



Together Against Online Hate Speech

TAOH

Together Against Online Hate Speech

Deliverable 2.2 Hate Speech Pattern Report

Dutch translation

Over TAOH

Together Against Online Hate Speech (TAOHS) ondersteunt maatschappelijke organisaties en media bij het voorkomen van en veilig en effectief reageren op online haat door te analyseren hoe haat zich manifesteert en verspreidt via sociale mediaplatforms en digitale omgevingen, en door het leveren van tools en training aan onze partners. Wij zijn een internationaal project dat wordt geleid door organisaties in Duitsland, Nederland en Spanje.

Consortium leden

- SPF – Stichting Sharing Perspectives - Nederland | Coördinator
- UvA – Universiteit van Amsterdam - Nederland
- Build Up - How to Build Up Germany - Duitsland
- IBH - ichbinhier e.V. - Duitsland
- FundiPau - Fundació par la Pau - Spanje



**Medegefinancierd door
de Europese Unie**

Gefinancierd door de Europese Unie. De hier geuite ideeën en meningen komen echter uitsluitend voor rekening van de auteur(s) en geven niet noodzakelijkerwijs die van de Europese Unie of het Europese Uitvoerende Agentschap onderwijs en cultuur (EACEA) weer. Noch de Europese Unie, noch het EACEA kan ervoor aansprakelijk worden gesteld.



Dit werk is gelicenseerd onder [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/)

“Online vijandigheid: een vergelijkende analyse van Instagram, X en TikTok in Spanje, Duitsland en Nederland”

Versie 2.2, 19 september 2026 (vertaald uit het English)

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

Abstract

Dit document rapporteert bevindingen uit een vergelijkende analyse van online vijandigheid voor het project ‘Together Against Online Hate Speech’, uitgevoerd door de volgende partners: Sharing Perspectives Foundation, BuildUp, ichbinhier e.V., FundiPau en de Universiteit van Amsterdam.

Het doel van dit rapport is patronen van online vijandigheid in Duitsland, Nederland en Spanje op Instagram, X en TikTok te vergelijken. Daarbij wordt onderscheid gemaakt tussen categorieën, zoals racisme en gendergerelateerde vijandigheid, en gedragspatronen wat betreft de gerichtheid op doelwitten en de mate van persistentie. Daartoe worden taalmodellen ontwikkeld om verschillende vormen van online vijandigheid te detecteren en classificeren. Deze worden open source beschikbaar gesteld.

De bevindingen bieden aanknopingspunten voor interventiestrategieën van maatschappelijke organisaties, activisten, rechtshandavingsinstanties en beleidsmakers die zich bezighouden met online vijandigheid. Ten eerste vragen gebeurtenisgedreven categorieën, namelijk ras, seksuele intimidatie, seksuele oriëntatie, religie, gewelddadige uitingen, criminaliteit, cultuur en politiek, om snelle reacties op gebeurtenissen. Ten tweede vraagt de concentratie van doelwitten binnen sommige categorieën, namelijk politiek, beledigingen, criminaliteit en cultuur, om een combinatie van monitoring en ondersteuning van personen die waarschijnlijk doelwit worden van online vijandigheid. Ten derde zijn verspreiders van vijandigheid gerelateerd aan ras, grof taalgebruik en beledigingen relatief persistent vergeleken met andere categorieën. Interventiestrategieën voor deze categorieën kunnen zich daarom richten op bekende personen of accounts die herhaaldelijk online vijandigheid uiten.

Citatie:

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.

Inhoudsopgave

Inleiding	6
Definitie	7
Methodologie	8
Dataverzameling.....	8
Taalmodel 1: Detectie van haatzaaiende uitlatingen.....	10
Taalmodel 2: classificatie van online vijandigheid	11
Annotatieprocedure	12
Constructie van lexicons.....	14
Modelprestaties	15
Bevindingen	17
Patronen per land.....	17
Patronen per platform	19
Patronen naar doelwitten en verspreiders	22
Gedragspatronen.....	23
Patronen per categorie	24
Beschrijving van categorieën.....	24
Criminaliteit	27
Cultuur	29
Gender	31
Beledigingen en persoonlijke aanvallen.....	33
Politiek.....	36
Ras	38
Religie	41
Seksuele intimidatie	43
Seksuele oriëntatie	45
Gewelddadige taal.....	47
Grof taalgebruik.....	50
Conclusies	52
Aanbevolen interventiestrategieën	53
Methodologische beperkingen	54
Dankwoord	57
Juridische disclaimer	57
Dataverklaring en licentie	58
Bibliografie	60

Inleiding

Dit rapport maakt deel uit van het project “Together Against Online Hate Speech” (TAOHS). Het overkoepelende doel van het project is om de weerbaarheid van maatschappelijke organisaties, mediaorganisaties en autoriteiten te vergroten, zodat zij online haatzaaiende uitlatingen op hun platforms kunnen begrijpen en aanpakken en kunnen bijdragen aan een veiligere en pluriforme online ruimte. Dit rapport richt zich op Duitsland, Nederland en Spanje.

Het project combineert taalclassificatiemodellen met dialooginterventies onder menselijke begeleiding en beleidsbeïnvloeding, om online vijandigheid integraal te begrijpen en aan te pakken. Hoewel contentmoderatie en rapportage belangrijk blijven voor het aanpakken van inhoud die illegaal is of platformbeleid schendt, schieten dergelijke strategieën tekort bij inhoud die niet illegaal maar wel problematisch is. Ook taalvormen zoals intolerantie en onbeschaafdheid vragen om interventie. Wij hopen personen, maatschappelijke organisaties, rechtshandavingsinstanties en beleidsmakers te ondersteunen die online met dergelijke taal te maken hebben.

Het doel van dit rapport is interventiestrategieën te informeren door patronen van online vijandigheid in kaart te brengen. Methodologisch is dit onderzoek opgezet om een breed scala aan online vijandigheid te detecteren, gericht tegen een representatieve verscheidenheid aan doelwitten, in verschillende sociale mediaomgevingen (Instagram, X en TikTok), en om deze patronen te vergelijken tussen de nationale contexten van Duitsland, Nederland en Spanje. Daartoe ontwikkelen we drie taalmodellen voor de classificatie van verschillende typen online vijandigheid in het Duits, Nederlands en Spaans. Deze taalmodellen zijn het resultaat van uitgebreid abductief coderen en prestatietesten en worden open source beschikbaar gesteld, zodat partnerorganisaties en andere belanghebbenden ze kunnen aanpassen voor toegespitste analyses.

In dit rapport passen we deze modellen toe voor een platform- en landenoverstijgende vergelijking, om patronen in de prevalentie, doelwitten en verspreiders van verschillende typen online vijandigheid te identificeren. Onze bevindingen suggereren in het algemeen dat het probleem wijdverspreid is en voornamelijk bestaat uit eenmalige vijandige reacties. Wanneer we gedrag verder uitsplitsen naar categorie, zien we opvallende verschillen tussen verschillende typen online vijandigheid.

Hieruit volgen drie bruikbare inzichten. Ten eerste vragen gebeurtenisgedreven categorieën, zoals ras, seksuele intimidatie, seksuele oriëntatie, religie, gewelddadige uitingen, criminaliteit, cultuur en politiek, om snelle reacties op gebeurtenissen. Ten tweede vraagt de concentratie van doelwitten in sommige categorieën, zoals politiek, beledigingen, criminaliteit en cultuur, om een combinatie van monitoring en ondersteuning van waarschijnlijke doelwitten van online vijandigheid. Ten derde zijn verspreiders van vijandigheid die verband houdt met ras, grof taalgebruik en beledigingen relatief persistent in vergelijking met andere categorieën. Interventiestrategieën voor deze categorieën kunnen zich richten op bekende personen of accounts die herhaaldelijk online vijandigheid genereren.

Het project “Together Against Online Hate Speech” bestaat uit de volgende partners: Sharing Perspectives Foundation, Build Up, IchBinHier, FundiPau en de Universiteit van Amsterdam. Het project wordt gefinancierd door de Europese Unie (CERV-2024-CHAR-LITI-SPEECH). De geuite standpunten en meningen zijn echter uitsluitend die van de auteur(s) en weerspiegelen niet noodzakelijkerwijs die van de Europese Unie. Noch de Europese Unie, noch de subsidieverlenende autoriteit kan hiervoor verantwoordelijk worden gehouden.

Definitie

Aan het begin van ons project formuleerden we de volgende definitie van “hate speech”: “het publiekelijk aanzetten tot geweld of haat gericht tegen een groep personen of een lid van een dergelijke groep, gedefinieerd op basis van ras, huidskleur, religie, afkomst of nationale of etnische herkomst.” We nemen ook verwijzingen naar gender, seksuele oriëntatie en lichamelijke beperking mee. Bij aanvang verwachten we dat dergelijke taal categorieën omvat zoals: “ontmenselijking, seksisme, politiek, racisme, religie, geweld en bedreigingen” (aangepast en vertaald van EOHH, 2026). In de praktijk bood deze definitie ons ‘sensitizing concepts’ en vormde zij de basis voor kalibratie tussen codeurs.

In dit rapport gebruiken we ook de overkoepelende term “online vijandigheid” wanneer we verwijzen naar taal die niet strikt illegaal is of niet valt onder haatzaaiende uitlatingen die betrekking hebben op de identiteit van personen, maar die desalniettemin schadelijk kan zijn voor de kwaliteit van online debat. Naast haatzaaiende uitlatingen omvatten andere typen online vijandigheid misbruik en onbeschaafdheid. Misbruik verwijst naar taal die bedoeld is

om een specifiek individu te beledigen, provoceren of van streek te maken (vertaling van Akhtar & Morrison, 2019; Ward & McLoughlin, 2020). Politieke onbeschaafdheid verwijst naar “een gebrek aan respect voor de sociale en culturele normen die persoonlijke interacties en het functioneren van democratische systemen reguleren” (vertaling van Bentivegna & Rega, 2024). Dit omvat normoverschrijdend discours, variërend van relatief milde onbeleefdheid tot meer expliciete vormen van vijandigheid, zonder dat er noodzakelijkerwijs sprake hoeft te zijn van de intentie om schade toe te brengen of van het viseren van specifieke sociale groepen (Rossini, 2019).

Methodologie

Het doel van dit onderzoek is interventiestrategieën te informeren door patronen van online vijandigheid in kaart te brengen. Methodologisch is dit onderzoek opgezet om een breed scala aan online vijandigheid te detecteren, gericht tegen een representatieve verscheidenheid aan doelwitten, in verschillende sociale mediaomgevingen (Instagram, X en TikTok), en om deze patronen te vergelijken tussen de nationale contexten van Duitsland, Nederland en Spanje.

Daartoe verzamelden we sociale mediadata van een representatieve selectie van meest waarschijnlijke doelwitten. We analyseerden deze dataset met behulp van twee taalmodellen: één model om vast te stellen of een bericht waarschijnlijk online vijandigheid bevat en een ander model om te bepalen op welk onderwerp het bericht betrekking heeft. Deze combinatie stelt ons in staat patronen tussen verschillende typen online vijandigheid te identificeren en te onderzoeken. We richten ons op patronen in de prevalentie, doelwitten en verspreiders van online vijandigheid. Dit hoofdstuk bespreekt deze procedure uitgebreider.

Dataverzameling

We verzamelden gegevens van Instagram, X en TikTok. Deze platforms zijn geselecteerd als waarschijnlijke plekken waar online vijandigheid voorkomt en omdat zij in de drie landen de meest gebruikte sociale media zijn. De selectie omvat daarnaast uiteenlopende gebruikersgroepen en functionaliteiten. Op X komen politieke opvattingen die nauw verbonden zijn met nieuwsgebeurtenissen veel voor, Instagram legt de nadruk op leefstijl en

TikTok kent relatief veel betrokkenheid van jongere gebruikersgroepen en open algoritmische ontdekking van content.

Berichten en reacties op X, voorheen Twitter, bestaan uit korte berichten en reacties van doorgaans 50 woorden of minder, die multimediamateriaal, externe URL-links en vermeldingen van andere gebruikers kunnen bevatten. Berichten en reacties op Instagram bestaan doorgaans uit korte bijschriften en reacties van ongeveer 50 woorden of minder, die externe URL-links en vermeldingen van andere gebruikers kunnen bevatten. Berichten en reacties op TikTok bestaan doorgaans uit korte bijschriften bij video's en reacties van ongeveer 50 woorden of minder, die externe URL-links en vermeldingen van andere gebruikers kunnen bevatten.

Het corpus is het resultaat van de volgende procedure. Eerst identificeerden we accounts van personen en organisaties die, voor elk van de drie landen Spanje, Duitsland en Nederland, naar verwachting het meest waarschijnlijk verschillende vormen van online haat, intolerantie en onbeschaafdheid ontvangen. De casusselectie binnen de Spaanse, Duitse en Nederlandse context volgde een doelgerichte steekproefstrategie, met als doel een diverse verzameling van socialemedia-accounts samen te stellen die het meest waarschijnlijk online haatzaaiende uitlatingen ontvangen.

Ons doel bij de dataverzameling was een dataset samen te stellen die een breed spectrum aan online vijandigheid omvat, ter ondersteuning van de ontwikkeling van een taalmodel dat gevoelig is voor zowel expliciete als subtielere taalvormen. Voor elk van de drie landen begon de selectie met het identificeren van prominente personen en organisaties die actief zijn in domeinen die worden gekenmerkt door grote publieke zichtbaarheid en betrokkenheid. Accounts werden geselecteerd uit domeinen zoals politiek, media, entertainment en cultuur, publieke dienstverlening, wetenschap, activisme, sport en maatschappelijke organisaties. Binnen elk domein werd interne diversiteit gewaarborgd door subgroepen op te nemen, zoals gekozen vertegenwoordigers, politieke partijen, journalisten, sporters en influencers. We maakten gebruik van onze bekendheid met de contexten en van aanvullende bronnen, zoals nieuwsberichten, om personen te identificeren van wie bekend was dat zij grote hoeveelheden online haat ontvangen. Zo was onze selectie gebaseerd op waarneembare patronen van intimidatie. In het bijzonder besteedden we aandacht aan personen die betrokken waren bij recente of lopende politieke, culturele of mediacontroverses, evenals aan

personen die geassocieerd werden met controversiële kwesties zoals migratie, geopolitieke conflicten of nationale debatten waarvan bekend is dat deze de blootstelling aan online vijandigheid vergroten. Deze steekproefstrategie sloot aan bij onze definitie van haatzaaiende uitlatingen en legde tegelijkertijd de uiteenlopende verschijningsvormen ervan binnen sociale groepen vast. We identificeerden 377 accounts voor Spanje, 217 accounts voor Duitsland en 360 accounts voor Nederland.

Voor elk van deze startaccounts verzamelden we gegevens van socialemediaprofielen op X, Instagram en TikTok. Reacties geplaatst op deze accounts, of berichten waarin het account wordt vermeld, in de periode van 1-1-2024 tot en met 31-12-2025, vallen binnen de selectiecriteria. Voor X betreft dit vermeldingen en retweets, voor Instagram vermeldingen en reposts, en voor TikTok vermeldingen en antwoorden. Accounts zonder een officiële of identificeerbare aanwezigheid op sociale media werden uitgesloten. Alleen openbaar beschikbare berichten en reacties zijn opgenomen. We nemen geen inhoud op die is verwijderd tussen het moment van plaatsing en de dataverzameling, in de periode september tot en met december 2025. Gegevens die als “privé” waren ingesteld of waren verwijderd tussen het moment van plaatsing en ons onderzoek, zijn niet opgenomen in onze dataverzameling.

Deze procedure leidde tot drie datasets:

- Duitse corpus: 1.452.268 berichten (598.865 van Instagram, 730.971 van TikTok, 122.432 van X).
- Spaanse corpus: 2.445.553 berichten (457.036 van Instagram, 1.953.819 van TikTok, 34.698 van X).
- Nederlandse corpus: 5.611.651 berichten (1.018.428 van Instagram, 680.296 van TikTok, 3.912.927 van X).

Taalmodel 1: Detectie van haatzaaiende uitlatingen

Om de waarschijnlijkheid te bepalen dat een bericht in onze dataset online vijandigheid bevat, maakten we eerst gebruik van bestaande taalclassificatiemodellen. Na een eerste testfase en een beoordeling van de manier waarop elk model is getraind, selecteerden we drie afzonderlijke opensourcemodelen (Aluru et al 2002; Pérez et al 2022; IMSyPP, 2022).

Deze modellen zijn getraind op eentalige socialemediadata met behulp van de BERTje-tokenizer, elk met als doel de aanwezigheid van haatzaaiende uitlatingen te detecteren (Basile et al, 2019; Pérez et al 2021; Cinelli et al, 2021; Kralj Novak, 2022; Baider et al, 2023). We passen deze modellen toe op onze datasets, met de volgende resultaten (Tabel 1).

Tabel 1. Prestaties van model 1 (doel: detectie van haatzaaiende uitlatingen)

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

Noot: Het Duitse model kende het label 'HATE' toe bij een drempelwaarde van 0,94, het Spaanse model het label 'hateful' bij een drempelwaarde van 0,052 en het Nederlandse model het label 'Label_2' bij een drempelwaarde van 0,48. Deze drempelwaarden zijn opgenomen met het oog op reproduceerbaarheid; de niveaus zijn niet vergelijkbaar tussen de modellen.

Het is belangrijk om hier te vermelden dat deze modellen zwak presteren. Deze modellen zijn geselecteerd omdat ze het beste presteerden van de beschikbare alternatieven tijdens de voorbereidende analyse ten tijde van het ontwerp van dit onderzoek. In de praktijk presteert een AUC-score van 0,564 slechts iets beter dan toeval, zoals uitgebreider wordt besproken in de sectie over methodologische beperkingen. Wij benadrukken hier dat wij afraden deze modellen in te zetten voor geautomatiseerde besluitvorming bij de detectie van haatzaaiende uitlatingen. Voor onze analyses vullen wij elk model aan met menselijke classificatie van online vijandigheid.

Taalmodel 2: classificatie van online vijandigheid

Hoewel model 1 is ontworpen om te detecteren of een bericht waarschijnlijk problematische taal bevat, helpt het niet om te begrijpen op welk onderwerp het betrekking heeft. Om verschillende typen vijandigheid te identificeren, ontwikkelen we een op regels gebaseerd taalmodel.

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

Annotatieprocedure

Om verschillende vormen van haatzaaiende uitlatingen te identificeren, hanteerden we een abductieve benadering van coderen (Timmermans & Tavory, 2012). Ten eerste waren drie deskundige codeurs betrokken bij opeenvolgende coderingsrondes van willekeurige steekproeven uit deze datasets, om categorieën van online vijandigheid te identificeren. Onze gedeelde definitie van online vijandigheid bood sensibiliserende concepten en vormde de basis voor discussies over semantiek en onze analyseobjecten. Om terugkerende taalkundige patronen binnen gebruikersinteracties te identificeren, interpreteerden we berichten van accounts binnen vergelijkbare domeinen of met gedeelde identiteitskenmerken. Daarbij documenteerden we veelvoorkomende trefwoorden, zinsdelen en uitdrukkingen van online vijandigheid. Dit proces omvatte ook het identificeren van subtielere vormen van schadelijke taal, zoals gecodeerde uitdrukkingen, scheldwoorden, slang en dogwhistles.

Ten tweede kwamen we na elke ronde waarin steekproeven van 500 berichten waren gecodeerd bijeen om toegekende labels te vergelijken, onze definitie van haatzaaiende uitlatingen te evalueren, bestaande literatuur te betrekken en ons codeboek aan te passen of uit te breiden. Op basis van onze uitgangsdefinitie verwachtten we aanvankelijk de volgende categorieën te vinden: ras, huidskleur, religie, afkomst of nationale of etnische herkomst, gender, seksuele oriëntatie en lichamelijke beperking. Vervolgens probeerden we deze categorieën inductief aan te vullen, wat leidde tot de opname van de categorieën “Onwaarheid”, “Anti-institutioneel wantrouwen”, “Complottheorieën”, “Polarisatie”, “Oekraïne”, “Rusland”, “Autocratie”, “Midden-Oosten”, “Klimaat”, “Criminaliteit”, “Sport”, “Beledigingen”, “Grof taalgebruik”, “Ironie en sarcasme” en “Seksuele intimidatie”. In deze fase behielden we, terwijl we alle inhoud nog inductief codeerden, ook de categorieën “Informatief”, “Kritisch” en “Positiviteit”, die later werden geschrapt om ons uitsluitend op online vijandigheid te richten.

Ten derde werd de resulterende lijst met termen vervolgens beoordeeld en verfijnd door personen die bekend waren met de online context van elk land, om de culturele en taalkundige relevantie te waarborgen. Naast het identificeren van verschillende categorieën van online vijandigheid en het onderwerp van het bericht, scoorden we aanvankelijk ook een intensiteitsmaat van 1 tot 10. Deze maat hebben we uiteindelijk losgelaten na kritische

reflectie op onze uiteenlopende interpretaties van intensiteit. In lijn met onze abductieve benadering wisselden we af tussen coderen en literatuuronderzoek. Dit leidde ertoe dat we onbeschaafdheid en onbeleefdheid onderscheidden van haatzaaiende uitlatingen (Nithyanand et al., 2017; Theocharis et al., 2020). Wij verstaan onder onbeschaafdheid “een gebrek aan respect voor de sociale en culturele normen die persoonlijke interacties en het functioneren van democratische systemen reguleren” (Bentivegna & Rega, 2024). Dergelijk gebrek aan respect kan zich uiten als interpersoonlijke onbeleefdheid of als discours dat democratische systemen en hun actoren delegitimeert, waaronder vormen zoals stigmatisering, aanvallen op basis van iemands rol, demonisering, scheldnamen, bedreigingen, aanzetten tot geweld, grof taalgebruik, beschuldigingen van onwettigheid en beledigende taal (Bentivegna & Rega, 2024). Na vier coderingsrondes waren we tevreden met het codeboek en dekten de geïdentificeerde categorieën de gegevens in de vierde steekproef afdoende. Op dat moment leverden onze codering en discussies geen nieuwe categorieën of labels meer op, wat wij als een verzadigingspunt beschouwden.

Ten vierde pasten we het codeboek toe om data buiten de oorspronkelijke steekproef te annoteren. Op dit punt beschouwden we het codeboek als voldoende uitputtend voor ons doel om online vijandigheid te identificeren. Twaalf deskundige codeurs, geworven via de consortiumpartners, kregen instructie over het toepassen van het codeboek op nieuwe willekeurige steekproeven. Dit werd gecombineerd met acht kalibratiesessies om onenigheden over labels te bespreken. Onenigheden werden opgelost wanneer na bespreking tussen de codeurs consensus was bereikt.

Uiteindelijk werd de volgende overeenstemming tussen codeurs bereikt:

- Duits: 7.852 berichten uit onze steekproef zijn door twee codeurs gecodeerd (0,54% van de 1.452.268 berichten in het Duitse corpus), en 5.335 zijn door drie codeurs gecodeerd (0,37% van het corpus), met een aanvaardbare overeenstemming tussen codeurs (Krippendorff's alpha 0,73).
- Spaans: 6.534 berichten uit onze steekproef zijn door twee codeurs gecodeerd (0,27% van de 2.445.553 berichten in het Spaanse corpus), en 4.719 zijn door drie codeurs gecodeerd (0,19% van het corpus), met een aanvaardbare overeenstemming tussen codeurs (Krippendorff's alpha 0,67).

- Nederlands: 12.933 berichten uit onze steekproef zijn door twee codeurs gecodeerd (0,24% van de 5.611.651 berichten in het Nederlandse corpus), en 9.091 zijn door drie codeurs gecodeerd (0,17% van het corpus), met een goede overeenstemming tussen codeurs (Krippendorff's alpha 0,83).

Constructie van lexicons

We voerden twee rondes van axiaal coderen uit, met als doel de categorieën te relateren aan ons begrip van haatzaaiende uitlatingen zoals dat zich gedurende het coderingsproces ontwikkelde (Green & Thorogood, 2005). Daartoe onderzochten we de relaties tussen categorieën. Uiteindelijk leidden deze overwegingen ertoe dat we de term “online vijandigheid” hanteerden. Deze term is breder dan “hate speech” en vertegenwoordigt daarom beter de verscheidenheid aan taal die we in onze dataset aantreffen. We gebruiken deze overkoepelende term in de hele analyse, tenzij we expliciet verwijzen naar op identiteit gebaseerde haat die gericht is tegen personen, overeenkomstig onze oorspronkelijke definitie van “hate speech”.

Als gevolg van het axiaal coderen werden verschillende categorieën samengevoegd. Van andere categorieën stelden we vast dat zij, in het licht van de doelstellingen van het project, onvoldoende prominent of relevant waren. De definitieve kerncategorieën in deze lexicons zijn: “Gender” [GEN], “Cultuur” [CUL], “Ras” [RAC], “Seksuele oriëntatie” [SOR], “Religie” [REL], “Validisme” [ABL], “Gewelddadige taal” [VIO], “Beledigingen en persoonlijke aanvallen” [INS], “Grof taalgebruik” [VUL], “Ironie en sarcasme” [IRO], “Seksuele intimidatie” [INT], “Politiek” [POL], “Onwaarheid” [UNT], “Criminaliteit” [CRI].

Vervolgens dienden drie rondes van selectief coderen om trefwoorden, in de vorm van n-grammen, te identificeren in de data waarvoor ten minste drie menselijke codeurs overeenstemming bereikten over de relevante categorie. Deze werden vergeleken met bestaande lexicons voor haatzaaiende uitlatingen: HADES (Tulkens et al, 2016) en HurtLex v1.2 (Bassignana et al, 2018).

De resulterende lijsten met trefwoorden per categorie vormen de lexicons waarop onze classificatiemodellen zijn gebaseerd. Deze zijn open source beschikbaar (Van Haperen et al, 2026).

Modelprestaties

We passen de classificatiemodellen toe op onze datasets, met de volgende resultaten (Tabel 2).

Table 2. Prestaties Model 2 (Doel: Classificatie)

	Duits					Spaans					Nederlands				
	N positives	Precision	Recall	F1	Prevalence	N positives	Precision	Recall	F1	Prevalence	N positives	Precision	Recall	F1	Prevalence
Ableism	16	0.05	0.62	0.09	.00	25	0.26	0.32	0.29	.01	62	0.24	0.9	0.38	.01
Crime	211	0.55	0.91	0.69	.01	28	0.44	1.00	0.62	.04	312	0.37	0.91	0.53	.05
Culture	162	0.57	0.93	0.70	.07	47	0.71	0.68	0.70	.02	348	0.38	0.95	0.54	.06
Gender	140	0.51	0.53	0.52	.03	63	0.54	0.52	0.53	.01	106	0.51	0.79	0.62	.02
Insults	1453	0.79	0.43	0.56	.29	528	0.77	0.42	0.54	.10	771	0.53	0.67	0.59	.12
Irony	291	0.07	0.09	0.08	.06	240	0.14	0.09	0.11	.04	418	0.11	0.02	0.04	.07
Politics	981	0.63	0.48	0.55	.43	522	0.84	0.46	0.59	.73	3894	0.77	0.44	0.56	.62
Race	54	0.79	0.89	0.83	.03	19	0.45	0.74	0.56	.00	44	0.39	0.86	0.54	.01
Religion	37	0.56	0.62	0.59	.02	25	0.79	0.44	0.56	.01	360	0.59	0.93	0.72	.06
S.Intim	242	0.88	0.77	0.82	.05	24	0.89	0.67	0.76	.01	178	0.79	0.98	0.87	.03
S.Orient	56	0.44	0.98	0.61	.02	20	0.53	0.90	0.67	.01	50	0.37	0.88	0.52	.01
Untruth	363	0.24	0.05	0.09	.07	70	0	0	0	.02	194	0.1	0.02	0.03	.03
Violence	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

Noot. Elk lexicon is uitsluitend opgesteld met trainingsdata. De hier gerapporteerde F1-score is geëvalueerd op een onafhankelijke testset die niet in de trainingsset was opgenomen. Om rekening te houden met grote verschillen in het aantal observaties en om klassenonevenwichtigheid aan te pakken, zijn strengere drempelwaarden gehanteerd om de recall te verbeteren voor de grotere categorieën Politiek en Beledigingen in alle talen, Criminaliteit in het Spaans en Duits, en Cultuur, Ras, Religie, Seksuele oriëntatie en Vulgair taalgebruik in het Duits.

We rapporteren F1 als de mate waarin het model berichten correct aan categorieën toewijst. We merken op dat de prevalentie van sommige van onze categorieën zeer laag is, bijvoorbeeld: 1 op de 1.000 berichten in onze Duitse dataset heeft het label “Validisme”. De prestaties van de modellen hangen sterk samen met prevalentie: zeldzamere categorieën leveren minder gegevens op om een model te trainen, waardoor betrouwbare classificatie moeilijker wordt. In het algemeen geldt dat we de categorisering zorgvuldiger moeten onderzoeken naarmate de F1-score lager is. Wij beschouwen F1-scores boven 0,70 als sterk

en F1-scores onder 0,50 als onaanvaardbaar. Dit leidde ertoe dat we de categorieën “Validisme”, “Ironie” en “Onwaarheid” in verdere analyses hebben weggelaten.

In de drie talen presteren de categorieën “seksuele intimidatie” en “gender” goed, doordat zij relatief duidelijke en eenduidige trefwoorden hebben. De lage prevalentie van “validisme”, “ras”, “seksuele oriëntatie” en “gewelddadige taal” kan worden verklaard doordat dit waarschijnlijk de categorieën zijn waarop het strengst wordt gemodereerd. De categorieën “politiek” en “beledigingen en persoonlijke aanvallen” zijn uitschieters met een hoge prevalentie als gevolg van coderingsbeslissingen. Tijdens het axiaal coderen werden veel sublabels samengevoegd onder “politiek”, waaronder ideologie, beleid en politici, waardoor dit de grootste en meest diffuse categorie werd. Om dit te ondervangen, is achteraf een strikte drempelwaarde toegepast op deze categorie.

In het licht van onze abductieve benadering van coderen hebben we gedurende het hele proces kritisch geëvalueerd of onze categorieën nog steeds weerspiegelden wat we in de gegevens aantreffen. We voerden uitgebreide handmatige inspecties uit van de gegevens en toegewezen labels voor alle categorieën en voor alle platforms en landen, door middel van handmatige inspectie en willekeurige steekproefsgewijze controles. In het bijzonder inspecteerden we fout-positieven en fout-negatieven handmatig en bespraken we grensgevallen tijdens kalibratiesessies. In de meeste gevallen leidde de handmatige inspectie tot wijzigingen, waarbij het ging om het uitsluiten van specifieke woorden met dubbelzinnige betekenissen. Een voorbeeld is het woord “Let”, dat in het Nederlands een persoon uit Letland aanduidt en daarom deel uitmaakte van de categorie “Cultuur”, maar dat vaker voorkomt in de uitdrukking “Let op”. Op vergelijkbare wijze komt het woord “Frans” in het Nederlands vaker voor als voornaam, bijvoorbeeld Frans Timmermans, dan als verwijzing naar de Franse cultuur. Als gevolg van soortgelijke dubbelzinnigheid, zwakke modelprestaties en overwegingen in het licht van de doelstellingen van het project, besloten we de categorieën “Onwaarheid” en “Ironie en sarcasme” weg te laten. De categorie “Validisme” kwam naar voren tijdens het inductief coderen, maar was in de daaropvolgende annotatie te zeldzaam om betrouwbare resultaten voor modelprestaties te produceren en werd daarom eveneens weggelaten. Deze categorieën blijven opgenomen in het lexicon en de methodologie, maar worden in de onderstaande analyse weggelaten ter bevordering van de validiteit.

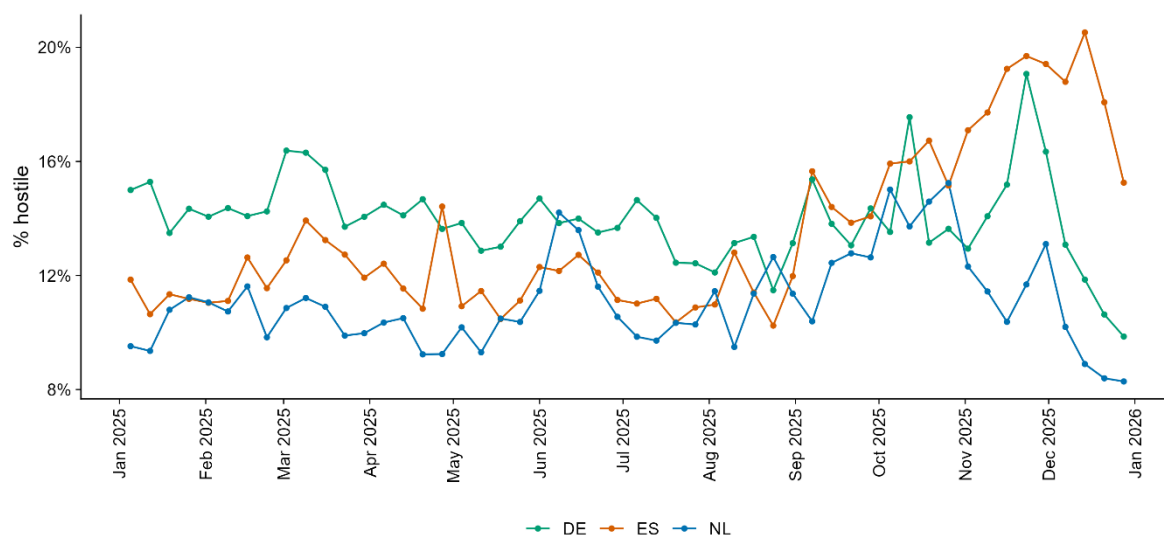
Bevindingen

Deze sectie rapporteert het volgende. Ten eerste bieden beschrijvingen van elk van de veertien categorieën kwalitatief inzicht in de verschillende typen online vijandigheid die zijn geïdentificeerd en geanalyseerd. Ten tweede bieden algemene beschrijvende statistieken een overzicht van algemene trends: verhoudingen door de tijd heen voor elk van de landen, elk van de platforms en elk van de categorieën online vijandigheid. Ten derde biedt de analyse van doelgerichtheid, persistentie en clustering inzicht in verschillende gedragstypen van verspreiders. Gezamenlijk bieden deze analyses bruikbare inzichten in welke typen interventies het meest geschikt zijn om specifieke typen online vijandigheid in context tegen te gaan.

Patronen per land

In de volgende sectie worden algemene patronen in de geïdentificeerde taal gerapporteerd, waarbij de drie landen worden vergeleken (Figuur 1).

Figuur 1. Volume online vijandigheid over tijd in Spanje, Duitsland, Nederland



Noot: Percentage berichten dat is geïdentificeerd als online vijandigheid, geaggregeerd per week en per land.

In onze dataset identificeerden we 13% van alle berichten als berichten die online vijandigheid bevatten, hoewel een gemiddelde dag ongeveer 8,8% kent. We benadrukken hierbij dat we accounts hebben geselecteerd die waarschijnlijk het vaakst vijandigheid ontvangen, wat ook richting gaf aan onze annotatie. In de praktijk is het hier gerapporteerde percentage niet representatief voor deze landen of platforms in het algemeen en vormt het waarschijnlijk een overschatting van online vijandigheid. Het verschil tussen het gemiddelde en de mediaan, evenals de pieken op specifieke dagen en de variatie in de gegevens (SD=1543, min=1, max=21.396), suggereren dat online vijandigheid wordt gedreven door plotselinge pieken in aandacht. Deze hangen waarschijnlijk samen met gebeurtenissen in de offline wereld, platformspecifieke dynamieken of gecoördineerde activiteit. Een groot deel van de haatdragende inhoud in onze dataset concentreert zich op specifieke dagen, naast een basisniveau van consistente activiteit op elk platform. Deze pieken in haatdragende inhoud clusteren rond bepaalde data, in plaats van willekeurige afzonderlijke dagen. Dit suggereert dat dergelijke activiteit verband houdt met specifieke nieuwsgebeurtenissen die de aandacht vasthouden.

Er bestaan opvallende verschillen tussen de landen wat betreft de verschillende categorieën van online vijandigheid (Tabel 3).

Tabel 3. Volume online vijandigheid per land (alle platforms)

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

Noot: beschrijvende statistieken van percentages per dag (het aantal berichten in een categorie als aandeel van het totale aantal dagelijkse berichten dat als haatdragend is geïdentificeerd).

Uit deze vergelijking volgen verschillende inzichten. Vijandigheid die verband houdt met politiek, beledigingen en criminaliteit is in elke taal de meest voorkomende categorie. In het Duits komen grof taalgebruik en, in mindere mate, seksuele oriëntatie relatief vaker voor dan in het Nederlands of Spaans. In het Nederlands komt vijandigheid die verband houdt met cultuur relatief vaker voor dan in de andere talen. Vijandigheid gerelateerd aan gender, ras, religie, seksuele intimidatie, gewelddadige taal en grof taalgebruik is in onze dataset voor elk van de talen relatief minder prominent.

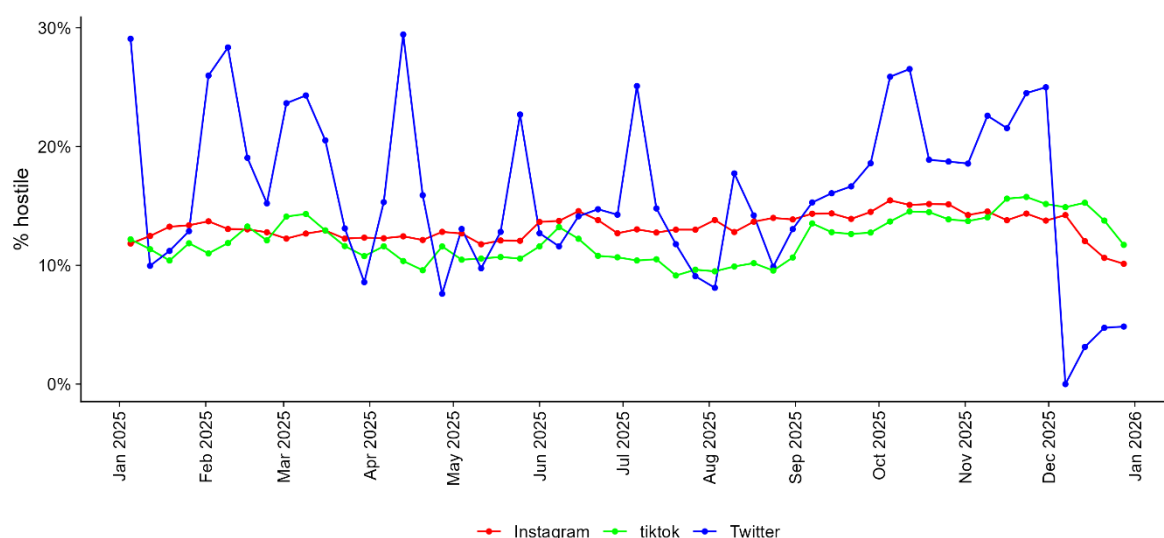
De medianen verschillen opvallend tussen de landen. Dit suggereert dat het aandeel vijandigheid voor de meeste categorieën in Nederland consistent door de tijd heen voorkomt, terwijl het in Duitsland en Spanje sterker wordt gedreven door activiteitspieken. Dit wordt verder ondersteund door relatief grote standaarddeviaties voor categorieën zoals politiek in Spanje. Categorieën zoals gewelddadige taal en seksuele intimidatie komen consistent door de tijd heen voor, terwijl politiek en beledigingen meer piekgewijs voorkomen, vooral in Spanje. Gezamenlijk duidt dit op activiteitspieken die mogelijk nauw verbonden zijn met nieuwsgebeurtenissen, een dynamiek die het sterkst zichtbaar is in Spanje.

In de volgende sectie van dit rapport onderzoeken we deze verschillen tussen landen voor elke categorie uitgebreider.

Patronen per platform

In de volgende sectie worden algemene patronen in de geïdentificeerde taal gerapporteerd, waarbij de drie platforms worden vergeleken (Figuur 2).

Figuur 2. Volume online vijandigheid over tijd per platform



Note: Percentage of posts identified as online hostility, aggregated by week and by platform.

Er zijn opvallende verschillen tussen de platforms. In onze dataset onderscheidt X zich door het hoogste aandeel online vijandigheid en door het grillige patroon door de tijd heen. Instagram heeft hogere aandelen online vijandigheid dan TikTok, maar beide platforms lijken relatief consistent door de tijd heen, zonder een duidelijke stijgende of dalende trend. Deze trends worden gedreven door activiteitspieken op specifieke dagen, in combinatie met een basisniveau van activiteit op elk platform. Er bestaan daarnaast opvallende verschillen tussen de platforms wat betreft de verschillende categorieën van online vijandigheid (Tabel 4).

Tabel 4. Percentages per platform (alle landen)

Category	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

Beschrijvende statistieken van percentages (het aantal berichten in een categorie als aandeel van het totale aantal dagelijkse berichten dat als haatdragend is geïdentificeerd). Let op: categorieën sluiten elkaar niet wederzijds uit. Een afzonderlijk bericht kan bijvoorbeeld zowel aan 'politiek' als aan 'beledigingen' worden toegewezen.

In onze dataset heeft X de hoogste aandelen online vijandigheid in alle categorieën, met name politiek en beledigingen. TikTok laat daarentegen consequent het laagste aandeel zien en Instagram neemt doorgaans een middenpositie in. Een opvallende uitzondering is gendergerelateerde inhoud, die het meest prominent is op Instagram.

Vergeleken met Instagram en TikTok heeft X voor 11 van de 13 categorieën de hoogste gemiddelde percentages, waaronder politiek (15,0%), beledigingen (10,3%), criminaliteit (4,7%), cultuur (4,1%), religie (2,3%), ras (2,1%) en grof taalgebruik (1,9%). Instagram neemt over het algemeen een middenpositie in. Hoewel de meeste categorieën minder uitgesproken zijn dan op X, overtreft Instagram X wat betreft gendergerelateerde vijandigheid (2,1% tegenover 1,3%). Daarmee is gender de enige categorie waarin Instagram duidelijk het hoogste gemiddelde aandeel heeft. TikTok heeft consequent het laagste aandeel in alle categorieën. De gemiddelde percentages zijn het laagst voor de meeste vormen van haatdragende inhoud, met name religie (0,5%), seksuele intimidatie (0,1%), gewelddadige taal (0,1%) en grof taalgebruik (0,3%). Politiek blijft de meest voorkomende categorie op TikTok, maar vertegenwoordigt minder dan de helft van het aandeel op X.

De medianen suggereren dat activiteitspieken op de meeste platforms een groot aandeel van de vijandige inhoud veroorzaken. De medianen op X voor politiek (16,4%), beledigingen (11,3%), cultuur (4,1%) en criminaliteit (4,0%) suggereren dat deze onderwerpen door de tijd heen een consistent basisniveau kennen. De variatie is ook het grootst voor politiek en beledigingen. De standaarddeviaties zijn voor deze categorieën het hoogst op alle platforms, vooral op X (SD = 12,3 voor politiek, SD = 9,4 voor beledigingen), wat wijst op aanzienlijke volatiliteit door de tijd heen.

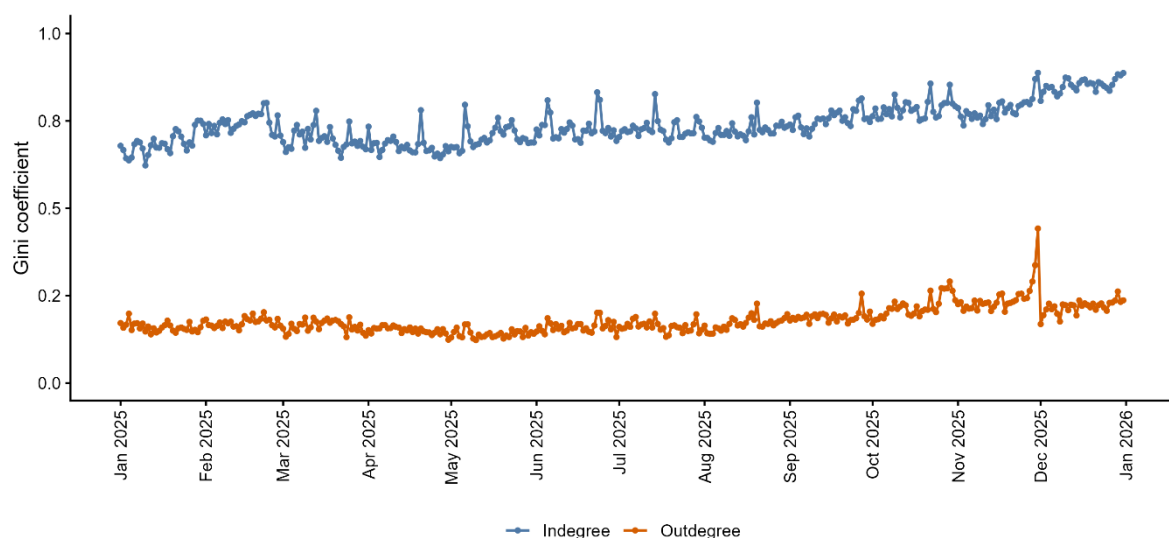
Over het geheel genomen wijzen deze bevindingen op duidelijke verschillen tussen platforms in de samenstelling van online vijandigheid. In onze dataset wordt X gekenmerkt door zowel

hogere aandelen als grotere volatiliteit tussen categorieën, vooral voor beledigingen en vijandigheid die verband houdt met politiek. Instagram laat relatief lagere aandelen zien, maar een groter aandeel gendergerelateerde vijandigheid, en meer consistentie door de tijd heen. TikTok vertoont de laagste aandelen voor de meeste categorieën vijandige inhoud die we hebben gedetecteerd.

Patronen naar doelwitten en verspreiders

In de volgende sectie worden algemene patronen onder doelwitten en verspreiders van online vijandigheid gerapporteerd. We richten ons eerst op de mate waarin de aandacht geconcentreerd is op specifieke doelwitten (Figuur 3).

Figuur 3. Concentratie online vijandigheid tegen doelen (indegree) en vanuit verspreiders(outdegree)



Deze figuur laat zien hoe de aandacht voor online vijandigheid is verdeeld over doelwitten en verspreiders. Een hogere Gini-coëfficiënt wijst op meer ongelijkheid tussen doelwitten of verspreiders: een relatief kleinere groep accounts ontvangt of verstuurt een onevenredig groot aandeel van de berichten. De belangrijkste bevinding is dat er perioden voorkomen waarin vijandige aandacht sterker geconcentreerd raakt, waarbij een klein aantal doelwitten een onevenredig groot aandeel van alle vijandige inhoud in onze dataset ontvangt. Dit wijst erop dat specifieke gebeurtenissen leiden tot een sterkere focus op specifieke doelwitten.

Figuur 3 laat ook zien in welke mate online vijandigheid geconcentreerd is onder verspreiders. Een hogere Gini-coëfficiënt betekent hier dat een kleiner aandeel van de verspreidende accounts verantwoordelijk is voor relatief veel van de vijandige inhoud die we in onze dataset hebben geïdentificeerd. De belangrijkste bevinding is de overwegend lage mate van concentratie. Dit wijst erop dat aanvallen afkomstig zijn van veel verschillende accounts, in plaats van dat een klein aantal accounts het grootste deel van de vijandige inhoud in onze dataset genereert.

Een verrassende bevinding is dat de concentratie bij zowel doelwitten als verspreiders, dus de mate waarin een klein aantal gebruikers verantwoordelijk is voor een onevenredig aandeel van ontvangen of verstuurd berichten, op vergelijkbare momenten toeneemt. We zien ook overeenkomsten in de geleidelijke toename van concentratie bij verspreiders en doelwitten, hoewel deze trend onder verspreiders minder uitgesproken is. We vermoeden dat wanneer een persoon prominent aanwezig wordt in de media, aanvallen sterker op die persoon geconcentreerd raken, terwijl tegelijkertijd een relatief kleiner aandeel accounts meer online vijandigheid genereert. Wij zien dit als een verdere aanwijzing dat vijandige inhoud verbonden is met nieuwsgebeurtenissen.

Gedragspatronen

Om verspreiders van online vijandigheid verder te onderscheiden, onderzoeken we hun gedrag wat betreft gerichte aanvallen en persistentie. Door verspreiders van online vijandigheid te vergelijken op “gerichtheid” (aan hoeveel doelwitten stuurt iemand haatdragende berichten, waarbij drie of minder doelwitten een maat voor focus vormen) en “persistentie” (op hoeveel dagen iemand haatdragende berichten tegen hetzelfde doelwit richt, waarbij drie dagen of meer gedurende een jaar een maat voor persistentie vormen), onderscheiden we vier verschillende gedragstypen (Tabel 5).

Tabel 5. Gedragsclusters onder verspreiders: focus en persistentie

Cluster	Type	Gemiddelde Doelconcentratie	Gemiddelde Persistentie	n
1	Generalisten	11.8	1.05	863
2	Gefixeerden	1	9.33	3
3	Recidivisten	2.98	2.21	174
4	Opportunisten	1.61	1.00	14669

Een eerste cluster bestaat uit een grote groep ‘opportunisten’: gebruikers met een lage intensiteit die 94% vertegenwoordigen van de accounts die problematische inhoud verspreiden in onze dataset. Het betreft accounts met minimale gerichtheid en zonder persistentie: slechts eenmalige haatdragende berichten gericht tegen één of twee doelwitten. Een tweede cluster is een middelgrote groep van generalisten (863 accounts): agressieve accounts die veel verschillende doelwitten aanvallen, maar met zeer weinig persistentie. Dit zijn eenmalige uitvallen die gelijktijdig tegen veel verschillende personen zijn gericht. Een derde cluster van ‘herhaalde overtredders’ is klein (174 accounts) en kenmerkt zich door een matige gerichtheid en matige persistentie. Deze groep bestaat uit personen die zich gedurende langere tijd met verschillende onderwerpen bezighouden en daarbij consequent problematische taal gebruiken. Een laatste cluster bestaat uit ‘fixeerders’: extreme uitschieters die langer dan een week gefixeerd zijn op één doelwit. In tegenstelling tot wat verwacht werd, zijn zulke accounts zeer zeldzaam in onze dataset. Een waarschijnlijke verklaring kan effectieve platformmoderatie zijn, waarbij de meest problematische accounts worden gerapporteerd en geblokkeerd. Dit zou ertoe leiden dat dergelijke accounts van onze dataverzameling zijn uitgesloten.

Patronen per categorie

Beschrijving van categorieën

De volgende categorieën zijn voortgekomen uit verschillende rondes van inductief coderen. Let op dat deze categorieën niet uitputtend zijn. Ook al dekten deze categorieën de steekproeven uit onze dataset die voor annotatie werden getrokken, gegevens die voor

andere landen, platforms of tijdsperioden worden verzameld, zullen waarschijnlijk uitbreiding van deze lijst vereisen. Daarnaast blijft een zekere mate van interpretatie bestaan en kunnen deze categorieën taal bevatten die onze codeurs als haatdragend hebben geïdentificeerd, terwijl een andere lezer deze mogelijk niet als haatdragend beschouwt. Deze categorieën sluiten elkaar ook niet uit: een bericht kan zowel als grof taalgebruik als als belediging worden gecategoriseerd.

De totale omvang van online vijandigheid kan als volgt verder worden uitgesplitst naar categorie (Tabel 6).

Tabel 6. Beschrijving per categorie (alle platforms, alle landen)

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

Noot: Beschrijvende statistieken van percentages (het aantal berichten in een categorie als aandeel van het totale aantal dagelijkse berichten dat als haatdragend is geïdentificeerd).

Ten eerste zijn verschillende categorieën voornamelijk gebeurtenisgedreven. Deze categorieën kennen een hoge volatiliteit: “Ras” (0,9%, SD = 0,968), “Seksuele intimidatie” (0,3%, SD = 0,291) en “Gewelddadige taal” (0,3%, SD = 0,307). Dergelijke vijandigheid is mogelijk relatief moeilijk te voorspellen en kan baat hebben bij snelle reacties, omdat de vijandigheid doorgaans sterk geconcentreerd is op enkele doelwitten en voornamelijk wordt verspreid door tijdelijke gebruikers.

Ten tweede combineren de categorieën “Politiek”, “Beledigingen”, “Criminaliteit” en “Cultuur” een consistent basisniveau met volatiliteit. Deze categorieën zijn verantwoordelijk

voor het grootste aandeel online vijandigheid in onze dataset. Politiek en beledigingen/persoonlijke aanvallen vertegenwoordigen dagelijks respectievelijk 8,3% en 5,1% van alle online vijandigheid. Criminaliteit (2,7%, $SD = 1,100$) kent een opvallend grote variabiliteit, wat erop wijst dat deze categorie voornamelijk gebeurtenisgedreven is. Voor cultuur (1,7%, $SD = 0,970$) is de variabiliteit relatief laag, wat suggereert dat dergelijke inhoud dagelijks consistent voorkomt. Omdat de mediaan en modus vergelijkbaar zijn met het gemiddelde, hebben zowel politiek als beledigingen een consistentere dagelijks basisniveau. Tegelijkertijd wijzen de standaarddeviaties op relatief grote variabiliteit in pieken. Dit suggereert dat haatdragende inhoud in deze categorieën wordt gedreven door zowel een stabiel basisniveau van activiteit als episodische activiteitspieken.

Ten derde worden categorieën met een stabiel basisniveau minder door gebeurtenissen gedreven en komen zij consistentere door de tijd heen voor. Het gaat om: “Gender” (1,4%, $SD = 0,823$), “Religie” (0,7%, $SD = 0,509$), “Seksuele oriëntatie” (0,8%, $SD = 0,803$) en “Grof taalgebruik” (0,8%, $SD = 0,540$). De standaarddeviaties wijzen hier op relatief geringe variabiliteit, wat suggereert dat dergelijke inhoud dagelijks consistent voorkomt in plaats van sterk gebeurtenisgedreven te zijn. Deze categorieën vormen dagelijks een structureel basisniveau van haatdragende inhoud. Deze categorieën kunnen baat hebben bij continue monitoring gericht op terugkerende doelwitten, in plaats van reactieve interventies of interventies gericht op herhaalde overtreeders.

Twee belangrijke kanttekeningen moeten hierbij worden benadrukt. Ten eerste hebben we, zoals besproken in de methodologie, socialemediagegevens verzameld die nog publiekelijk beschikbaar waren. Berichten waren daardoor al onderworpen aan platformmoderatie voordat zij in onze dataset terechtkwamen. Het is aannemelijk dat de categorieën in de eerste groep het zwaarst worden gemodereerd, waardoor zij in onze analyse relatief minder vaak voorkomen en volatieler zijn. Ten tweede benadrukken we dat prevalentie niet gelijkstaat aan impact. Berichten in deze categorieën bevatten de meest virulente vormen van online vijandigheid in onze dataset. Het is redelijk om aan te nemen dat deze categorieën een onevenredige impact hebben, ook als zij minder vaak voorkomen.

Over het geheel genomen zijn politiek en persoonlijke beledigingen verantwoordelijk voor het grootste aandeel berichten dat door onze modellen als problematisch is aangemerkt. Dit suggereert dat online vijandigheid in de verschillende landen en op de verschillende platforms

niet willekeurig is, maar structurele kenmerken heeft: sommige typen zijn sterk gebeurtenisgedreven, terwijl andere een consistentere basisniveau hebben. Dit kan helpen bepalen waarin middelen moeten worden geïnvesteerd voor interventies tegen online vijandigheid.

Vervolgens onderzoeken we de doelwitten en het gedrag van verspreiders voor elk van de categorieën.

Criminaliteit

Deze categorie heeft betrekking op taal over crimineel gedrag, illegale activiteiten en de politie.

- Spaans voorbeeld: “quien venga a delinquir fuera”.
- Duits voorbeeld: “Das sind kriminell die wo andere Mitmenschen gefährden die müssen alle in Therapie und weggesperrt werden !!!”.
- Nederlands voorbeeld: “Die gore crimineel zou harder gestraft moeten worden”.

Inhoud gerelateerd aan criminaliteit komt matig vaak voor in onze dataset, met een geleidelijke toename op Instagram en meer variabele niveaus met een lichte algemene daling na vroege pieken op TikTok. De categorie is volatiel op X, zonder een aanhoudende trend. Tussen de landen is dit type inhoud relatief prominent in Duitsland, matig prominent in Spanje en, vergeleken met andere categorieën, een van de minder prominente typen inhoud in Nederland.

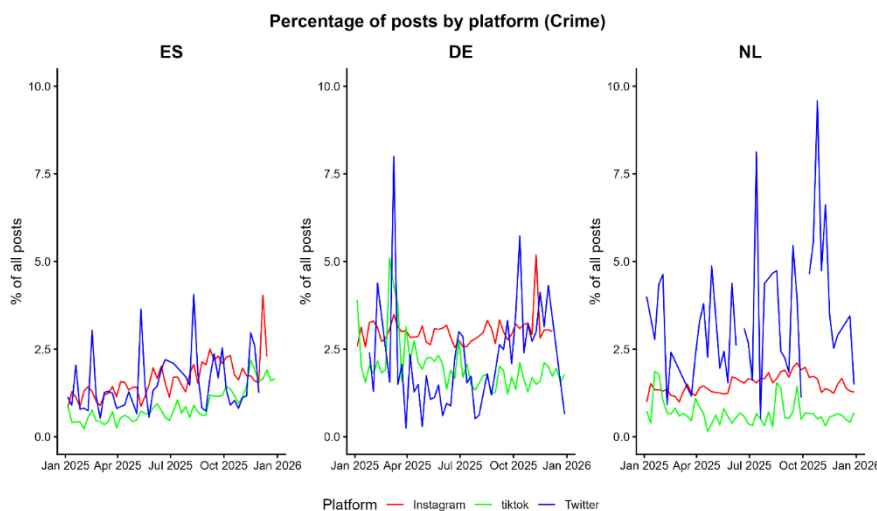
Wat doelwitten betreft, zien we een matige concentratie in combinatie met een relatief stabiele set doelwitten (gemiddelde Gini = 0,23, HHI = 0,25). De concentratie is consistent door de tijd heen, met weinig systematische stijging of daling. Tegen het einde van het jaar daalt de HHI parallel aan een toename van de activiteit. Dit wijst op fluctuerende aandacht door de tijd heen, maar de doelwitten van deze categorie blijven door de tijd heen relatief consistent geconcentreerd. Dit suggereert een patroon van verschuivende aandacht binnen een veranderende groep doelwitten, in plaats van een aanhoudende focus op specifieke accounts.

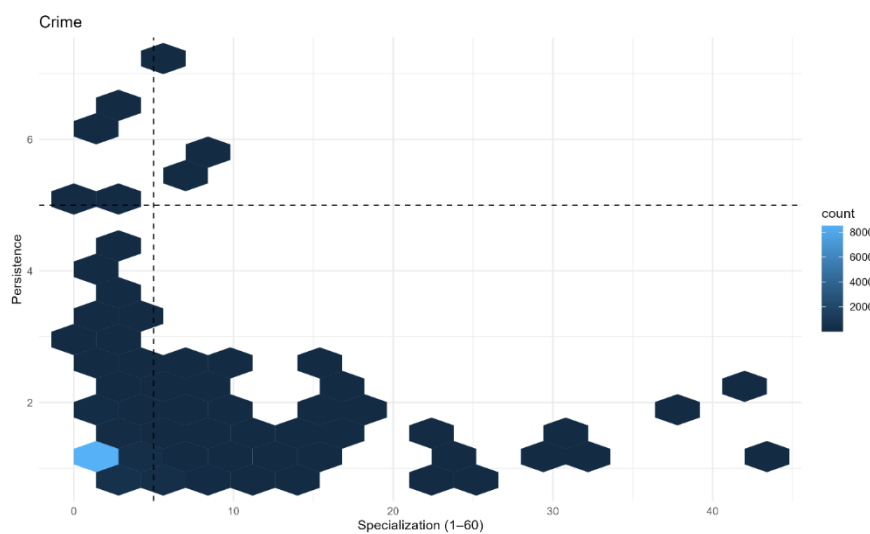
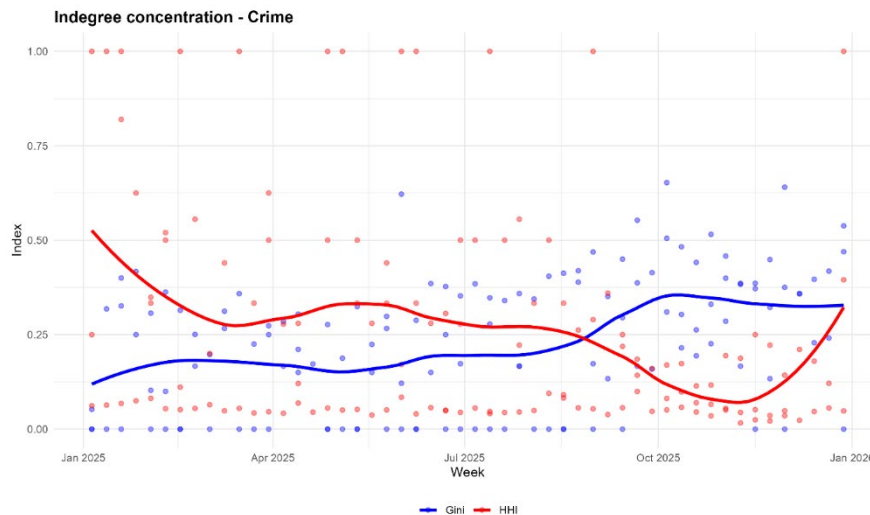
Verspreiders van criminaliteitsgerelateerde inhoud zijn, vergeleken met andere categorieën in onze dataset, in vergelijkbare mate geconcentreerd: 95% heeft een gerichte focus en lage

persistentie. Deze categorie heeft echter een van de hoogste aandelen meer gediversifieerde gerichtheid, namelijk 4,9% diffuse gerichtheid met lage persistentie, en slechts 0,1% gerichte focus met hoge persistentie. Dit suggereert dat online vijandigheid in criminaliteitsgerelateerde inhoud zelden aanhoudend is en gepaard gaat met relatief brede gerichtheid.

Wat interventies betreft, suggereren deze bevindingen dat online vijandigheid gerelateerd aan criminaliteit relatief vaak voorkomt, consistent door de tijd heen optreedt, gericht is op een relatief grote maar stabiele groep doelwitten en afkomstig is van een tijdelijke en weinig gerichte groep verspreiders. Dit maakt het een van de meer diffuse categorieën, waarvoor interventie mogelijk uitdagender is en er minder aanwijzingen zijn voor gebeurtenissen die kunnen worden voorzien.

Figuur 4. Vergelijking van misdaadgerelateerde vijandigheid naar aandeel, doelwitten en verspreiders





Cultuur

Deze categorie heeft betrekking op taal over specifieke culturen en culturele stereotypen.

- Spaans voorbeeld: “Menos clase que un gitano recién llegado de Rumanía”.
- Duits voorbeeld: “Sofort auch raus mit den Migranten !!!! Keiner will euch !!!!”.
- Nederlands voorbeeld: “smerige halve Marokkaan”.

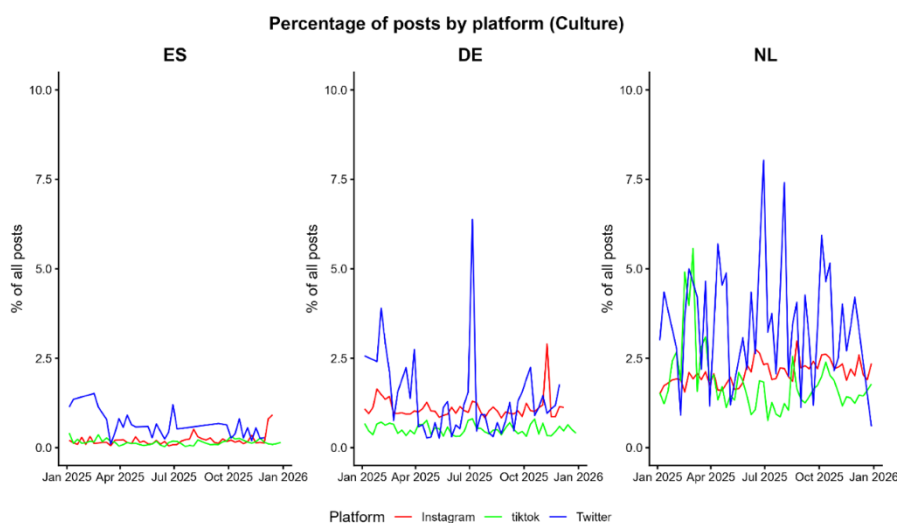
Cultuurgerelateerde inhoud komt zeer vaak voor in onze Nederlandstalige dataset, maar opvallend veel minder in onze Spaanse en Duitse datasets. In Nederland komt deze inhoud consistent voor op alle drie de platforms en is zij, vergeleken met andere categorieën, relatief prominent op zowel Instagram als TikTok. Op X komt online vijandigheid die verband houdt met cultuur vaak voor en in activiteitspieken.

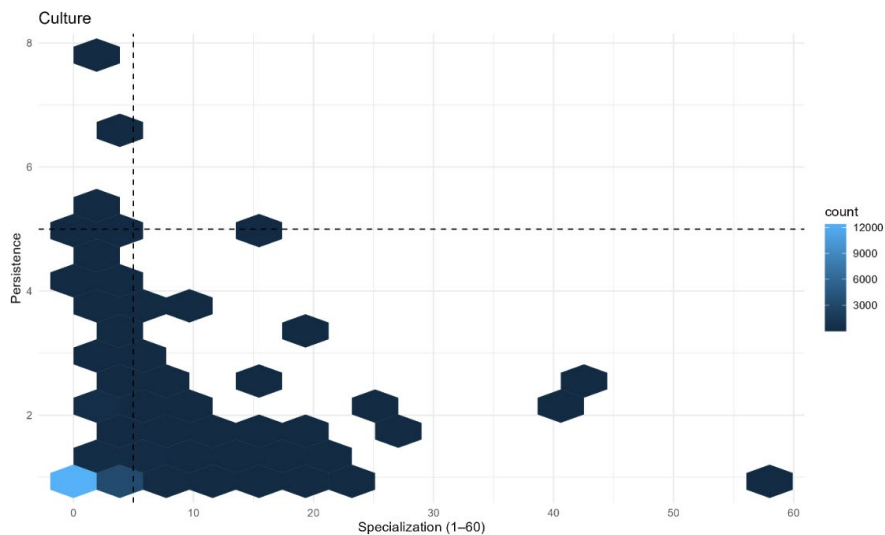
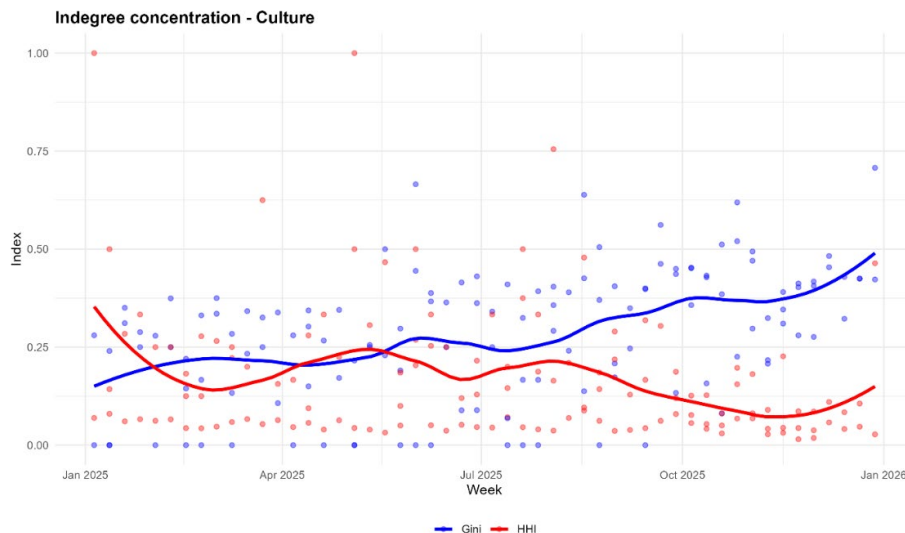
Wat doelwitten betreft, is online vijandigheid gerelateerd aan cultuur een van de meest ongelijk verdeelde categorieën, in combinatie met relatief lage dominantie onder doelwitten (gemiddelde Gini = 0,28, HHI = 0,17). Dit wijst op een geconcentreerde groep doelwitten die relatief vaak verandert. Beide patronen zijn consistent door de tijd heen. Dit suggereert dat dergelijke vijandigheid ongelijk is verdeeld, maar zonder blijvende centrale doelwitten.

Verspreiders van cultuurgerelateerde online vijandigheid behoren, vergeleken met andere categorieën in onze dataset, tot de meest gericht opererende en tijdelijke gebruikers (97,4% gerichte focus met lage persistentie). Slechts 2,6% van de gebruikers richt zich op verschillende doelwitten, zonder aanwijzingen voor persistentie door de tijd heen. Dit suggereert dat dergelijke verspreiders sterk gefragmenteerd zijn en slechts bij één enkele uiting betrokken zijn.

Wat interventies betreft, suggereren deze bevindingen dat online vijandigheid gerelateerd aan cultuur relatief vaak voorkomt in het Nederlands en consistent optreedt op alle platforms, maar veel minder vaak voorkomt in het Spaans en Duits. Over de landen heen is de aandacht in deze categorie ongelijk verdeeld, maar met grotere variatie tussen doelwitten, en afkomstig van een zeer tijdelijke maar relatief gerichte groep verspreiders. Dit maakt het een van de meer diffuse categorieën, waarvoor interventie mogelijk uitdagender is en er minder aanwijzingen zijn voor gebeurtenissen die kunnen worden voorzien.

Figuur 5. Vergelijking van cultuurgerelateerde vijandigheid naar aandeel, doelwitten en verspreiders





Gender

Deze categorie heeft betrekking op intolerante taal die gericht is tegen specifieke aspecten van identiteit in relatie tot gender en genderstereotypen. Zij omvat gendernormen en misogynie.

- Spaans voorbeeld: “sin maquillaje,es una bruja”.
- Duits voorbeeld: “Ich kann diese dumme Hexe nicht mehr ertragen.”.
- Nederlands voorbeeld: “Alle vrouwen zijn dom”.

Gendergerelateerde online vijandigheid komt in onze dataset weinig voor op elk van de platforms en in elk van de landen. Zij is opvallend zeldzaam in het Nederlands en duidelijk

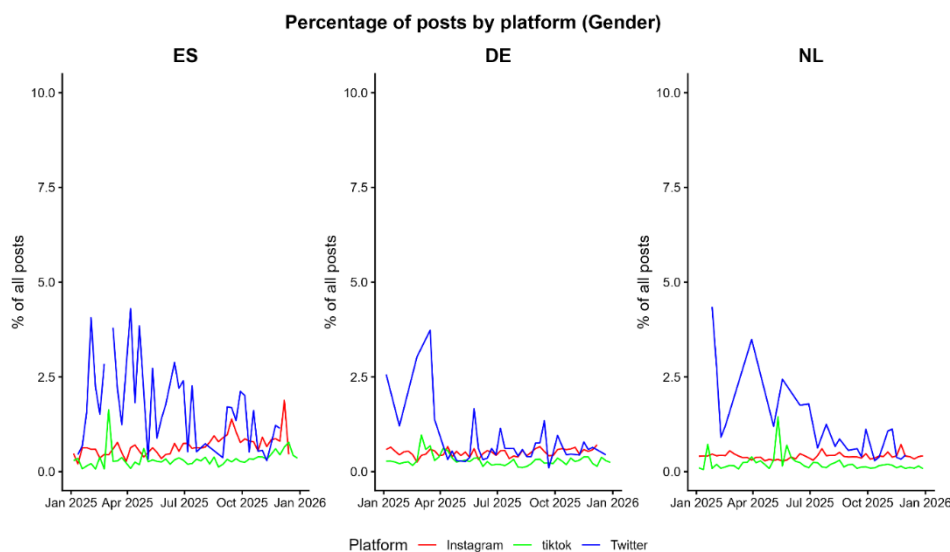
vaker aanwezig in het Spaans. Vergeleken met andere categorieën zijn de schommelingen op X bovendien opvallend gematigd. Deze relatief stabiele trend is consistent door de tijd heen.

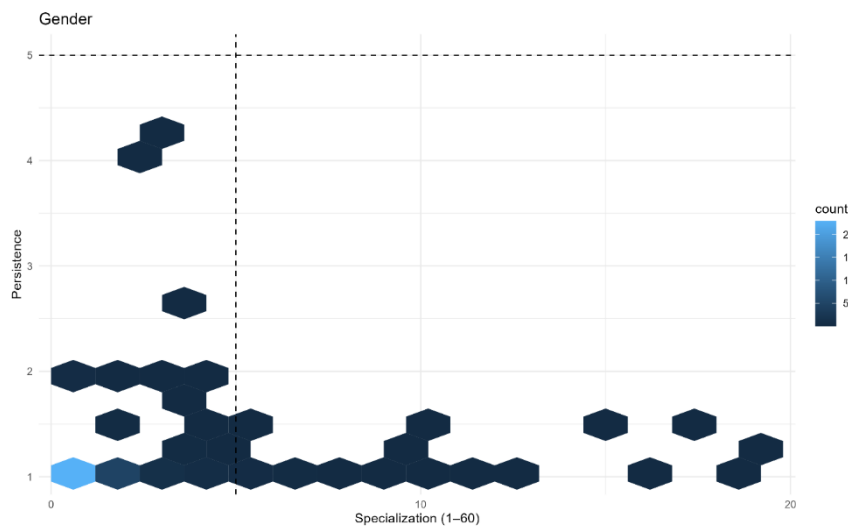
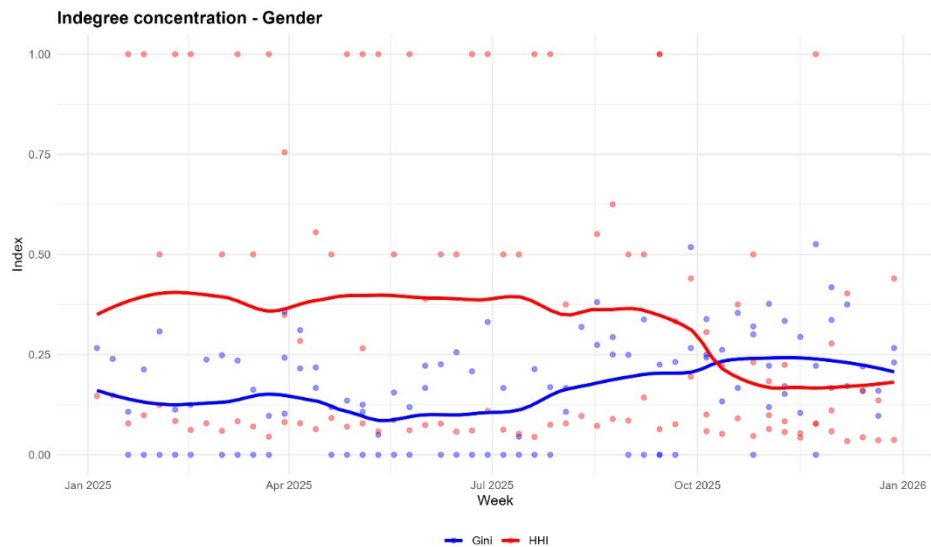
Wat doelwitten betreft, is online vijandigheid gerelateerd aan gender relatief diffuus, in combinatie met relatief hoge dominantie (gemiddelde Gini = 0,16, HHI = 0,32). Dit suggereert dat, hoewel de ongelijkheid beperkt is, de aandacht geconcentreerd is op enkele belangrijke doelwitten. Deze trend is consistent door de tijd heen.

Wat verspreiders van online vijandigheid betreft, is gender in onze dataset over het geheel genomen de meest geconcentreerde categorie (98,3% gerichte focus met lage persistentie), met minimale spreiding over verschillende doelwitten (1,7%) en geen persistente gebruikers. Dit suggereert dat dergelijke vijandigheid sterk is gericht op specifieke doelwitten en bestaat uit eenmalige betrokkenheid met vrijwel geen persistentie.

Wat interventies betreft, is gender een categorie met een consistent basisniveau van activiteit, die in het Spaans meer aandacht verdient dan in de andere talen. Over de landen heen is de aandacht binnen deze categorie relatief diffuus, maar met een relatief stabiele kerngroep van doelwitten. De vijandigheid is afkomstig van een tijdelijke en gerichte groep verspreiders. Interventies in deze categorie worden minder geïnformeerd door gebeurtenisgedreven aandacht, maar kunnen worden gericht op prominente accounts die consistent dergelijke vijandigheid aantrekken.

Figuur 6. Vergelijking van gendergerelateerde vijandigheid naar aandeel, doelwitten en verspreiders





Beledigingen en persoonlijke aanvallen

Deze categorie heeft betrekking op een breed scala aan taal die bedoeld is om een individu te beledigen of persoonlijk aan te vallen, zonder expliciete groepsidentiteiten als doelwit te nemen. Het is een brede categorie die niet alleen online vijandigheid omvat, maar ook onbeschaafdheid en onbeleefdheid in het algemeen, waaronder grove, aanstootgevende en denigrerende taal.

- Spaans voorbeeld: “Eres amiga de este no hay mas preguntas miserable”.
- Duits voorbeeld: “Du bist so erbärmlich.”.
- Nederlands voorbeeld: “Wat ben jij een ontzettende eikel”.

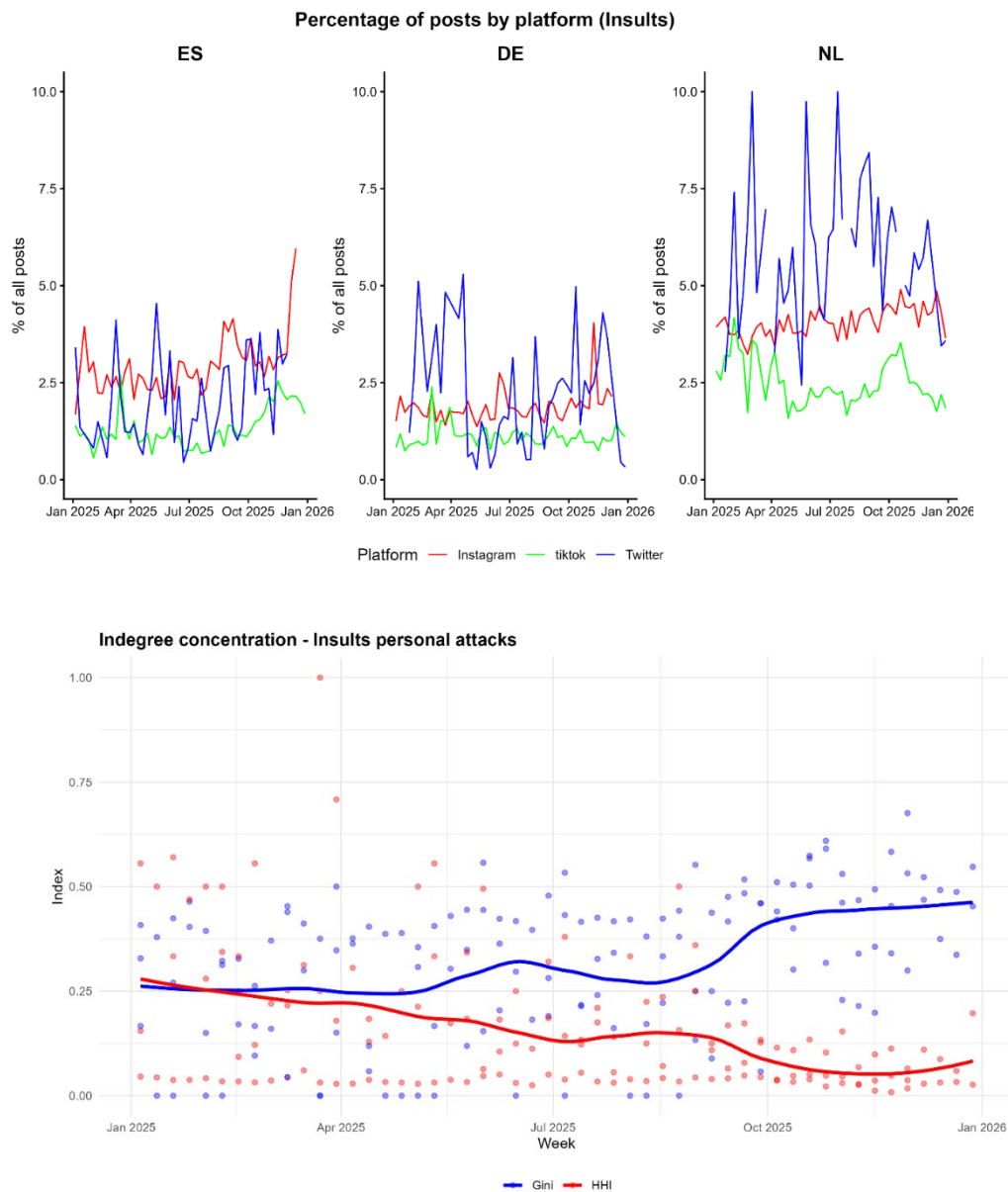
Online vijandigheid in de vorm van beledigingen of persoonlijke aanvallen is de op een na meest voorkomende categorie in onze dataset op elk van de platforms en in elk van de landen, en is bijzonder prominent in het Nederlands. De categorie is, vergeleken met andere categorieën, opvallend aanwezig op Instagram en iets minder op TikTok. Op X is zij consistent volatiel. Instagram kent consistent hoge niveaus met bescheiden variatie door de tijd heen, terwijl TikTok een vergelijkbare stabiliteit vertoont met incidentele pieken.

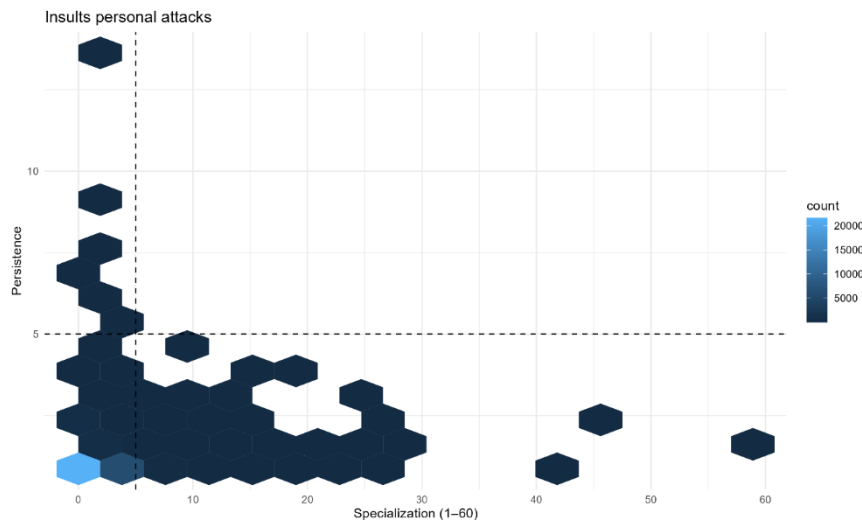
Wat doelwitten betreft, is dit de op een na meest ongelijk verdeelde categorie in onze dataset, gecombineerd met relatief lage dominantie van doelwitten (gemiddelde Gini = 0,32, HHI = 0,15). Dit suggereert dat beledigingen sterk geconcentreerd zijn, zonder consistente centrale doelwitten. Hoewel we in onze dataset een toenemende divergentie zien tussen meer concentratie en minder dominantie, lijkt de trend door de tijd heen relatief stabiel. Over het geheel genomen wijzen deze patronen op diffuus maar ongelijk gericht gedrag.

Wat verspreiders van beledigingen betreft, zien we in onze dataset meer variatie dan bij andere categorieën. Hoewel de verspreiding nog steeds sterk geconcentreerd is en weinig persistentie kent (97,2% gerichte focus met lage persistentie), is er een groter aandeel accounts dat een bredere groep doelwitten beledigt (2,7% diffuse gerichtheid) en een verhoudingsgewijs groter aandeel persistente aanvallers (0,1%).

Wat interventies betreft, is de aandacht in deze categorie voor elk van de talen sterk ongelijk verdeeld, met grote variatie tussen doelwitten. Hoewel het moeilijk is te voorspellen welk type gebeurtenissen dergelijke inhoud kan aanjagen en de aandacht grotendeels verspreid is over veel verschillende doelwitten, suggereren deze patronen ook dat er kerngroepen van verspreiders zijn die specifieke doelwitten aanhoudend beledigen.

Figuur 7. Beledigingen en persoonlijke aanvallen: vergelijking naar aandeel, doelwitten en verspreiders





Politiek

Deze categorie heeft betrekking op een breed scala aan taal over politiek, politici, politieke partijen, specifiek beleid, oorlogen en conflicten, ideologie en links-rechtspolarisatie.

- Spaans voorbeeld: “Basura woke”.
- Duits voorbeeld: “Selten so gelacht Hochverschuldung haben wir durch das Russen Drecksvolk die Freude von dem AFD Gesindel”.
- Nederlands voorbeeld: “Geert Wilders is een engnek”.

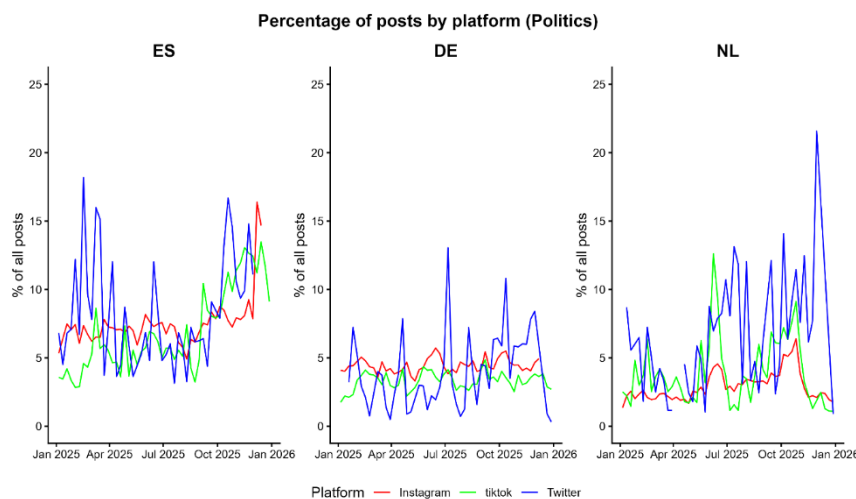
Online vijandigheid gerelateerd aan politiek is de meest voorkomende categorie in onze dataset op elk van de platforms en in elk van de landen. Zij is op alle platforms relatief volatiel en kent consistent grote variatie door de tijd heen, het sterkst op X. Dit suggereert dat dergelijke inhoud kan worden gedreven door gebeurtenissen en nieuwscycli.

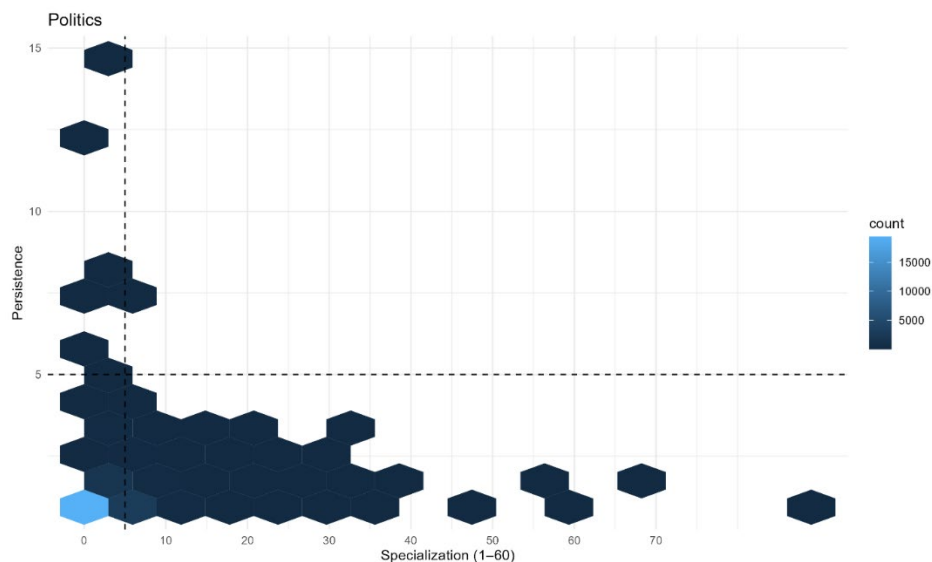
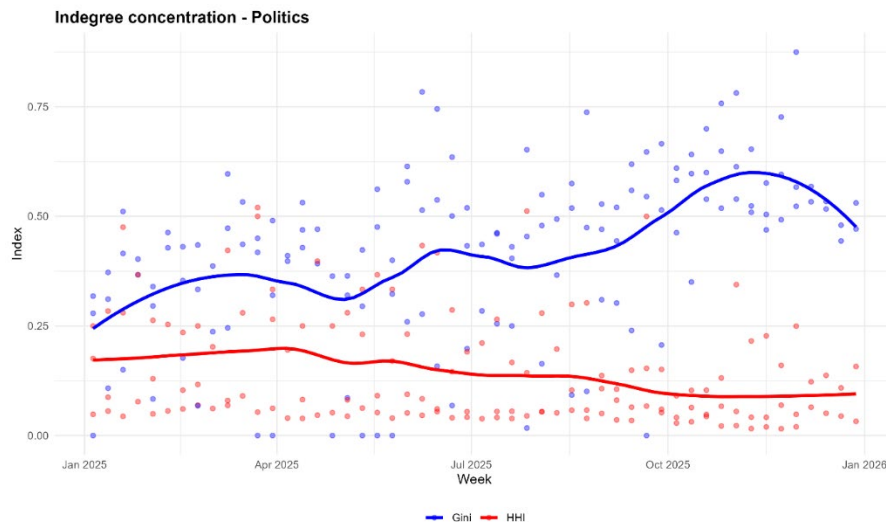
Wat doelwitten betreft, is dit de meest ongelijk verdeelde categorie in onze dataset, gecombineerd met relatief lage dominantie van doelwitten (gemiddelde Gini = 0,42, HHI = 0,14). Dit suggereert dat politieke vijandigheid sterk geconcentreerd is, maar dat er weinig consolidatie rond centrale doelwitten plaatsvindt. Hoewel we in onze dataset een toenemende divergentie zien tussen meer concentratie en minder dominantie, lijkt de trend door de tijd heen relatief stabiel. Over het geheel genomen wijzen deze patronen op diffuse gerichtheid zonder consistente focus.

Wat verspreiders van politiekgerelateerde vijandigheid betreft, zien we in onze dataset meer variatie dan bij de meeste andere categorieën. Hoewel de verspreiding nog steeds sterk geconcentreerd is en weinig persistentie kent (94,1% gerichte focus met lage persistentie), heeft deze categorie het grootste aandeel accounts dat een bredere groep doelwitten beledigt (5,8% diffuse gerichtheid), evenals een verhoudingsgewijs groter aandeel persistente aanvallers (0,1%). Dit suggereert dat politieke vijandigheid vaker gericht is op meerdere doelwitten en dat sommige groepen gebruikers deze vijandigheid persistent verspreiden.

Wat interventies betreft, is de aandacht in deze categorie voor elk van de talen sterk ongelijk verdeeld, met grote variatie tussen doelwitten. Hoewel het moeilijk is te voorspellen welk type gebeurtenissen dergelijke inhoud kan aanjagen en de aandacht diffuus over veel verschillende doelwitten is verspreid, worden deze patronen waarschijnlijk gedreven door nieuwscycli. Zij suggereren ook dat er kerngroepen van verspreiders bestaan die specifieke doelwitten aanhoudend beledigen.

Figuur 8. Vergelijking van politiekgerelateerde vijandigheid naar aandeel, doelwitten en verspreiders





Ras

Deze categorie heeft betrekking op intolerante taal die gericht is tegen specifieke aspecten van identiteit in relatie tot ras en etniciteit, evenals raciale stereotypen.

- Spaans voorbeeld: “Qué les pasa a los hombres n*? Por qué son la mayoría así?”.
- Duits voorbeeld: “Was hat der N* mit deutscher Jugend zu tun?”
- Nederlands voorbeeld: “Terug naar je eigen LAND VIEZE AAP”.

Online vijandigheid gerelateerd aan ras komt niet vaak voor in onze dataset. In onze dataset vinden we weinig resterend bewijs in het Spaans en Duits, of op Nederlands Instagram en TikTok, hoewel deze vorm in het Nederlands wel voorkomt met aanzienlijke volatiliteit op X. We verwachtten verschillen tussen platforms te vinden, aangezien X in de onderzochte

periode naar verluidt was gestopt met strikte moderatie. We constateren aanvaardbare modelprestaties, die in combinatie met een sterke overeenstemming tussen codeurs suggereren dat deze taal nauwkeurig kan worden gedetecteerd. Alles bij elkaar erkennen we dat de beperkte hoeveelheid bewijs in deze categorie in onze dataset een gevolg kan zijn van de methodologie, en niet noodzakelijkerwijs bewijs is van afwezigheid.

Wat aandacht en doelwitten betreft, heeft racisme in onze dataset een zeer lage gemiddelde Gini-coëfficiënt (0,08), maar een hoge HHI (0,42). Dit suggereert dat, hoewel de ongelijkheid laag is, de aandacht sterk geconcentreerd is op een kleine groep doelwitten. De trend is consistent door de tijd heen, maar we merken op dat de volatiliteit relatief hoog is als gevolg van het beperkte aantal waarnemingen. In combinatie met de activiteitspieken suggereert dit episodische gerichtheid, met uitbarstingen van geconcentreerde aandacht op een klein aantal doelwitten. Dit kan verband houden met nieuwsgebeurtenissen die specifieke personen tot prominente doelwitten maken.

Wat verspreiders betreft, zien we een sterk geconcentreerd patroon (96,9% gerichte focus met lage persistentie). Vergeleken met andere categorieën in onze dataset kent racisme een relatief groot aandeel gebruikers met een bredere gerichtheid (2,9% diffuse gerichtheid met lage persistentie). We zien een klein aandeel persistente gebruikers (0,1%).

Wat interventies betreft, suggereren deze bevindingen dat racisme mogelijk al strikt wordt gemodereerd. Op basis van wat in onze dataset resteert, lijkt racisme grotendeels episodisch, met weinig aanhoudend of breed gericht gedrag, maar met een kern van persistente aanvallers, vooral op X.

Figuur 9. Vergelijking van rasgerelateerde vijandigheid naar aandeel, doelwitten en verspreiders



Religie

Deze categorie heeft betrekking op intolerante taal die gericht is tegen specifieke aspecten van identiteit in relatie tot religie, evenals religieuze stereotypen.

- Spaans voorbeeld: “Yo voy a entrar en una Mezquita y me voy a cargar 😏😏”.
- Duits voorbeeld: “Ein verlogener Jude, das ist ja nichts Neues”.
- Nederlands voorbeeld: “Weer een dag dat Soennieten zich niet normaal kunnen gedragen”.

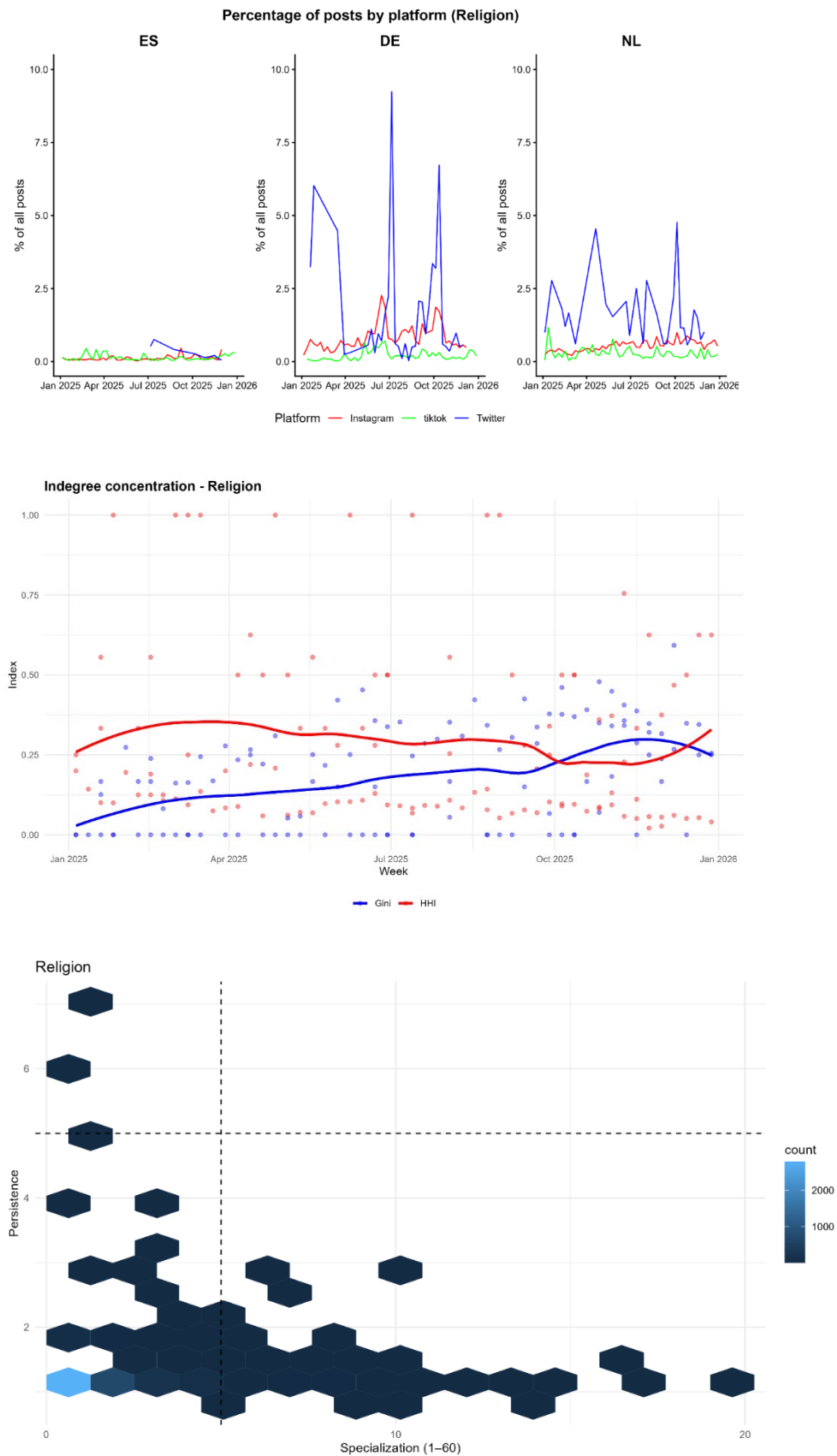
Online vijandigheid gerelateerd aan religie komt in onze dataset niet vaak voor in het Duits en Nederlands, op alle platforms, maar is opvallend zeldzaam in het Spaans op alle platforms in onze dataset. Waar religiegerelateerde vijandigheid wel voorkomt in onze dataset, gebeurt dit relatief consistent door de tijd heen op Instagram en TikTok, terwijl zij op X zeer volatiel is. Deze trends zijn consistent door de tijd heen en suggereren een combinatie van zowel een basisniveau van religiegerelateerde haat als activiteitspieken die door gebeurtenissen worden gedreven.

Wat doelwitten betreft, zien we vergeleken met andere categorieën een matige concentratie in combinatie met matige dominantie van doelwitten (gemiddelde Gini = 0,18, HHI = 0,29). Deze trend lijkt relatief stabiel door de tijd heen, met een lichte toename in concentratie. Deze patronen wijzen op een relatief diffuse groep doelwitten.

Wat verspreiders van religiegerelateerde vijandigheid betreft, is dit een relatief homogene categorie in onze dataset. De meeste accounts richten zich op één enkel doelwit en kennen weinig persistentie (97,1% gerichte focus met lage persistentie), met een relatief groot aandeel diffuse gerichtheid (2,8%). We zien zeer weinig persistentie. Dit suggereert sterk episodisch maar relatief diffuus gericht gedrag.

Wat interventies betreft, is de aandacht in deze categorie relatief diffuus, maar bestaat er enige consistentie tussen doelwitten. We zien opvallende verschillen tussen de talen die waarschijnlijk verband houden met methodologische beperkingen. Het zal moeilijk zijn om tegen dit type inhoud te interveniëren, aangezien er weinig aanwijzingen zijn over welk type gebeurtenissen dergelijke inhoud kan aanjagen, er geen gebruikelijke verdachten zijn en er door de tijd heen zeer weinig persistentie onder verspreiders bestaat. Afhankelijk van de nationale context kunnen er consistente doelwitten zijn.

Figuur 10. Vergelijking van religiegerelateerde vijandigheid naar aandeel, doelwitten en verspreiders



Seksuele intimidatie

Deze categorie heeft betrekking op taal die seksueel suggestief of vernederend is en gericht is tegen een individu, zonder verwijzing naar een groep of genderrol.

- Spaans voorbeeld: “Por qué no te abres más de patas? Qué modales tan exquisitos. Tonta del culo”.
- Duits voorbeeld: “wer dich N*tte fickt ist zu faul zum wichsen”.
- Nederlands voorbeeld: “Ga je beschimilde hoere vrendin neuke die zwijn eet elke dag overwichte kkr slet”.

Seksuele intimidatie komt niet vaak voor in onze dataset. We vinden weinig resterend bewijs in het Spaans en Duits of op Nederlands Instagram en TikTok, hoewel deze vorm in het Nederlands wel voorkomt met aanzienlijke volatiliteit op X. We verwachtten verschillen tussen platforms te vinden, aangezien X in de onderzochte periode naar verluidt was gestopt met strikte moderatie. We constateren ook aanvaardbare modelprestaties, die in combinatie met een sterke overeenstemming tussen codeurs suggereren dat deze taal nauwkeurig kan worden gedetecteerd. Alles bij elkaar erkennen we, net als bij racisme, dat de beperkte hoeveelheid bewijs in deze categorie in onze dataset een gevolg kan zijn van de methodologie en niet noodzakelijkerwijs bewijs is van afwezigheid. De patronen die we wel aantreffen zijn stabiel door de tijd heen: volatiliteit op X, met een consistent basisniveau van seksuele intimidatie op Instagram en TikTok, in elk van de landen.

Wat aandacht en doelwitten betreft, heeft seksuele intimidatie in onze dataset een zeer lage gemiddelde Gini-coëfficiënt (0,11), maar een hoge HHI (0,44). Dit suggereert dat, hoewel de ongelijkheid over het algemeen laag is, de aandacht sterk geconcentreerd is op een kleine groep doelwitten. De trend is consistent door de tijd heen, maar we merken op dat de volatiliteit relatief hoog is als gevolg van het beperkte aantal waarnemingen. In combinatie met activiteitspieken op X, en in mindere mate op Instagram, suggereert dit episodische gerichtheid met uitbarstingen van geconcentreerde aandacht op een klein aantal doelwitten. Net als bij racisme kan dit samenhangen met nieuwsgebeurtenissen die specifieke personen tot prominente doelwitten maken.

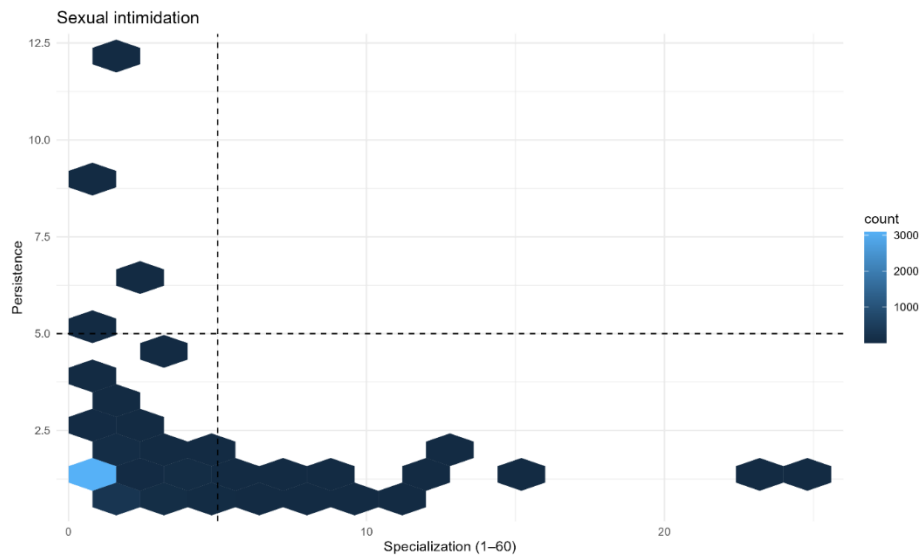
Wat verspreiders betreft, zien we een sterk geconcentreerd patroon (97,7% gerichte focus met lage persistentie). Net als bij racisme, en in tegenstelling tot andere categorieën in onze

dataset, kent seksuele intimidatie een relatief groot aandeel gebruikers met een bredere gerichtheid (2,3% diffuse gerichtheid met lage persistentie). We vinden geen betekenisvolle persistentie in onze dataset.

Wat interventies betreft, lijkt seksuele intimidatie op basis van het resterende bewijs in onze dataset grotendeels episodisch te zijn, met enig consistent gericht gedrag maar beperkte persistentie en geen kerngroep van daders die kan worden aangesproken. Deze bevindingen suggereren dat seksuele intimidatie mogelijk al strikt wordt gemodereerd, mogelijk inclusief het rapporteren en vervolgens blokkeren van accounts, op Instagram en TikTok, en in mindere mate op X.

Figuur 11. Vergelijking van vijandigheid gerelateerd aan seksuele intimidatie naar aandeel, doelwitten en verspreiders





Seksuele oriëntatie

Deze categorie heeft betrekking op intolerante taal die gericht is tegen specifieke aspecten van identiteit in relatie tot seksuele oriëntatie, evenals gerelateerde stereotypen.

- Spaans voorbeeld: “la persona gay en muuuy promiscua y guarra en el aspecto sexual. Tu eres gay? Pues disfrutalo”.
- Duits voorbeeld: “Dreckige Schwuchtel”.
- Nederlands voorbeeld: “Donderop man vieze flikker”.

Online vijandigheid gerelateerd aan seksuele oriëntatie is geen zeer prominente categorie. In onze dataset komt zij het sterkst naar voren in het Duits op elk platform, op X in het Nederlands en slechts zelden in het Spaans. De categorie is volatiel op X en relatief stabiel op TikTok en Instagram. Dit wijst op een basisniveau van dergelijke inhoud, in combinatie met gebeurtenissen en nieuwscycli die pieken in aandacht veroorzaken.

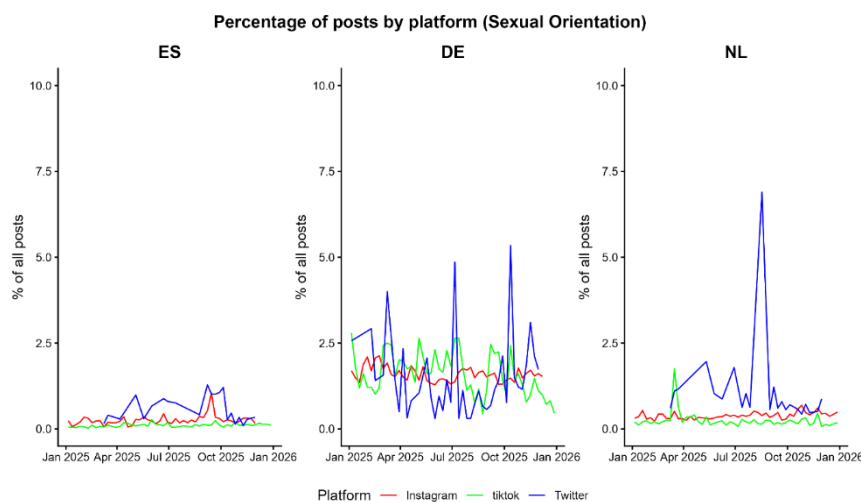
Wat doelwitten betreft, is dit een van de minst geconcentreerde categorieën in onze dataset, in combinatie met matige dominantie van doelwitten (gemiddelde Gini = 0,11, HHI = 0,32). Dit suggereert dat vijandigheid met betrekking tot seksuele oriëntatie enigszins diffuus is, maar dat er sprake is van een zekere consolidatie rond centrale doelwitten. Hoewel we in onze dataset enige convergentie zien naar meer concentratie en minder dominantie, lijkt de trend

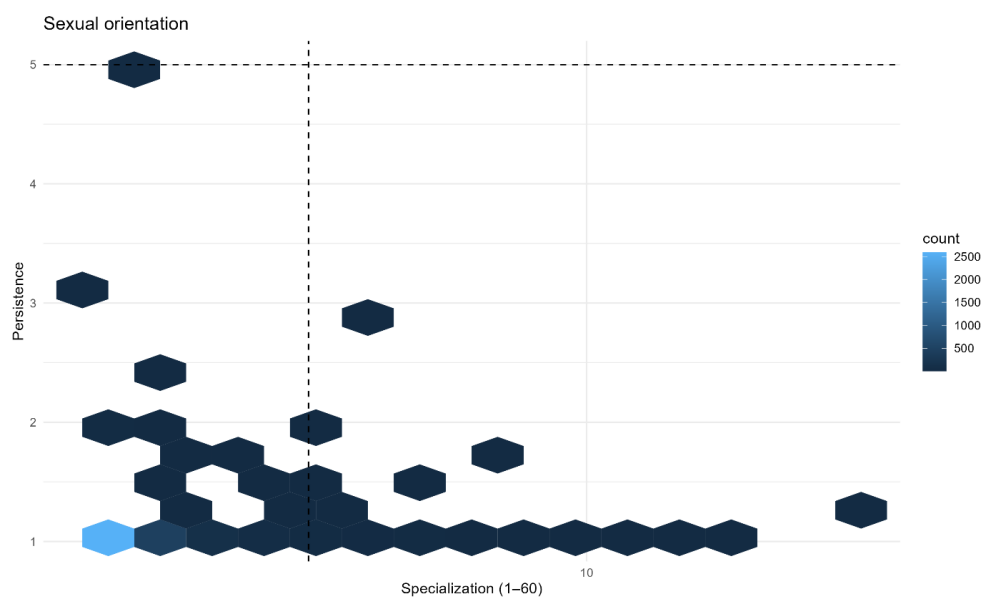
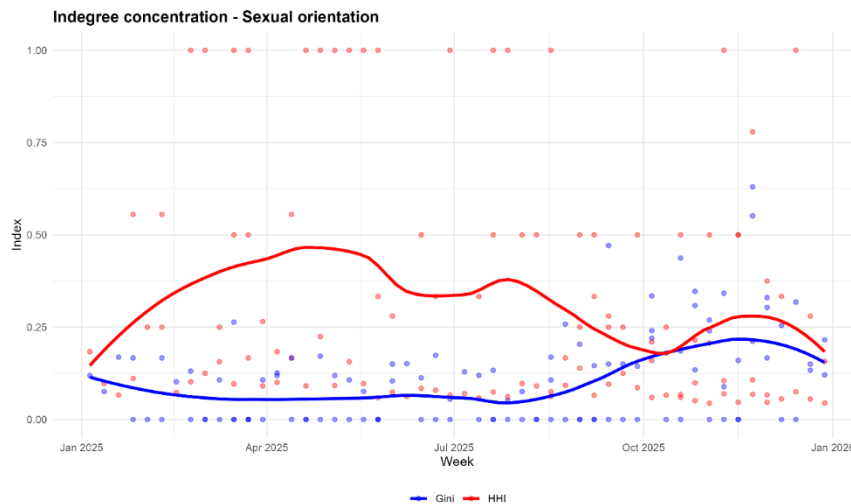
door de tijd heen relatief stabiel. Over het geheel genomen wijzen deze patronen op een overwegend diffuse concentratie, maar met matig consistente doelwitten.

Wat verspreiders van vijandigheid gerelateerd aan seksuele oriëntatie betreft, zien we een relatief groot aandeel tijdelijke aanvallers (98,4% gerichte focus met lage persistentie), met verwaarloosbare diffuse gerichtheid of persistentie. Dit suggereert dat vijandigheid op basis van seksuele oriëntatie wordt gedomineerd door eenmalige aanvallen.

Wat interventies betreft, suggereren deze bevindingen dat vijandigheid op basis van seksuele oriëntatie mogelijk al effectief wordt gemodereerd. Op basis van wat in onze dataset resteert, lijkt seksuele oriëntatie voornamelijk gekenmerkt te worden door tijdelijke aanvallen, met zeer weinig aanhoudend of breed gericht gedrag.

Figuur 12. Vergelijking van vijandigheid gerelateerd aan seksuele oriëntatie naar aandeel, doelwitten en verspreiders





Gewelddadige taal

Deze categorie heeft betrekking op taal die aanzet tot geweld of geweld verheerlijkt, waaronder gerichte bedreigingen met geweld of de dood.

- Spaans voorbeeld: “Esa soberbia con la que aplaude y niega es merecedora de una guillotina”.
- Duits voorbeeld: “Der gehört an die Wand”.
- Nederlands voorbeeld: “Tering jong, zal jou doodtrappen.”.

Gewelddadige taal komt niet vaak voor in onze dataset. We vinden weinig resterend bewijs in het Spaans en Duits, of op Nederlands Instagram en TikTok, hoewel deze vorm in het

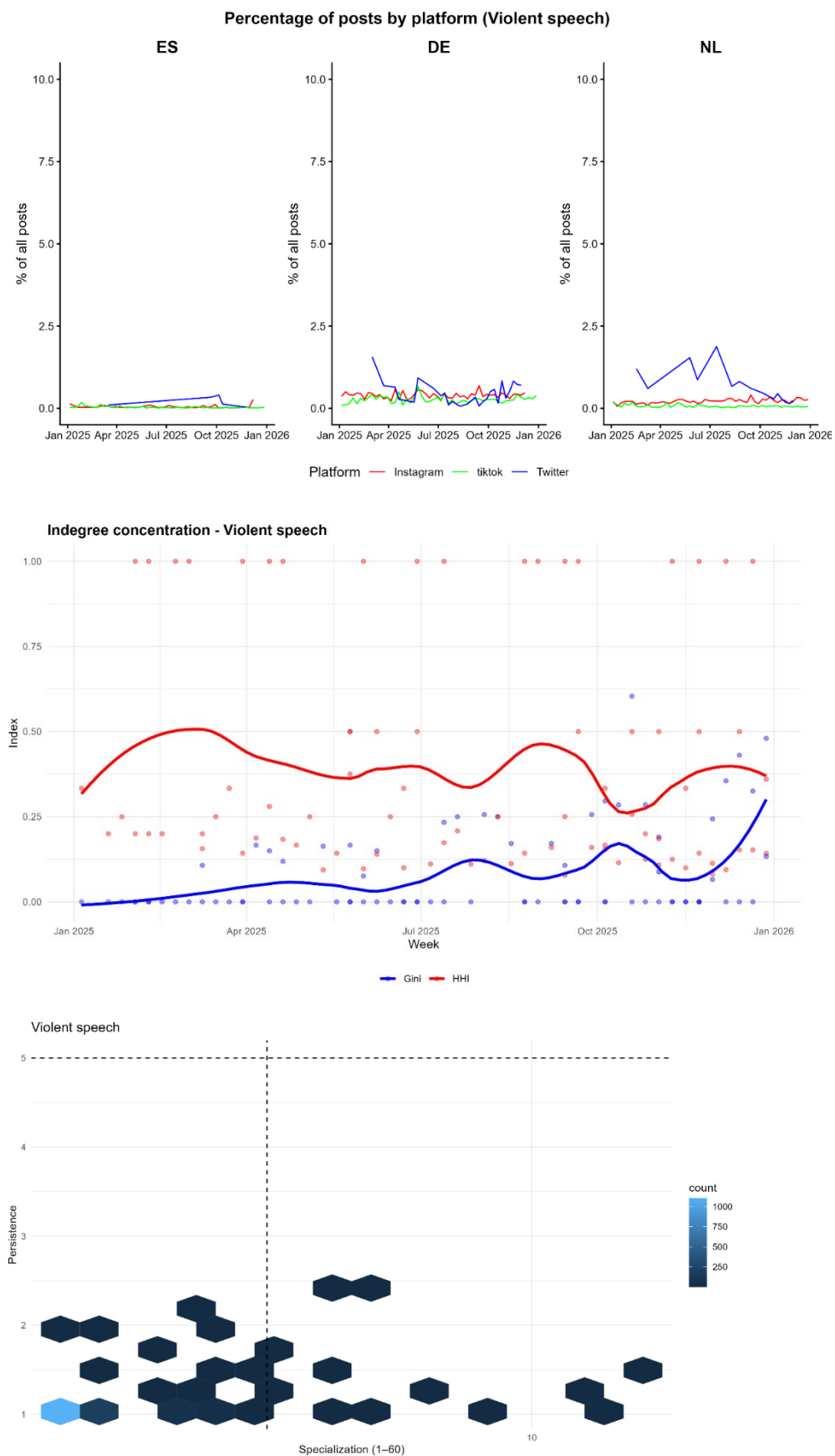
Nederlands met enige volatiliteit op X voorkomt. We verwachtten verschillen tussen platforms te vinden, aangezien X in de onderzochte periode naar verluidt was gestopt met strikte moderatie. We constateren aanvaardbare modelprestaties, die in combinatie met een sterke overeenstemming tussen codeurs suggereren dat deze taal nauwkeurig kan worden gedetecteerd. Alles bij elkaar erkennen we, net als bij ras en seksuele intimidatie, dat de beperkte hoeveelheid bewijs in deze categorie in onze dataset een gevolg kan zijn van de methodologie en niet noodzakelijkerwijs bewijs is van afwezigheid.

Wat aandacht en doelwitten betreft, heeft gewelddadige taal in onze dataset een zeer lage gemiddelde Gini-coëfficiënt (0,08), maar een hoge HHI (0,40). Dit suggereert dat, hoewel de ongelijkheid laag is, de aandacht sterk geconcentreerd is op een kleine groep doelwitten. De trend is consistent door de tijd heen, maar we merken het beperkte aantal waarnemingen op. In combinatie met slechts gematigde activiteitspieken suggereert dit een basisniveau van gewelddadige taal, in plaats van episodische uitbarstingen van aandacht. Deze categorie lijkt dus minder afhankelijk van nieuwscycli.

Wat verspreiders betreft, zien we een sterk geconcentreerd patroon (97,7% gerichte focus met lage persistentie). Vergeleken met andere categorieën in onze dataset kent gewelddadige taal een relatief groot aandeel gebruikers met een bredere gerichtheid (2,3% diffuse gerichtheid met lage persistentie). In onze dataset is zeer weinig bewijs voor persistente aanvallers die zich door de tijd heen op dezelfde doelwitten richten.

Wat interventies betreft, suggereren deze bevindingen op basis van het resterende bewijs van gewelddadige taal in onze dataset dat deze mogelijk al effectief wordt gemodereerd. Voor zover zichtbaar in onze dataset lijkt gewelddadige taal voornamelijk gekenmerkt te worden door tijdelijke aanvallen, met zeer weinig aanhoudend of breed gericht gedrag.

Figuur 13. Vergelijking van gewelddadige uitingen naar aandeel, doelwitten en verspreiders



Grof taalgebruik

Deze categorie heeft betrekking op taal die als onbeleefd kan worden geïnterpreteerd en aanstootgevend kan zijn, maar niet gericht is tegen een specifieke groep of een individu. Hieronder vallen vloeken, schelden en krachttermen.

- Spaans voorbeeld: “Coño señora entonces que hace usted y todo su equipo de gobierno”.
- Duits voorbeeld: “drecks Politik!”.
- Nederlands voorbeeld: “God verdomme vuile moordenaars.”.

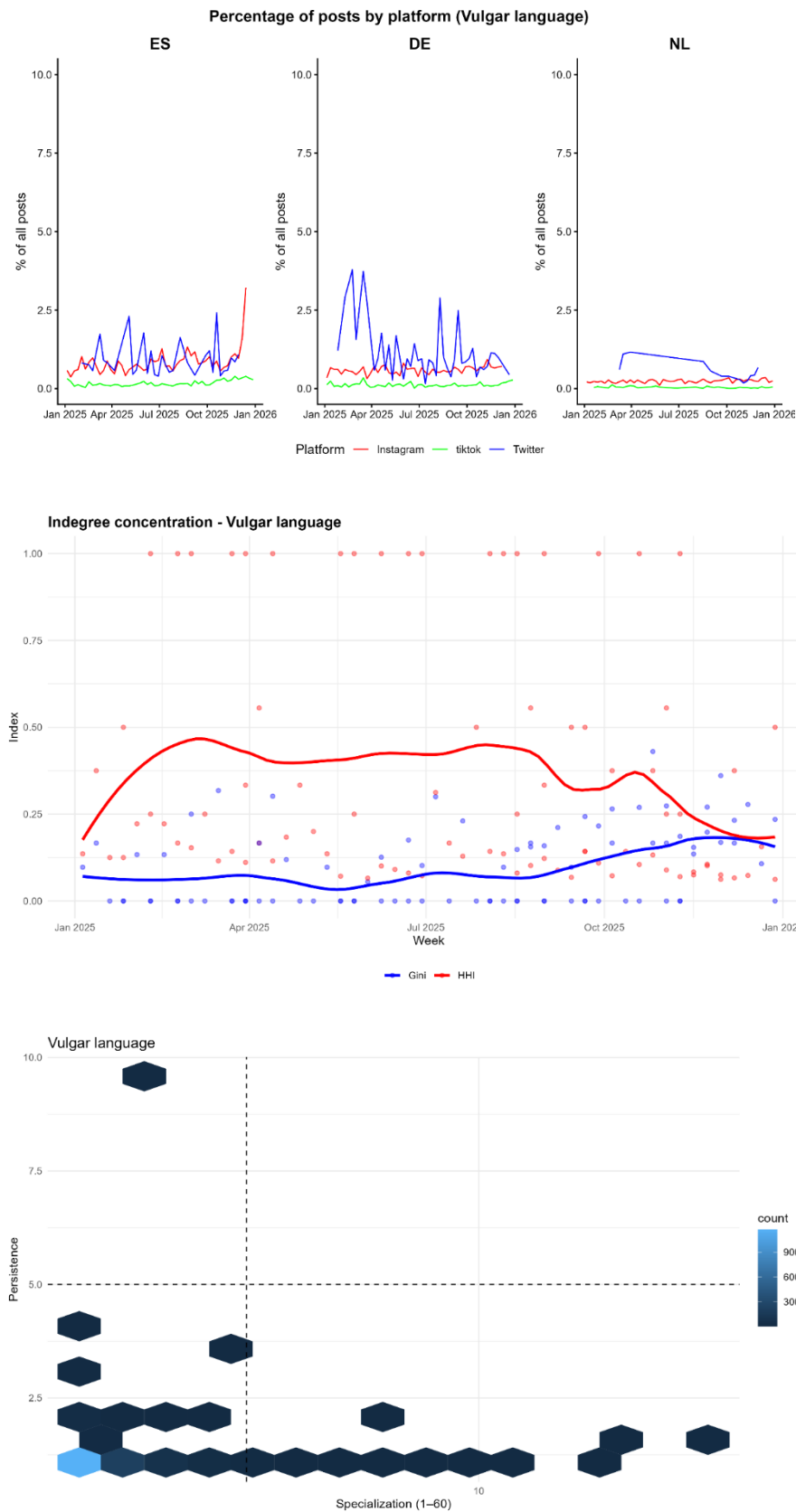
Grof taalgebruik is geen zeer prominente categorie in onze dataset. In onze gegevens komt het relatief vaker voor in het Spaans op elk platform, op Duits X en minder vaak in het Nederlands. Het komt relatief consistent voor op elk platform, ook op X vergeleken met andere categorieën. Dit suggereert een basisniveau van activiteit in plaats van gebeurtenisgedreven pieken.

Wat aandacht en doelwitten betreft, heeft grof taalgebruik in onze dataset een zeer lage gemiddelde Gini-coëfficiënt (0,11) en een hoge HHI (0,36). Dit suggereert dat, hoewel de ongelijkheid laag is, de aandacht sterk geconcentreerd is op een kleine groep doelwitten. De trend is consistent door de tijd heen, hoewel we het beperkte aantal waarnemingen opmerken. In combinatie met slechts gematigde activiteitspieken suggereert dit een basisniveau van grof taalgebruik in plaats van episodische uitbarstingen van aandacht. Deze categorie lijkt dus minder afhankelijk van nieuwscycli.

Wat verspreiders betreft, zien we een sterk geconcentreerd patroon (98,5% gerichte focus met lage persistentie). Vergeleken met andere categorieën in onze dataset kent grof taalgebruik een matig aandeel gebruikers met een bredere gerichtheid (1,4% diffuse gerichtheid met lage persistentie). We zien een klein aandeel persistente gebruikers (0,1%).

Wat interventies betreft, suggereren deze bevindingen dat grof taalgebruik op Instagram en TikTok mogelijk al in grote mate wordt gemodereerd. Op basis van wat in onze dataset resteert, lijkt grof taalgebruik grotendeels episodisch, met enig breed gericht gedrag en een kern van persistente aanvallers, vooral op X.

Figuur 14. Vergelijking van vulgair taalgebruikgerelateerde vijandigheid naar aandeel, doelwitten en verspreiders



Conclusies

Over het geheel genomen suggereren onze bevindingen dat online vijandigheid wijdverspreid is en voornamelijk bestaat uit eenmalige vijandige reacties. Wat interventies betreft, suggereert dit dat de voornaamste uitdaging niet ligt in het tegengaan van een klein aantal hardnekkige veelplegers, maar van een groot aantal eenmalige vijandige reacties afkomstig van veel verschillende accounts.

We benadrukken dat de beperkte hoeveelheid bewijs in onze dataset voor categorieën zoals “Ras” of “Gewelddadige taal” het gevolg kan zijn van vertekeningen in de dataverzameling en methodologie, en niet noodzakelijkerwijs bewijs is van afwezigheid. We wijzen ook op de beperkte nauwkeurigheid en recall van de taalmodellen die hier zijn gebruikt om verschillende typen vijandigheid te detecteren. Door de vergelijking tussen categorieën, landen en platforms leiden we de volgende inzichten af.

In het algemeen constateren we dat verspreiders van vijandigheid diffuser zijn dan verwacht. Online vijandigheid in onze dataset is voornamelijk afkomstig van veel verschillende accounts, in plaats van geconcentreerd te zijn bij enkele accounts. We vinden weinig overlap tussen categorieën onder verspreiders van online vijandigheid. Dit suggereert dat de meeste accounts zich bezighouden met één type vijandigheid, in plaats van hun repertoire te verbreden. Onder verspreiders van online vijandigheid vinden we zeer weinig bewijs voor gerichte, persistente aanvallen. In alle categorieën is gerichte focus met lage persistentie het dominante patroon, wat suggereert dat de meeste gebruikers zich, ongeacht het onderwerp, kortstondig en op een beperkt aantal doelwitten richten. We concluderen dat de meeste online vijandigheid in onze dataset bestaat uit eenmalige berichten, met zeer weinig herhaalde overtreders.

We zien ook dat doelwitten van online vijandigheid sterk geconcentreerd zijn. Ditzelfde patroon zien we wanneer we onderscheid maken naar type vijandigheid. In alle categorieën vinden we een hoge mate van concentratie van doelwitten, waarbij het grootste deel van de aandacht op relatief weinig doelwitten is gericht, en uiteenlopende niveaus van consistentie, waarbij dezelfde accounts doelwit blijven. Wanneer we gerichtheid, vijandig berichtgedrag en volatiliteit verder differentiëren, vinden we opvallende verschillen tussen categorieën die verschillende interventiestrategieën onderbouwen.

Aanbevolen interventiestrategieën

Wat interventies betreft, stellen we op basis van onze bevindingen drie toepasbare inzichten voor, samengevat in Tabel 7.

Tabel 7. Aanbevolen interventiestrategie per categorietype

Categorie	Doel-concentratie	Type verspreider	Volatiliteit	Focus interventie
Ras	Geconcentreerd + consistent	Recidivisten (Diffuus + consistent)	Episodisch	Specifieke doelwitten + specifieke verspreiders, + events
Seksuele intimidatie	Diffuus + consistent	Generalisten (Diffuus + tijdelijk)	Consistent + episodisch	Specifieke doelwitten + events
Religie	Diffuus + wisselend	Generalisten (Diffuus + tijdelijk)	Consistent + episodisch	Specifieke doelwitten + events
Seksuele oriëntatie	Diffuuse + consistent	Generalisten (Diffuus + tijdelijk)	Consistent + episodisch	Specific doelwitten + events
Vulgaire taal	Goncentreerd + consistent	Recidivisten (Diffuus + consistent)	Consistent	Specific doelwitten + Specifieke verspreiders
Criminaliteit	Diffuus + consistent	Generalisten (Diffuus + tijdelijk)	Consistent + episodisch	Events
Cultuur	Geconcentreerd + wisseld	Opportunisten (Gefocust + tijdelijk)	Consistent + episodisch	Events
Politiek	Geconcentreerd + wisseld	Recidivisten (Diffuus + persistent)	Episodisch	Events
Gender	Diffuus + consistent	Opportunisten (Gefocust + tijdelijk)	Consistent	Specifieke doelwitten
Gewelddadige taal	Geconcentreerd + consistent	Generalisten (Diffuus + tijdelijk)	Consistent	Specifieke doelwitten
Beledigingen	Geconcentreerd + wisseld	Recidivisten (Diffuus + persistent)	Consistent	Specifieke verspreiders

Een eerste groep categorieën wordt gekenmerkt door episodische activiteit: ras, seksuele intimidatie, seksuele oriëntatie, religie, gewelddadige taal, criminaliteit, cultuur en politiek.

Vergeleken met andere categorieën zijn deze relatief volatieler en waarschijnlijk sterker gebeurtenisgedreven. Interventiestrategieën voor deze categorieën kunnen zich richten op het anticiperen op en reageren op gebeurtenissen die online vijandigheid genereren.

Een tweede groep categorieën wordt gekenmerkt door concentratie van doelwitten: ras, seksuele intimidatie, religie, seksuele oriëntatie, gender, grof taalgebruik en gewelddadige taal. Vergeleken met andere categorieën zijn de doelwitten van deze vormen van vijandigheid relatief geconcentreerd en/of stabiel. Interventiestrategieën voor deze categorieën kunnen zich richten op het monitoren van accounts die waarschijnlijk doelwit worden van online vijandigheid.

Een derde groep categorieën wordt gekenmerkt door consistentie onder verspreiders: ras, grof taalgebruik en beledigingen. Vergeleken met andere categorieën zijn verspreiders van deze vormen van vijandigheid relatief persistent. Interventiestrategieën voor deze categorieën kunnen zich richten op bekende accounts die online vijandigheid genereren.

We benadrukken dat de lage recall en precisie van de taalmodellen een belangrijke beperking vormen. In de praktijk zullen de hier gebruikte modellen veel berichten met online vijandigheid missen en veel gevallen ten onrechte als positief classificeren. Gezien ons doel om moderatie te ondersteunen, achten wij “valse alarmen” verkieslijk boven het missen van veel problematische berichten. Deze nauwkeurigheid zal waarschijnlijk in de loop van de tijd en bij toepassing op andere platforms verder verslechteren. Wij raden af deze modellen te gebruiken voor geautomatiseerde besluitvorming. De praktische toepasbaarheid van deze modellen ligt in het markeren van inhoud voor beoordeling door experts.

Methodologische beperkingen

De dataverzameling en methodologie die aan dit rapport ten grondslag liggen, kennen belangrijke kanttekeningen.

Ten eerste merken we op dat de meest voorkomende vormen van vijandigheid in onze dataset mogelijk ondervertegenwoordigd zijn als gevolg van onze dataverzamelingsstrategie. Hoewel onze datasets omvangrijk zijn, zijn zij slechts gedeeltelijk representatief. Door, om privacyredenen en gezien de praktische beperkingen van socialemediagegevens, alleen publiek beschikbare gegevens te verzamelen, zijn de strengst gemodereerde vormen van

inhoud mogelijk verwijderd voordat we deze konden verzamelen. Daarom is het waarschijnlijk dat racisme en gewelddadige taal in het algemeen, en gendergerelateerd geweld in het bijzonder, in de praktijk prominenter zijn dan onze analyse suggereert. Zelfs wanneer moderatie dergelijke inhoud na enige tijd verwijderd, kan de schade al zijn aangericht.

Ten tweede blijven er, eveneens gerelateerd aan de dataverzameling, beperkingen bestaan als gevolg van het prioriteren van breedte en diversiteit. Onze aanvankelijke focus op het identificeren van prominente personen impliceert dat minder prominente of zichtbare actoren mogelijk ondervertegenwoordigd zijn. We hebben geprobeerd hiervoor te corrigeren door doelgerichte steekproeftrekking op basis van identiteit. Kenmerken zoals ras, etniciteit, genderidentiteit, seksuele oriëntatie, religie, migratieachtergrond, taal en politieke ideologie zijn meegenomen om ervoor te zorgen dat de dataset groepen omvat die onevenredig vaak doelwit zijn van online vijandigheid, maar mogelijk minder prominent aanwezig zijn in het publieke debat. We streefden ernaar minderheidsgroepen te vertegenwoordigen die specifiek zijn voor elke nationale context, zoals taalminderheden in Spanje en personen die deelnemen aan debatten over historisch gevoelige onderwerpen in Duitsland. Een gerelateerde beperking van onze aanpak is de overlap tussen sectoren, zoals politiek en activisme, die pogingen om evenwicht tussen categorieën aan te brengen bemoeilijkte. Uiteindelijk leidde dit ertoe dat sommige categorieën sterk vertegenwoordigd zijn, zoals politiek, ten koste van andere, zoals sport. Wij zijn van mening dat onze aanpak uitgebreide dekking biedt over sociale domeinen en identiteitgroepen, terwijl rekening wordt gehouden met contextuele en culturele variatie in de uiting van online haat.

Ten derde erkennen we dat socialemediaberichten niet neutraal of natuurlijk voorkomend zijn, en dat socialemedia-accounts geen natuurlijke personen vertegenwoordigen. Onze dataset omvat waarschijnlijk verschillende vormen van automatisering, spam en andere vormen van kwaadwillende intentie. Het identificeren van bots en vormen van coördinatie valt buiten de reikwijdte van dit project, maar we moedigen nader onderzoek naar de relatie tussen gecoördineerde vormen van manipulatie en de verschillende typen vijandigheid die in dit rapport worden geïdentificeerd sterk aan.

Ten vierde erkennen we dat de vergelijking van gegevens van verschillende socialemediaplatforms, Instagram, X en TikTok, en verschillende nationale contexten, Duitsland, Spanje en Nederland, sterk beperkt is. Elk socialemediaplatform heeft

uiteenlopende gebruikersgroepen, specifieke omgangstalen en eigen mogelijkheden die betekenisvol verweven zijn met het taalgebruik op deze platforms. Daarnaast erkennen we dat betekenis sterk afhankelijk is van context. Een vergelijkende analyse van de betekenis van taal en gerelateerde verschillen tussen landen en platforms vereist kwalitatief onderzoek. Wij zijn van mening dat de meer kwantitatief georiënteerde vergelijkingen in dit rapport gerechtvaardigd zijn door ons doel om patronen in de prevalentie van taal tussen landen en platforms te identificeren. We beschouwen deze vergelijkingen als indicatieve trends, en niet als nauwkeurige weergaven.

Ten vijfde is geautomatiseerde taalclassificatie inherent beperkt in termen van prestaties. We erkennen dat de lage recall en precisie van modellen voor haatspraakdetectie, en in mindere mate van de classificatiemodellen, een belangrijke beperking vormen van de hier beschreven methode. Dit geldt in het bijzonder voor categorieën die minder vaak voorkomen in onze dataset en waarvoor dus minder waarnemingen beschikbaar zijn om de modellen te trainen. Voor praktische doeleinden betekent dit dat de modellen veel fout-positieven zullen produceren. Gezien ons doel om moderatie te ondersteunen, achten wij valse alarmen verkieslijk boven het missen van veel problematische berichten. De drie bestaande modellen voor haatspraakdetectie zijn geselecteerd na een voorlopige analyse van hun prestaties tijdens de onderzoeksofzet. Ten tijde van de onderzoeksofzet is besloten om LLM's, als beter presterende alternatieven voor op regels gebaseerde lexicons, niet te gebruiken vanwege praktische haalbaarheid en beperkingen op het gebied van databeheer en privacy. Het finetunen van BERT-gebaseerde modellen met onze geannoteerde gegevens blijft een ambitie voor toekomstig werk. Voor onze analyse hebben we deze beperking aangepakt met een semi-superviseerde methode, waarin geautomatiseerde detectie wordt gecombineerd met menselijke annotatie en classificatie. We hebben deze beperking ook proberen te ondervangen door voortdurende kritische reflectie, beoordeling van overeenstemming tussen codeurs, kalibratie, handmatige inspectie en steekproefsgewijze controles om de validiteit van onze bevindingen te waarborgen. Desondanks erkennen we dat de classificatiemodellen waarop onze analyses zijn gebaseerd tekortschieten op het gebied van nauwkeurigheid, precisie en recall. Naarmate taal en taalgebruik zich verder ontwikkelen en online vijandigheid verschuift naar andere woorden en betekenissen, zal de nauwkeurigheid van deze modellen waarschijnlijk verder afnemen. Wij raden af deze modellen te gebruiken voor

geautomatiseerde besluitvorming. De praktische toepasbaarheid van deze modellen ligt in het markeren van inhoud voor beoordeling door experts.

Ten zesde merken we op dat we weinig kunnen zeggen over de impact van online vijandigheid. Onze analyses bieden inzicht in patronen van verschillende typen vijandigheid, maar kunnen de gevolgen ervan voor het leven van de mensen die deze ontvangen niet onderscheiden. We erkennen dat sommige typen vijandigheid meer impact kunnen hebben dan andere, ook wanneer categorieën zoals racisme niet het sterkst naar voren komen in onze analyse. Daarnaast kan de combinatie van verschillende categorieën, zoals ras en gender, schadelijke effecten verder versterken.

Over het geheel genomen biedt deze methodologie een omvattend en aanpasbaar kader voor grensoverschrijdende dataverzameling. Door systematisch accounts te identificeren die actief zijn in het publieke domein en worden blootgesteld aan uiteenlopende online aanvallen, omvat de resulterende dataset een breed scala aan online vijandigheid. De lexicons die voor deze analyse zijn opgesteld, zijn open source beschikbaar (van Haperen et al. 2026). Met passende aandacht voor de lokale context is deze aanpak reproduceerbaar en kan zij moderatie ondersteunen van zowel openlijke als meer genuanceerde vormen van online vijandigheid.

Dankwoord

Dank aan degenen die hebben bijgedragen aan de procedure en analyse die in dit rapport worden gepresenteerd: 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 en Julie Hawke. Dank voor jullie bijdragen en waardevolle inzichten.

Juridische disclaimer

Het doel van dit rapport is het identificeren van trends en patronen. Het rapport vormt geen juridische kwalificatie van de opvattingen of uitingen van individuen. Waar mogelijk worden

persoonsgegevens vóór de analyse geaggregeerd en gepseudonimiseerd of geanonimiseerd, in overeenstemming met de AVG en relevante Europese regelgeving. De analyses in dit rapport zijn gebaseerd op geautomatiseerde detectie en interpretatie van openbare onlinegegevens en zijn onderhevig aan methodologische en contextuele beperkingen.

Dataverklaring en licentie

De lexicons die in het kader van dit project zijn opgesteld en die zijn gebruikt voor de huidige analyse, zijn open source beschikbaar in de repository van de Universiteit van Amsterdam.

Naam dataset: *Violence, Intolerance and Impoliteness thematic Corpus v1.0*

Ontwikkelaars dataset: 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.

Auteur dataverklaring: Sander van Haperen, Universiteit van Amsterdam

Licentie: Naamsvermelding, NietCommercieel, GelijkDelen 4.0 Internationaal (CC BY 4.0)

U bent vrij om:

- **Delen:** het materiaal te kopiëren en opnieuw te verspreiden, in elk medium en elke vorm.
- **Aanpassen:** het materiaal te remixen, te veranderen en erop voort te bouwen.

De licentiegever kan deze vrijheden niet intrekken, zolang u zich aan de licentievoorwaarden houdt.

Onder de volgende voorwaarden:

- **Naamsvermelding:** u moet passende naamsvermelding geven, een link naar de licentie opnemen en aangeven of wijzigingen zijn aangebracht. U mag dit op elke redelijke wijze doen, maar niet op een manier die de indruk wekt dat de licentiegever u of uw gebruik onderschrijft.
- **NietCommercieel:** u mag het materiaal niet voor commerciële doeleinden gebruiken.
- **GelijkDelen:** als u het materiaal remixt, verandert of erop voortbouwt, moet u uw bijdragen onder dezelfde licentie als het origineel verspreiden.

- **Geen aanvullende beperkingen:** u mag geen juridische voorwaarden of technologische maatregelen toepassen die anderen juridisch beperken in het doen van hetgeen de licentie toestaat.

Kennisgevingen: u hoeft niet aan de licentievoorwaarden te voldoen voor onderdelen van het materiaal die tot het publieke domein behoren of wanneer uw gebruik is toegestaan op grond van een toepasselijke uitzondering of beperking. Er worden geen garanties gegeven. De licentie verleent mogelijk niet alle toestemmingen die voor uw beoogde gebruik nodig zijn. Andere rechten, zoals portretrechten, privacyrechten of morele rechten, kunnen bijvoorbeeld beperken hoe u het materiaal gebruikt.

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

Bibliografie

- 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.