina con una focalización relativamente amplia.

En términos de intervención, estos hallazgos sugieren que la hostilidad en línea relacionada con la delincuencia es relativamente frecuente, aparece de forma constante a lo largo del tiempo, se centra en un conjunto relativamente amplio pero estable de objetivos y procede

de un conjunto transitorio y poco focalizado de difusores. Esto la convierte en una de las categorías más difusas y posiblemente más difíciles de abordar, con menos indicios de acontecimientos que puedan anticiparse.

Figura 4. Comparación de la hostilidad relacionada con la delincuencia según proporción, objetivos y difusores





Cultura

Esta categoría se refiere al lenguaje relacionado con culturas específicas y estereotipos culturales.

- **Ejemplo en español:** «Menos clase que un gitano recién llegado de Rumanía».
- **Ejemplo en alemán:** «Sofort auch raus mit den Migranten !!!! Keiner will euch !!!».
- **Ejemplo en neerlandés:** «smerige halve Marokkaan».

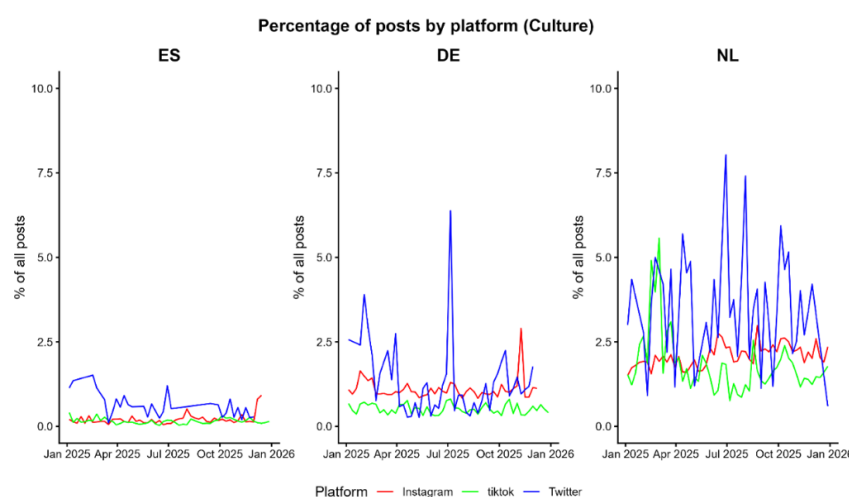
El contenido relacionado con la cultura es muy frecuente en nuestro conjunto de datos en neerlandés, pero notablemente menos frecuente en los conjuntos de datos español y alemán. Aparece de forma constante en las tres plataformas de los Países Bajos y es relativamente destacado tanto en Instagram como en TikTok, en comparación con otras categorías. En X, la hostilidad en línea relacionada con la cultura aparece con frecuencia y en picos de actividad.

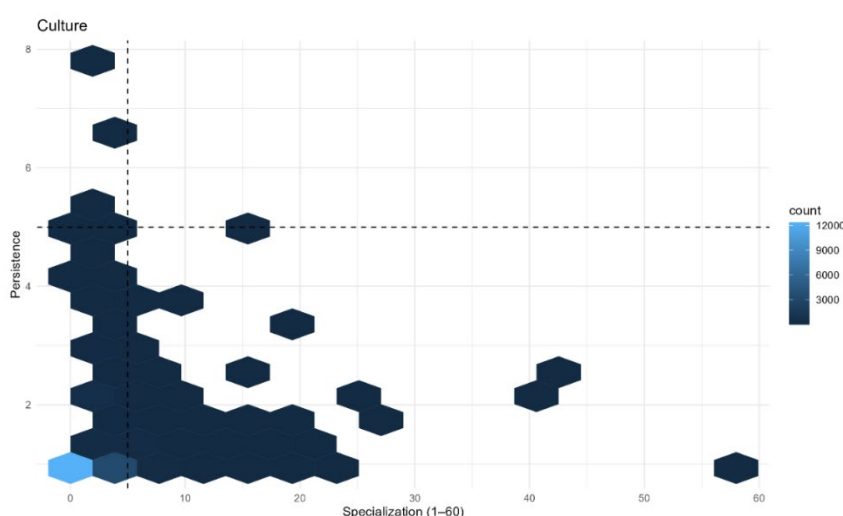
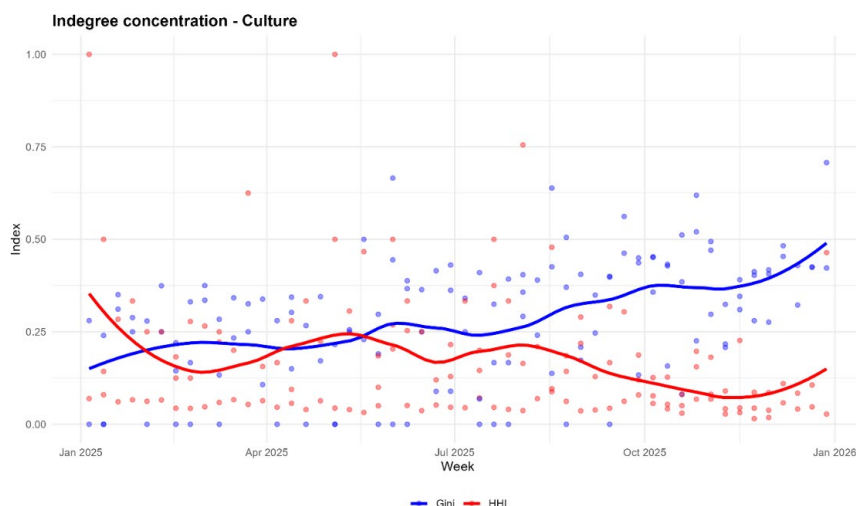
En cuanto a los objetivos, la hostilidad en línea relacionada con la cultura constituye una de las categorías más desiguales, combinada con un predominio relativamente bajo entre los objetivos, con un coeficiente de Gini medio de 0,28 y un HHI de 0,17. Esto indica un conjunto concentrado de objetivos que cambia con relativa frecuencia. Ambos patrones se mantienen constantes a lo largo del tiempo. Esto sugiere que esta hostilidad se distribuye de manera desigual entre los objetivos, pero sin objetivos persistentes y claramente predominantes.

Los difusores de hostilidad en línea relacionada con la cultura se encuentran entre los más focalizados y transitorios en comparación con otras categorías de nuestro conjunto de datos: el 97,4 % presenta una focalización concentrada y baja persistencia. Solo el 2,6 % de las personas usuarias diversifica sus objetivos, sin evidencia de persistencia a lo largo del tiempo. Esto sugiere que dichos difusores están muy fragmentados y se implican en un único mensaje.

En términos de intervención, estos hallazgos sugieren que la hostilidad en línea relacionada con la cultura es relativamente frecuente en neerlandés y aparece de forma constante a lo largo del tiempo en todas las plataformas, pero es mucho menos frecuente en español y alemán. En los distintos países, la atención en esta categoría se distribuye de forma desigual y muestra una mayor variación entre los objetivos; procede de un conjunto de difusores muy transitorio, pero relativamente focalizado. Esto la convierte en una de las categorías más difusas y posiblemente más difíciles de abordar, con menos indicios de acontecimientos que puedan anticiparse.

Figura 5. Comparación de la hostilidad relacionada con la cultura según proporción, objetivos y difusores





Género

Esta categoría se refiere al lenguaje intolerante dirigido contra aspectos específicos de la identidad relacionados con el género y los estereotipos de género. Incluye normas de género y misoginia.

- **Ejemplo en español:** «sin maquillaje, es una bruja».
- **Ejemplo en alemán:** «Ich kann diese dumme Hexe nicht mehr ertragen.».
- **Ejemplo en neerlandés:** «Alle vrouwen zijn dom».

La hostilidad en línea relacionada con el género es poco frecuente en nuestro conjunto de datos en todas las plataformas y países. Es especialmente rara en neerlandés y notablemente más común en español. En comparación con otras categorías, la fluctuación en X también es

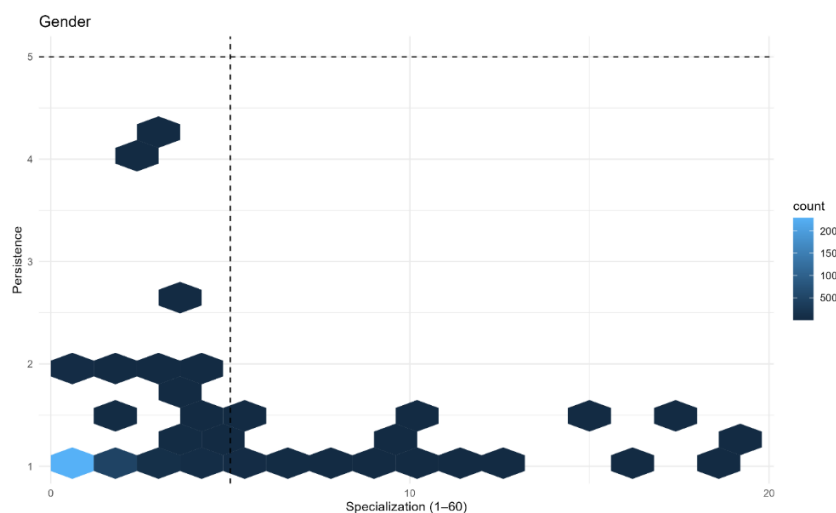
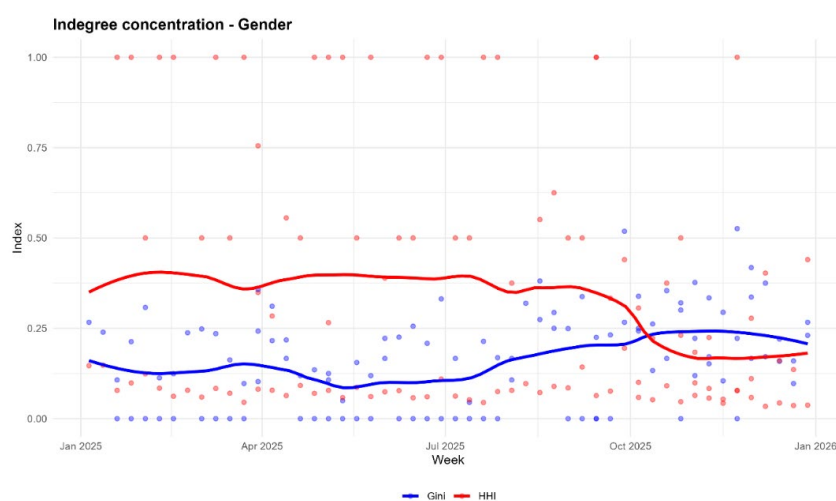
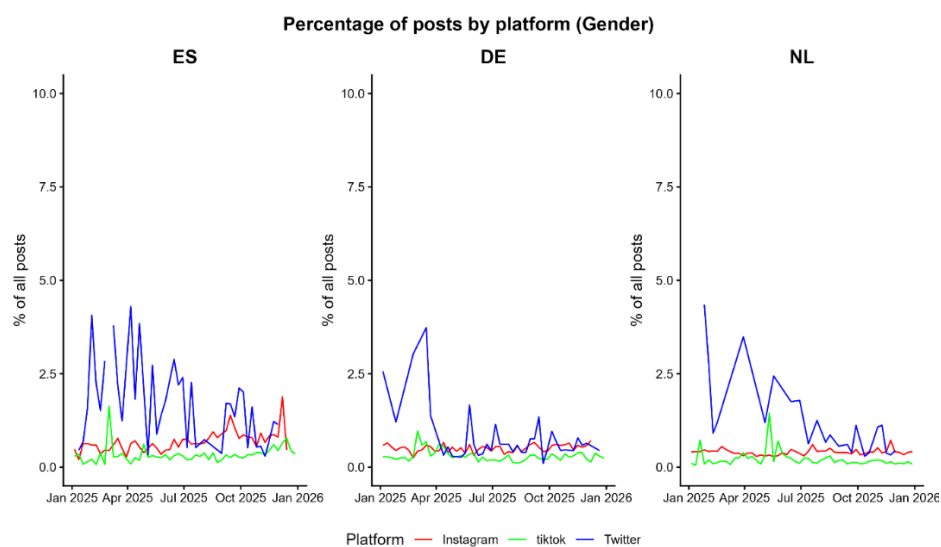
notablemente moderada. Esta tendencia relativamente estable se mantiene a lo largo del tiempo.

En cuanto a los objetivos, la hostilidad en línea relacionada con el género es relativamente difusa y se combina con un predominio relativamente elevado, con un coeficiente de Gini medio de 0,16 y un HHI de 0,32. Esto sugiere que, aunque la desigualdad es limitada, la atención se concentra en unos pocos objetivos clave. Esta tendencia se mantiene constante a lo largo del tiempo.

En cuanto a los difusores de hostilidad en línea, en nuestro conjunto de datos la categoría «Género» es la más concentrada en términos generales: el 98,3 % presenta una focalización concentrada y baja persistencia, con una diversificación mínima, 1,7 %, y sin personas usuarias persistentes. Esto sugiere que esta hostilidad está muy dirigida a objetivos específicos y que consiste en interacciones aisladas con una persistencia prácticamente inexistente.

En términos de intervención, «Género» es una categoría con un nivel de actividad de base constante, que requiere mayor atención en español que en los demás idiomas. Entre los países, la atención en esta categoría es relativamente difusa, pero presenta un núcleo relativamente estable de objetivos y procede de un conjunto de difusores transitorio y focalizado. Las intervenciones en esta categoría se orientan menos por una atención impulsada por acontecimientos, pero pueden dirigirse a cuentas destacadas que reciben de forma constante este tipo de hostilidad.

Figura 6. Comparación de la hostilidad relacionada con el género según proporción, objetivos y difusores



Insultos y ataques personales

Esta categoría se refiere a una amplia variedad de lenguaje destinado a insultar o atacar personalmente a una persona, sin dirigirse explícitamente contra identidades grupales. Es una categoría amplia que abarca no solo la hostilidad en línea, sino también la falta de civismo y la descortesía en términos más generales, incluido el lenguaje grosero, ofensivo y denigrante.

- **Ejemplo en español:** «Eres amiga de este no hay mas preguntas miserable».
- **Ejemplo en alemán:** «Du bist so erbärmlich.».
- **Ejemplo en neerlandés:** «Wat ben jij een ontzettende eikel».

La hostilidad en línea en forma de insultos o ataques personales es la segunda categoría más frecuente de nuestro conjunto de datos en todas las plataformas y países, y resulta particularmente destacada en neerlandés. Tiene una presencia notable en Instagram en comparación con otras categorías y es algo menos frecuente en TikTok. En X presenta una volatilidad constante. Instagram mantiene niveles altos de forma consistente, con una variación moderada a lo largo del tiempo, mientras que TikTok muestra una estabilidad similar con picos ocasionales.

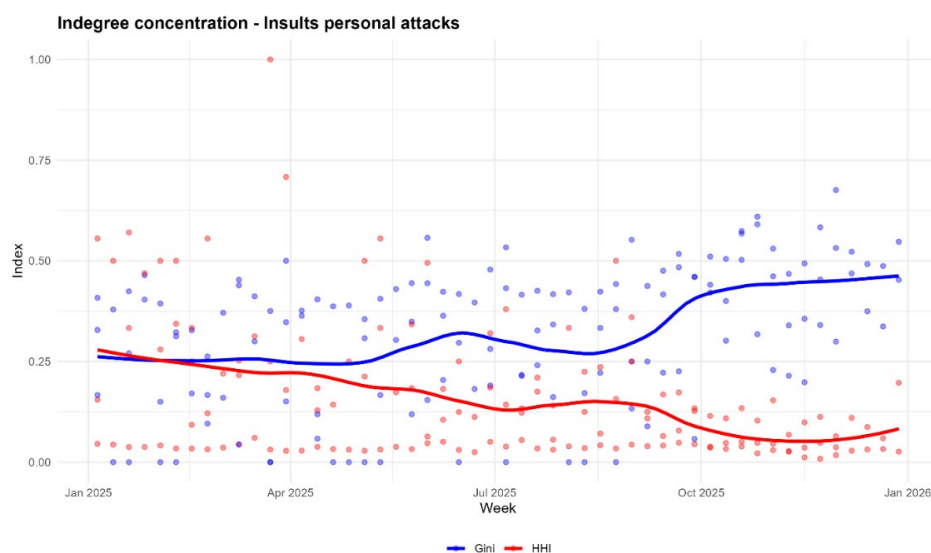
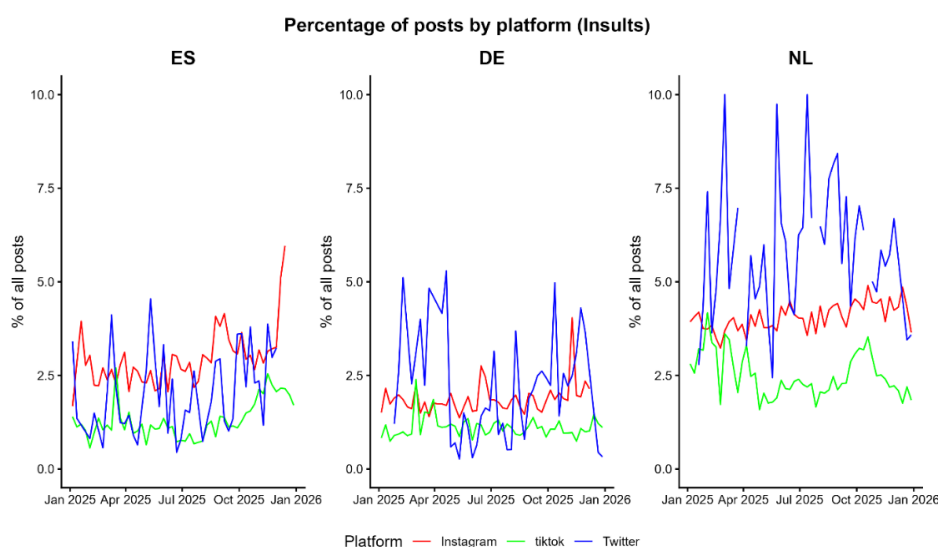
En cuanto a los objetivos, esta es la segunda categoría más desigual de nuestro conjunto de datos, combinada con un predominio relativamente bajo de los objetivos, con un coeficiente de Gini medio de 0,32 y un HHI de 0,15. Esto sugiere que los insultos están muy concentrados, pero sin objetivos focales constantes. Aunque en nuestro conjunto de datos observamos una divergencia creciente entre una mayor concentración y un menor predominio, la tendencia parece relativamente estable a lo largo del tiempo. En conjunto, estos patrones sugieren un comportamiento de focalización difuso, aunque desigual.

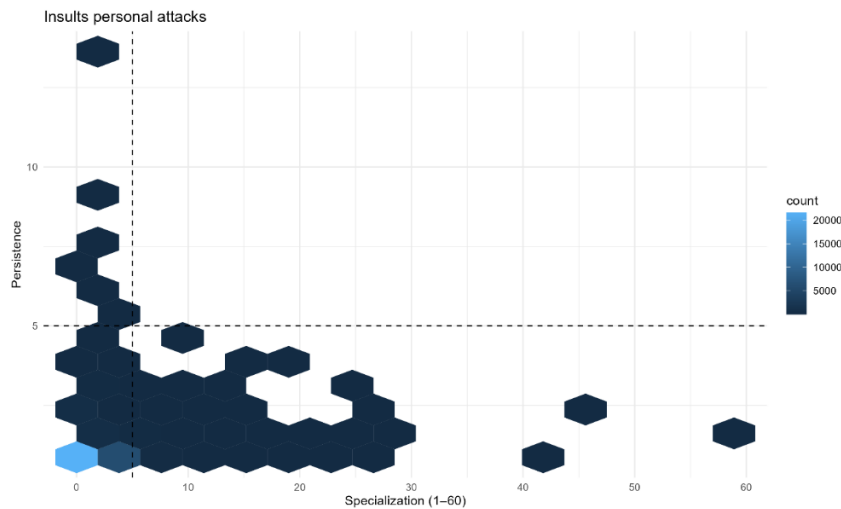
En cuanto a los difusores de insultos, encontramos una mayor variación en nuestro conjunto de datos que en otras categorías. Aunque sigue existiendo una fuerte concentración y poca persistencia, con un 97,2 % de focalización concentrada y baja persistencia, hay una proporción mayor de cuentas que insultan a un conjunto más amplio de objetivos, 2,7 % de focalización difusa, y una proporción comparativamente mayor de atacantes persistentes, 0,1 %.

En términos de intervención, la atención en esta categoría es muy desigual y presenta una amplia variación entre los objetivos en cada uno de los idiomas. Aunque es difícil anticipar qué

tipo de acontecimientos pueden impulsar este contenido y la atención se distribuye en gran medida entre muchos objetivos diferentes, estos patrones también sugieren la existencia de grupos centrales de difusores que insultan de forma persistente a objetivos específicos.

Figura 7. Insultos y ataques personales: comparación según proporción, objetivos y difusores





Política

Esta categoría se refiere a una amplia variedad de lenguaje relacionado con la política, las figuras políticas, los partidos políticos, políticas públicas específicas, guerras y conflictos, la ideología y la polarización entre la izquierda y la derecha.

- **Ejemplo en español:** «Basura woke».
- **Ejemplo en alemán:** «Selten so gelacht Hochverschuldung haben wir durch das Russen Drecksvolk die Freude von dem AfD Gesindel».
- **Ejemplo en neerlandés:** «Geert Wilders is een engnek».

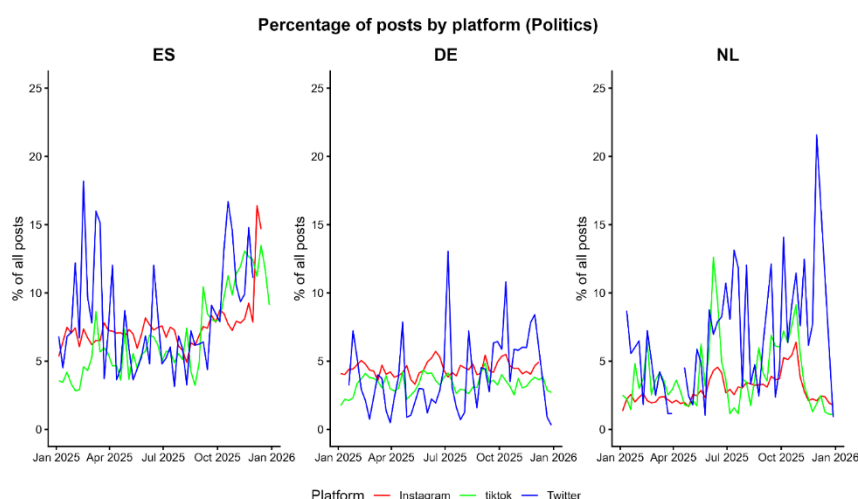
La hostilidad en línea relacionada con la política es la categoría más frecuente de nuestro conjunto de datos en todas las plataformas y países. Es relativamente volátil en todas las plataformas y presenta de forma constante altos niveles de variación a lo largo del tiempo, especialmente pronunciados en X. Esto sugiere que este contenido puede estar impulsado por acontecimientos y ciclos informativos.

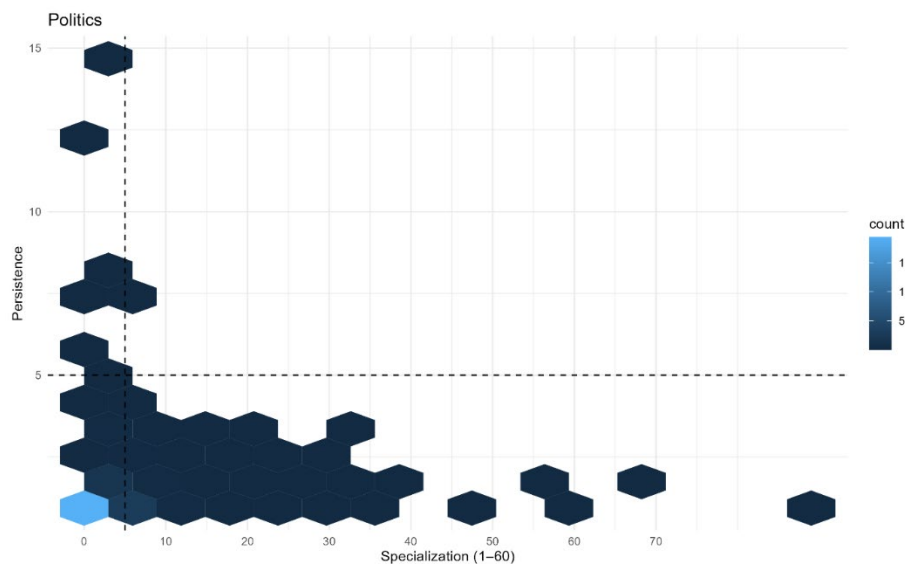
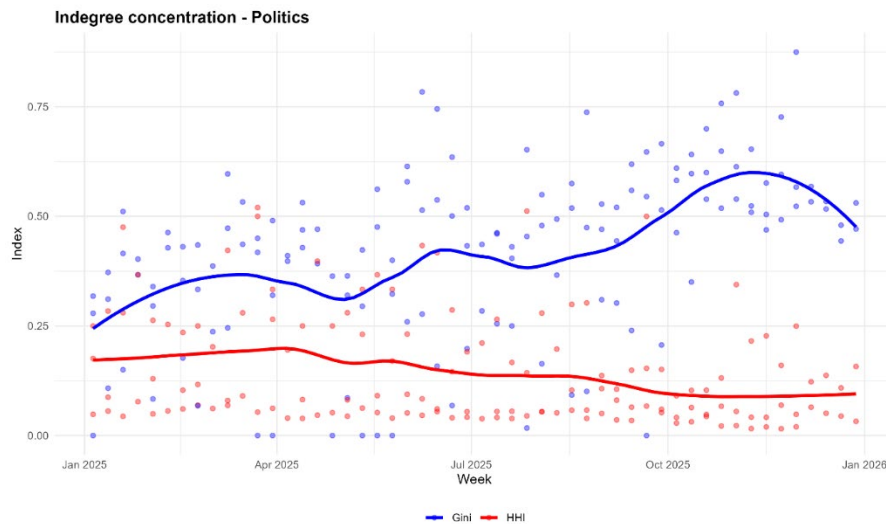
En cuanto a los objetivos, esta es la categoría más desigual de nuestro conjunto de datos, combinada con un predominio relativamente bajo de los objetivos, con un coeficiente de Gini medio de 0,42 y un HHI de 0,14. Esto sugiere que la hostilidad política está muy concentrada, pero presenta poca consolidación en torno a objetivos focales. Aunque en nuestro conjunto de datos observamos una divergencia creciente entre una mayor concentración y un menor predominio, la tendencia parece relativamente estable a lo largo del tiempo. En conjunto, estos patrones sugieren una focalización difusa, sin un foco constante.

En cuanto a los difusores de hostilidad relacionada con la política, encontramos una mayor variación en nuestro conjunto de datos que en la mayoría de las demás categorías. Aunque sigue existiendo una fuerte concentración y poca persistencia, con un 94,1 % de focalización concentrada y baja persistencia, esta categoría presenta la mayor proporción de cuentas que insultan a un conjunto más amplio de objetivos, 5,8 % de focalización difusa, y una proporción comparativamente mayor de atacantes persistentes, 0,1 %. Esto sugiere que la hostilidad política tiene una mayor probabilidad de dirigirse contra múltiples objetivos y de proceder persistentemente de algunos grupos de personas usuarias.

En términos de intervención, la atención en esta categoría es muy desigual y presenta una amplia variación entre los objetivos en cada uno de los idiomas. Aunque es difícil anticipar qué tipo de acontecimientos pueden impulsar este contenido y la atención se distribuye de forma difusa entre muchos objetivos diferentes, es probable que estos patrones estén impulsados por ciclos informativos y sugieren la existencia de grupos centrales de difusores que insultan de forma persistente a objetivos específicos.

Figura 8. Comparación de la hostilidad relacionada con la política según proporción, objetivos y difusores





Raza

Esta categoría se refiere al lenguaje intolerante dirigido contra aspectos específicos de la identidad relacionados con la raza y la etnia, así como contra estereotipos raciales.

- **Ejemplo en español:** «Qué les pasa a los hombres n*? Por qué son la mayoría así?».
- **Ejemplo en alemán:** «Was hat der N* mit deutscher Jugend zu tun?»
- **Ejemplo en neerlandés:** «Terug naar je eigen LAND VIEZE AAP».

La hostilidad en línea relacionada con la raza no aparece con frecuencia en nuestro conjunto de datos. Encontramos poca evidencia restante en español y alemán, así como en Instagram y TikTok en neerlandés, aunque sí aparece en neerlandés con una volatilidad considerable en

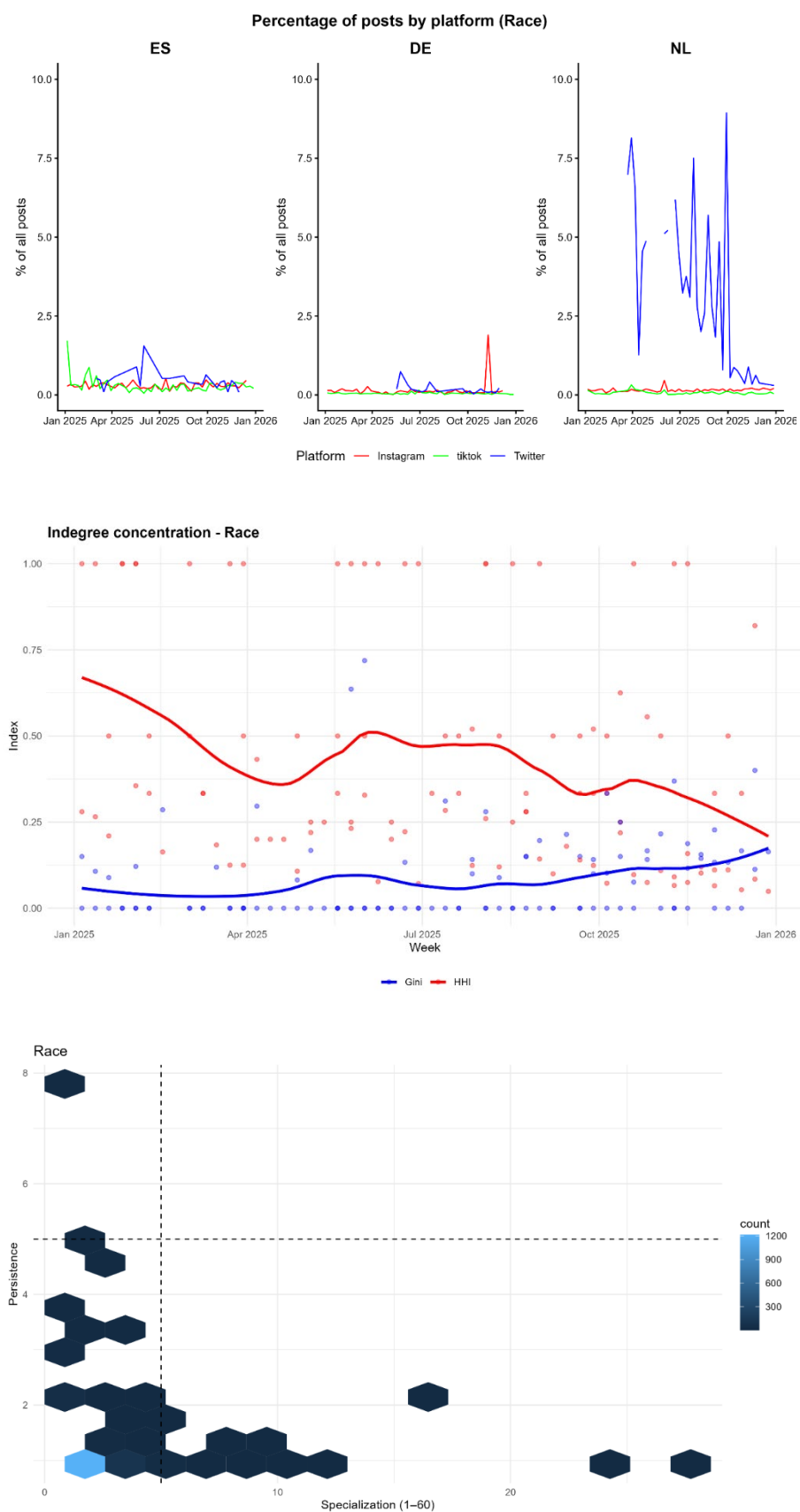
X. Esperábamos encontrar diferencias entre plataformas, ya que se afirma que X dejó de aplicar una moderación estricta durante el periodo observado. Observamos un rendimiento aceptable del modelo que, combinado con un fuerte acuerdo entre codificadores, sugiere que este lenguaje puede detectarse con precisión. En conjunto, reconocemos que la escasez de evidencia en esta categoría dentro de nuestro conjunto de datos puede ser consecuencia de la metodología, más que evidencia de ausencia.

En cuanto a la atención y los objetivos, el racismo en nuestro conjunto de datos presenta un coeficiente de Gini medio muy bajo, 0,08, pero un HHI elevado, 0,42. Esto sugiere que, aunque la desigualdad es baja, la atención se concentra fuertemente en un conjunto reducido de objetivos. La tendencia se mantiene constante a lo largo del tiempo, aunque observamos una volatilidad relativamente alta como resultado del número limitado de observaciones. Junto con los picos de actividad, esto sugiere una focalización episódica, con incrementos repentinos de atención concentrada sobre un número reducido de objetivos. Esto puede estar relacionado con acontecimientos informativos que convierten a determinadas personas en objetivos relevantes.

En cuanto a los difusores, encontramos un patrón muy concentrado: el 96,9 % presenta una focalización concentrada y baja persistencia. En comparación con otras categorías de nuestro conjunto de datos, el racismo presenta una proporción relativamente elevada de personas usuarias con una focalización más amplia, 2,9 % de focalización difusa y baja persistencia. Encontramos una pequeña proporción de personas usuarias persistentes, 0,1 %.

En términos de intervención, estos hallazgos sugieren que el racismo ya puede estar sometido a una moderación estricta. A partir de lo que permanece en nuestro conjunto de datos, el racismo parece ser en gran medida episódico, con poco comportamiento de focalización sostenida o amplia, pero con un núcleo de atacantes persistentes, especialmente en X.

Figura 9. Comparación de la hostilidad relacionada con la raza según proporción, objetivos y difusores



Religión

Esta categoría se refiere al lenguaje intolerante dirigido contra aspectos específicos de la identidad relacionados con la religión, así como contra estereotipos religiosos.

- **Ejemplo en español:** «Yo voy a entrar en una Mezquita y me voy a cargar 😂😂».
- **Ejemplo en alemán:** «Ein verlogener Jude, das ist ja nichts Neues».
- **Ejemplo en neerlandés:** «Weer een dag dat Soennieten zich niet normaal kunnen gedragen».

La hostilidad en línea relacionada con la religión no es muy frecuente en nuestros datos en alemán y neerlandés, en ninguna de las plataformas, pero es especialmente rara en español en todas las plataformas de nuestro conjunto de datos. Cuando la hostilidad relacionada con la religión aparece en nuestro conjunto de datos, lo hace de forma relativamente constante a lo largo del tiempo en Instagram y TikTok, mientras que en X presenta una elevada volatilidad. Estas tendencias se mantienen estables a lo largo del tiempo, lo que sugiere una combinación de un nivel de base de odio relacionado con la religión y picos de actividad impulsados por acontecimientos.

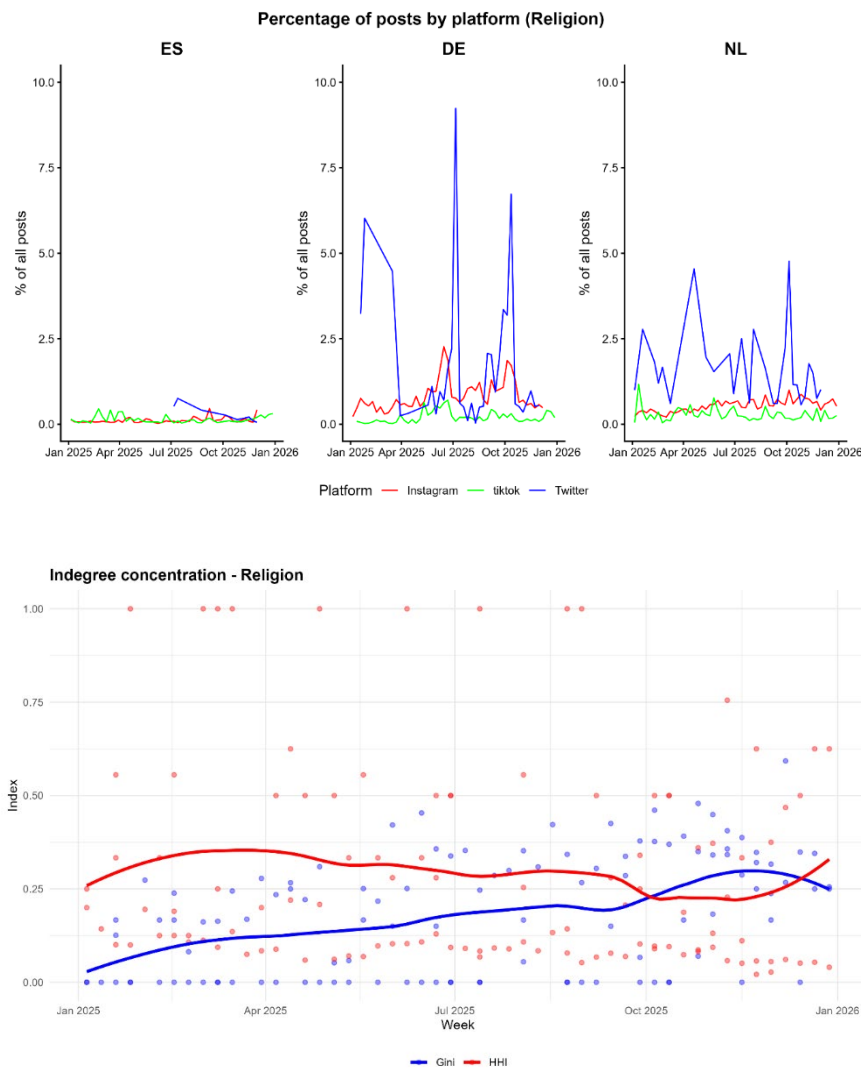
En cuanto a los objetivos, en comparación con otras categorías, observamos una concentración moderada combinada con un predominio moderado de los objetivos, con un coeficiente de Gini medio de 0,18 y un HHI de 0,29. Esta tendencia parece relativamente estable a lo largo del tiempo, con un ligero aumento hacia una mayor concentración. Estos patrones sugieren un conjunto de objetivos relativamente difuso.

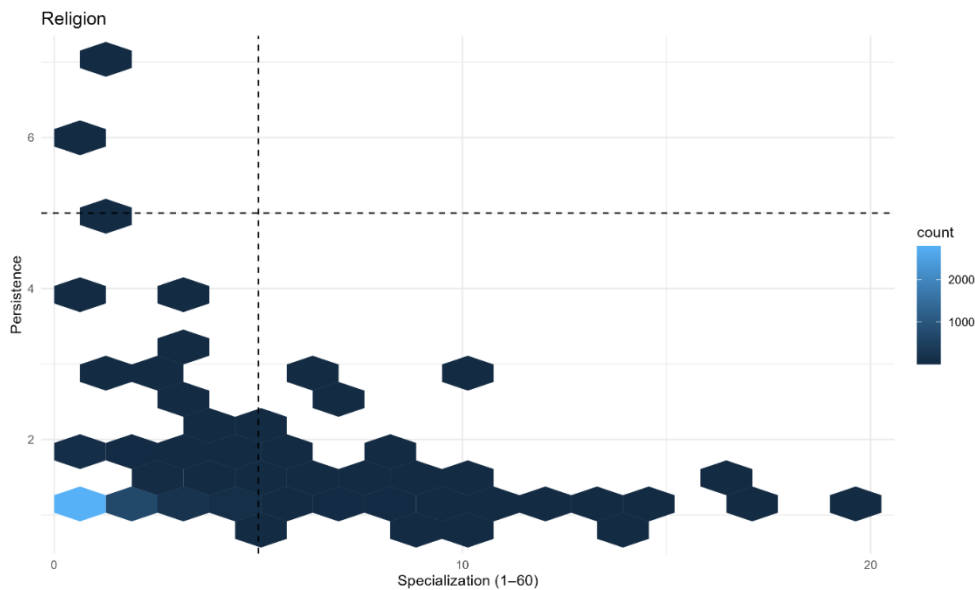
En cuanto a los difusores de hostilidad relacionada con la religión, se trata de una categoría relativamente homogénea en nuestro conjunto de datos. La mayoría de las cuentas se centra en un único objetivo y presenta poca persistencia, con un 97,1 % de focalización concentrada y baja persistencia, junto con una proporción relativamente elevada de focalización difusa, 2,8 %. Encontramos muy poca persistencia. Esto sugiere un comportamiento de focalización muy episódico, pero relativamente difuso.

En términos de intervención, la atención en esta categoría es relativamente difusa, aunque existe cierta constancia entre los objetivos. Observamos diferencias notables entre los idiomas que probablemente están relacionadas con limitaciones metodológicas. Este tipo de contenido será difícil de abordar, ya que existen pocos indicios sobre qué tipo de

acontecimientos pueden impulsarlo, no se identifican actores recurrentes y hay muy poca persistencia entre los difusores a lo largo del tiempo. Dependiendo del contexto nacional, puede haber objetivos constantes.

Figura 10. Comparación de la hostilidad relacionada con la religión según proporción, objetivos y difusores





Intimidación sexual

Esta categoría se refiere a lenguaje sexualmente sugerente o denigrante dirigido contra una persona, sin hacer referencia a un grupo ni a un rol de género.

- **Ejemplo en español:** «Por qué no te abres más de patas? Qué modales tan exquisitos. Tonta del culo».
- **Ejemplo en alemán:** «wer dich N*tte fickt ist zu faul zum wichen».
- **Ejemplo en neerlandés:** «Ga je beschimilde hoere vrendin neuke die zwijn eet elke dag overwichte kkr slet».

La intimidación sexual no aparece con frecuencia en nuestro conjunto de datos. Encontramos poca evidencia restante en español y alemán, así como en Instagram y TikTok en neerlandés, aunque sí aparece en neerlandés con una volatilidad considerable en X. Esperábamos encontrar diferencias entre plataformas, ya que se afirma que X dejó de aplicar una moderación estricta durante el periodo observado. También observamos un rendimiento aceptable del modelo que, combinado con un fuerte acuerdo entre codificadores, sugiere que este lenguaje puede detectarse con precisión. En conjunto, y de manera similar al racismo, reconocemos que la escasez de evidencia en esta categoría dentro de nuestro conjunto de datos puede ser consecuencia de la metodología, más que evidencia de ausencia.

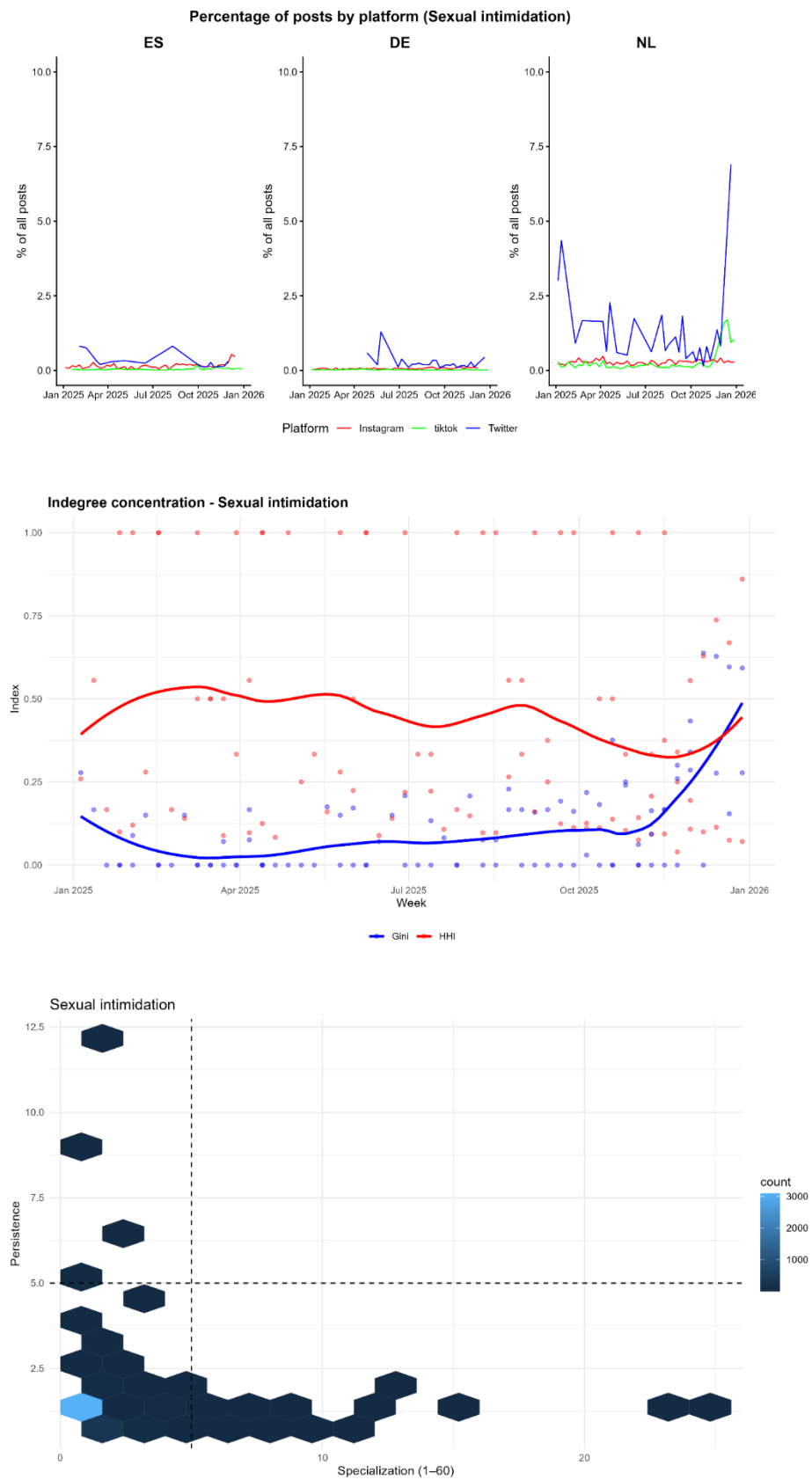
Los patrones que sí identificamos se mantienen estables a lo largo del tiempo: volatilidad en X y un nivel de base constante de intimidación sexual en Instagram y TikTok en cada uno de los países.

En cuanto a la atención y los objetivos, la intimidación sexual en nuestro conjunto de datos presenta un coeficiente de Gini medio muy bajo, 0,11, pero un HHI elevado, 0,44. Esto sugiere que, aunque la desigualdad es generalmente baja, la atención se concentra fuertemente en un conjunto reducido de objetivos. La tendencia se mantiene constante a lo largo del tiempo, aunque observamos una volatilidad relativamente alta como resultado del número limitado de observaciones. Junto con los picos de actividad en X, y en menor medida en Instagram, esto sugiere una focalización episódica, con incrementos repentinos de atención concentrada sobre un número reducido de objetivos. Al igual que en el caso del racismo, esto puede estar relacionado con acontecimientos de actualidad que convierten a personas específicas en objetivos relevantes.

En cuanto a los difusores, encontramos un patrón muy concentrado: el 97,7 % presenta una focalización concentrada y baja persistencia. Al igual que el racismo, y en contraste con otras categorías de nuestro conjunto de datos, la intimidación sexual presenta una proporción relativamente elevada de personas usuarias con una focalización más amplia, 2,3 % de focalización difusa y baja persistencia. No encontramos una persistencia significativa en nuestro conjunto de datos.

En términos de intervención, a partir de la evidencia de intimidación sexual que permanece en nuestro conjunto de datos, esta parece ser en gran medida episódica, con cierto comportamiento de focalización constante, pero con una persistencia limitada y sin un grupo central de perpetradores al que dirigirse. Estos hallazgos sugieren que la intimidación sexual puede ya estar sometida a una moderación estricta, posiblemente incluida la denuncia y posterior suspensión de cuentas, en Instagram y TikTok, y en menor medida en X.

Figura 11. Comparación de la hostilidad relacionada con la intimidación sexual según proporción, objetivos y difusores



Orientación sexual

Esta categoría se refiere al lenguaje intolerante dirigido contra aspectos específicos de la identidad relacionados con la orientación sexual, así como contra estereotipos asociados.

- **Ejemplo en español:** «la persona gay es muuuy promiscua y guarra en el aspecto sexual. Tu eres gay? Pues disfrutalo».
- **Ejemplo en alemán:** «Dreckige Schwuchtel».
- **Ejemplo en neerlandés:** «Donderop man vieze flicker».

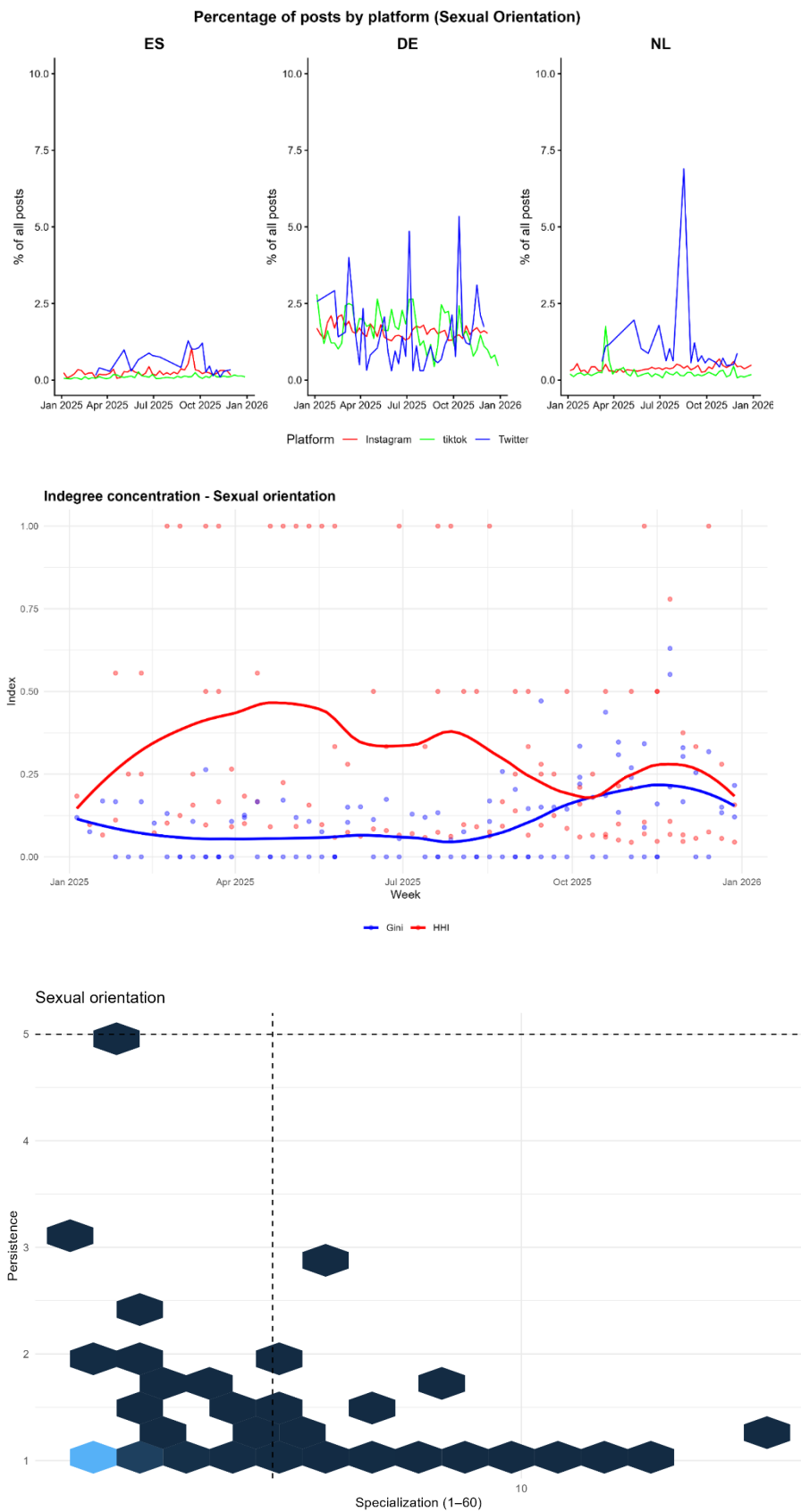
La hostilidad en línea relacionada con la orientación sexual no es una categoría muy destacada. En nuestro conjunto de datos, aparece de forma más prominente en alemán en todas las plataformas, en neerlandés en X y rara vez en español. Presenta volatilidad en X y una relativa estabilidad en TikTok e Instagram. Esto sugiere un nivel de base de este tipo de contenido, combinado con acontecimientos y ciclos informativos que impulsan incrementos repentinos de atención.

En cuanto a los objetivos, esta es una de las categorías menos concentradas de nuestro conjunto de datos, combinada con un predominio moderado de los objetivos, con un coeficiente de Gini medio de 0,11 y un HHI de 0,32. Esto sugiere que la hostilidad relacionada con la orientación sexual es relativamente difusa, pero presenta cierto grado de consolidación en torno a objetivos focales. Aunque en nuestro conjunto de datos observamos cierta convergencia hacia una mayor concentración y un menor predominio, la tendencia parece relativamente estable a lo largo del tiempo. En conjunto, estos patrones sugieren una concentración generalmente difusa, pero con objetivos moderadamente constantes.

En cuanto a los difusores de hostilidad relacionada con la orientación sexual, encontramos una proporción relativamente elevada de atacantes transitorios, con un 98,4 % de focalización concentrada y baja persistencia, y una focalización difusa o persistencia insignificantes. Esto sugiere que la hostilidad basada en la orientación sexual está dominada por ataques aislados.

En términos de intervención, estos hallazgos sugieren que la hostilidad basada en la orientación sexual ya puede estar moderada de forma efectiva. A partir de lo que permanece en nuestro conjunto de datos, la hostilidad relacionada con la orientación sexual parece predominantemente transitoria, con muy poco comportamiento de focalización sostenida o amplia.

Figura 12. Comparación de la hostilidad relacionada con la orientación sexual según proporción, objetivos y difusores



Lenguaje violento

Esta categoría se refiere al lenguaje que incita a la violencia o la glorifica, incluidas las amenazas directas, también de muerte.

- **Ejemplo en español:** «Esa soberbia con la que aplaude y niega es merecedora de una guillotina».
- **Ejemplo en alemán:** «Der gehört an die Wand».
- **Ejemplo en neerlandés:** «Tering jong, zal jou doodtrappen.».

El lenguaje violento no aparece con frecuencia en nuestro conjunto de datos. Encontramos poca evidencia restante en español y alemán, así como en Instagram y TikTok en neerlandés, aunque sí aparece en neerlandés con cierta volatilidad en X. Esperábamos encontrar diferencias entre plataformas, ya que se afirma que X dejó de aplicar una moderación estricta durante el periodo observado. Observamos un rendimiento aceptable del modelo que, combinado con un fuerte acuerdo entre codificadores, sugiere que este lenguaje puede detectarse con precisión. En conjunto, y de forma similar a las categorías «Raza» e «Intimidación sexual», reconocemos que la escasez de evidencia en esta categoría dentro de nuestro conjunto de datos puede ser consecuencia de la metodología, más que evidencia de ausencia.

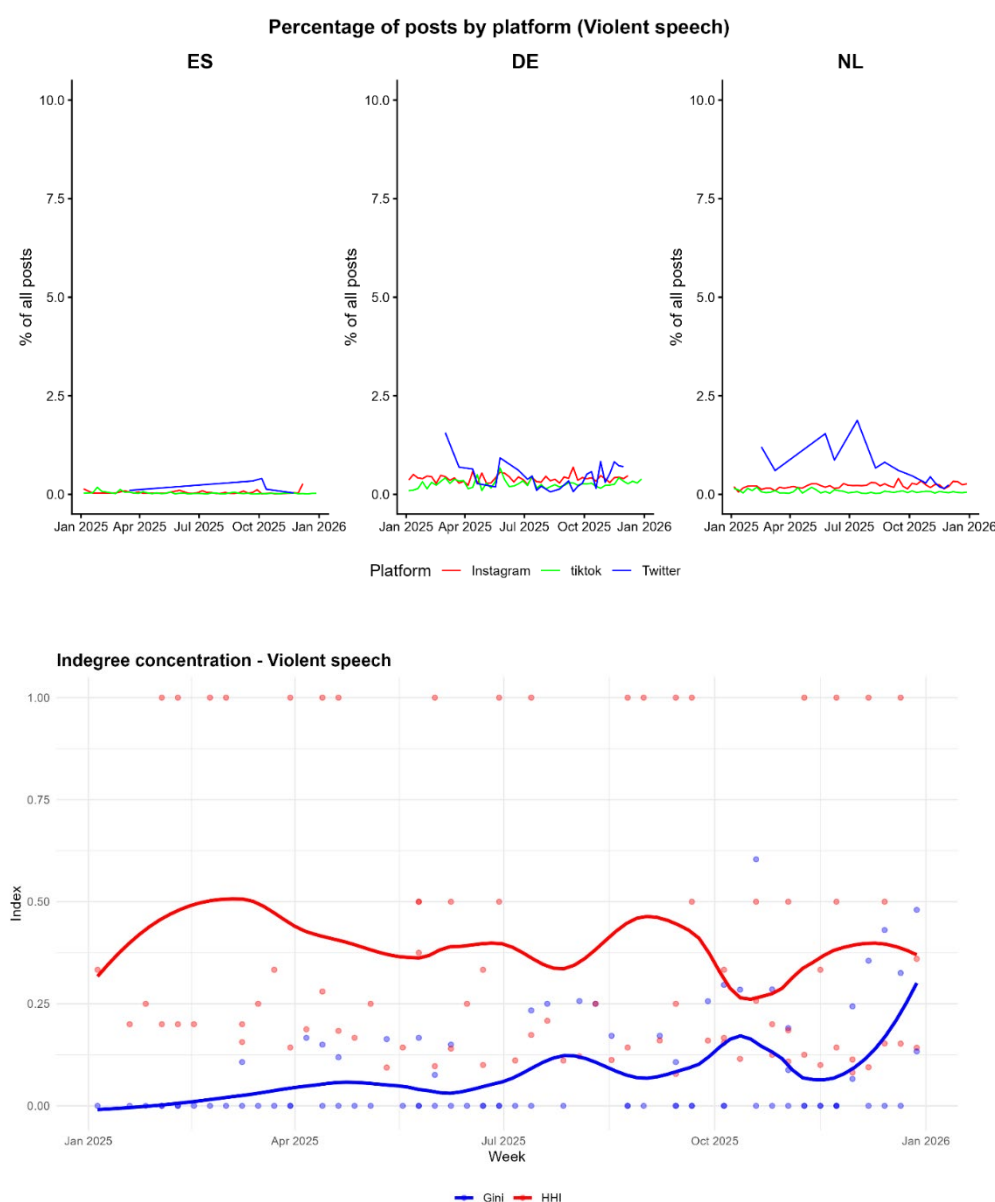
En cuanto a la atención y los objetivos, el lenguaje violento en nuestro conjunto de datos presenta un coeficiente de Gini medio muy bajo, 0,08, pero un HHI elevado, 0,40. Esto sugiere que, aunque la desigualdad es baja, la atención se concentra fuertemente en un conjunto reducido de objetivos. La tendencia se mantiene estable a lo largo del tiempo, aunque señalamos el número limitado de observaciones. Junto con unos picos de actividad solo moderados, esto sugiere un nivel de base de lenguaje violento, en lugar de incrementos episódicos de atención y, por tanto, una menor dependencia aparente de los ciclos informativos.

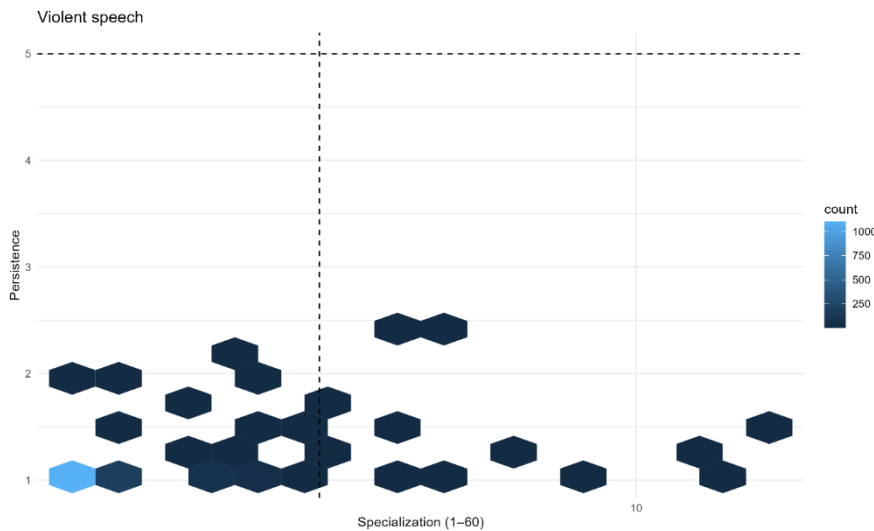
En cuanto a los difusores, encontramos un patrón muy concentrado: el 97,7 % presenta una focalización concentrada y baja persistencia. En comparación con otras categorías de nuestro conjunto de datos, el lenguaje violento presenta una proporción relativamente elevada de personas usuarias con una focalización más amplia, 2,3 % de focalización difusa y baja

persistencia. Hay muy poca evidencia en nuestro conjunto de datos de atacantes persistentes que se centren en los mismos objetivos a lo largo del tiempo.

En términos de intervención, a partir de la evidencia de lenguaje violento que permanece en nuestro conjunto de datos, estos hallazgos sugieren que el lenguaje violento ya puede estar moderado de forma efectiva. A partir de lo que permanece en nuestro conjunto de datos, el lenguaje violento parece predominantemente transitorio, con muy poco comportamiento de focalización sostenida o amplia.

Figura 13. Comparación del discurso violento según proporción, objetivos y difusores





Lenguaje vulgar

Esta categoría se refiere al lenguaje que puede interpretarse como descortés y puede resultar ofensivo, pero que no se dirige contra un grupo o una persona específicos. Incluye palabrotas, insultos y expresiones malsonantes.

- **Ejemplo en español:** «Coño señora entonces que hace usted y todo su equipo de gobierno».
- **Ejemplo en alemán:** «drecks Politik!».
- **Ejemplo en neerlandés:** «God verdomme vuile moordenaars.».

El lenguaje vulgar no es una categoría muy destacada en nuestro conjunto de datos. Aparece con relativa mayor frecuencia en español en todas las plataformas, en X en alemán y con poca frecuencia en neerlandés. Se presenta de manera relativamente constante en todas las plataformas, también en X en comparación con otras categorías. Esto sugiere un nivel de actividad de base, en lugar de picos impulsados por acontecimientos.

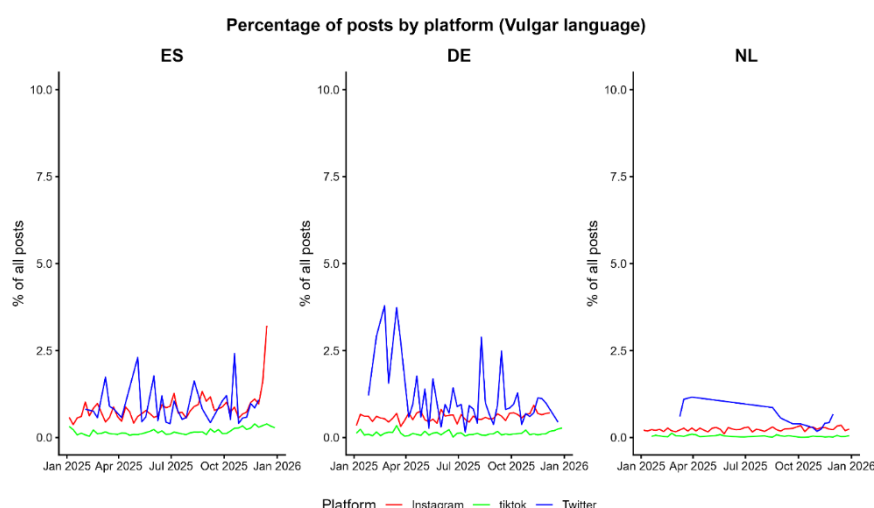
En cuanto a la atención y los objetivos, el lenguaje vulgar en nuestro conjunto de datos presenta un coeficiente de Gini medio muy bajo, 0,11, y un HHI elevado, 0,36. Esto sugiere que, aunque la desigualdad es baja, la atención se concentra fuertemente en un conjunto reducido de objetivos. La tendencia se mantiene constante a lo largo del tiempo, aunque señalamos el número limitado de observaciones. Junto con unos picos de actividad solo moderados, esto sugiere un nivel de base de lenguaje vulgar, en lugar de incrementos

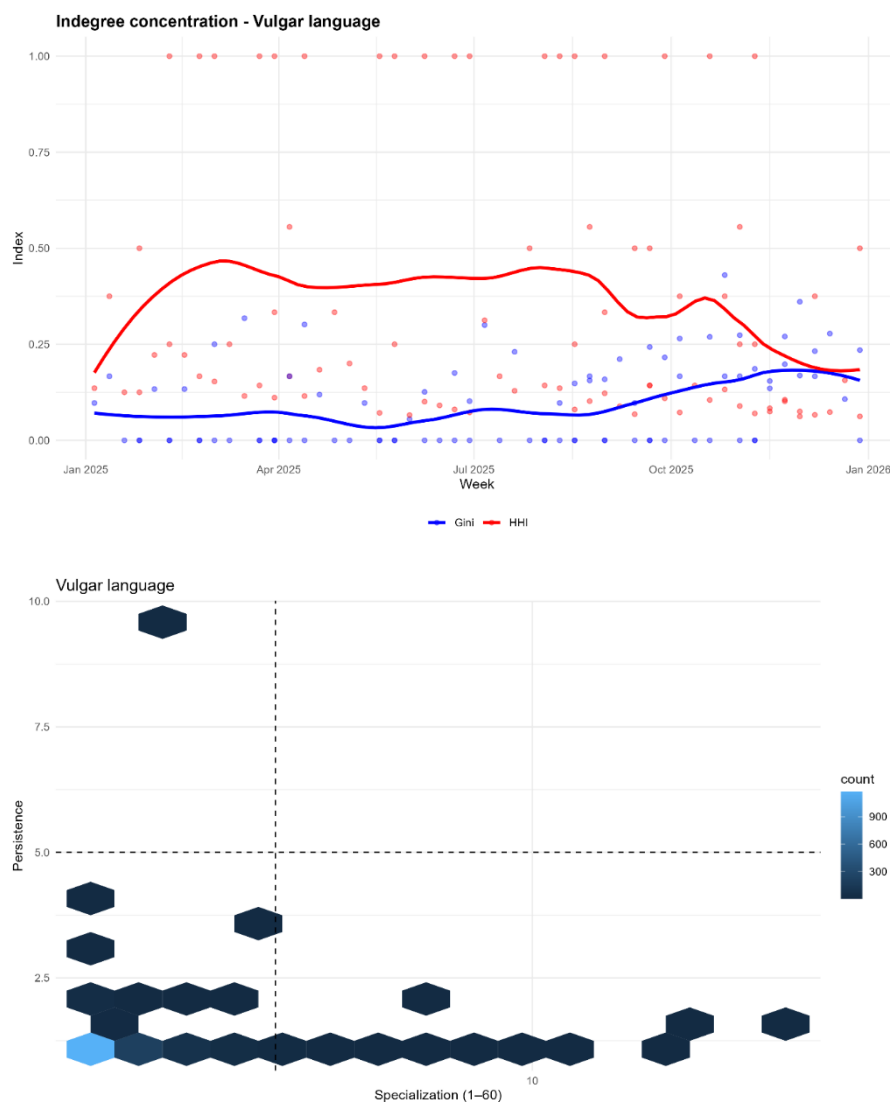
episódicos de atención y, por tanto, una menor dependencia aparente de los ciclos informativos.

En cuanto a los difusores, encontramos un patrón muy concentrado: el 98,5 % presenta una focalización concentrada y baja persistencia. En comparación con otras categorías de nuestro conjunto de datos, el lenguaje vulgar presenta una proporción moderada de personas usuarias con una focalización más amplia, 1,4 % de focalización difusa y baja persistencia. Encontramos una pequeña proporción de personas usuarias persistentes, 0,1 %.

En términos de intervención, estos hallazgos sugieren que el lenguaje vulgar ya puede estar moderado en gran medida en Instagram y TikTok. A partir de lo que permanece en nuestro conjunto de datos, el lenguaje vulgar parece ser en gran medida episódico, con cierto comportamiento de focalización amplia y con un núcleo de atacantes persistentes, especialmente en X.

Figura 14. Comparación de la hostilidad relacionada con la vulgaridad según proporción, objetivos y difusores





Conclusiones

En conjunto, nuestros hallazgos sugieren que la hostilidad en línea está muy extendida y consiste predominantemente en comentarios hostiles aislados. En términos de intervención, esto sugiere que el principal desafío no consiste en contrarrestar a un pequeño número de personas reincidentes, sino a un gran número de comentarios hostiles aislados procedentes de muchas cuentas diferentes.

Subrayamos que la escasez de evidencia en nuestro conjunto de datos de categorías como «Raza» o «Lenguaje violento» puede ser el resultado de sesgos en la recopilación de datos y la metodología, más que evidencia de ausencia. También señalamos la precisión y la

exhaustividad limitadas de los modelos lingüísticos utilizados para detectar los distintos tipos de hostilidad. A partir de la comparación entre categorías, países y plataformas, extraemos las siguientes conclusiones.

En general, observamos que los difusores de hostilidad están más dispersos de lo esperado. La hostilidad en línea de nuestro conjunto de datos procede predominantemente de muchas cuentas diferentes, en lugar de concentrarse en unas pocas. Encontramos poco solapamiento entre categorías en los difusores de hostilidad en línea, lo que sugiere que la mayoría de las cuentas participa en un único tipo de hostilidad, en lugar de diversificar sus repertorios. Entre los difusores de hostilidad en línea, encontramos muy poca evidencia de ataques focalizados y persistentes. En todas las categorías, el patrón dominante es una focalización concentrada y baja persistencia, lo que sugiere que la mayoría de las personas usuarias participa de forma breve y limitada, independientemente del tema.

Concluimos que la mayor parte de la hostilidad en línea de nuestro conjunto de datos consiste en publicaciones aisladas, con muy pocas personas reincidentes. Asimismo, observamos que los objetivos de la hostilidad en línea están fuertemente concentrados. Identificamos este mismo patrón al diferenciar por tipo de hostilidad. En todas las categorías, encontramos grados elevados de concentración de los objetivos, es decir, la mayor parte de la atención se dirige a un número relativamente reducido de objetivos, así como distintos grados de constancia, con las mismas cuentas siendo objeto de ataques. Al diferenciar más detalladamente la focalización, el comportamiento de publicación hostil y la volatilidad, encontramos diferencias notables entre categorías que orientan estrategias de intervención distintas.

Estrategias de intervención recomendadas

En términos de intervención, proponemos tres conclusiones aplicables basadas en nuestros hallazgos, resumidas en la Tabla 7.

Tabla 7. Estrategia de intervención recomendada según los tipos de categorías

Categoría	Concentración	Tipo destacado de difusor	Volatilidad	Foco de intervención
Race	Concentrado + estable	Reincidentes (Difuso + persistente)	Episódico	Objetivos específicos + difusores específicos + acontecimientos
Sexual intimidation	Difuso + estable	Generalistas (Difuso + transitorio)	Consistente + episódico	Objetivos específicos + acontecimientos
Religion	Difuso + inestable	Generalistas (Difuso + transitorio)	Consistente + episódico	Objetivos específicos + acontecimientos
Sexual orientation	Difuso + estable	Generalistas (Difuso + transitorio)	Consistente + episódico	Objetivos específicos + acontecimientos
Vulgar language	Concentrado + estable	Reincidentes (Difuso + persistente)	Consistente	Objetivos específicos + difusores específicos
Crime	Difuso + estable	Generalistas (Difuso + transitorio)	Consistente + episódico	Acontecimientos
Culture	Concentrado + inestable	Oportunistas (Focalizado + transitorio)	Consistente + episódico	Acontecimientos
Politics	Concentrado + inestable	Reincidentes (Difuso + persistente)	Episódico	Acontecimientos
Gender	Difuso + estable	Oportunistas (Focalizado + transitorio)	Consistente	Objetivos específicos
Violent language	Concentrado + estable	Generalistas (Difuso + transitorio)	Consistente	Objetivos específicos
Insults	Concentrado + inestable	Reincidentes (Difuso + persistente)	Consistente	Difusores específicos

Un primer grupo de categorías se caracteriza por una actividad episódica: raza, intimidación sexual, orientación sexual, religión, lenguaje violento, delincuencia, cultura y política. En comparación con otras categorías, estas presentan una mayor volatilidad relativa y probablemente dependen en mayor medida de acontecimientos. Las estrategias de intervención para estas categorías podrían centrarse en anticipar acontecimientos que generan hostilidad en línea y responder a ellos.

Un segundo grupo de categorías se caracteriza por la concentración de los objetivos: raza, intimidación sexual, religión, orientación sexual, género, lenguaje vulgar y lenguaje violento. En comparación con otras categorías, los objetivos de estos tipos de hostilidad están relativamente concentrados o son relativamente estables. Las estrategias de intervención para estas categorías podrían centrarse en monitorizar las cuentas de posibles objetivos de hostilidad en línea.

Un tercer grupo de categorías se caracteriza por la constancia entre los difusores: raza, lenguaje vulgar e insultos. En comparación con otras categorías, los difusores de estos tipos de hostilidad son relativamente más persistentes. Las estrategias de intervención para estas categorías podrían centrarse en actores recurrentes que generan hostilidad en línea.

Subrayamos que la baja exhaustividad y precisión de los modelos lingüísticos constituyen una limitación fundamental. En la práctica, los modelos empleados aquí no detectarán muchas publicaciones que contienen hostilidad en línea y marcarán muchos casos como falsos positivos. Teniendo en cuenta nuestro objetivo de apoyar la moderación, consideramos preferibles las falsas alarmas a dejar sin detectar muchas publicaciones problemáticas. Es probable que esta precisión empeore con el tiempo y al aplicarse a otras plataformas. Desaconsejamos el uso de estos modelos para la toma de decisiones automatizada. La aplicabilidad práctica de estos modelos consiste en marcar contenido para que sea examinado por personas expertas.

Limitaciones metodológicas

La recopilación de datos y la metodología que fundamentan este informe presentan salvedades importantes.

En primer lugar, señalamos que las formas más generalizadas de hostilidad pueden estar infrarrepresentadas en nuestro conjunto de datos como resultado de nuestra estrategia de recopilación. Aunque son extensos, nuestros conjuntos de datos son solo parcialmente representativos. La recopilación exclusiva de datos disponibles públicamente, por razones de privacidad y dadas las limitaciones prácticas de los datos de redes sociales, implica que las formas de contenido sometidas a una moderación más estricta ya se habían eliminado antes de que pudiéramos recopilarlas. Como resultado, es probable que el racismo y el lenguaje violento en general, y la violencia basada en el género de manera más específica, sean más relevantes en la práctica de lo que sugiere nuestro análisis. Incluso si la moderación elimina dicho contenido después de un tiempo, el daño ya puede haberse producido.

En segundo lugar, también en relación con la recopilación de datos, persisten limitaciones derivadas de priorizar la amplitud y la diversidad. Nuestro enfoque inicial en identificar figuras destacadas implica que los actores menos destacados o visibles pueden estar infrarrepresentados. Intentamos tener esto en cuenta mediante un muestreo intencional basado en la identidad. Se consideraron características como la raza, la etnia, la identidad de género, la orientación sexual, la religión, el origen migratorio, el idioma y la ideología política, para asegurar que el conjunto de datos incluyera a grupos desproporcionadamente atacados por la hostilidad en línea, pero que pueden ser menos prominentes en el debate público. Procuramos representar a poblaciones minoritarias específicas de cada contexto nacional, como las minorías lingüísticas en España y las personas implicadas en debates sobre temas históricamente sensibles en Alemania.

Una limitación relacionada con nuestro enfoque es el solapamiento entre sectores, como la política y el activismo, que complicó los esfuerzos por equilibrar las categorías. En última instancia, esto llevó a que algunas categorías estuvieran fuertemente representadas, como la política, a expensas de otras, como el deporte. Consideramos que nuestro enfoque garantiza una cobertura amplia de ámbitos sociales y grupos identitarios, al tiempo que tiene en cuenta la variación contextual y cultural en la expresión de odio en línea.

En tercer lugar, reconocemos que las publicaciones en redes sociales no son neutrales ni se producen de forma natural, y que las cuentas de redes sociales no representan necesariamente a personas físicas. Nuestro conjunto de datos probablemente incluye diversas formas de automatización, spam y otras manifestaciones de intencionalidad maliciosa. La

identificación de bots y de formas de coordinación queda fuera del alcance de este proyecto, pero recomendamos firmemente seguir examinando la relación entre las formas coordinadas de manipulación y los distintos tipos de hostilidad identificados en este informe.

En cuarto lugar, reconocemos que la comparación de datos procedentes de distintas plataformas de redes sociales, Instagram, X y TikTok, y de diferentes contextos nacionales, Alemania, España y los Países Bajos, está considerablemente limitada. Cada plataforma de redes sociales cuenta con bases de personas usuarias diversificadas, formas de expresión particulares y funcionalidades específicas que se entrelazan de manera significativa con el uso del lenguaje en estas plataformas. Asimismo, reconocemos que el significado del lenguaje está estrechamente relacionado con el contexto. El análisis comparativo del significado del lenguaje y de las diferencias relacionadas entre países y plataformas requiere investigación cualitativa. Consideramos que las comparaciones de orientación más cuantitativa ofrecidas en este informe se justifican por nuestro objetivo de identificar patrones en la prevalencia del lenguaje entre países y plataformas. Entendemos estas comparaciones como tendencias indicativas, más que como representaciones precisas.

En quinto lugar, la clasificación automatizada del lenguaje está intrínsecamente limitada en términos de rendimiento. Reconocemos que la baja exhaustividad y precisión de los modelos de detección de discurso de odio, y en menor medida de los modelos de clasificación, constituyen una limitación fundamental del método descrito aquí. Este es particularmente el caso de las categorías menos frecuentes en nuestro conjunto de datos y que, por tanto, cuentan con menos observaciones para entrenar los modelos. A efectos prácticos, esto significa que los modelos producirán muchos falsos positivos. Teniendo en cuenta nuestro objetivo de apoyar la moderación, consideramos preferibles las falsas alarmas a dejar sin detectar muchas publicaciones problemáticas.

Los tres modelos preexistentes para la detección de discurso de odio se seleccionaron tras un análisis preliminar de rendimiento en la fase de diseño de la investigación. En ese momento, se decidió no utilizar modelos de lenguaje de gran tamaño, que podrían ofrecer un mejor rendimiento que los léxicos basados en reglas, debido a razones de viabilidad práctica y a restricciones relacionadas con la gestión de datos y la privacidad. El ajuste fino de modelos basados en BERT utilizando nuestros datos anotados sigue siendo una aspiración para trabajos futuros.

A efectos de nuestro análisis, abordamos esta limitación mediante un método semisupervisado que combina detección automatizada, anotación humana y clasificación. También procuramos abordar esta limitación mediante reflexión crítica continua, evaluación del acuerdo entre codificadores, calibración, inspección manual y verificaciones puntuales, con el fin de garantizar la validez de nuestros hallazgos. Sin embargo, reconocemos que los modelos de clasificación que sustentan nuestros análisis presentan deficiencias en cuanto a exactitud, precisión y exhaustividad. A medida que el lenguaje y sus usos sigan evolucionando, y que la hostilidad en línea se desplace hacia diferentes palabras y significados, es probable que la precisión de estos modelos disminuya aún más con el tiempo. Desaconsejamos el uso de estos modelos para la toma de decisiones automatizada. Su aplicabilidad práctica consiste en marcar contenido para que sea examinado por personas expertas.

En sexto lugar, señalamos que no podemos afirmar mucho sobre el impacto de la hostilidad en línea. Nuestros análisis permiten comprender los patrones de los distintos tipos de hostilidad, pero no pueden diferenciar sus impactos en las vidas de las personas que la reciben. Reconocemos que algunos tipos de hostilidad pueden tener un impacto mayor que otros, incluso si categorías como el racismo no son las más destacadas en nuestro análisis. Asimismo, la acumulación de distintas categorías, como raza y género, puede intensificar aún más los efectos perjudiciales.

En conjunto, esta metodología proporciona un marco amplio y adaptable para la recopilación transnacional de datos. Mediante la identificación sistemática de cuentas que operan en el ámbito público y que están expuestas a una variedad de ataques en línea, el conjunto de datos resultante cubre una amplia gama de formas de hostilidad en línea. Los léxicos elaborados para este análisis están disponibles como recursos de código abierto, véase van Haperen et al., 2026. Con una atención adecuada al contexto local, este enfoque es replicable y puede contribuir a la moderación tanto de formas manifiestas como de formas más sutiles de hostilidad en línea.

Agradecimientos

Gracias a quienes contribuyeron al procedimiento y al análisis presentados en este informe: Casper van der Heijden, Bart van der Velden, Kris Ruijgrok, Joyce van der Lans, Anna Claire

Shepherd, Nick van Loo, Caitlin Gray, Pilar Richter, Kamil Nakad, Lea Bund, Paula Magnanimo, Claudia Castro, Maria Olle, Anonymous Annotator 1, Christina Hübers, Rita Costa Cots, Mahmoud Bastati, Pablo Aguiar, Maria Raduà, Helena Puig Larrauri, Benjamin Cerigo, and Julie Hawke. Gracias por sus contribuciones y valiosas aportaciones.

Descargo de responsabilidad legal

El objetivo de este informe es identificar tendencias y patrones, y no constituye una calificación de las opiniones o declaraciones de individuos con fines legales. Cuando es posible, los datos personales se agregan y se seudonimizan o anonimizan antes del análisis, de conformidad con el RGPD y la normativa europea pertinente. Los análisis de este informe se basan en la detección e interpretación automatizadas de datos públicos en línea, sujetas a limitaciones metodológicas y contextuales.

Declaración de datos y licencia

Los lexicones elaborados en el marco de este proyecto, y utilizados para fundamentar el análisis actual, están disponibles en código abierto en el repositorio de la Universidad de Ámsterdam.⁴

Nombre del conjunto de datos: Violence, Intolerance and Impoliteness thematic Corpus v1.0

Desarrolladores del conjunto de datos: Sander van Haperen, Kris Ruijgrok, Joyce van der Lans, Anna Claire Shepherd, Nick van Loo, Caitlin Gray, Pilar Richter, Kamil Nakad, Lea Bund, Paula Magnanimo, Claudia Castro, Maria Olle, Anonymous Annotator 1.

Autor de la declaración de datos: Sander van Haperen (University of Amsterdam)

Licencia: Attribution-NonCommercial-ShareAlike 4.0 International (CC BY 4.0)

Usted es libre de:

- **Compartir:** copiar y redistribuir el material en cualquier medio o formato.
- **Adaptar:** remezclar, transformar y crear a partir del material.

⁴ <https://doi.org/10.21942/uva.31876402>

El otorgante de la licencia no puede revocar estas libertades siempre que usted cumpla los términos de la licencia.

Bajo los siguientes términos:

- **Atribución:** debe otorgar el crédito correspondiente, proporcionar un enlace a la licencia e indicar si se realizaron cambios. Puede hacerlo de cualquier manera razonable, siempre que no sugiera que el otorgante de la licencia le respalda a usted o al uso que hace del material.
- **No Comercial:** no puede utilizar el material con fines comerciales.
- **Compartir Igual:** si remezcla, transforma o crea a partir del material, debe distribuir sus contribuciones bajo la misma licencia que el original.
- **Sin restricciones adicionales:** no puede aplicar términos legales ni medidas tecnológicas que restrinjan legalmente a otras personas de hacer cualquier cosa que la licencia permita.

Avisos: No tiene que cumplir la licencia respecto de los elementos del material que sean de dominio público o cuando su uso esté permitido por una excepción o limitación aplicable. No se ofrecen garantías. Es posible que la licencia no le otorgue todos los permisos necesarios para el uso previsto. Por ejemplo, otros derechos, como los derechos de publicidad, privacidad o morales, pueden limitar la forma en que utiliza el material.

<https://creativecommons.org/licenses/by-nc-sa/4.0/>

Bibliografía

- Akhtar, S., & Morrison, C. M. (2019). The prevalence and impact of online trolling of UK members of parliament. *Computers in Human Behavior*, 99, 322–327. <https://doi.org/10.1016/j.chb.2019.05.015>
- Aluru, S. S., Mathew, B., Saha, P., & Mukherjee, A. (2020). Deep learning models for multilingual hate speech detection. *arXiv Preprint arXiv:2004.06465*. <https://arxiv.org/abs/2004.06465>
- Baider, F. (2023). Accountability issues, online covert hate speech, and the efficacy of counter-speech. *Politics and Governance*, 11(2), 249-260.
- Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F. M. R., ... & Sanguinetti, M. (2019, June). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation* (pp. 54-63).
- Bassignana, E., Basile, V., & Patti, V. (2018, December). Hurtlex: A multilingual lexicon of words to hurt. In *Proceedings of the fifth Italian conference on computational linguistics (CLiC-it 2018)* (pp. 52-57).
- Bentivegna, S., & Rega, R. (2024). Politicians under fire: Citizens' incivility against political leaders on social media. *Social Media + Society*, 10(4), 20563051241298415. <https://doi.org/10.1177/20563051241298415>
- Cinelli, M., Pelicon, A., Mozetič, I., Quattrocioni, W., Kralj Novak, P., & Zollo, F. (2021). Dynamics of online hate and misinformation. *Scientific Reports*, 11(1), 22083.
- European Observatory of Online Hate. (2026). Definitions: Hate speech. Retrieved September 19, 2026, from <https://eooi.eu/faq>
- Green, J., & Thorogood, N. (2004). Analysing qualitative data. *Qualitative methods for health research*, 2, 173-200.
- IMSyPP. (2022). hate_speech_nl. Hugging Face. https://huggingface.co/IMSyPP/hate_speech_nl
- Kralj Novak, P., Scantamburlo, T., Pelicon, A., Cinelli, M., Mozetič, I., & Zollo, F. (2022). Handling disagreement in hate speech modelling. In B. Bouchon-Meunier, R. R. Yager, A. P. Carvalho, & R. R. Fullér (Eds.), *Information processing and management of uncertainty in knowledge-based systems* (pp. 681–695). Springer. https://doi.org/10.1007/978-3-031-08974-9_54
- Nithyanand, R., Schaffner, B., & Gill, P. (2017). Online political discourse in the Trump era (Version 1). *arXiv*. <https://doi.org/10.48550/ARXIV.1711.05303>
- Pérez, J. M., Furman, D. A., Alonso Alemany, L., & Luque, F. M. (2022). RoBERTuito: A pre-trained language model for social media text in Spanish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 7235–7243). European Language Resources Association. <https://aclanthology.org/2022.lrec-1.785>
- Pérez, J. M., Giudici, J. C., & Luque, F. (2021). pysentimiento: A Python toolkit for sentiment analysis and SocialNLP tasks. *arXiv*. <https://arxiv.org/abs/2106.09462>

- Rossini, P. (2019). Disentangling uncivil and intolerant discourse in political talk. In R. G. Boatright, T. J. Shaffer, S. Sobieraj, & D. Goldthwaite Young (Eds.), *A crisis of civility* (1st ed., pp. 142–157). Routledge.
- Theocharis, Y., Barberá, P., Fazekas, Z., & Popa, S. A. (2020). The dynamics of political incivility on Twitter. *SAGE Open*, 10(2), 2158244020919447. <https://doi.org/10.1177/2158244020919447>
- Timmermans, S., & Tavory, I. (2012). Theory construction in qualitative research: From grounded theory to abductive analysis. *Sociological Theory*, 30(3), 167–186.
- Tulkens, S., Hilte, L., Lodewyckx, E., Verhoeven, B., & Daelemans, W. (2016). The automated detection of racist discourse in Dutch social media. *Computational Linguistics in the Netherlands Journal*, 6, 3–20.
- van Haperen, S., Ruijgrok, K., van der Lans, J., Gray, C., van Loo, N., & Shepherd, A. C. (2026). Violence, intolerance and impoliteness thematic corpus. University of Amsterdam. <https://doi.org/10.21942/uva.31876402>
- Ward, S., & McLoughlin, L. (2020). Turds, traitors and tossers: The abuse of UK MPs via Twitter. *The Journal of Legislative Studies*, 26(1), 47–73.