



Together Against Online Hate Speech

TAOH

Together Against Online Hate Speech

Deliverable 2.2 Hate Speech Pattern Report

About TAOH

Together Against Online Hate Speech (TAOHS) aims to support civil society organisations, media outlets and institutions to prevent and respond safely and effectively to online hate speech by analysing how it manifests and spreads across social media platforms and digital environments and delivering tools and training to our partners. We are an international project led by organisations in Germany, the Netherlands and Spain.

Consortium Members

- SPF - Sharing Perspectives Foundation - Netherlands | Coordinator
- UvA - University of Amsterdam - Netherlands
- Build Up - How to Build Up Germany - Germany
- IBH - ichbinhier e.V. - Germany
- FundiPau - Fundació par la Pau - Spain



**Co-funded by
the European Union**

Funded by the European Union. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.



This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/)

“Online Hostility: Comparative analysis of Instagram, X, and TikTok in Spain, Germany, and the Netherlands”

Version 2.2 September 19th, 2026

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

Abstract

This document reports findings from a comparative analysis of online hostility for the project “Together Against Online Hate Speech”, by the following partners: Sharing Perspectives Foundation, BuildUp, ichbinhier e.V., FundiPau, and the University of Amsterdam.

The goal of this report is to compare patterns of online hostility in Germany, the Netherlands and Spain on Instagram, X and TikTok, differentiating categories (e.g. racism, gender-based hostility) and behavioral patterns in terms of targeting and persistence. To that end, language models are developed to detect and classify different forms of online hostility, and made available open source.

Findings inform intervention strategies for civil society organizations, activists, law enforcement, and policy-makers engaging with online hostility. First, event-driven categories (Race, sexual intimidation, sexual orientation, religion, violent speech, crime, culture, and politics) call for rapid responses to events. Second, target concentration in some categories (politics, insults, crime and culture) call for a combination of monitoring and supporting likely targets of online hostility. Third, spreaders of hostility related to race, vulgar language, and insults are relatively persistent compared to other categories. Intervention strategies for these categories might focus on usual suspects generating online hostility.

Citation:

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.

Contents

Introduction	5
Definition	6
Methodology	7
Data collection	7
Language Model 1: Hate speech detection	9
Language Model 2: Online hostility classifier	10
Annotation procedure	10
Construction of lexicons	12
Model performance	13
Findings	15
Patterns by country	15
Patterns by platform	17
Patterns by targets and spreaders	19
Behavioral patterns	21
Patterns by category	22
Description of categories	22
Crime	24
Culture	26
Gender	28
Insults and personal attacks	30
Politics	32
Race	34
Religion	37
Sexual intimidation	39
Sexual orientation	41
Violent speech	43
Vulgar language	45
Conclusions	48
Recommended intervention strategies	48
Methodological limitations	50
Acknowledgements	53
Legal disclaimer	53
Data statement and license	53
Bibliography	55

Introduction

This report is part of the project “Together Against Online Hate Speech” (TAOHS). The overarching goal of the project is to increase the resilience of civil society organizations, media outlets and authorities to understand and tackle online hate speech on their platforms and contribute to a safer and plural online space. This report focuses on Germany, the Netherlands, and Spain.

The project combines language classification models with human-led dialogue interventions and policy influence to understand and engage online hostility comprehensively. While content moderation and reporting remain important for addressing content that is illegal or violates platform policies, such strategies struggle with content that is not illegal but still problematic, and forms of language such as intolerance and incivility warrant intervention. Our hope is to support individuals, civil society organizations, law enforcement, and policy-makers who engage with such language online.

The goal of this report is to inform intervention strategies by identifying patterns of online hostility. Methodologically, this study is designed to detect a broad range of online hostility, directed at a representative variety of targets, in different social media spaces (Instagram, X, and TikTok), and comparing these patterns across the different national contexts of Germany, the Netherlands, and Spain. To that end, we develop three language models for classification of different types of online hostility in German, Dutch and Spanish. The result of extensive abductive coding and performance testing, these language models are made available open source to be adapted for tailored analyses by partner organizations and other stakeholders.

We employ these models in this report for a cross-platform and cross-country comparison to identify patterns in terms of prevalence, targets, and spreaders of different types of online hostility. Generally, our findings suggest the problem is widespread, consisting predominantly of one-off hostile comments. Further differentiating behavior by category, we find striking differences between different types of online hostility.

From this follow three actionable insights. First, event-driven categories (race, sexual intimidation and violent speech) may be difficult to anticipate but may benefit from rapid responses to events because hostility is concentrated on a few targets and spread mostly by transient users. Second, diffuse categories (politics, insults, crime and culture) can to some

degree be anticipated for being more consistent over time, and although attention shifts continually across many targets, these categories include small and relatively persistent repeat offenders. Third, stable baseline categories (gender, religion, sexual orientation, vulgar language) are less driven by events and may benefit from continuous monitoring focused on recurring targets rather than reactive interventions or repeat offenders.

The project “Together Against Online Hate Speech” consists of the following partners: Sharing Perspectives Foundation, Build Up, IchBinHier, FundiPau, and the University of Amsterdam. The project is funded by the European Union (CERV-2024-CHAR-LITI-SPEECH). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Definition

At the outset of our project, we formulated the following definition of “hate speech”: “publicly incite to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin.’. We also include reference to gender, sexual orientation, and physical ability. At the outset, we expect such language includes categories such as: “Dehumanization, sexism, politics, racism, religion, violence, and threats” (adapted from EOHH, 2026). Practically, this definition provided us with sensitizing concepts and informed calibration among annotators.

In this report we also use the umbrella term “online hostility”, when referring to language that is not strictly illegal or hate speech relating to the identity of individuals, but may still be detrimental to the quality of online spaces. Beyond hate speech, other types of online hostility include abuse and incivility. Abuse refers to language intended to insult, provoke, or distress a specific individual (Akhtar & Morrison, 2019; Ward & McLoughlin, 2020). Political incivility refers to “a lack of respect for the social and cultural norms that govern personal interactions and the functioning of democratic systems” (Bentivegna & Rega, 2024), which covers norm-violating discourse ranging from relatively mild impoliteness to more overt forms of hostility, without necessarily requiring intent to harm or the targeting of specific social groups (Rossini, 2019).

Methodology

The goal of this study is to inform intervention strategies by identifying patterns of online hostility. Methodologically, this study is designed to detect a broad range of online hostility, directed at a representative variety of targets, in different social media spaces (Instagram, X, and TikTok), and comparing these patterns across the different national contexts of Germany, the Netherlands, and Spain.

To that end, we collected data from social media for a representative selection of most-likely targets. We analyzed this dataset using two language models: one model to determine whether a post likely contains online hostility and another model to determine what topic the post relates to. This combination allows us to identify and examine patterns across different types of online hostility. We focus on patterns in terms of prevalence, targets, and spreaders of online hostility. This chapter discusses this procedure in more detail.

Data collection

We collected data from Instagram, X, and TikTok. These platforms were selected as likely places where online hostility would occur, and for being the social media most widely used across the three countries. The selection also covers a variety of different user bases and affordances. Political opinions closely tied to news events are prevalent on X, Instagram emphasizes lifestyle, and TikTok features more engagement with younger demographics and open-ended algorithmic discovery.

Posts and comments on X (formerly Twitter) include short messages and reactions typically of 50 words or fewer, which may contain multimedia materials, external URL links, and mentions of other users. Posts and comments on Instagram are typically short captions and reactions of approximately 50 words or fewer which may contain external URL links and mentions of other users. Posts and comments on TikTok are typically short captions accompanying videos, and reactions of approximately 50 words or fewer which may contain external URL links and mentions of other users.

The corpus is the result of the following procedure. First, we identified accounts of individuals and organizations deemed most likely to receive various forms of online hate, intolerance and incivility, for each of the three countries: Spain, Germany, and the Netherlands. Case selection within the Spanish, German and Dutch settings followed a purposive sampling strategy with the aim of compiling a diverse set of social media accounts that are most likely to receive online hate speech.

Our objective for data collection was to construct a dataset that captures a wide spectrum of online hostility, to support the development of a language model sensitive to both explicit and more subtle forms of language. For each of the three countries, selection began with the identification of prominent individuals and organizations operating in domains characterized by high public visibility and engagement. Accounts were selected from domains including politics, media, entertainment and culture, public service, academia, activism, sports, and civil society organizations. Within each domain internal diversity was ensured by incorporating subgroups such as elected officials, political parties, journalists, athletes, and influencers. We relied on familiarity with the settings and supplementary sources such as news reports to identify individuals known to receive high volumes of online hate, ensuring that our selection was grounded in observable patterns of harassment. In particular, attention was paid to figures involved in recent or ongoing political, cultural, or media controversies, as well as those associated with controversial issues such as migration, geopolitical conflicts or national debates which are known to increase exposure to online hostility. This sampling strategy aligned with our definition of hate speech while capturing its diverse manifestations across social groups. We identified 377 accounts for Spain, 217 accounts for Germany, and 360 accounts for the Netherlands.

For each of these seed accounts we collected data from social media profiles on X, Instagram, and TikTok. Comments and reactions posted on these accounts, or posts mentioning the account, in the period between 1-1-2024 to 31-12-2025 are included in the selection criteria. For X this includes mentions and retweets, for Instagram mentions and reposts, for TikTok mentions and replies. Accounts without an official or identifiable social media presence were excluded. Only publicly available posts and comments are included. We do not include content which was deleted between posting and data collection (in the period September-December

2025), data that was set to “private” or had been deleted between the time of posting and our study are not included in our data collection.

This procedure led to three datasets:

- German corpus: 1.452.268 posts (598.865 from Instagram, 730.971 from TikTok, 122.432 from X).
- Spanish corpus: 2.445.553 posts (457.036 from Instagram, 1.953.819 from TikTok, 34.698 from X).
- Dutch corpus: 5.611.651 posts (1.018.428 from Instagram, 680.296 from TikTok, 3.912.927 from X).

Language Model 1: Hate speech detection

To determine the probability of a post in our dataset containing online hostility, we first used existing language classifiers. After preliminary testing and consideration of how each model was trained, we selected three separate opensource models (Aluru et al 2002; Pérez et al 2022; IMSyPP, 2022).

These models are trained on monolingual social media data using the BERTje tokenizer, each with the goal of detecting the presence of hate speech (Basile et al, 2019; Pérez et al 2021; Cinelli et al, 2021; Kralj Novak, 2022; Baider et al, 2023). We employ these models on our datasets with the following results (Table 1).

Table 1. Performance Model 1 (Goal: Hate speech detection)

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

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

Note: German model applied label “HATE” at threshold 0.94, Spanish model applied label “hateful” at threshold 0.052, Dutch model applied “Label_2” at threshold 0.48. Note these thresholds are added here for purposes of replicability, the levels are not comparable between the models.

It is important to note here the weak performance of these models. These models were selected as performing best from among alternatives during preliminary analysis at the time of designing this study. In practice, an AUC score of 0.564 performs only slightly better than random chance, which is discussed in more detail in the section on methodological limitations. Here, we emphasize that we advise against employing these models for automated decision-making in hate-speech detection. For our purposes of our analyses, we complement each with human classification of online hostility.

Language Model 2: Online hostility classifier

While model 1 is designed to detect whether a post likely contains problematic language, it does not help to understand what topic it relates to. To identify different types of hostility we develop a rule-based language model.

Annotation procedure

To identify different forms of hate speech, we adopted an abductive approach to coding (Timmermans & Tavory, 2012). First, three expert annotators were involved in consecutive iterations of coding random samples from these datasets to identify categories of online hostility. Our shared definition of online hostility provided sensitizing concepts and informed discussions about semantics and our objects of analysis. To identify recurring linguistic patterns within user interactions we interpreted posts across accounts within similar domains or sharing identity markers, documenting common keywords, phrases, and expressions of online hostility. This process also included the identification of more subtle forms of harmful language such as coded expressions, slurs, slang, and dogwhistles.

Second, after each round of coding samples of 500 posts, we convened to compare assigned labels, evaluate our definition of hate speech, engage with prior literature, and adjust our code book or add to it. Based on our starting definition we initially expected to find the following categories: race, color, religion, descent or national or ethnic origin, gender, sexual orientation, and physical ability. We then sought to complement these categories inductively,

which led to inclusion of the categories “Untruth”, “Anti-institutional distrust”, “Conspiracies”, “Polarization”, “Ukraine”, “Russia”, “Autocracy”, “Middle-East”, “Climate”, “Crime”, “Sports”, “Insults”, “Vulgarity”, “Irony and Sarcasm”, and “Sexual intimidation”. At this stage, while still inductively coding any content, we still maintained the categories “Informative”, “Critical”, and “Positivity”, which were dropped later to focus exclusively on online hostility.

Third, the resulting list of terms was subsequently reviewed and refined by individuals familiar with each country’s online context to ensure cultural and linguistic relevance. Aside from identifying different categories of online hostility and topic of the post, we initially also scored a measure of intensity from 1 to 10, which was ultimately abandoned after critical reflection on our differing interpretations of intensity. In line with our abductive approach, we iterated between coding and literature review, which led us to distinguish incivility and impoliteness from hate speech (Nithyanand et al., 2017; Theocharis et al., 2020). We think of incivility defined as “a lack of respect for the social and cultural norms that govern personal interactions and the functioning of democratic systems” (Bentivegna & Rega, 2024). Such disrespect may surface as interpersonal impoliteness or as discourses that delegitimize democratic systems and their actors, including forms such as stigmatization, role-based attacks, demonization, name-calling, threats, incitement to violence, vulgarity, allegations of illegality, and offensive language (Bentivegna & Rega, 2024). After four iterations of coding, we were satisfied with the code book and the identified categories adequately covered the data encountered in the fourth sample. At that point, our coding and discussions did not generate new categories or labels, which we took as a saturation point.

Fourth, we applied the code book to annotate out-of-sample data. At this point we considered the code book exhaustive enough for our goal of identifying online hostility. Twelve expert annotators (recruited through consortium partners) were instructed in applying the code book on new random samples, combined with eight calibration sessions to discuss disagreements over labels. Disagreements were resolved when consensus was reached following discussion among annotators.

Ultimately, the following intercoder agreement was achieved:

- German: 7.852 posts from our sample are coded by two annotators (0,54% of 1.452.268 in German corpus), and 5.335 are coded by three annotators (0,37% of corpus), with acceptable intercoder agreement (Krippendorff's alpha 0,73).
- Spanish: 6.534 posts from our sample are coded by two annotators (0,27% of 2.445.553 in Spanish corpus), and 4.719 are coded by three annotators (0,19% of corpus), with acceptable intercoder agreement (Krippendorff's alpha 0,67).
- Dutch: 12.933 posts from our sample are coded by two annotators (0,24% of 5.611.651 in Dutch corpus), and 9.091 are coded by three annotators (0,17% of corpus), with good intercoder agreement (Krippendorff's alpha 0,83).

Construction of lexicons

We undertook two rounds of axial coding with the goal of relating categories back to our understanding of hate speech as it developed throughout the coding process (Green & Thorogood, 2004). To do so, we considered relationships between categories. Ultimately, these reflections led us to adopt the term “online hostility”, which is broader than “hate speech” and thus better represents the variety of language we encountered in our dataset. We report this umbrella term throughout the analysis unless explicitly referring to identity-based hate directed at individuals, as per our original definition of “hate speech”.

As a result of axial coding, several categories were merged. Of others, we determined they were not sufficiently salient or relevant in light of the project's goals. The final core categories represented in these lexicons are: “Gender” [GEN], “Culture” [CUL], “Race” [RAC], “Sexual orientation” [SOR], “Religion” [REL], “Ableism” [ABL], “Violent language” [VIO], “Insults and personal attacks” [INS], “Vulgar language” [VUL], “Irony and sarcasm” [IRO], “Sexual intimidation” [INT], “Politics” [POL], “Untruth” [UNT], “Crime” [CRI].

Subsequently, three rounds of selective coding served to identify keywords (n-gram) from the data where at least three human annotators agreed on the relevant category, and cross-referenced with existing hate speech lexicons: HADES (Tulkens et al, 2016) and HurtLex v1.2 (Bassignana et al, 2018).

The resulting list of keywords per category are the lexicons at the basis of our classification models, which are available open source (Van Haperen *et al*, 2026).

Model performance

We employ the classification models on our datasets with the following results (Table 2).

Table 2. Performance Model 2 (Goal: Classification)

	German					Spanish					Dutch				
	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

Note. Each lexicon was constructed using training data only, F1 reported here was evaluated on independent test set excluded from training set. To account for large differences in number of observations and address class imbalances, stricter thresholds were adopted to improve recall for the larger categories Politics and Insults in all languages; Crime in Spanish and German; Culture, Race, Religion, Sexual Orientation, and Vulgar language in German .

We report F1 as a measure of how well the model correctly assigns posts to categories. We note that prevalence for some of our categories is very low (e.g. 1 in every 1000 posts in our German dataset is labelled as “Ableism”). Performance of the models is strongly associated with prevalence: rarer categories provide less data to train a model making reliable classification more difficult. In general, the lower the F1 the more careful we need to examine categorization. We deem F1 scores above 0.70 as strong, and F1 scores below 0.50 unacceptable which led us to drop the categories (Ableism, Irony, Untruth) in further analyses.

Across the three languages, categories performing well are: “sexual intimidation” and “gender”, with relatively clear and unambiguous keywords. The low prevalence of “ableism”, “race”, “sexual orientation” and “violent speech” may be explained by likely being the most strictly moderated categories. The categories “politics” and “insults and personal attacks” are outliers with high prevalence as the result of coding decisions. During axial coding, many sublabels were combined under “politics” (ideology, policies, politicians, et cetera), making it the largest and most diffuse category. To address this, a strict threshold was applied post-hoc to this category.

In light of our abductive approach to coding, throughout the process we critically evaluated whether our categories still reflected what we found in the data. We performed extensive manual inspection of the data and assigned labels for all of the categories and across the platforms and countries, through manual inspection and random spot checks. In particular, we manually inspected false positives and false negatives, and discussed edge cases in calibration sessions. In most cases manual inspection led us to changes, it concerned the exclusion of specific words carrying ambiguous meanings. An example is the word “Let”, which in Dutch signifies a person from Latvia and was therefore part of the “Culture” category, but also more commonly occurs as “Let op” [watch out]. Similarly, the word “Frans” [French] occurs more commonly in Dutch as a first name [e.g. Frans Timmermans] than as a reference to French culture. As a result of similar ambiguity, poor model performance and deliberation in light of the project’s aims, we decided to drop the categories “Untruth” and “Irony and Sarcasm”. The category “Ableism” came up during inductive coding but was too rare in subsequent annotation to produce reliable results for model performance and was thus also dropped. These categories remain in the lexicon and methodology but are omitted from the analysis below in the interest of validity.

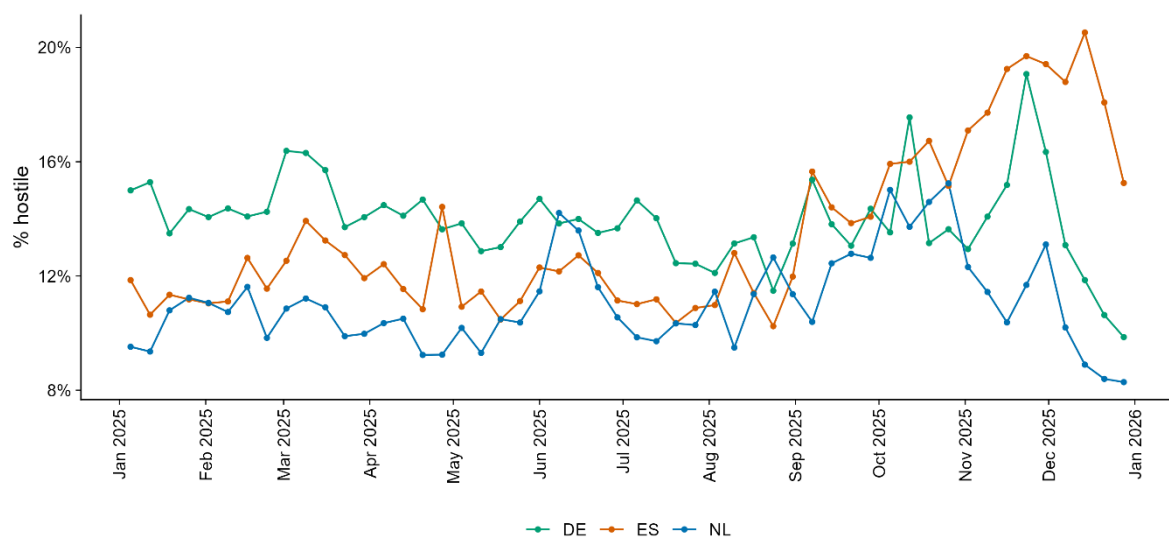
Findings

This section reports the following. First, descriptions for each of the fourteen categories provide qualitative insight into the different types of online hostility identified and analyzed. Second, general descriptives provide an overview of overall trends: proportions over time for each of the countries, each of the platforms, and each of the categories of online hostility. Third, analysis of targeting, persistence, and clustering provides insight into different types of behavior of spreaders. Together, these analyses provide actionable insight into what types of interventions are best suited to counter specific types of online hostility in context.

Patterns by country

The following section reports overall patterns of language identified comparing the three countries (Figure 1).

Figure 1. Volume of online hostility over time for Spain, Germany, Netherlands



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

In our dataset, we identified 13% of all posts in our dataset as containing online hostility (although the typical day has roughly 8.8%). We emphasize here that we sampled on accounts most likely to receive hostility, which also informed our annotation. Practically, the percentage

reported here is non-representative for these countries or platforms generally and likely an overestimation of online hostility. The gap between mean and median, as well as the spikes on specific days and the variability in the data (SD=1543, min=1, max=21.396), suggest that online hostility is driven by bursts of attention which are likely related to real-world events, platform specific dynamics, or coordinated activity. Much of the hateful content in our dataset is concentrated on specific days, combined with a baseline of consistent activity on each platform. These bursts of hateful content are clustered around particular dates (rather than random isolated days), suggesting such activity is related to specific news events sustaining attention.

Notable differences exist for the different categories of online hostility between the countries (Table 3).

Table 3. Volume of online hostility by country (all 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

Note: Descriptives of percentages by day (number of posts in a category as part of total number of daily posts identified as hateful).

Several insights follow from this comparison. Hostility related to politics, insults, and crime are the most common category in each language. In German, vulgar language and sexual orientation to a lesser degree are relatively more common than in Dutch or Spanish. In Dutch, hostility related to culture is relatively more common than it is in the other languages. Hostility

related to gender, race, religion, sexual intimidation, violent speech and vulgar language are relatively less salient in our dataset for each of the languages.

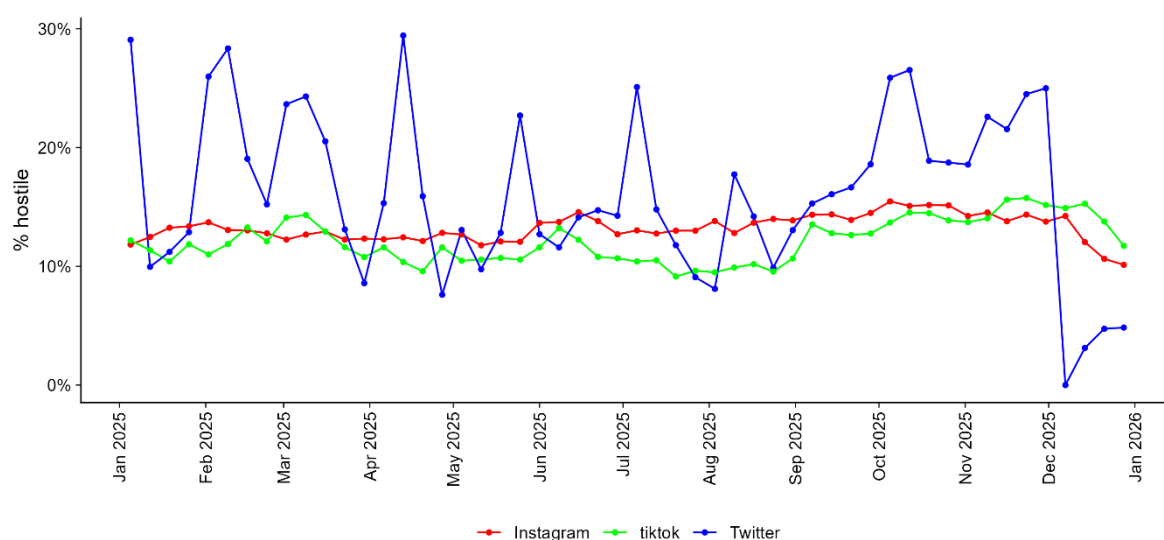
Medians are notably different between the countries, suggesting that the share of hostility for most categories more consistently occurs over time in the Netherlands, while being more strongly driven by peaks of activity in Germany and Spain. This is further supported by relatively large standard deviations for categories such as Politics in Spain. Categories such as violent speech and sexual intimidation occur more consistently over time, whereas politics and insults are more bursty, most notably so in Spain. Together, this suggests peaks of activity that may be closely tied to (news) events, a dynamic most pronounced in Spain.

We explore these differences between countries by each category in more detail in the following section of this report.

Patterns by platform

The following section reports overall patterns of language identified comparing the three platforms (Figure 2).

Figure 2. Volume of online hostility over time by platform



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

Notable differences appear between platforms. In our dataset, X stands out for having the highest share of online hostility and for its volatile pattern over time. Instagram has higher shares of online hostility than TikTok, but both appear relatively consistent over time without a clear increasing or decreasing trend. These trends are driven by bursts of activity on specific days in combination with a baseline of activity on each platform.

Notable differences exist for the different categories of online hostility between the countries (Table 4).

Table 4. Proportions by platform (all countries)

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 intim.	7863	626	0.4	0	3.4	7047	0.8	0.4	5.7	190	0.1	0	0.3
Sexual orient.	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

Note: Descriptives of percentages (number of posts in a category as part of total number of daily posts identified as hateful). Note that categories are not mutually exclusive (e.g. a single post may be assigned to both "politics" and "insults.")

In our dataset, X has the highest shares of online hostility across categories, particularly politics and insults, whereas TikTok consistently showed the lowest share and Instagram generally occupied a moderate position, with the notable exception of gender-related content being most salient on Instagram.

Compared with Instagram and TikTok, X has the highest mean percentages for 11 of the 13 categories, including politics (15.0%), insults (10.3%), crime (4.7%), culture (4.1%), religion (2.3%), race (2.1%), and vulgar language (1.9%). Instagram generally occupies an intermediate position. While most categories are less pronounced compared to X, Instagram exceeds X for

gender-related hostility (2.1% vs. 1.3%) making gender the only category where Instagram clearly has the highest average share. TikTok consistently has the lowest share across categories. Mean percentages are lowest for most forms of hateful content, particularly religion (0.5%), sexual intimidation (0.1%), violent speech (0.1%), and vulgar language (0.3%). Politics remains the most common category on TikTok but at less than half the proportion on X.

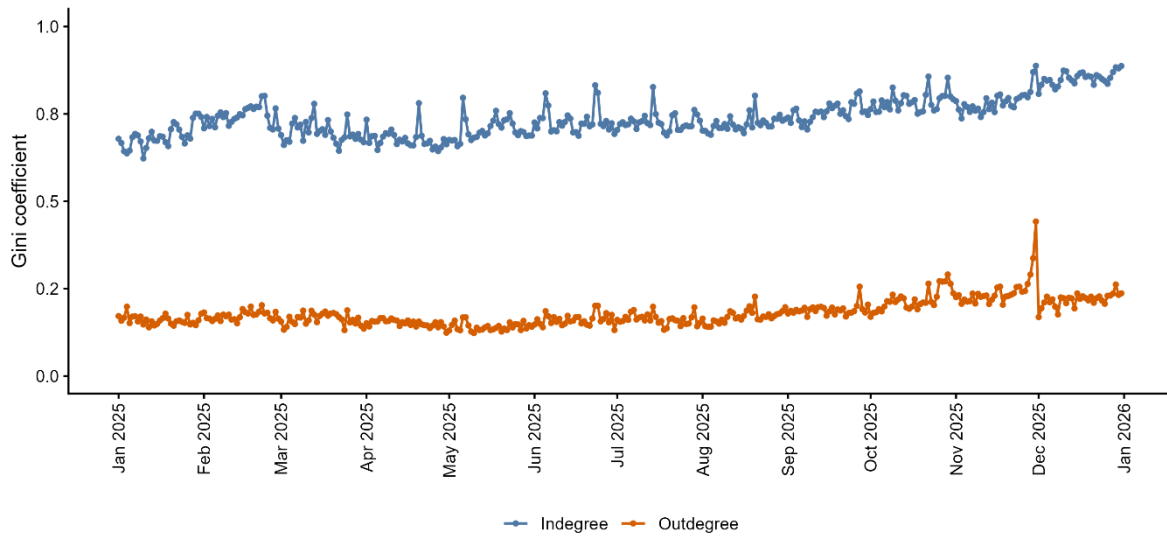
Medians suggest bursts of activity generate a large share of the hostile content across most platforms. The medians on X for politics (16.4%), insults (11.3%), culture (4.1%), and crime (4.0%), suggest these topics have a consistent baseline over time. Variation is also greatest for politics and insults. Standard deviations are highest for these categories across all platforms, especially on X (SD = 12.3 for politics, SD = 9.4 for insults), which indicates substantial volatility over time.

Overall, these findings suggest clear differences between platforms in the composition of online hostility. In our dataset, X is characterized by both higher shares and greater volatility across categories, particularly for insults and hostility related to politics, while Instagram shows relatively lower shares (but a greater share of gender-related hostility) and more consistency over time, and TikTok exhibits the lowest shares for most categories of hostile content that we detected.

Patterns by targets and spreaders

The following section reports overall patterns among targets and spreaders of online hostility. We first focus on how concentrated the attention to specific targets is (Figure 3).

Figure 3. Concentration of online hostility to targets (indegree) and from spreaders (outdegree)



This figure shows how attention is distributed among targets and spreaders of online hostility. A higher gini-coefficient indicates more inequality among targets or spreaders: a relatively smaller group of accounts receiving or sending a disproportionately large share of posts. The key takeaway here is the occurrence of periods in which hostile attention becomes more concentrated, with a small number of targets receiving a disproportionately large share of all hostile content in our dataset. This indicates specific events generate stronger focus on specific targets.

Figure 3 also shows how concentrated online hostility is from spreaders. A higher gini-coefficient here indicates a smaller share of spreaders accounts for relatively much of the hostile content we identified in our dataset. The key takeaway is the overall low degree of concentration, indicating that attacks originate from many different accounts, rather than a small number of accounts generating most of the hostile content in our dataset.

A surprising finding here is that concentration among both targets and spreaders (whether a small number of users accounts for a disproportionate share of received or sent posts) increases at similar times. We also note similarities in the gradual increase of concentration between spreaders and targets, although this trend is less pronounced among spreaders. We suspect that when an individual becomes prominent in media coverage, attacks become more concentrated on them, while at the same time a relatively smaller share of accounts generate

more of the online hostility. We take this as further indication that hostile content is tied to (news) events.

Behavioral patterns

To further differentiate spreaders of online hostility, we examine their targeting behavior in terms of focused targeting and persistence. By comparing spreaders of online hostility on “targeting” (how many targets does someone send hateful posts to, with three or fewer targets indicating a measure of focus) and “persistence” (how many days someone directs hateful posts at the same target, with three days or more during a year indicating a measure of persistence), we discern four distinct types of behavior (Table 5).

Table 5. Behavioral Clusters among spreaders: focus and persistence

Cluster	Type	Average Target focus	Average Persistence	n
1	Generalists	11.8	1.05	863
2	Fixators	1	9.33	3
3	Repeat offenders	2.98	2.21	174
4	Opportunists	1.61	1.00	14669

A first cluster is a mass of ‘opportunists’, low-intensity users representing 94% of accounts spreading problematic content in our dataset. This comes from accounts with minimal targeting and no persistence: just one-off hateful posts directed at one or two targets. A second cluster is a medium sized group (863 accounts) of generalists, aggressive accounts targeting a many different targets, but with very little persistence. These are one-off tangents directed at many different individuals simultaneously. A third cluster of ‘repeater offenders’ is small (174 accounts) with moderate targeting and moderate persistence. These represent individuals who engage with a variety of topics over longer periods and do so consistently with problematic language. A final cluster consists of ‘fixators’: extreme outliers who are fixated on one targets for more than a week. Other than expected, such accounts are very rare in our dataset . One likely explanation might be effective platform moderation, with the most

problematic accounts being reported and banned, which would lead such accounts to be excluded from our data collection.

Patterns by category

Description of categories

The following categories resulted from several iterations of inductive coding. Note that these categories are not exhaustive. Even if these categories covered samples drawn from our dataset for annotation, assume that data collected for other countries, platforms, or time periods will require extension of this list. Furthermore, a degree of interpretation remains, and these categories may contain language identified by our annotators as hateful, while another reader may not think it is hateful. Also note that these categories are not mutually exclusive: a post may be categorized both as Vulgar language and as an Insult.

The overall volumes of online hostility can be further differentiated by category as follows (Table 6).

Table 6. Descriptives by category (all platforms, all countries)

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

Note: Descriptives of percentages (number of posts in a category as part of total number of daily posts identified as hateful).

First, several categories are primarily event-driven. These categories have high volatility “Race” (0.9%, SD = 0.968), “Sexual intimidation” (0.3%, SD = 0.291) and “Violent speech” (0.3%, 0.307). Such hostility may be relatively difficult to anticipate and may benefit from rapid responses because hostility tends to be strongly concentrated on a few targets and spread mostly by transient users.

Second, the categories “Politics”, “Insults”, “Crime”, and “Culture” pair a consistent baseline with volatility. These categories account for the largest share of online hostility in our dataset. Politics and Insults/personal attacks account for respectively 8.3% and 5.1% of all online hostility daily. Crime (2.7%, SD = 1.100) has notable high variability, suggesting this is primarily event-driven. For Culture (1.7%, SD = 0.970), variability is relatively low, suggesting such content consistently appears on a daily basis. With median and mode similar to the mean, both Politics and Insults have stronger daily baseline consistency, while the standard deviations indicate relatively high variability in peaks. This suggests hateful content in these categories are driven by both a steady baseline of activity and episodic peaks of activity.

Third, stable baseline categories are less driven by events and appear more consistently over time. These are: “Gender” (1.4%, SD = 0.823), “Religion” (0.7%, SD = 0.509), “Sexual orientation” (0.8%, SD = 0.803), and “Vulgar language” (0.8%, SD = 0.540). Here, standard deviations indicate variability is relatively low, suggesting such content consistently appears on a daily basis rather than being strongly event-driven. These categories form a structural baseline of hateful content daily. These categories may benefit from continuous monitoring focused on recurring targets rather than reactive interventions or repeat offenders.

Two important caveats need to be emphasized here. First, as discussed in the methodology, we collected data from social media that was still available publicly. By then, posts had already been subjected to platform moderation before ending up in our dataset. It is plausible that categories in the first tier are most heavily moderated, thus making them relatively less prevalent and more volatile in our analysis. Second, we underline here that prevalence does not equate impact. Posts in these categories contain the most virulent forms of online hostility in our dataset and it may be reasonably assumed that these categories have disproportionate impact, even if less prevalent.

Overall, politics and personal insults account for the largest share of posts flagged as problematic by our models. This suggests online hostility across the countries and platforms is not random but has structural features: some are highly event-driven, others have a more consistent baseline. This may help to decide where to invest resources for interventions against online hostility.

We next turn to examination of targets and spreader behavior for each of the categories.

Crime

This category relates to language about criminal behavior, illegal activities, and the police.

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

Crime-related content is moderately common in our dataset, with a gradual increase on Instagram and more variable levels with a slight overall decline after early peaks on TikTok. The category is volatile on X, with no sustained trend. Across countries, this is a relatively prominent type of content in Germany, moderately prominent in Spain, and one of the less prominent types of content in the Netherlands, compared to other categories.

In terms of targets, we find moderate concentration combined with a relatively stable set of targets (mean Gini = 0.23, HHI = 0.25). Concentration is consistent over time with little systematic increase or decrease in trend. There is a decrease in HHI parallel to an increase in activity towards the end of the year, indicating fluctuating attention across time, but targets of this category are relatively consistently concentrated over time. This suggests a pattern of shifting attention among a changing set of targets rather than sustained focus on specific accounts.

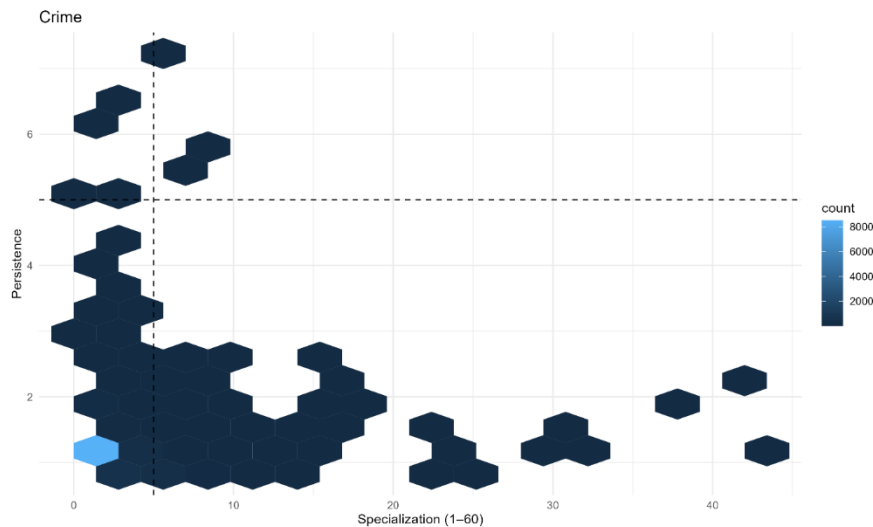
Spreaders of crime-related content are, compared to other categories in our dataset, equally concentrated (95% focused targeting–low persistence), but it has one of the highest shares of more diversified targeting (4.9% diffuse targeting–low persistence) among the categories and

a negligible 0.1% focused targeting–high persistence. This suggests online hostility in crime related content is rarely sustained, combined with relatively broad targeting.

In terms of interventions, these findings suggests that online hostility related to Crime is relatively common, occurs consistently over time, is focused on a relatively large but stable set of targets, coming from a transient and unfocused set of spreaders. This makes it one of the more diffuse categories that may be more challenging to intervene in, with fewer indications of events that may be anticipated.

Figure 4. Crime-related hostility comparison of share, targets, spreaders





Culture

This category relates to language concerning specific cultures and cultural stereotypes.

- Spanish example: “Menos clase que un gitano recién llegado de Rumanía”.
- German example: “ Sofort auch raus mit den Migranten !!!! Keiner will euch !!!”.
- Dutch example: “smerige halve Marokkaan”.

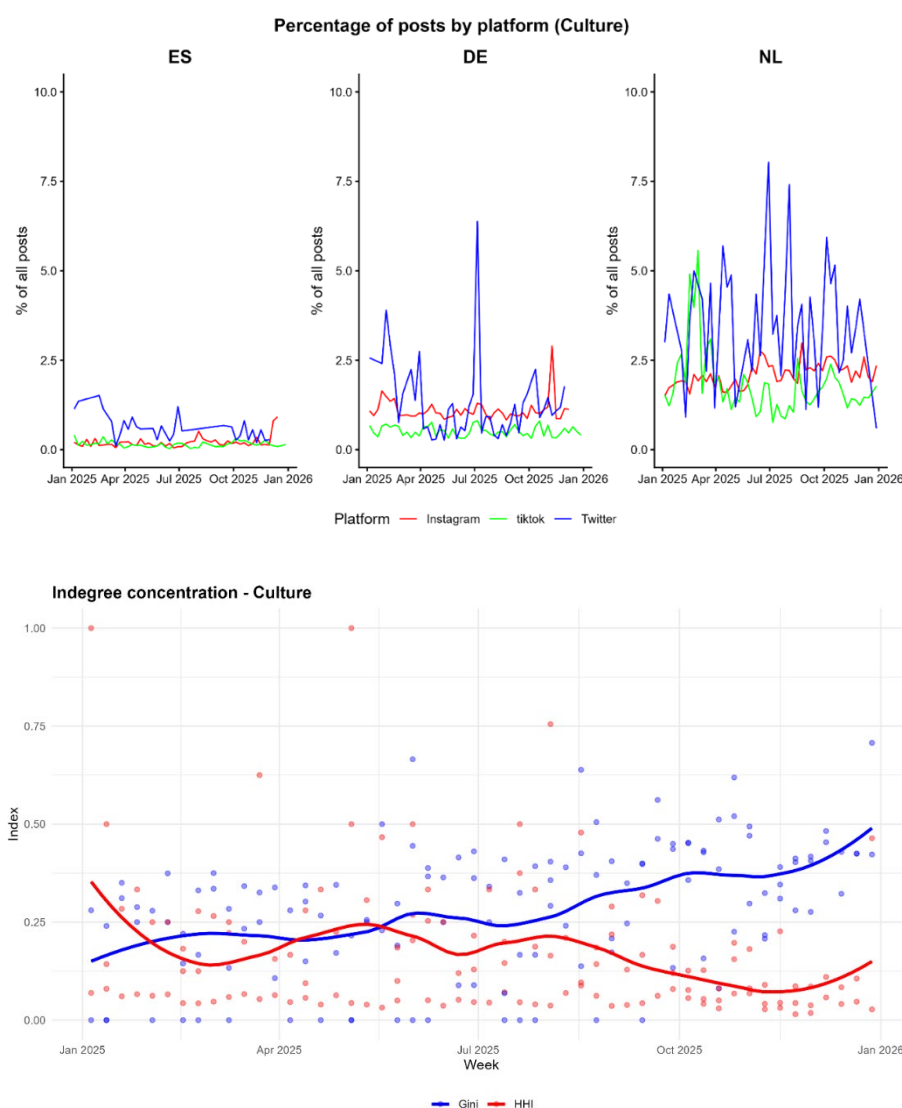
Culture-related content is very common in our Dutch language dataset, but strikingly less so in our Spanish and German datasets. It occurs consistently on all three platforms in the Netherlands and is relatively prominent on both Instagram and TikTok compared to other categories. On X, online hostility related to culture occurs often and in peaks of activity.

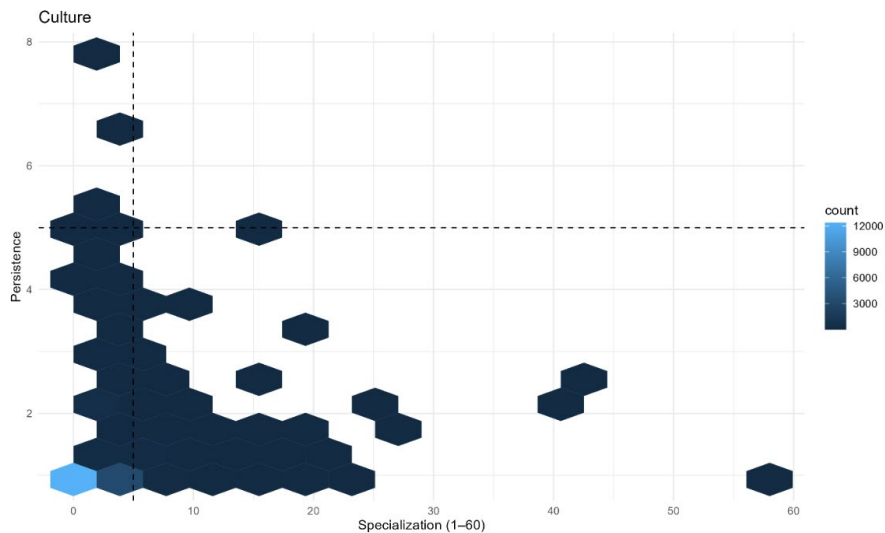
In terms of targets, online hostility related to Culture is one of the most unequal categories combined with relatively low dominance among targets (mean Gini = 0.28, HHI = 0.17), which indicates a concentrated set of targets that changes relatively often. Both patterns are consistent over time. This suggests such hostility is distributed across relatively unequal but without persistent focal targets.

Spreaders of culture-related online hostility are among the most focused and transient compared to other categories in our dataset (97.4% focused targeting–low persistence)., with only 2.6% of users diversifying their targets with no evidence of persistence over time. This suggests such spreaders are highly fragmented and engage with just a single message.

In terms of interventions, these findings suggests that online hostility related to Culture is relatively common in Dutch, occurring consistently over time on all platforms, but much less so in Spanish and German. Across the countries, attention in this category unequal but with a larger variance among targets, coming from a very transient but relatively focused set of spreaders. This makes it one of the more diffuse categories that may be more challenging to intervene in, with fewer indications of events that may be anticipated.

Figure 5. Culture-related hostility comparison of share, targets, spreaders





Gender

This category relates to intolerant language targeting specific aspects of identity concerning gender and gender stereotypes. It includes gender norms and misogyny.

- Spanish example: “sin maquillaje,es una bruja”.
- German example: “Ich kann diese dumme Hexe nicht mehr ertragen.”.
- Dutch example: “Alle vrouwen zijn dom”.

Gender-related online hostility is uncommon in our dataset across each of the platforms and countries, strikingly rare in Dutch and notably more common in Spanish. Compared to other categories, fluctuation on X is also notably tempered. This relatively stable trend is consistent over time.

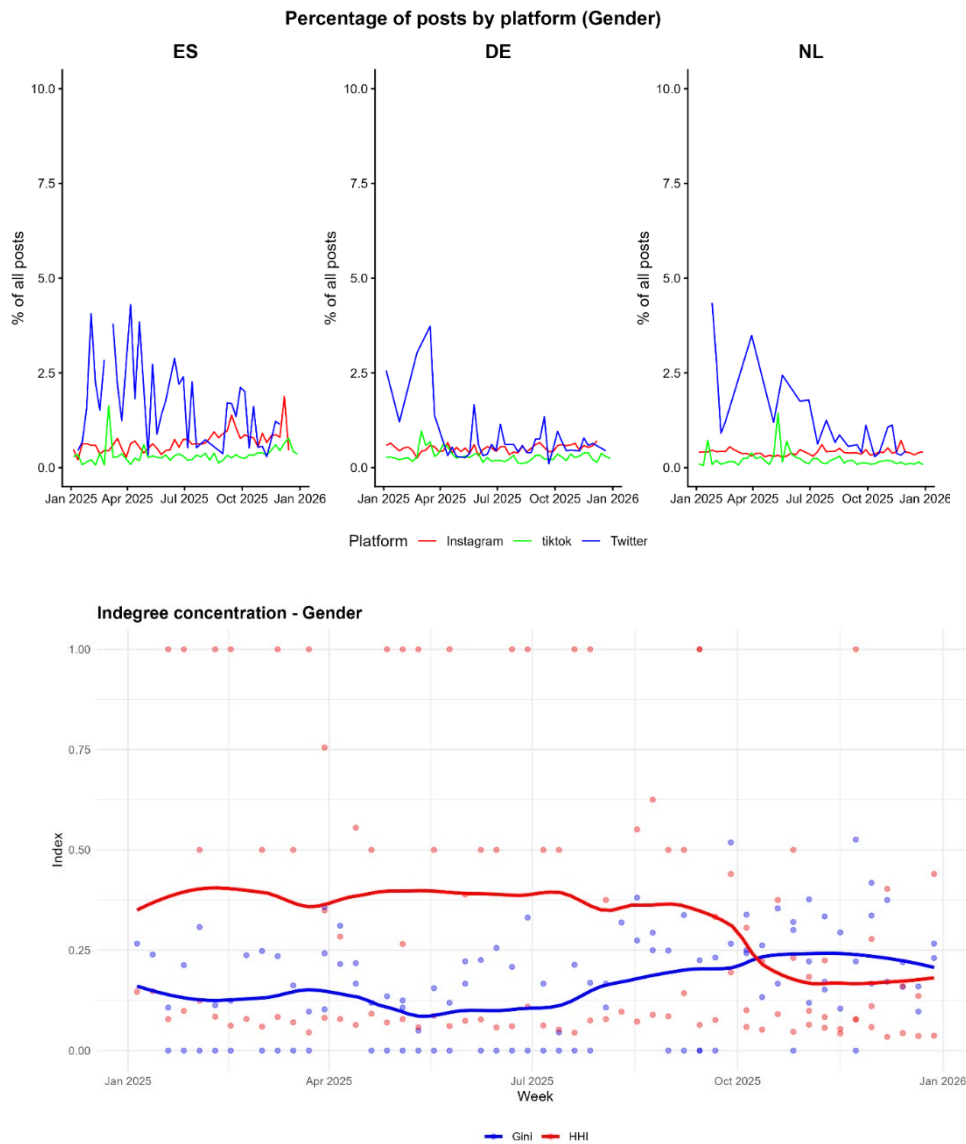
In terms of targets, online hostility related to gender is relatively diffuse combined with relatively high dominance (mean Gini = 0.16, HHI = 0.32). This suggests that while inequality is limited, attention is concentrated on a few key targets. This trend is consistent over time.

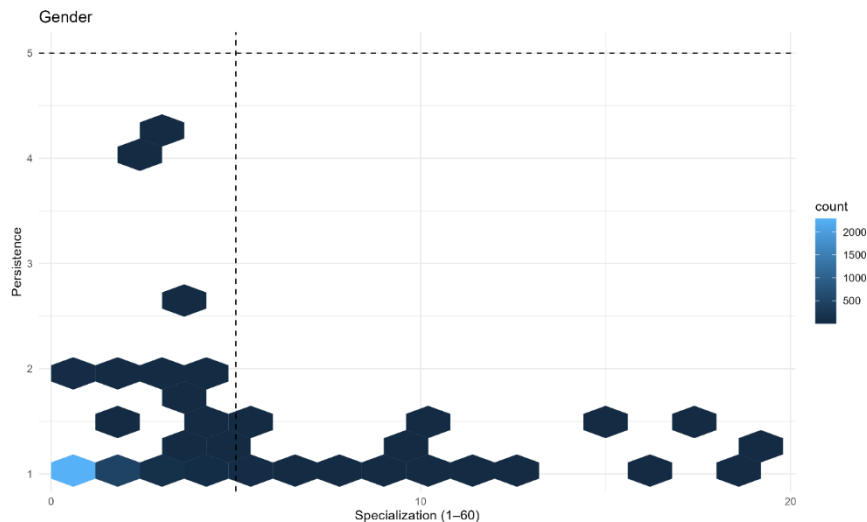
In terms of spreaders of online hostility, in our dataset Gender is the most concentrated category overall (98.3% focused targeting–low persistence), with minimal diversification (1.7%) and no persistent users. This suggests such hostility is very focused on specific targets, one-off engagement with almost no persistence.

In terms of interventions, Gender is a category with a consistent baseline of activity, warranting attention in Spanish more than in the other languages. Across the countries,

attention in this category is relatively diffuse, but with a relatively stable core of targets, coming from a transient and focused set of spreaders. Interventions in this category are less informed by event-driven attention but may be directed at salient accounts that consistently draw such hostility.

Figure 6. Gender-related hostility comparison of share, targets, spreaders





Insults and personal attacks

This category relates to a broad array of language meant to insult or personally attack an individual, without targeting explicit group identities. It is a broad category covering not only online hostility, but incivility and impoliteness more generally, including coarse, offensive and derogatory language.

- Spanish example: “Eres amiga de este no hay mas preguntas miserable”.
- German example: “Du bist so erbärmlich.”.
- Dutch example: “Wat ben jij een ontzettende eikel”.

Online hostility in the form of insults or personal attacks is the second most common category in our dataset across each of the platforms and countries, and particularly prominent in Dutch. It is notably present on Instagram compared to other categories and somewhat less so on TikTok. It appears consistently volatile on X, Instagram has consistently high levels with modest variation over time, while TikTok showed similar stability with occasional peaks.

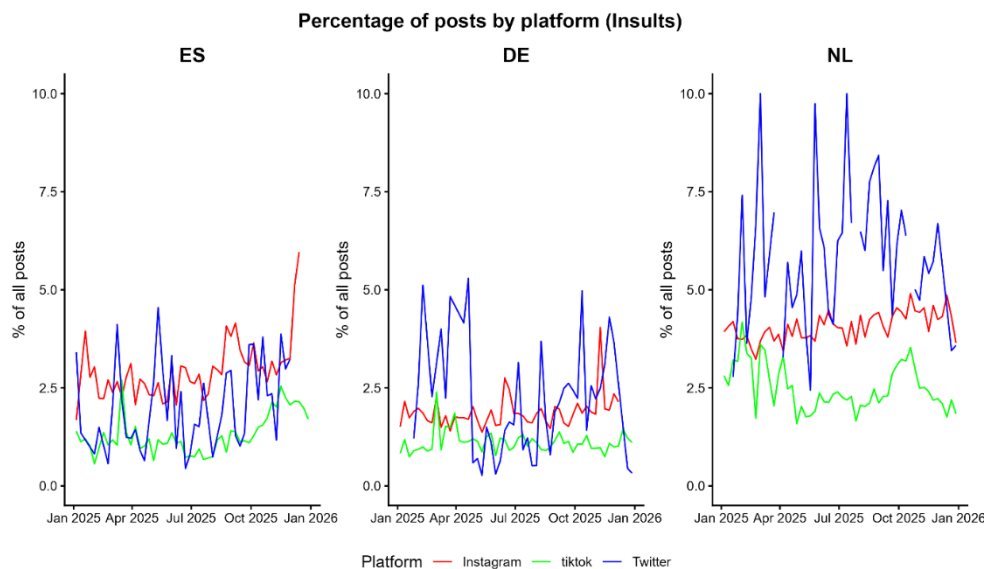
In terms of targets, this is the second-most unequal category in our dataset, combined with relatively low dominance of targets (mean Gini = 0.32, HHI = 0.15). This suggests insults are very concentrated, without consistent focal targets. Although in our dataset we see some increasing divergence between more concentration and less dominance, the trend appears

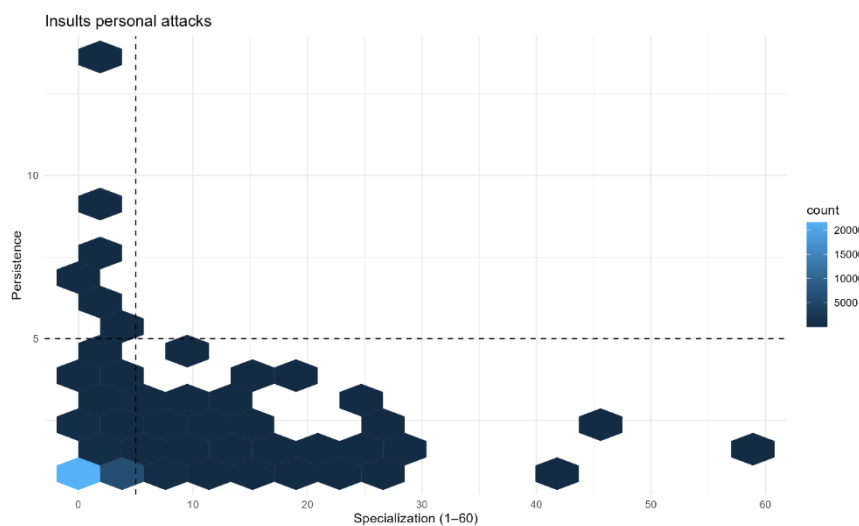
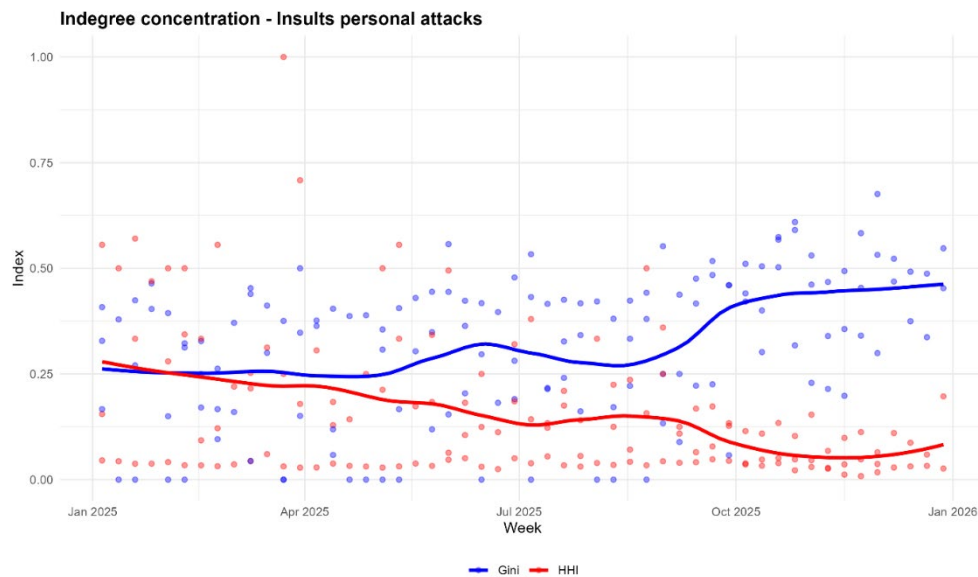
relatively stable over time. Overall, these patterns suggest diffuse but uneven targeting behavior.

In terms of spreaders of insults, we find more variation in our dataset compared to other categories. While still strongly concentrated with little persistence (*97.2% focused targeting–low persistence*), there is a larger share of accounts insulting a broader set of targets (*2.7% diffuse targeting*), and a comparably larger share of persistent attackers (*0.1%*).

In terms of interventions, attention in this category is highly unequal with large variance among targets for each of the languages. While it is difficult to anticipate what type of events may drive such content, and attention is largely diffused across many different targets, these patterns also suggest there are core groups of spreaders that persistently insult specific targets.

Figure 7. Insults and personal attacks, comparison of share, targets, spreaders





Politics

This category relates to a broad array of language concerning politics, politicians, political parties, specific policies, wars and conflict, ideology, and left wing/right wing polarization.

- Spanish example: “Basura woke”.
- German example: “Selten so gelacht Hochverschuldung haben wir durch das Russen Drecksvolk die Freude von dem AFD Gesindel”.
- Dutch example: “Geert Wilders is een engnek”.

Online hostility related to politics is the most common category in our dataset across each of the platforms and countries. It is relatively volatile on all platforms and consistently so with

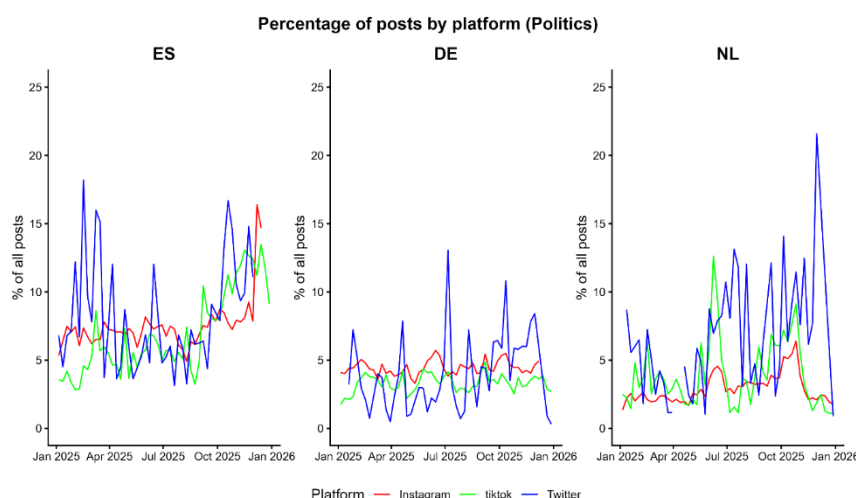
high levels of variation over time, which are most pronounced on X. This suggests such content may be driven by events and news cycles.

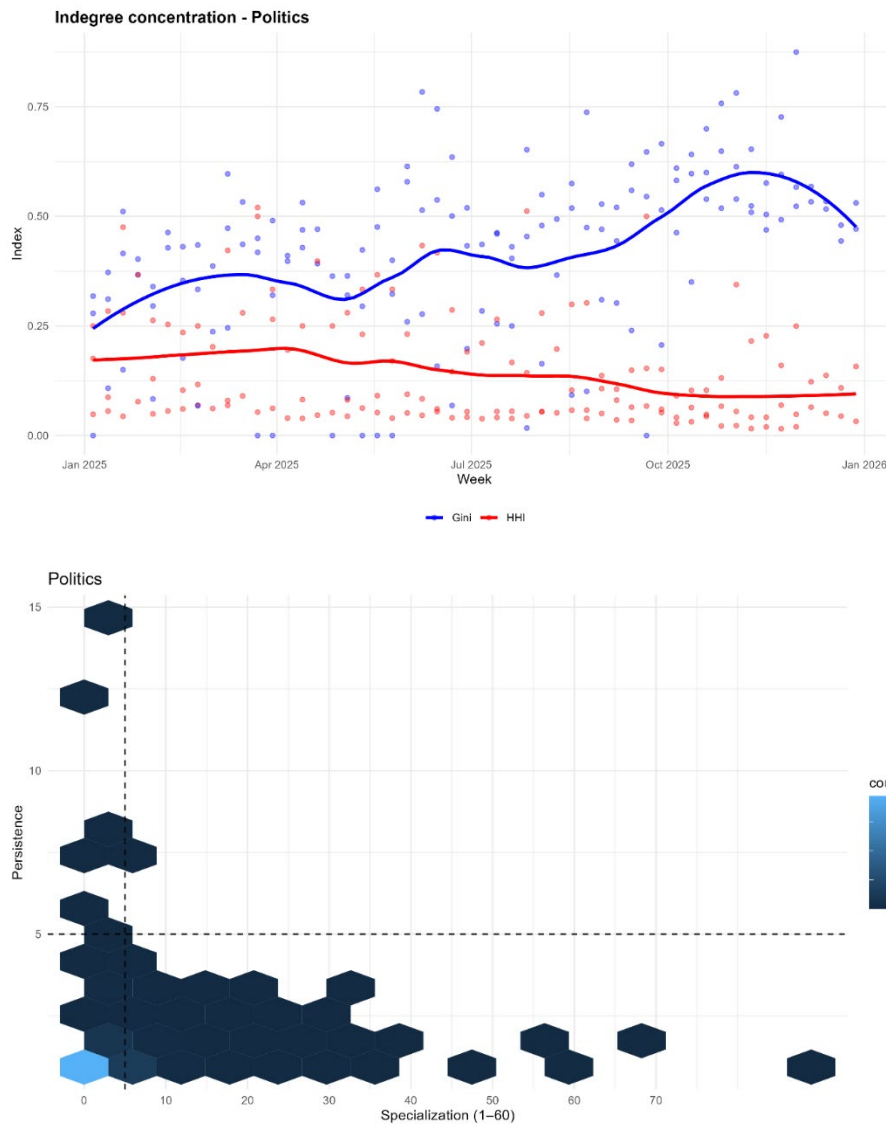
In terms of targets, this is the most unequal category in our dataset, combined with relatively low dominance of targets (mean Gini = 0.42, HHI = 0.14). This suggests political hostility is very concentrated but with little consolidation of focal targets. Although in our dataset we see some increasing divergence between more concentration and less dominance, the trend appears relatively stable over time. Overall, these patterns suggest diffuse targeting with no consistent focus.

In terms of spreaders of politics-related hostility, we find more variation in our dataset compared to most other categories. While still strongly concentrated with little persistence (94.1% focused targeting–low persistence), this category has the largest share of accounts insulting a broader set of targets (5.8% diffuse targeting), and a comparably larger share of persistent attackers (0.1%). This suggests that political hostility is more likely directed at multiple targets and from some groups of users persistently.

In terms of interventions, attention in this category is highly unequal with large variance among targets for each of the languages. While difficult to anticipate what type of events may drive such content and attention is diffuse across many different targets, these patterns are likely driven by news cycles and suggest there are core groups of spreaders that persistently insult specific targets.

Figure 8. Politics-related hostility comparison of share, targets, spreaders





Race

This category relates to intolerant language targeting specific aspects of identity concerning race and ethnicity as well as racial stereotypes.

- Spanish example: “Qué les pasa a los hombres n*? Por qué son la mayoría así?”⁴.
- German example: “Was hat der N* mit deutscher Jugend zu tun?”
- Dutch example: “Terug naar je eigen LAND VIEZE AAP”.

Race-related online hostility does not appear often in our dataset. In our dataset we find little remaining evidence in Spanish and German or on Dutch Instagram and TikTok, although it

⁴ Harmful slurs have been abbreviated to avoid unnecessarily reproducing discriminatory language

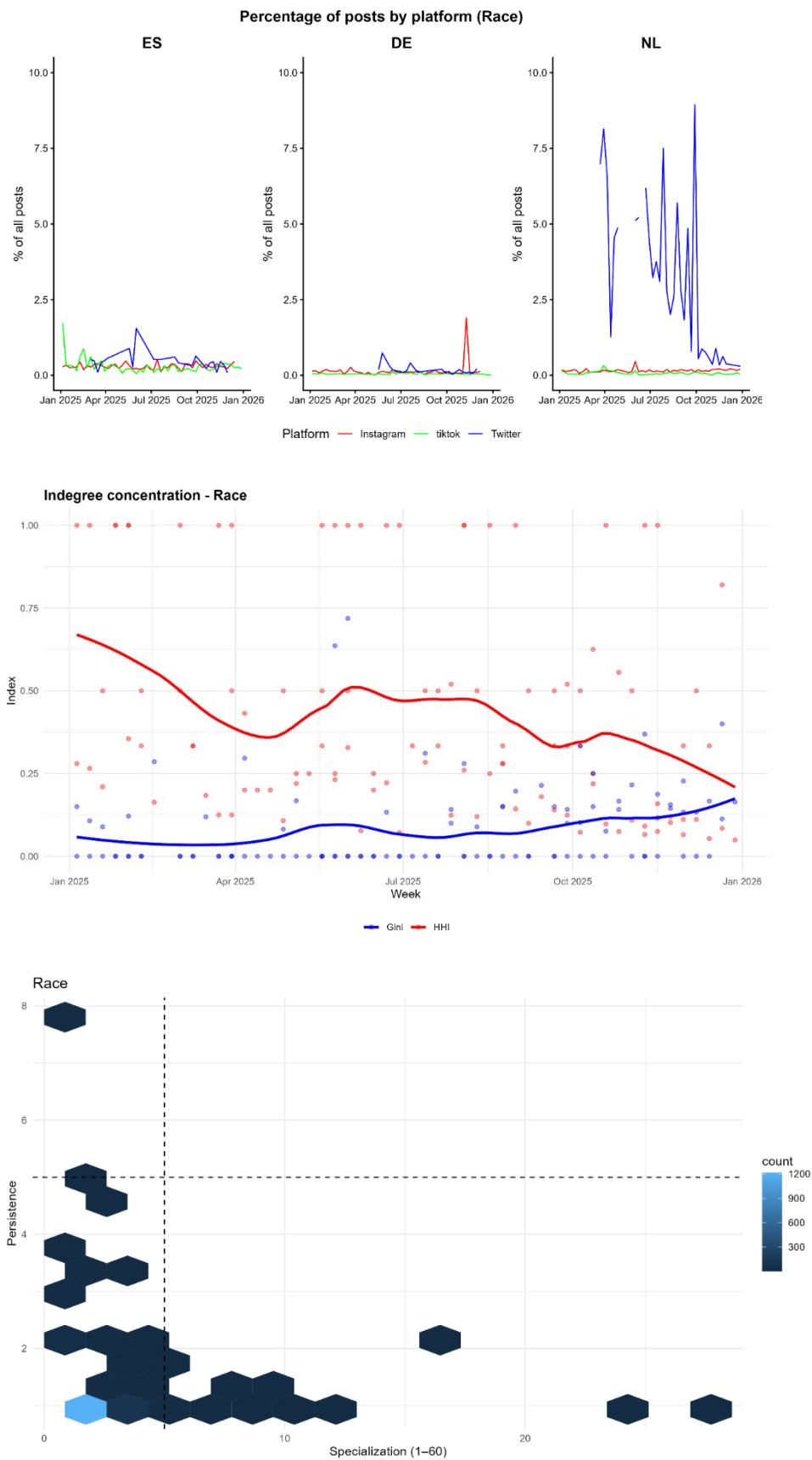
does appear in Dutch with substantial volatility on X. We expected to find differences across platforms with X allegedly discontinuing strict moderation in the period under observation. We note acceptable model performance which, combined with strong intercoder agreement, suggests this language can be accurately detected. Taken together, we acknowledge the rarity of evidence in this category in our dataset may be a result of methodology rather than evidence of absence.

In terms of attention and targets, racism in our dataset has a very low mean Gini (0.08) but high HHI (0.42), suggesting that while inequality is low, attention is strongly concentrated on a small set of targets. The trend is consistent over time, but we do note volatility is relatively high as a result of the limited number of observations. Together with the peaks of activity this suggests episodic targeting with bursts of focused attention on a small number of targets. This may be related to news events making specific individuals salient targets.

In terms of spreaders, we find a highly concentrated pattern (96.9% focused targeting–low persistence). Compared to other categories in our dataset, Racism has a relatively large share of users with broader targeting (2.9% diffuse targeting–low persistence). We find a small share of persistent users (0.1%).

In terms of interventions, these findings suggests that Racism may be already strictly moderated. From what remains in our dataset, racism appears largely episodic, with little sustained or broad targeting behavior, but with a core of persistent attackers, most notably so on X.

Figure 9. Race-related hostility comparison of share, targets, spreaders



Religion

This category relates to intolerant language targeting specific aspects of identity concerning religion as well as religious stereotypes.

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

Religion-related online hostility is not very common in our dataset for German and Dutch across all platforms but is strikingly rare in Spanish on all platforms in our dataset. Where religion-related hostility does appear in our dataset, it does relatively consistently over time on Instagram and TikTok, while appearing highly volatile on X. These trends are consistent over time, suggesting a combination of both baseline religion-related hate combined with peaks of activity driven by events.

In terms of targets, compared to other categories we find moderate concentration combined with moderate dominance of targets (mean Gini = 0.18, HHI = 0.29), this trend appears relatively stable over time with a slight increase towards more concentration. These patterns suggest a relatively diffuse set of targets.

In terms of spreaders of religion-related hostility, this is a relatively homogeneous category in our dataset. Most accounts focus on a single target and with little persistence (97.1% focused targeting–low persistence), with a relatively large share of diffuse targeting (2.8%). We find very little persistence. This suggests highly episodic but relatively diffuse focusing behavior.

In terms of interventions, attention in this category is relatively diffuse but with some consistency among targets. We find conspicuous differences across the languages which are likely related to methodological constraints. This type of content will be difficult to intervene in, with little indication for what type of events may drive such content, no usual suspects and very little persistence among spreaders over time. Depending on country context, there may be consistent targets.

Figure 10. Religion-related hostility comparison of share, targets, spreaders



Sexual intimidation

This category relates to language that is sexually suggestive or demeaning and directed at an individual, without reference to a group or gender role.

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

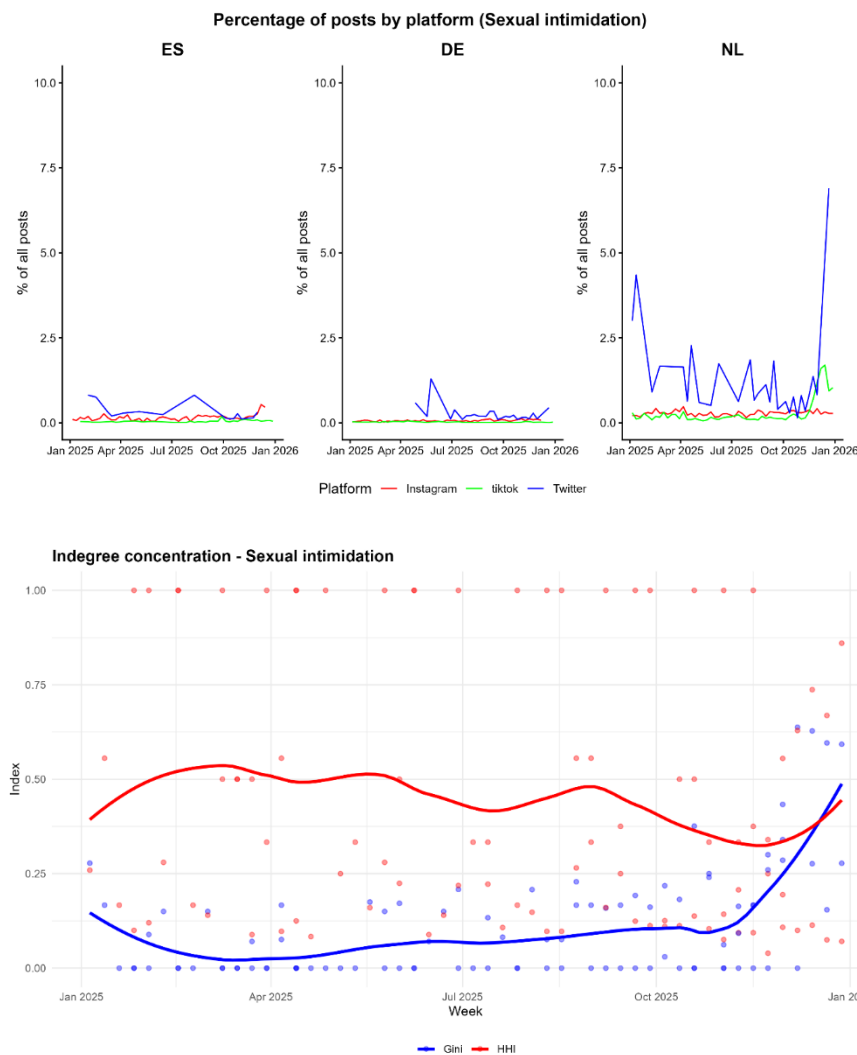
Sexual intimidation does not appear often in our dataset. In our dataset we find little remaining evidence in Spanish and German or on Dutch Instagram and TikTok, although it does appear in Dutch with substantial volatility on X. We expected to find differences across platforms with X allegedly discontinuing strict moderation in the period under observation. We also note acceptable model performance which combined with strong intercoder agreement suggests this language can be accurately detected. Taken together, and similarly to racism, we acknowledge the rarity of evidence in this category in our dataset may be a result of methodology rather than evidence of absence. The patterns that we do find are stable over time: volatility on X with a consistent baseline of sexual intimidation on Instagram and TikTok, in each of the countries.

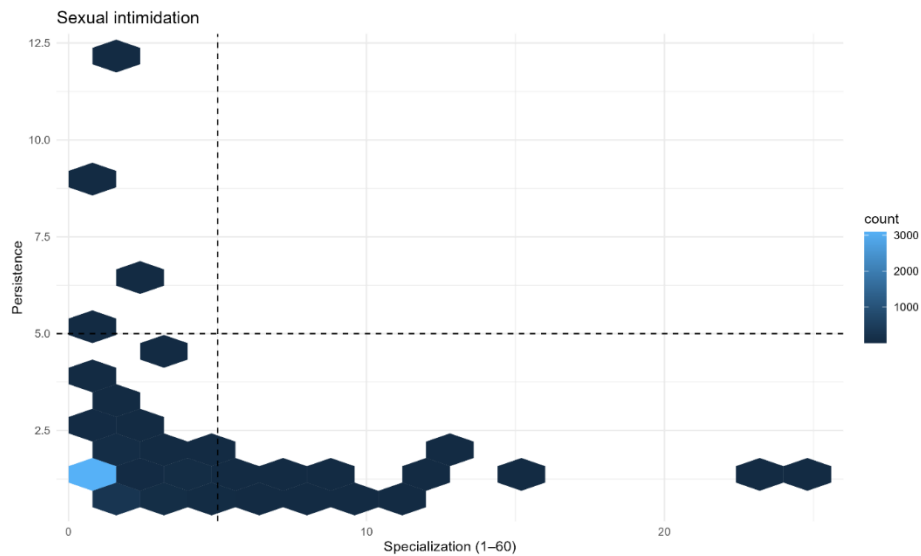
In terms of attention and targets, sexual intimidation in our dataset has a very low mean Gini (0.11) but high HHI (0.44), suggesting that while inequality is generally low attention is strongly concentrated among a small set of targets. The trend is consistent over time, but we do note volatility is relatively high as a result of the limited number of observations. Together with the peaks of activity on X (as well as Instagram to a lesser degree) this suggests episodic targeting with bursts of focused attention on a small number of targets. Similar to racism, this may be related to (news) events making specific individuals salient targets.

In terms of spreaders, we find a highly concentrated pattern (97.7% focused targeting–low persistence). Similar to Racism and in contrast to other categories in our dataset, Sexual intimidation has a relatively large share of users with broader targeting (2.3% diffuse targeting–low persistence). We find no meaningful persistence in our dataset.

In terms of interventions, from what evidence of sexual intimidation remains in our dataset it appears largely episodic, with some consistent targeting behavior, but limited persistence and no core group of perpetrators to address. These findings suggest that sexual intimidation may be already strictly moderated (possibly including reporting and subsequent banning of accounts) on Instagram and TikTok, and less so on X.

Figure 11. Sexual intimidation-related hostility comparison of share, targets, spreaders





Sexual orientation

This category relates to intolerant language targeting specific aspects of identity concerning sexual orientation as well as related stereotypes.

- Spanish example: “ la persona gay en muuuy promiscua y guarra en el aspecto sexual. Tu eres gay? Pues disfrutalo”.
- German example: “Dreckige Schwuchtel”.
- Dutch example: “Donderop man vieze flikker”.

Online hostility related to sexual orientation is not a very salient category. In our dataset it appears most prominently in German on each platform, on X in Dutch, and rarely in Spanish. It occurs volatily on X platforms and relatively stable on TikTok and Instagram. This suggests a baseline of such content combined with events and news cycles driving bursts of attention.

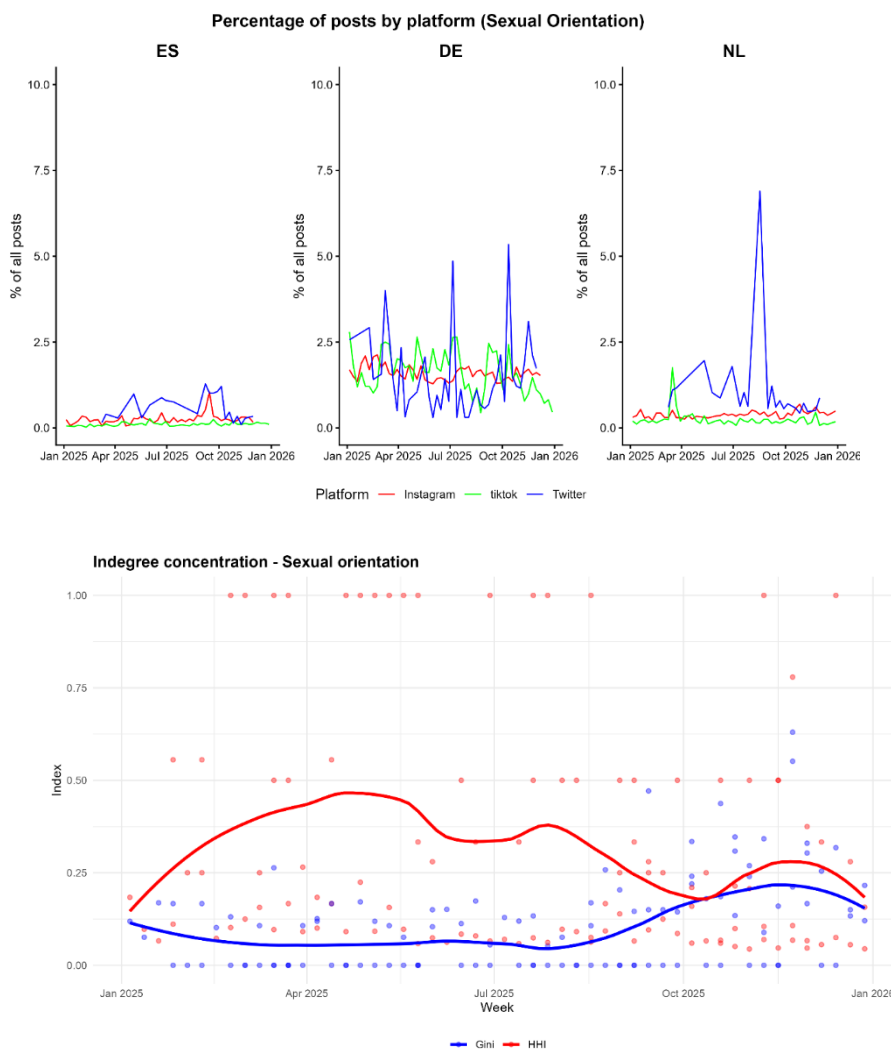
In terms of targets, this is one of the least concentrated categories in our dataset, combined with moderate dominance of targets (mean Gini = 0.11, HHI = 0.32). This suggests hostility relating to sexual orientation is somewhat diffuse but with a degree of consolidation of focal targets. Although in our dataset we see some convergence towards more concentration and less dominance, the trend appears relatively stable over time. Overall, these patterns suggest diffuse concentration generally but moderately consistent targets.

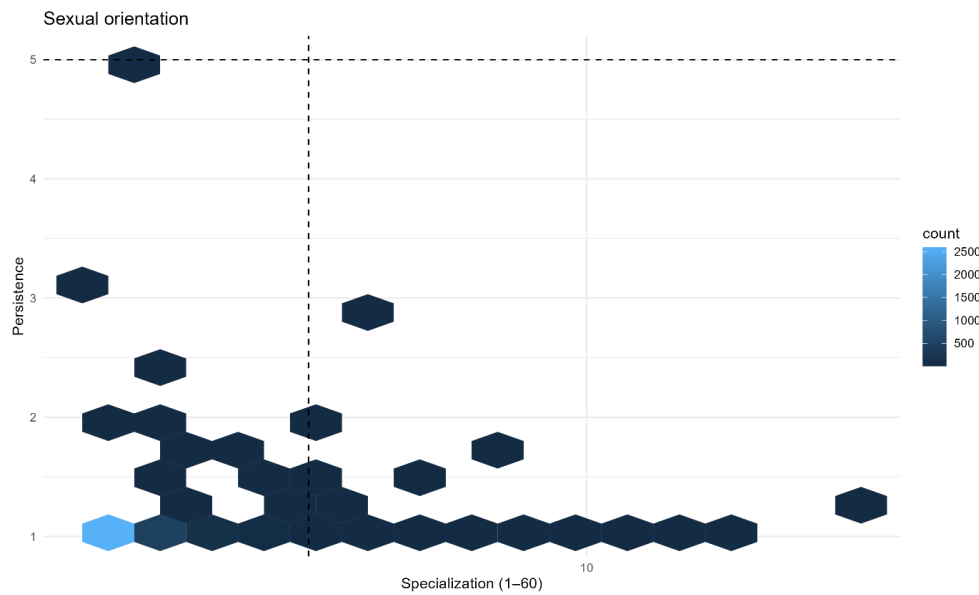
In terms of spreaders of hostility related to sexual orientation, we find a relatively large share of transient attackers (98.4% focused targeting–low persistence), with negligible diffuse

targeting or persistence. This suggests that hostility based on sexual orientation is dominated by one-off attacks.

In terms of interventions, these findings suggest that hostility based on sexual orientation may be already effectively moderated. From what remains in our dataset, sexual orientation appears predominantly transient, with very little sustained or broad targeting behavior.

Figure 12. Sexual orientation-related hostility comparison of share, targets, spreaders





Violent speech

This category relates to language inciting violence or glorifying violence, including directed (death) threats.

- Spanish example: “Esa soberbia con la que aplaude y niega es merecedora de una guillotina”.
- German example: “Der gehört an die Wand”.
- Dutch example: “Tering jong, zal jou doodtrappen.”.

Violent speech does not appear often in our dataset. In our dataset we find little remaining evidence in Spanish and German or on Dutch Instagram and TikTok, although it does appear in Dutch with some volatility on X. We expected to find differences across platforms with X allegedly discontinuing strict moderation in the period under observation. We note acceptable model performance which, combined with strong intercoder agreement suggests this language can be accurately detected. Taken together, and similar to Race and Sexual intimidation, we acknowledge the rarity of evidence in this category in our dataset may be a result of methodology rather than evidence of absence.

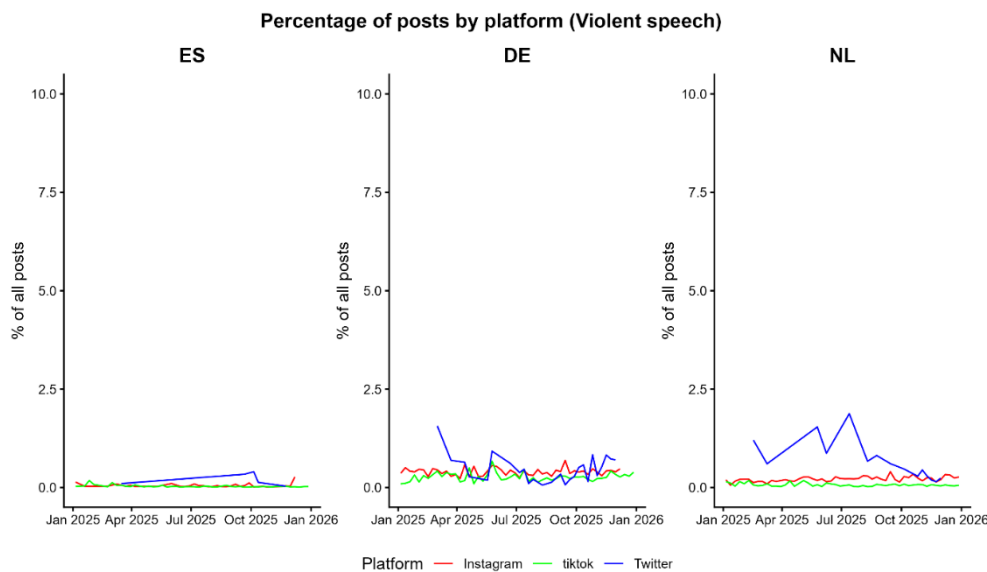
In terms of attention and targets, violent speech in our dataset has a very low mean Gini (0.08) but high HHI (0.40), suggesting that while inequality is low, attention is strongly concentrated on a small set of targets. The trend is consistent over time, but we do note the limited number of observations. Together with only moderate peaks of activity this suggests a baseline of

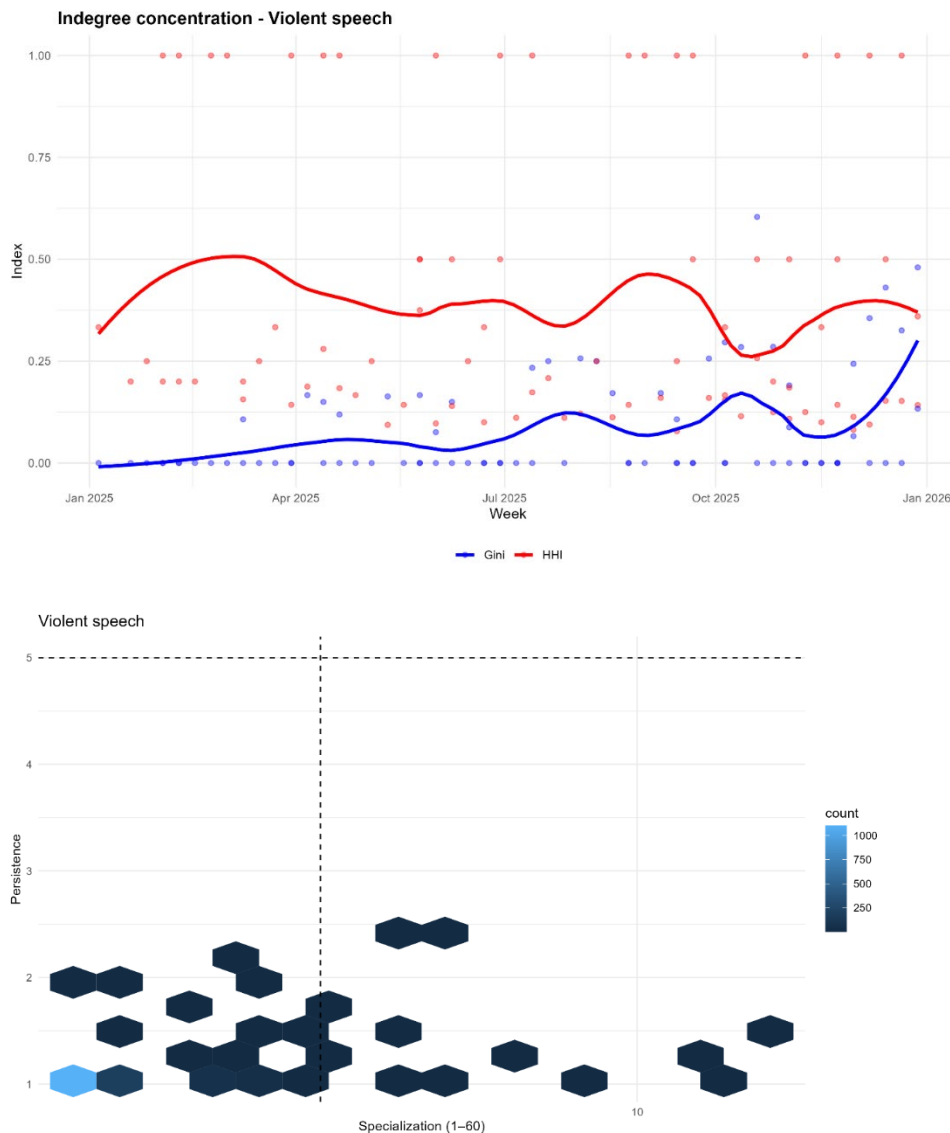
violent speech, instead of episodic bursts of attention, thus seemingly less reliant on news cycles.

In terms of spreaders, we find a highly concentrated pattern (97.7% focused targeting–low persistence). Compared to other categories in our dataset, Violent speech has a relatively large share of users with broader targeting (2.3% diffuse targeting–low persistence). There is very little evidence in our dataset of persistent attackers focusing on the same targets over time.

In terms of interventions, from what evidence of violent language remains in our dataset these findings suggests that violent speech may be already effectively moderated. From what remains in our dataset, violent speech appears predominantly transient, with very little sustained or broad targeting behavior.

Figure 13. Violent speech comparison of share, targets, spreaders





Vulgar language

This category relate to language that may be interpreted as impolite and may be offensive but is not directed at a specific group or an individual. This includes cussing, cursing, and swearing.

- Spanish example: “Coño señora entonces que hace usted y todo su equipo de gobierno”.
- German example: “drecks Politik!”.
- Dutch example: “God verdomme vuile moordenaars.”.

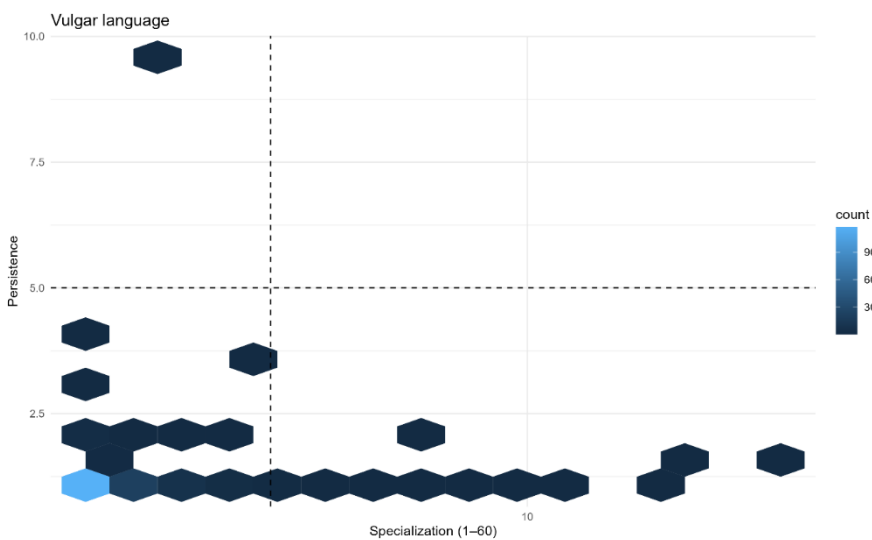
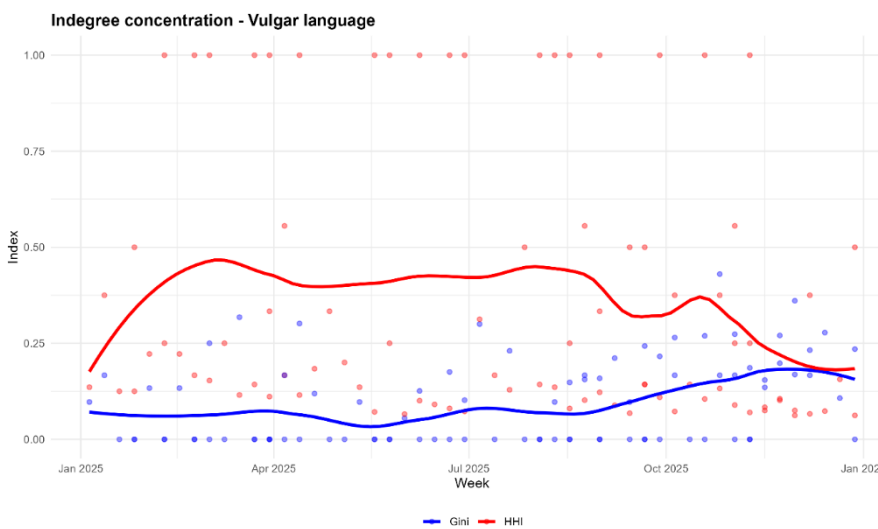
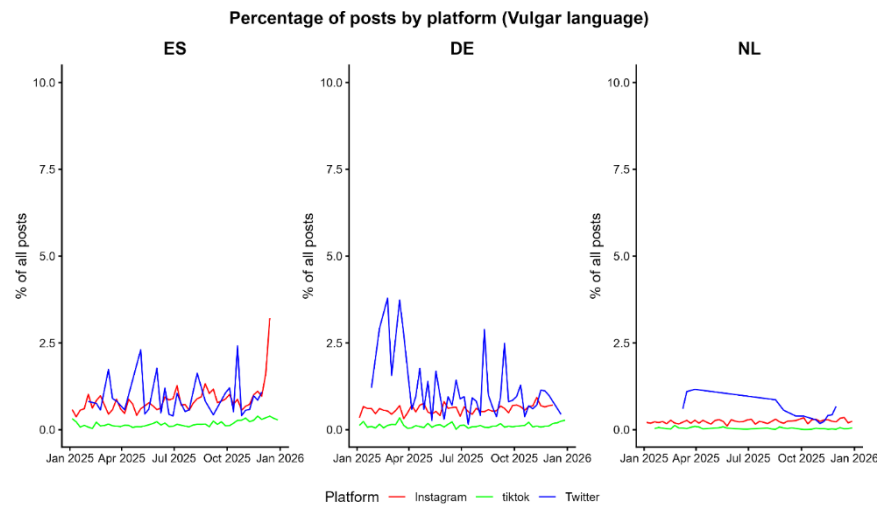
Vulgar language is not a very prominent category in our dataset. In our data it appears relatively more often in Spanish on each platform, on German X, and uncommonly in Dutch. It occurs relatively consistently on each platform, also X in comparison to other categories. This suggests a baseline of activity instead of event-driven peaks.

In terms of attention and targets, vulgar language in our dataset has a very low mean Gini (0.11) and high HHI (0.36), suggesting that while inequality is low, attention is strongly concentrated on a small set of targets. The trend is consistent over time although we do note the limited number of observations. Together with only moderate peaks of activity this suggests a baseline of vulgar language instead of episodic bursts of attention, thus seemingly less reliant on news cycles.

In terms of spreaders, we find a highly concentrated pattern (98.5% focused targeting–low persistence). Compared to other categories in our dataset, vulgarity has a moderate share of users with broader targeting (1.4% diffuse targeting–low persistence). We find a small share of persistent users (0.1%).

In terms of interventions, these findings suggest that vulgarity may be already moderated to a large extent on Instagram and TikTok. From what remains in our dataset, vulgarity appears largely episodic, with some broad targeting behavior, and with a core of persistent attackers, most notably so on X.

Figure 14. Vulgarity-related hostility comparison of share, targets, spreaders



Conclusions

Overall, our findings suggest online hostility is widespread and consists predominantly of one-off hostile comments. In terms of interventions, this suggests the main challenge is not to counteract a small number of repeat offenders, but a large number of one-off hostile comments originating from many different accounts.

We underline here that the rarity of evidence in our dataset of categories such as “Race” or “Violent language” may be a result of biases in data collection and methodology rather than evidence of absence. We also note the limited accuracy and recall of the language models used here to detect different types of hostility. It is from the comparison across categories, countries and platforms that we draw the following insights.

In general, we note that spreaders of hostility are more diffuse than expected. Online hostility in our dataset predominantly originates from many different accounts, rather than being concentrated from a few accounts. We find little overlap across categories among spreaders of online hostility, suggesting most accounts engage in a single type of hostility rather than diversifying repertoires. Among spreaders of online hostility, we find very little evidence of focused persistent attacks. Across all categories, the dominant pattern is focused targeting and low persistence, suggesting that most users engage briefly and narrowly regardless of topic. We conclude most online hostility in our dataset are one-off posts with very limited repeat offenders. We also note targets of online hostility are strongly concentrated. We see this same pattern when differentiating by type of hostility. Across all categories we find high degrees of concentration of targets (most attention directed at relatively few targets), and varying degrees of consistency among targets with the same accounts being targeted. Further differentiating targeting, hostile posting behavior and volatility, we find striking differences between categories which inform different intervention strategies.

Recommended intervention strategies

In terms of interventions, we suggest three actionable insights based on our findings, summarized in Table 7.

Table 7. Recommended intervention strategy by category types

Category	Target concentration	Notable spreader type	Volatility	Intervention focus
Race	Concentrated + stable	Repeat offenders (Diffuse + persistent)	Episodic	Specific targets + specific spreaders, + events
Sexual intimidation	Diffuse + stable	Generalists (Diffuse + transient)	Consistent + episodic	Specific targets + events
Religion	Diffuse + unstable	Generalists (Diffuse + transient)	Consistent + episodic	Specific targets + events
Sexual orientation	Diffuse + stable	Generalists (Diffuse + transient)	Consistent + episodic	Specific targets + events
Vulgar language	Concentrated + stable	Repeat offenders (Diffuse + persistent)	Consistent	Specific targets + Specific spreaders
Crime	Diffuse + stable	Generalists (Diffuse + transient)	Consistent + episodic	Events
Culture	Concentrated + unstable	Opportunists (Focused + transient)	Consistent + episodic	Events
Politics	Concentrated + unstable	Repeat offenders (Diffuse + persistent)	Episodic	Events
Gender	Diffuse + stable	Opportunists (Focused + transient)	Consistent	Specific targets
Violent language	Concentrated + stable	Generalists (Diffuse + transient)	Consistent	Specific targets
Insults	Concentrated + unstable	Repeat offenders (Diffuse + persistent)	Consistent	Specific spreaders

A first group of categories is characterized by episodic activity (race, sexual intimidation, sexual orientation, religion, violent speech, crime, culture, and politics). Compared to other categories, these are relatively more volatile and are likely more strongly event-driven. Intervention strategies for these categories might focus on anticipating and reacting to events generating online hostility.

A second group of categories is characterized by target concentration (race, sexual intimidation, religion, sexual orientation, gender, and vulgar language and violent language).

Compared to other categories, targets of these types of hostility are relatively concentrated and/or stable. Intervention strategies for these categories might focus on monitoring the accounts of likely targets of online hostility.

A third group of categories is characterized by consistency among spreaders (Race, vulgar language, and insults). Compared to other categories, spreaders of these types of hostility are relatively persistent. Intervention strategies for these categories might focus on usual suspects generating online hostility.

We emphasize here the weak recall and precision of the language models are a key limitation. In practice, the models employed here will miss many posts containing online hostility, and flag many cases as false positive. With an eye on our goal of supporting moderation, we deem “false alarms” preferable to missing many problematic posts. This accuracy will likely become worse over time and when applied to other platforms. We advise against the use of these models for automated decision-making. The practical applicability of these models is to flag content for inspection by experts.

Methodological limitations

The data collection and methodology informing this report come with important caveats.

First, we note that the most pervasive forms of hostility may be underrepresented in our dataset as a result of our data collection strategy. While extensive, our datasets are only partly representative. Collecting only publicly available data, for reasons of privacy and in light of the practical constraints of social media data, means that the most stringently moderated forms of content were deleted before we had a chance to collect it. As a result, it is likely that racism and violent speech generally and gender-based violence more specifically are more salient in practice than our analysis suggests. Even if moderation deletes such content after some time, damage may already be done.

Second, also related to data collection, limitations remain as a result of prioritizing breadth and diversity. Our initial focus on identifying prominent figures implies that less prominent or visible actors may be underrepresented. We sought to account for this by purposive sampling on identity. Characteristics such as race, ethnicity, gender identity, sexual orientation, religion,

migration background, language, and political ideology were taken into account to ensure the dataset includes groups disproportionately targeted by online hostility but who may be less prominent in the public debate. We sought to represent minority populations specific to each national context, such as linguistic minorities in Spain and individuals engaged in debates surrounding historically sensitive topics in Germany. A related limitation of our approach is overlap between sectors (such as politics and activism) which complicated efforts to balance across categories. Ultimately, this led to some categories being strongly represented (such as politics) at the expense of others (such as sports). We believe our approach ensures extensive coverage across social domains and identity groups, while accounting for contextual and cultural variation in online hate expression.

Third, we acknowledge that social media posts are not neutral or naturally occurring, and social media accounts do not represent natural persons. Our dataset likely includes different forms of automation, spam, and other forms of malicious intent. Identification of bots and forms of coordination are beyond the scope of this project, but we strongly encourage further examination of the relationship between coordinated forms of manipulation and the different types of hostility identified in this report.

Fourth, we acknowledge that comparison of data from different social media platforms (Instagram, X, TikTok) *and* different national contexts (Germany, Spain, the Netherlands) is severely restricted. Each social media platform has diversified user bases, particular vernaculars, and specific affordances that are meaningfully interwoven with the use of language on these platforms. Furthermore, we acknowledge the meaning of is very strongly related to context. Comparative analysis of the meaning of language and related differences across countries and platforms warrants qualitative inquiry. We believe the more quantitatively oriented comparisons offered in this report are justified by our goal of identifying patterns in the prevalence of language across countries and platforms. We think of these comparisons as indicative trends rather than accurate representations.

Fifth, automated language classification is inherently limited in terms of performance. We acknowledge the weak recall and precision of hate speech detection models and to a lesser degree the classification models is a key limitation of the method described here. This is the case in particular for categories that are less common in our dataset and thus have fewer observations to train the models. For practical purposes, this means the models will produce

many false positives. With an eye on our goal of supporting moderation, we deem false alarms preferable to missing many problematic posts. The three pre-existing models for hate speech detection were selected after preliminary performance analysis at the stage of research design. At the time of research design, LLMs as better performing alternatives to rule-based lexicons were decided against for reasons of practical feasibility and restrictions relating to data management and privacy. Finetuning of BERT-based models using our annotated data remains an ambition for future work. For purposes of our analysis, we addressed this limitation through a semi-supervised method, combining automated detection and human annotation and classification. We also sought to address this limitation through ongoing critical reflection, assessment of intercoder agreement, calibration, manual inspection and spot checking to ensure the validity of our findings. Nevertheless, we acknowledge the classification models underpinning our analyses are flawed in terms of accuracy, precision and recall. As language and its uses further develop, online hostility shifting to different words and meanings, accuracy of these models is likely to further decline over time. We advise against the use of these models for automated decision-making. The practical applicability of these models is to flag content for inspection by experts.

Sixth, we note that we cannot say much about the impact of online hostility. Our analyses provide insight into the patterns of different types of hostility but cannot differentiate impacts in the lives of people receiving it. We acknowledge some types of hostility may be more impactful than others, even if categories such as racism are not the most pronounced in our analysis, and that the compounding of different categories such as race and gender may further intensify harmful effects.

Overall, this methodology provides a comprehensive and adaptable framework for cross-national data collection. By systematically identifying accounts operating in the public domain and who are exposed to a variety of online attacks, the resulting dataset covers a broad array of online hostility. The lexicons constructed for this analysis are available open source (van Haperen *et al*, 2026). With appropriate attention to local context, this approach is replicable and may support moderation of both overt and more nuanced forms of online hostility.

Acknowledgements

Thank you to those who contributed to the procedure and analysis presented in this report: Casper van der Heijden, Bart van der Velden, Kris Ruijgrok, Joyce van der Lans, Anna Claire Shepherd, Nick van Loo, Caitlin Gray, Pilar Richter, Kamil Nakad, Lea Bund, Paula Magnanimo, Claudia Castro, Maria Olle, Anonymous Annotator 1, Christina Hübers, Rita Costa Cots, Mahmoud Bastati, Pablo Aguiar, Maria Raduà, Helena Puig Larrauri, Benjamin Cerigo, and Julie Hawke. Thank you for your contributions and valuable insights.

Legal disclaimer

The goal of this report is to identify trends and patterns, and is not a qualification of opinions or statements of individuals for legal purposes. Where possible, personal data are aggregated and pseudonymized or anonymized before analysis in accordance with GDPR and relevant European regulations. Analyses in this report are based on automated detection and interpretation of public online data, which are subject to methodological and contextual limitations.

Data statement and license

The lexicons crafted as part of this project, and which were used to inform the current analysis, are available open source in the repository of the University of Amsterdam⁵.

Data set name: Violence, Intolerance and Impoliteness thematic Corpus v1.0

Data set developer(s): 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.

Data statement author(s): Sander van Haperen (University of Amsterdam)

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

You are free to:

⁵ <https://doi.org/10.21942/uva.31876402>

- **Share** - copy and redistribute the material in any medium or format
- **Adapt** - remix, transform, and build upon the material

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

- **Attribution** - You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
- **NonCommercial** - You may not use the material for commercial purposes.
- **ShareAlike** - If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.
- **No additional restrictions** - You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

Notices: You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable exception or limitation. No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as publicity, privacy, or moral rights may limit how you use the material.

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

Bibliography

- 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://eoooh.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.