



**Together Against
Online Hate Speech**

TAOH

Together Against Online Hate Speech

Deliverable 2.2 Hate Speech Pattern Report

German translation

Über TAOH

Together Against Online Hate Speech (TAOHS) ist ein internationales Projekt von Organisationen aus Deutschland, den Niederlanden und Spanien. Gemeinsam untersuchen wir, wie Hass im Netz entsteht, sich verbreitet und in sozialen Medien sowie in digitalen Räumen wirkt. Darauf aufbauend entwickeln wir Werkzeuge, Trainings und praxisnahe Empfehlungen, mit denen zivilgesellschaftliche Organisationen, Medien und Institutionen Hass im Netz vorbeugen und ihm sicher sowie wirksam begegnen können.

Konsortialpartner

- SPF - Sharing Perspectives Foundation - Niederlande | Koordinator
- UvA - University of Amsterdam - Niederlande
- Build Up - How to Build Up Germany - Deutschland
- IBH - ichbinhier e.V. - Deutschland
- FundiPau - Fundació par la Pau - Spanien



**Kofinanziert von der
Europäischen Union**

gemacht werden.

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich



Dieses Werk ist lizenziert unter einer [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/)

„Feindseligkeit im Internet: Vergleichende Analyse von Instagram, X und TikTok in Spanien, Deutschland und den Niederlanden“

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

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

Abstract

Dieses Dokument berichtet über die Ergebnisse einer vergleichenden Analyse von Feindseligkeit im Internet im Rahmen des Projekts „Gemeinsam gegen Hassrede im Internet“. Das Projekt wird von folgenden Partnern durchgeführt: Sharing Perspectives Foundation, BuildUp, ichbinhier e.V., FundiPau und der Universität Amsterdam.

Ziel dieses Berichts ist es, Modelle von Feindseligkeit im Internet in Deutschland, den Niederlanden und Spanien auf Instagram, X und TikTok zu vergleichen. Dabei werden Kategorien, beispielsweise Rassismus und geschlechtsbezogene Feindseligkeit, sowie Verhaltensmuster hinsichtlich der Auswahl von Zielpersonen und der Beständigkeit unterschieden. Zu diesem Zweck werden Sprachmodelle entwickelt, um unterschiedliche Formen von Feindseligkeit im Internet zu erkennen und zu klassifizieren. Diese Modelle werden als Open Source verfügbar gemacht.

Die Ergebnisse bilden eine Grundlage für Interventionsstrategien von zivilgesellschaftlichen Organisationen, Aktivistinnen und Aktivisten, Strafverfolgungsbehörden sowie politischen Entscheidungsträgerinnen und Entscheidungsträgern, die sich mit Feindseligkeit im Internet befassen. Erstens erfordern ereignisbezogene Kategorien wie Rassismus, sexuelle Einschüchterung, sexuelle Orientierung, Religion, gewaltvolle Sprache, Kriminalität, Kultur und Politik schnelle Reaktionen auf Ereignisse. Zweitens legt die Konzentration auf bestimmte Zielpersonen oder Gruppen in einigen Kategorien, insbesondere Politik, Beleidigungen, Kriminalität und Kultur, eine Kombination aus Beobachtung und Unterstützung wahrscheinlicher Betroffener von Feindseligkeit im Internet nahe. Drittens sind Personen, die Feindseligkeit im Zusammenhang mit Rassismus, vulgärer Sprache und Beleidigungen verbreiten, im Vergleich zu anderen Kategorien relativ beständig aktiv.

Interventionsstrategien für diese Kategorien könnten sich daher auf wiederholt auffällige Akteure konzentrieren, die Feindseligkeit im Internet verbreiten.

Zitat:

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.

Inhaltsverzeichnis

Einleitung.....	6
Definition	7
Methodik	8
Datenerhebung	9
Modell 1: Erkennung von Hassrede	11
Modell 2: Klassifikator für Feindseligkeit im Internet	12
Annotierungsverfahren	12
Erstellung der Lexika	15
Modellleistung.....	16
Ergebnisse	18
Muster nach Ländern	18
Muster nach Plattform	21
Muster nach Zielpersonen und Verbreitenden	23
Verhaltensmuster	25
Muster nach Kategorien.....	26
Beschreibung der Kategorien	26
Kriminalität	29
Kultur	31
Geschlecht	33
Beleidigungen und persönliche Angriffe	36
Politik.....	38
Rasse.....	40
Religion	43
Sexuelle Einschüchterung.....	45
Sexuelle Orientierung.....	48
Gewaltvolle Sprache.....	50
Vulgäre Sprache.....	52
Schlussfolgerungen.....	55
Empfohlene Interventionsstrategien	56
Methodische Einschränkungen	58
Danksagungen	61
Rechtlicher Hinweis	61
Datenerklärung und Lizenz.....	62
Literaturverzeichnis.....	64

Einleitung

Dieser Bericht ist Teil des Projekts „Together Against Online Hate Speech“ (TAOHS). Das übergeordnete Ziel des Projekts besteht darin, die Widerstandsfähigkeit zivilgesellschaftlicher Organisationen, von Medien und Behörden zu stärken, damit sie Hassrede im Internet auf ihren Plattformen besser verstehen und bekämpfen und zu einem sichereren und pluralistischen digitalen Raum beitragen können. Dieser Bericht konzentriert sich auf Deutschland, die Niederlande und Spanien.

Das Projekt verbindet sprachbasierte Klassifikationsmodelle mit dialogorientierten Interventionen unter menschlicher Leitung und politischer Einflussnahme, um Feindseligkeit im Internet umfassend zu verstehen und ihr entgegenzuwirken. Während Inhaltsmoderation und Meldesysteme weiterhin wichtig sind, um rechtswidrige Inhalte oder Inhalte, die gegen Plattformrichtlinien verstoßen, zu behandeln, stoßen solche Strategien bei Inhalten an Grenzen, die nicht rechtswidrig, aber dennoch problematisch sind. Auch Sprachformen wie Intoleranz und Unhöflichkeit erfordern Interventionen. Unser Ziel ist es, Einzelpersonen, zivilgesellschaftliche Organisationen, Strafverfolgungsbehörden sowie politische Entscheidungsträgerinnen und Entscheidungsträger zu unterstützen, die mit solcher Sprache im Internet umgehen.

Ziel dieses Berichts ist es, Interventionsstrategien durch die Identifikation von Modellen der Feindseligkeit im Internet zu informieren. Methodisch ist diese Studie darauf ausgerichtet, ein breites Spektrum von Feindseligkeit im Internet zu erkennen, das sich gegen eine repräsentative Vielfalt von Zielpersonen und Zielgruppen richtet, in verschiedenen sozialen Medienräumen wie Instagram, X und TikTok auftritt und in den unterschiedlichen nationalen Kontexten Deutschlands, der Niederlande und Spaniens verglichen wird. Zu diesem Zweck entwickeln wir drei Sprachmodelle zur Klassifikation verschiedener Formen von Feindseligkeit im Internet auf Deutsch, Niederländisch und Spanisch. Die Ergebnisse umfangreicher abduktiver Kodierung und Leistungstests werden als Open Source zur Verfügung gestellt, damit Partnerorganisationen und weitere Interessengruppen die Sprachmodelle für spezifische Analysen anpassen können.

In diesem Bericht verwenden wir diese Modelle für einen plattformübergreifenden und länderübergreifenden Vergleich, um Modelle hinsichtlich der Verbreitung, der Zielpersonen

und Zielgruppen sowie der Verbreitenden unterschiedlicher Formen von Feindseligkeit im Internet zu identifizieren. Insgesamt legen unsere Ergebnisse nahe, dass das Problem weit verbreitet ist und überwiegend aus einzelnen feindseligen Kommentaren besteht. Bei einer nach Kategorien differenzierten Betrachtung zeigen sich jedoch deutliche Unterschiede zwischen verschiedenen Formen von Feindseligkeit im Internet.

Daraus ergeben sich drei umsetzbare Erkenntnisse. Erstens erfordern ereignisbezogene Kategorien wie Rassismus, sexuelle Einschüchterung, sexuelle Orientierung, Religion, gewaltvolle Sprache, Kriminalität, Kultur und Politik rasche Reaktionen auf aktuelle Ereignisse. Zweitens legt die Konzentration auf bestimmte Zielpersonen oder Zielgruppen in einigen Kategorien, insbesondere Politik, Beleidigungen, Kriminalität und Kultur, eine Kombination aus Beobachtung und Unterstützung wahrscheinlicher Betroffener von Feindseligkeit im Internet nahe. Drittens sind Personen, die Feindseligkeit im Zusammenhang mit Rassismus, vulgärer Sprache und Beleidigungen verbreiten, im Vergleich zu anderen Kategorien relativ dauerhaft aktiv. Interventionsstrategien für diese Kategorien könnten sich daher auf wiederholt auffällige Akteure konzentrieren, die Feindseligkeit im Internet verbreiten.

Das Projekt „Together Against Online Hate Speech“ umfasst die folgenden Partner: Sharing Perspectives Foundation, Build Up, IchBinHier, FundiPau und die Universität Amsterdam. Das Projekt wird von der Europäischen Union finanziert (CERV-2024-CHAR-LITI-SPEECH). Die geäußerten Ansichten und Meinungen stammen jedoch ausschließlich von den Autorinnen und Autoren und spiegeln nicht zwangsläufig die Ansichten der Europäischen Union wider. Weder die Europäische Union noch die Bewilligungsbehörde können dafür verantwortlich gemacht werden.

Definition

Zu Beginn unseres Projekts formulierten wir folgende Definition von „Hassrede“: „öffentliche Aufstachelung zu Gewalt oder Hass gegen eine Gruppe von Personen oder gegen ein Mitglied einer solchen Gruppe, die durch Bezugnahme auf Rasse, Hautfarbe, Religion, Abstammung oder nationale oder ethnische Herkunft definiert wird.“ Darüber hinaus beziehen wir Geschlecht, sexuelle Orientierung und körperliche Beeinträchtigung ein. Zu Beginn gingen wir davon aus, dass solche Sprachkategorien wie „Entmenschlichung, Sexismus, Politik,

Rassismus, Religion, Gewalt und Drohungen“ umfasst, angelehnt an EOHH, 2026. Praktisch lieferte uns diese Definition sensibilisierende Konzepte und diente der Abstimmung zwischen den annotierenden Personen.

In diesem Bericht verwenden wir zudem den Oberbegriff „Feindseligkeit im Internet“. Damit bezeichnen wir Sprache, die nicht notwendigerweise rechtswidrig ist oder Hassrede darstellt, die sich auf die Identität von Einzelpersonen bezieht, aber dennoch die Qualität digitaler Räume beeinträchtigen kann. Neben Hassrede zählen auch Beschimpfungen und Unhöflichkeit zu anderen Formen der Feindseligkeit im Internet. Beschimpfungen beziehen sich auf Sprache, die darauf abzielt, eine bestimmte Person zu beleidigen, zu provozieren oder zu verletzen (Akhtar & Morrison, 2019; Ward & McLoughlin, 2020). Politische Unhöflichkeit bezeichnet „einen Mangel an Respekt gegenüber den sozialen und kulturellen Normen, die persönliche Interaktionen und das Funktionieren demokratischer Systemregeln“ (Bentivegna & Rega, 2024). Dies umfasst normverletzende Diskurse, die von vergleichsweise milder Unhöflichkeit bis zu offeneren Formen der Feindseligkeit reichen, ohne notwendigerweise eine Schädigungsabsicht oder die gezielte Ansprache bestimmter sozialer Gruppen vorauszusetzen (Rossini, 2019).

Methodik

Ziel dieser Studie ist es, Interventionsstrategien durch die Identifikation von Modellen der Feindseligkeit im Internet zu informieren. Methodisch ist die Studie darauf ausgerichtet, ein breites Spektrum von Feindseligkeit im Internet zu erkennen, das sich gegen eine repräsentative Vielfalt von Zielpersonen und Zielgruppen richtet, in verschiedenen sozialen Medienräumen wie Instagram, X und TikTok auftritt und in den unterschiedlichen nationalen Kontexten Deutschlands, der Niederlande und Spaniens verglichen wird.

Zu diesem Zweck haben wir Daten aus sozialen Medien für eine repräsentative Auswahl besonders wahrscheinlicher Zielpersonen und Zielgruppen gesammelt. Wir analysierten diesen Datensatz mithilfe zweier Sprachmodelle: Ein Modell bestimmt, ob ein Beitrag wahrscheinlich Feindseligkeit im Internet enthält, während das andere ermittelt, auf welches Thema sich der Beitrag bezieht. Diese Kombination ermöglicht es uns, Modelle verschiedener Formen von Feindseligkeit im Internet zu identifizieren und zu untersuchen. Wir konzentrieren

uns auf Modelle hinsichtlich der Verbreitung, der Zielpersonen und Zielgruppen sowie der Verbreitenden von Feindseligkeit im Internet. Dieses Kapitel erklärt das Vorgehen ausführlicher.

Datenerhebung

Wir haben Daten von Instagram, X und TikTok gesammelt. Diese Plattformen wurden ausgewählt, weil sie als wahrscheinliche Orte für das Auftreten von Feindseligkeit im Internet gelten und zu den in allen drei Ländern am häufigsten genutzten sozialen Medien gehören. Die Auswahl umfasst zudem unterschiedliche Nutzergruppen und technische Möglichkeiten. Auf X sind politische Meinungsäußerungen, die eng mit aktuellen Nachrichtenereignissen verknüpft sind, besonders verbreitet. Instagram legt den Schwerpunkt auf Lebensstil, während TikTok eine stärkere Beteiligung jüngerer Bevölkerungsgruppen und eine offene, algorithmisch gesteuerte Entdeckung von Inhalten ermöglicht.

Beiträge und Kommentare auf X, vormals Twitter, umfassen kurze Nachrichten und Reaktionen von in der Regel 50 Wörtern oder weniger. Sie können multimediale Materialien, externe URL Verweise und Erwähnungen anderer Nutzerinnen und Nutzer enthalten. Beiträge und Kommentare auf Instagram bestehen typischerweise aus kurzen Bildunterschriften und Reaktionen von etwa 50 Wörtern oder weniger und können externe URL Verweise sowie Erwähnungen anderer Nutzerinnen und Nutzer enthalten. Beiträge und Kommentare auf TikTok bestehen üblicherweise aus kurzen Bildunterschriften zu Videos und Reaktionen von etwa 50 Wörtern oder weniger. Auch sie können externe URL Verweise und Erwähnungen anderer Nutzerinnen und Nutzer enthalten.

Der Korpus ist das Ergebnis des folgenden Verfahrens. Zunächst identifizierten wir für jedes der drei Länder Spanien, Deutschland und die Niederlande Konten von Einzelpersonen und Organisationen, die mit hoher Wahrscheinlichkeit verschiedenen Formen von Hassrede, Intoleranz und Unhöflichkeit im Internet ausgesetzt sind. Die Fallauswahl in den spanischen, deutschen und niederländischen Kontexten folgte einer gezielten Stichprobenstrategie. Ihr Ziel war es, eine vielfältige Auswahl von Konten in sozialen Medien zusammenzustellen, die mit hoher Wahrscheinlichkeit Hassrede im Internet erhalten.

Unser Ziel bei der Datensammlung war die Erstellung eines Datensatzes, der ein breites Spektrum von Feindseligkeit im Internet erfasst, um die Entwicklung eines Sprachmodells zu unterstützen, das sowohl für explizite als auch für subtilere Sprachformen empfindlich ist. Für jedes der drei Länder begann die Auswahl mit der Identifikation prominenter Einzelpersonen und Organisationen, die in Bereichen mit hoher öffentlicher Sichtbarkeit und Beteiligung tätig sind. Konten wurden aus Bereichen wie Politik, Medien, Unterhaltung und Kultur, öffentlichem Dienst, Wissenschaft, Aktivismus, Sport und zivilgesellschaftlichen Organisationen ausgewählt. Innerhalb jedes Bereichs wurde Vielfalt sichergestellt, indem Untergruppen wie gewählte Amtsträgerinnen und Amtsträger, politische Parteien, Journalistinnen und Journalisten, Sportlerinnen und Sportler sowie Influencerinnen und Influencer einbezogen wurden.

Wir nutzten unsere Vertrautheit mit den jeweiligen Kontexten und ergänzende Quellen wie Nachrichtenberichte, um Personen zu identifizieren, die bekanntermaßen einem hohen Ausmaß an Hass im Internet ausgesetzt sind. Dadurch sollte gewährleistet werden, dass unsere Auswahl auf beobachtbaren Mustern von Belästigung beruht. Besonderes Augenmerk lag auf Personen, die in aktuelle oder fortdauernde politische, kulturelle oder mediale Kontroversen eingebunden waren, sowie auf Personen, die mit kontroversen Themen wie Migration, geopolitischen Konflikten oder nationalen Debatten in Verbindung stehen. Diese Themen erhöhen bekanntermaßen die Wahrscheinlichkeit, Feindseligkeit im Internet ausgesetzt zu sein. Diese Stichprobenstrategie entsprach unserer Definition von Hassrede und erfasste zugleich deren vielfältige Erscheinungsformen in unterschiedlichen sozialen Gruppen. Wir identifizierten 377 Konten für Spanien, 217 Konten für Deutschland und 360 Konten für die Niederlande.

Für jedes dieser Ausgangskonten sammelten wir Daten aus den Profilen in sozialen Medien auf X, Instagram und TikTok. In die Auswahlkriterien wurden Kommentare und Reaktionen einbezogen, die auf diesen Konten veröffentlicht wurden, sowie Beiträge, in denen das jeweilige Konto im Zeitraum vom 01.01.2024 bis zum 31.12.2025 erwähnt wurde. Für X umfasst dies Erwähnungen und Retweets, für Instagram Erwähnungen und erneute Veröffentlichungen und für TikTok Erwähnungen und Antworten. Konten ohne offizielle oder identifizierbare Präsenz in sozialen Medien wurden ausgeschlossen. Es wurden ausschließlich öffentlich verfügbare Beiträge und Kommentare einbezogen. Inhalte, die zwischen ihrer

Veröffentlichung und der Datensammlung im Zeitraum von September bis Dezember 2025 gelöscht wurden, wurden nicht berücksichtigt. Ebenso wurden Daten, die als privat eingestellt waren oder zwischen dem Zeitpunkt der Veröffentlichung und unserer Studie gelöscht wurden, nicht in die Datensammlung aufgenommen.

Dieses Vorgehen führte zu drei Datensätzen:

- **Deutscher Korpus:** 1.452.268 Beiträge, davon 598.865 von Instagram, 730.971 von TikTok und 122.432 von X.
- **Spanischer Korpus:** 2.445.553 Beiträge, davon 457.036 von Instagram, 1.953.819 von TikTok und 34.698 von X.
- **Niederländischer Korpus:** 5.611.651 Beiträge, davon 1.018.428 von Instagram, 680.296 von TikTok und 3.912.927 von X.

Modell 1: Erkennung von Hassrede

Um die Wahrscheinlichkeit zu bestimmen, dass ein Beitrag in unserem Datensatz Feindseligkeit im Internet enthält, verwendeten wir zunächst bestehende Sprachklassifikatoren. Nach vorläufigen Tests und einer Prüfung der jeweiligen Trainingsgrundlagen wählten wir drei separate Open Source Modelle aus (Aluru et al., 2002; Pérez et al., 2022; IMSyPP, 2022).

Diese Modelle wurden jeweils mit einsprachigen Daten aus sozialen Medien und unter Verwendung des BERTje Tokenizers trainiert. Ihr gemeinsames Ziel besteht darin, das Vorhandensein von Hassrede zu erkennen (Basile et al., 2019; Pérez et al., 2021; Cinelli et al., 2021; Kralj Novak, 2022; Baider et al., 2023). Wir wenden diese Modelle auf unsere Datensätze an und erhalten die folgenden Ergebnisse, siehe Tabelle 1.

Tabelle 1. Leistung des Modells 1, Ziel: Erkennung von Hassrede

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

Anmerkung: Das deutsche Modell vergab bei einem Schwellenwert von 0,94 das Label „HATE“, das spanische Modell bei einem Schwellenwert von 0,052 das Label „hateful“ und das niederländische Modell bei einem Schwellenwert von 0,48 das Label „Label_2“. Diese Schwellenwerte werden hier zur Replizierbarkeit angegeben. Die Werte sind zwischen den Modellen nicht vergleichbar.

Es ist wichtig, auf die geringe Leistungsfähigkeit dieser Modelle hinzuweisen. Diese Modelle wurden ausgewählt, weil sie bei der vorläufigen Analyse während der Konzeption dieser Studie unter den verfügbaren Alternativen die besten Ergebnisse erzielten. In der Praxis ist ein AUC Wert von 0,564 nur geringfügig besser als eine zufällige Zuordnung. Dies wird im Abschnitt zu den methodischen Einschränkungen ausführlicher erläutert. Wir raten ausdrücklich davon ab, diese Modelle für automatisierte Entscheidungen bei der Erkennung von Hassrede einzusetzen. Für unsere Analysen ergänzen wir die Ergebnisse jedes Modells durch eine menschliche Klassifikation von Feindseligkeit im Internet.

Modell 2: Klassifikator für Feindseligkeit im Internet

Während Modell 1 darauf ausgelegt ist zu erkennen, ob ein Beitrag wahrscheinlich problematische Sprache enthält, ermöglicht es kein Verständnis dafür, auf welches Thema sich der Beitrag bezieht. Um verschiedene Formen der Feindseligkeit zu identifizieren, entwickeln wir ein regelbasiertes Sprachmodell.

Annotierungsverfahren

Um verschiedene Formen von Hassrede zu identifizieren, verwendeten wir einen abduktiven Kodierungsansatz (Timmermans & Tavory, 2012). Zunächst kodierten drei fachkundige annotierende Personen in aufeinanderfolgenden Durchgängen Zufallsstichproben aus diesen Datensätzen, um Kategorien der Feindseligkeit im Internet zu identifizieren. Unsere

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

gemeinsame Definition von Feindseligkeit im Internet lieferte sensibilisierende Konzepte und strukturierte Diskussionen über Semantik und unsere Untersuchungsgegenstände. Um wiederkehrende sprachliche Modelle in Nutzerinteraktionen zu bestimmen, interpretierten wir Beiträge von Konten aus ähnlichen Bereichen oder mit gemeinsamen Identitätsmerkmalen. Dabei dokumentierten wir häufig verwendete Schlüsselwörter, Formulierungen und Ausdrucksweisen der Feindseligkeit im Internet. Dieser Prozess umfasste auch die Identifikation subtilerer Formen schädigender Sprache, etwa kodierter Ausdrücke, abwertender Bezeichnungen, Slang und Hundepfeifen.

Zweitens kamen wir nach jeder Kodierungsrunde mit Stichproben von 500 Beiträgen zusammen, um die vergebenen Kategorien zu vergleichen, unsere Definition von Hassrede zu bewerten, frühere Literatur einzubeziehen und unseren Kodierleitfaden anzupassen oder zu erweitern. Auf Grundlage unserer Ausgangsdefinition erwarteten wir zunächst die folgenden Kategorien: Rasse, Hautfarbe, Religion, Abstammung oder nationale beziehungsweise ethnische Herkunft, Geschlecht, sexuelle Orientierung und körperliche Beeinträchtigung. Anschließend ergänzten wir diese Kategorien induktiv. Dies führte zur Aufnahme der Kategorien „Unwahrheit“, „antinstitutionelles Misstrauen“, „Verschwörungserzählungen“, „Polarisierung“, „Ukraine“, „Russland“, „Autokratie“, „Naher Osten“, „Klima“, „Kriminalität“, „Sport“, „Beleidigungen“, „Vulgärität“, „Ironie und Sarkasmus“ sowie „sexuelle Einschüchterung“. In dieser Phase behielten wir während der weiterhin induktiven Kodierung aller Inhalte auch die Kategorien „informativ“, „kritisch“ und „positiv“ bei. Diese wurden später verworfen, um ausschließlich Feindseligkeit im Internet zu untersuchen.

Drittens wurde die entstandene Begriffsliste anschließend von Personen überprüft und verfeinert, die mit dem jeweiligen digitalen Kontext der einzelnen Länder vertraut waren. Damit sollte ihre kulturelle und sprachliche Angemessenheit sichergestellt werden. Neben der Identifikation verschiedener Kategorien der Feindseligkeit im Internet und des Themas eines Beitrags bewerteten wir zunächst auch die Intensität auf einer Skala von 1 bis 10. Nach einer kritischen Reflexion unserer unterschiedlichen Auslegungen von Intensität wurde dieses Vorgehen jedoch verworfen. Im Einklang mit unserem abduktiven Ansatz wechselten wir iterativ zwischen Kodierung und Literaturrecherche. Dies führte dazu, dass wir Unhöflichkeit und Respektlosigkeit von Hassrede unterschieden (Nithyanand et al., 2017; Theocharis et al., 2020).

Wir verstehen Unhöflichkeit als „einen Mangel an Respekt gegenüber den sozialen und kulturellen Normen, die persönliche Interaktionen und das Funktionieren demokratischer Systeme regeln“ (Bentivegna & Rega, 2024). Dieser Mangel an Respekt kann sich als zwischenmenschliche Unhöflichkeit oder als Diskurs äußern, der demokratische Systeme und ihre Akteure delegitimiert. Dazu gehören etwa Stigmatisierung, rollenbezogene Angriffe, Dämonisierung, Beschimpfungen, Drohungen, Aufrufe zu Gewalt, Vulgärität, Vorwürfe der Rechtswidrigkeit und beleidigende Sprache (Bentivegna & Rega, 2024). Nach vier Kodierungsdurchgängen waren wir mit dem Kodierleitfaden zufrieden, und die identifizierten Kategorien deckten die in der vierten Stichprobe vorgefundenen Daten angemessen ab. Zu diesem Zeitpunkt führten unsere Kodierung und Diskussionen nicht mehr zu neuen Kategorien oder Bezeichnungen. Dies betrachteten wir als Sättigungspunkt.

Viertens wandten wir den Kodierleitfaden auf Daten an, die nicht Teil der zuvor kodierten Stichproben waren. Zu diesem Zeitpunkt betrachteten wir den Kodierleitfaden als hinreichend umfassend für unser Ziel, Feindseligkeit im Internet zu identifizieren. Zwölf fachkundige annotierende Personen, die über vom jeweiligen Konsortiumspartner identifiziert wurden, erhielten Anweisungen zur Anwendung des Kodierleitfadens auf neue Zufallsstichproben. Ergänzend fanden acht Kalibrierungssitzungen statt, in denen Uneinigkeiten bei der Zuordnung von Kategorien diskutiert wurden. Uneinigkeiten wurden aufgelöst, sobald die annotierenden Personen nach der Diskussion einen Konsens erreichten.

Letztlich wurde die folgende Übereinstimmung zwischen den Kodierenden erreicht:

- **Deutsch:** 7.852 Beiträge aus unserer Stichprobe wurden von zwei annotierenden Personen kodiert, entsprechend 0,54 % der 1.452.268 Beiträge im deutschen Korpus. 5.335 Beiträge wurden von drei annotierenden Personen kodiert, entsprechend 0,37 % des Korpus. Die Übereinstimmung zwischen den Kodierenden war akzeptabel (Krippendorffs Alpha 0,73).
- **Spanisch:** 6.534 Beiträge aus unserer Stichprobe wurden von zwei annotierenden Personen kodiert, entsprechend 0,27 % der 2.445.553 Beiträge im spanischen Korpus. 4.719 Beiträge wurden von drei annotierenden Personen kodiert, entsprechend 0,19 % des Korpus. Die Übereinstimmung zwischen den Kodierenden war akzeptabel (Krippendorffs Alpha 0,67).

- **Niederländisch:** 12.933 Beiträge aus unserer Stichprobe wurden von zwei annotierenden Personen kodiert, entsprechend 0,24 % der 5.611.651 Beiträge im niederländischen Korpus. 9.091 Beiträge wurden von drei annotierenden Personen kodiert, entsprechend 0,17 % des Korpus. Die Übereinstimmung zwischen den Kodierenden war gut (Krippendorffs Alpha 0,83).

Erstellung der Lexika

Wir führten zwei Durchgänge axialer Kodierung durch, um die Kategorien wieder mit unserem Verständnis von Hassrede zu verknüpfen, das sich im Verlauf des Kodierungsprozesses entwickelt hatte (Green & Thorogood, 2005). Hierzu betrachteten wir die Beziehungen zwischen den Kategorien. Letztlich führten diese Überlegungen dazu, dass wir den Begriff „Feindseligkeit im Internet“ verwendeten. Dieser ist weiter gefasst als „Hassrede“ und bildet daher die Vielfalt der in unserem Datensatz vorgefundenen Sprache besser ab. Wir verwenden diesen Oberbegriff in der gesamten Analyse, sofern wir uns nicht ausdrücklich auf identitätsbezogenen Hass gegen Einzelpersonen im Sinne unserer ursprünglichen Definition von „Hassrede“ beziehen.

Als Ergebnis der axialen Kodierung wurden mehrere Kategorien zusammengeführt. Bei anderen stellten wir fest, dass sie angesichts der Projektziele nicht ausreichend deutlich ausgeprägt oder relevant waren. Die endgültigen Kernkategorien dieser Lexika sind: „Geschlecht“ [GEN], „Kultur“ [CUL], „Rasse“ [RAC], „sexuelle Orientierung“ [SOR], „Religion“ [REL], „Ableismus“ [ABL], „gewaltvolle Sprache“ [VIO], „Beleidigungen und persönliche Angriffe“ [INS], „vulgäre Sprache“ [VUL], „Ironie und Sarkasmus“ [IRO], „sexuelle Einschüchterung“ [INT], „Politik“ [POL], „Unwahrheit“ [UNT] und „Kriminalität“ [CRI].

Anschließend dienten drei Durchgänge selektiver Kodierung dazu, Schlüsselwörter und Wortfolgen von n Wörtern aus den Daten zu identifizieren, bei denen mindestens drei menschliche annotierende Personen in der Zuordnung zur jeweiligen Kategorie übereinstimmten. Diese Begriffe wurden mit bestehenden Lexika für Hassrede abgeglichen: HADES (Tulkens et al., 2016) und HurtLex v1.2 (Bassignana et al., 2018).

Die daraus resultierenden Listen von Schlüsselwörtern je Kategorie bilden die Grundlage unserer Klassifikationsmodelle und sind als Open Source verfügbar (Van Haperen et al., 2026).

Modellleistung

Wir wenden die Klassifikationsmodelle auf unsere Datensätze an und erhalten die folgenden Ergebnisse, siehe Tabelle 2.

Tabelle 2. Leistung des Modells 2, Ziel: Klassifikation

	Deutsch					Spanisch					Niederländisch				
	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

Anmerkung. Jedes Lexikon wurde ausschließlich anhand von Trainingsdaten erstellt. Der hier berichtete F1 Wert wurde anhand eines unabhängigen Testsatzes evaluiert, der nicht im Trainingsatz enthalten war. Um große Unterschiede in der Zahl der Beobachtungen zu berücksichtigen und Klassenungleichgewichte zu behandeln, wurden strengere Schwellenwerte festgelegt, um die Recall Werte für die größeren Kategorien Politik und Beleidigungen in allen Sprachen, Kriminalität im Spanischen und Deutschen sowie Kultur, Ethnizität, Religion, sexuelle Orientierung und vulgäre Sprache im Deutschen zu verbessern.

Wir berichten F_1 als Maß dafür, wie gut das Modell Beiträge den jeweiligen Kategorien korrekt zuordnet. Die Prävalenz einiger Kategorien ist sehr niedrig. Beispielsweise wird in unserem deutschen Datensatz nur einer von 1.000 Beiträgen als „Ableismus“ kategorisiert. Die Leistungsfähigkeit der Modelle hängt stark mit der Prävalenz zusammen: Seltene Kategorien

liefern weniger Daten für das Training eines Modells, wodurch eine zuverlässige Klassifikation erschwert wird. Generell gilt: Je niedriger der F_1 Wert, desto sorgfältiger müssen wir die Kategorisierung prüfen. F_1 Werte über 0,70 betrachten wir als gut, während Werte unter 0,50 als nicht akzeptabel gelten. Dies führte dazu, dass wir die Kategorien „Ableismus“, „Ironie“ und „Unwahrheit“ aus den weiteren Analysen ausschlossen.

Über die drei Sprachen hinweg erzielten die Kategorien „sexuelle Einschüchterung“ und „Geschlecht“ gute Ergebnisse, da sie über vergleichsweise klare und eindeutige Schlüsselwörter verfügen. Die geringe Prävalenz von „Ableismus“, „Rasse“, „sexueller Orientierung“ und „gewaltvoller Sprache“ könnte dadurch erklärt werden, dass diese Kategorien vermutlich am strengsten moderiert werden. Die Kategorien „Politik“ und „Beleidigungen und persönliche Angriffe“ stellen aufgrund von Kodierungsentscheidungen Ausreißer mit hoher Prävalenz dar. Während der axialen Kodierung wurden viele Unterkategorien unter „Politik“ zusammengeführt, darunter Ideologie, politische Maßnahmen, Politikerinnen und Politiker sowie weitere Themen. Dadurch wurde diese Kategorie zur größten und zugleich am stärksten diffusen Kategorie. Um dies zu berücksichtigen, wurde nachträglich ein strenger Schwellenwert auf diese Kategorie angewendet.

Im Einklang mit unserem abduktiven Kodierungsansatz prüften wir während des gesamten Prozesses kritisch, ob unsere Kategorien weiterhin das widerspiegeln, was wir in den Daten vorfanden. Wir führten umfangreiche manuelle Prüfungen der Daten und der zugeordneten Kategorien für sämtliche Kategorien sowie für alle Plattformen und Länder durch. Dies erfolgte durch manuelle Sichtung und zufällige Stichprobenkontrollen. Insbesondere untersuchten wir falsche positive und falsche negative Klassifikationen manuell und diskutierten Grenzfälle in Kalibrierungssitzungen. In den meisten Fällen führte die manuelle Prüfung zu Anpassungen, insbesondere zum Ausschluss bestimmter Wörter mit mehrdeutiger Bedeutung.

Ein Beispiel ist das Wort „Let“, das im Niederländischen eine Person aus Lettland bezeichnet und deshalb Teil der Kategorie „Kultur“ war, jedoch häufiger in der Formulierung „Let op“ vorkommt, was „Achtung“ bedeutet. Ebenso kommt das Wort „Frans“ im Niederländischen häufiger als Vorname vor, beispielsweise bei Frans Timmermans, als als Bezug auf die französische Kultur. Aufgrund ähnlicher Mehrdeutigkeiten, geringer Modellleistung und Überlegungen im Hinblick auf die Ziele des Projekts beschlossen wir, die Kategorien

„Unwahrheit“ und „Ironie und Sarkasmus“ auszuschließen. Die Kategorie „Ableismus“ trat während der induktiven Kodierung auf, war jedoch in der nachfolgenden Annotierung zu selten, um zuverlässige Ergebnisse zur Modellleistung zu erzielen. Sie wurde daher ebenfalls ausgeschlossen. Diese Kategorien bleiben im Lexikon und in der Methodik enthalten, werden jedoch im Interesse der Validität in der nachfolgenden Analyse nicht berücksichtigt.

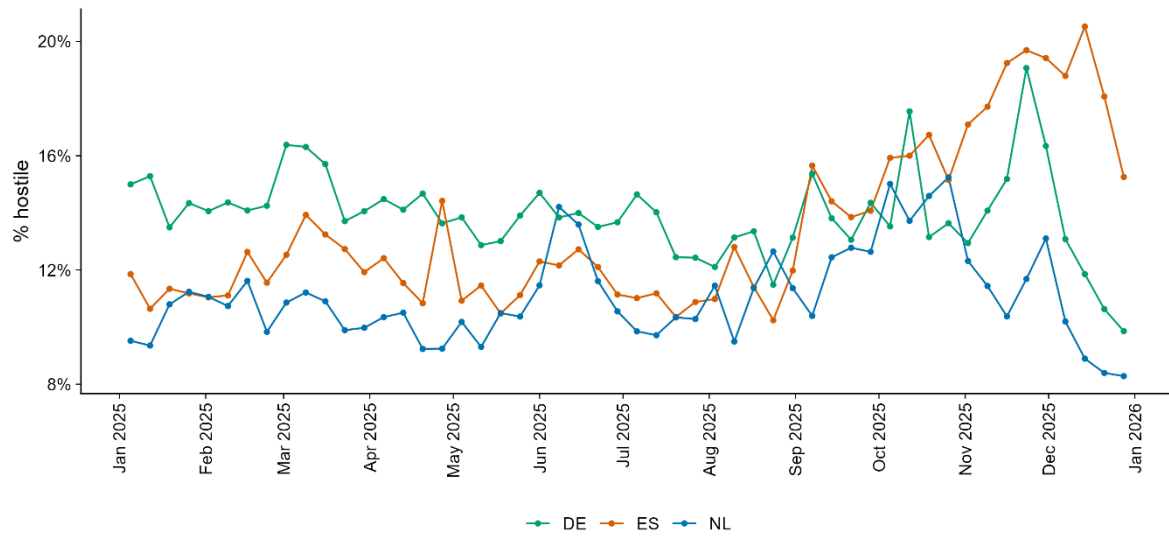
Ergebnisse

Dieser Abschnitt berichtet Folgendes. Erstens liefern Beschreibungen der vierzehn Kategorien qualitative Einblicke in die verschiedenen identifizierten und analysierten Formen von Feindseligkeit im Internet. Zweitens bieten allgemeine deskriptive Ergebnisse einen Überblick über übergreifende Trends, nämlich die Anteile im Zeitverlauf für jedes der Länder, jede der Plattformen und jede Kategorie von Feindseligkeit im Internet. Drittens ermöglicht die Analyse von Adressierung, Beständigkeit und Clusterbildung Einblicke in unterschiedliche Verhaltensweisen der Verbreitenden. Zusammengenommen liefern diese Analysen umsetzbare Erkenntnisse darüber, welche Interventionsformen geeignet sind, um bestimmten Formen von Feindseligkeit im Internet im jeweiligen Kontext entgegenzuwirken.

Muster nach Ländern

Der folgende Abschnitt berichtet über die identifizierten allgemeinen Sprachmuster im Vergleich der drei Länder, siehe Abbildung 1.

Abbildung 1. Umfang der Feindseligkeit im Internet im Zeitverlauf für Spanien, Deutschland und die Niederlande



Anmerkung: Prozentsatz der als Online Feindseligkeit identifizierten Beiträge, aggregiert nach Woche und Land.

In unserem Datensatz identifizierten wir 13 % aller Beiträge als Beiträge, die Feindseligkeit im Internet enthalten, wobei der typische Tageswert bei etwa 8,8 % liegt. Wir betonen, dass wir Konten ausgewählt haben, die mit hoher Wahrscheinlichkeit Feindseligkeit erhalten. Dies beeinflusste auch unsere Annotierung. Der hier berichtete Anteil ist daher weder für die drei Länder noch für die Plattformen insgesamt repräsentativ und überschätzt das Ausmaß von Feindseligkeit im Internet wahrscheinlich.

Die Differenz zwischen Mittelwert und Median, die Ausschläge an bestimmten Tagen sowie die Variabilität der Daten mit $SD = 1543$, $min = 1$ und $max = 21.396$ legen nahe, dass Feindseligkeit im Internet durch Phasen erhöhter Aufmerksamkeit geprägt ist. Diese stehen wahrscheinlich mit Ereignissen außerhalb des Internets, plattformspezifischen Dynamiken oder koordinierten Aktivitäten in Zusammenhang. Ein großer Teil der feindseligen Inhalte in unserem Datensatz konzentriert sich auf bestimmte Tage, während auf jeder Plattform zugleich ein beständiges Grundniveau an Aktivität besteht. Diese Phasen feindseliger Inhalte bündeln sich um bestimmte Zeitpunkte und nicht um zufällig auftretende Einzeltage. Dies legt nahe, dass diese Aktivität mit spezifischen Nachrichtenereignissen zusammenhängt, die anhaltende Aufmerksamkeit erzeugen.

Zwischen den Ländern bestehen deutliche Unterschiede in den verschiedenen Kategorien der Feindseligkeit im Internet, siehe Tabelle 3.

Tabelle 3. Umfang der Feindseligkeit im Internet nach Land, alle Plattformen

Category	Total n	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

Deskriptive Statistiken der täglichen Prozentwerte (Anzahl der Beiträge in einer Kategorie als Anteil an der Gesamtzahl der täglich als hasserfüllt identifizierten Beiträge).

Aus diesem Vergleich ergeben sich mehrere Erkenntnisse. Feindseligkeit im Zusammenhang mit Politik, Beleidigungen und Kriminalität stellt in jeder Sprache die häufigste Kategorie dar. Im Deutschen kommen vulgäre Sprache und, in geringerem Maße, Feindseligkeit im Zusammenhang mit sexueller Orientierung relativ häufiger vor als im Niederländischen oder Spanischen. Im Niederländischen ist Feindseligkeit im Zusammenhang mit Kultur relativ häufiger als in den anderen Sprachen. Feindseligkeit im Zusammenhang mit Geschlecht, Rasse, Religion, sexueller Einschüchterung, gewaltvoller Sprache und vulgärer Sprache tritt in unserem Datensatz in allen Sprachen relativ weniger deutlich hervor.

Die Mediane unterscheiden sich merklich zwischen den Ländern. Dies legt nahe, dass der Anteil feindseliger Inhalte in den meisten Kategorien in den Niederlanden über die Zeit beständiger auftritt, während er in Deutschland und Spanien stärker durch Aktivitätsspitzen geprägt ist. Dies wird zusätzlich durch die vergleichsweise hohen Standardabweichungen in

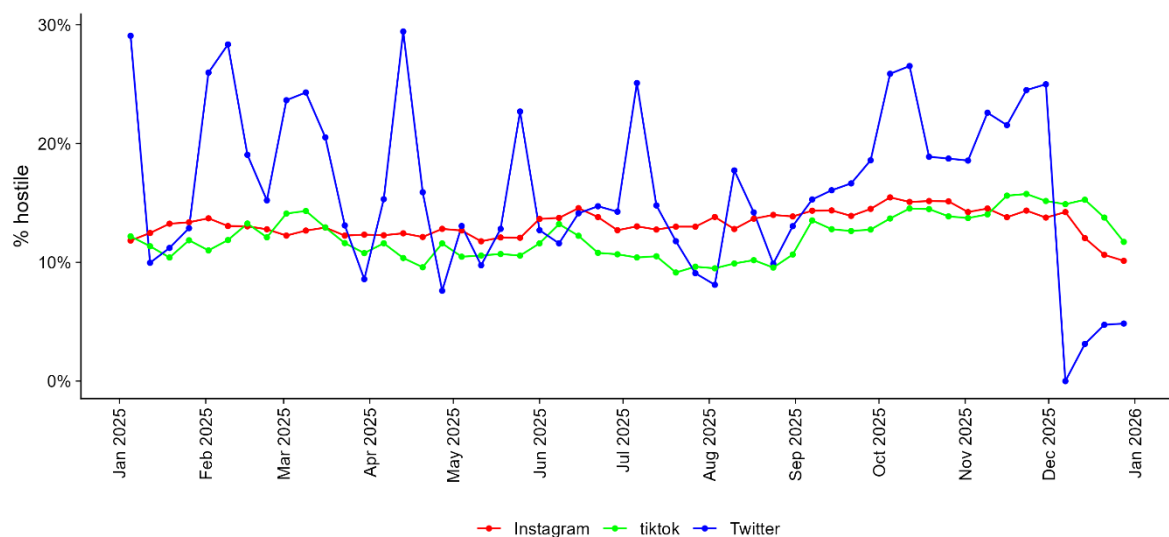
Kategorien wie Politik in Spanien gestützt. Kategorien wie gewaltvolle Sprache und sexuelle Einschüchterung treten über die Zeit beständiger auf, während Politik und Beleidigungen stärker durch zeitlich gebündelte Aktivität gekennzeichnet sind. Dies gilt besonders für Spanien. Insgesamt deutet dies auf Aktivitätsspitzen hin, die eng mit aktuellen Ereignissen, insbesondere Nachrichtenereignissen, verbunden sein könnten. Diese Dynamik ist in Spanien am stärksten ausgeprägt.

Im folgenden Abschnitt dieses Berichts untersuchen wir diese Unterschiede zwischen den Ländern für jede Kategorie genauer.

Muster nach Plattform

Der folgende Abschnitt berichtet über die identifizierten allgemeinen Sprachmuster im Vergleich der drei Plattformen, siehe Abbildung 2.

Abbildung 2. Umfang der Feindseligkeit im Internet im Zeitverlauf nach Plattform



Anmerkung: Prozentsatz der als Online Feindseligkeit identifizierten Beiträge, aggregiert nach Woche und Plattform.

Zwischen den Plattformen zeigen sich deutliche Unterschiede. In unserem Datensatz sieht man, dass X den höchsten Anteil an Feindseligkeit im Internet zeigt, und durch ein im Zeitverlauf stark schwankendes Modell hervor. Instagram hat höhere Anteile von

Feindseligkeit im Internet als TikTok. Beide Plattformen erscheinen jedoch über die Zeit relativ stabil, ohne einen klaren zunehmenden oder abnehmenden Trend. Diese Entwicklungen werden durch Aktivitätsschübe an bestimmten Tagen in Verbindung mit einem beständigen Aktivitätsniveau auf jeder Plattform geprägt.

Zwischen den Ländern bestehen deutliche Unterschiede in den verschiedenen Kategorien der Feindseligkeit im Internet, siehe Tabelle 4.

Tabelle 4. Anteile nach Plattform, alle Länder

Kategorie	Total n	Instagram				X				TikTok			
		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

Deskriptive Statistiken der Prozentwerte (Anzahl der Beiträge in einer Kategorie als Anteil an der Gesamtzahl der täglich als hasserfüllt identifizierten Beiträge). Beachten Sie, dass die Kategorien nicht gegenseitig ausschließend sind, da ein einzelner Beitrag beispielsweise sowohl der Kategorie „Politik“ als auch der Kategorie „Beleidigungen“ zugeordnet werden kann.

In unserem Datensatz zeigt X die höchsten Anteile von Feindseligkeit im Internet über all Kategorien auf, insbesondere in den Kategorien Politik und Beleidigungen. TikTok zeigt dagegen durchgängig die niedrigsten Anteile, während Instagram im Allgemeinen eine mittlere Position einnimmt. Eine bemerkenswerte Ausnahme bilden geschlechtsbezogene Inhalte, die auf Instagram am deutlichsten hervortreten.

Im Vergleich zu Instagram und TikTok weist X in 11 der 13 Kategorien die höchsten durchschnittlichen Anteile auf. Dazu gehören Politik mit 15,0 %, Beleidigungen mit 10,3 %, Kriminalität mit 4,7 %, Kultur mit 4,1 %, Religion mit 2,3 %, Rasse mit 2,1 % und vulgäre Sprache

mit 1,9 %. Instagram nimmt im Allgemeinen eine mittlere Position ein. Während die meisten Kategorien im Vergleich zu X weniger ausgeprägt sind, übersteigt Instagram X bei geschlechtsbezogener Feindseligkeit mit 2,1 % gegenüber 1,3 %. Geschlecht ist damit die einzige Kategorie, in der Instagram eindeutig den höchsten durchschnittlichen Anteil aufweist. TikTok hat über alle Kategorien hinweg durchgängig die niedrigsten Anteile. Die durchschnittlichen Anteile sind bei den meisten Formen feindseliger Inhalte am niedrigsten, insbesondere bei Religion mit 0,5 %, sexueller Einschüchterung mit 0,1 %, gewaltvoller Sprache mit 0,1 % und vulgärer Sprache mit 0,3 %. Politik bleibt auf TikTok die häufigste Kategorie, liegt jedoch bei weniger als der Hälfte des Anteils auf X.

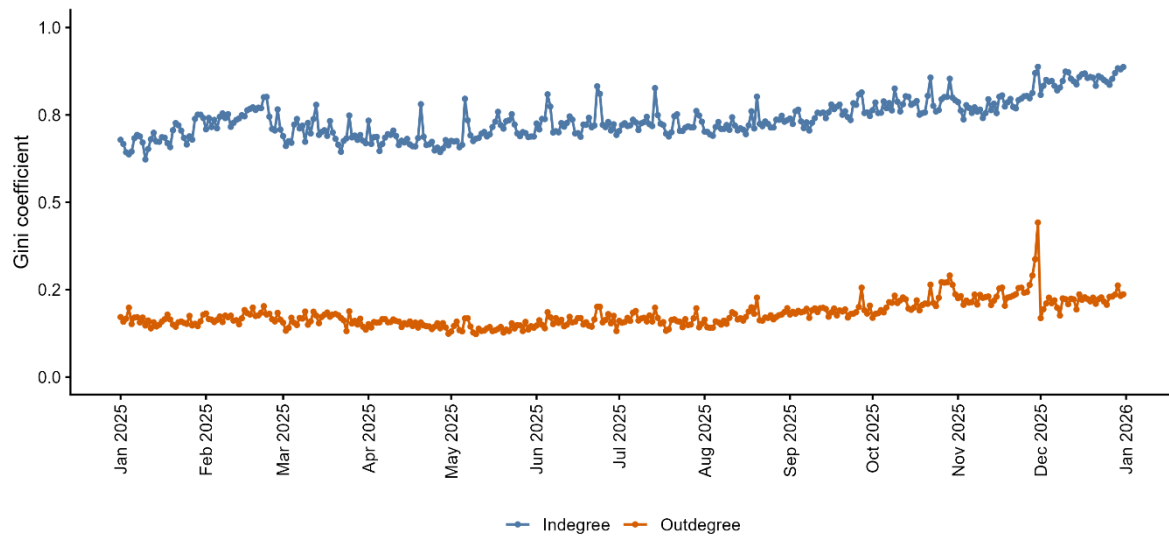
Die Mediane deuten darauf hin, dass Aktivitätsschübe auf den meisten Plattformen einen großen Anteil der feindseligen Inhalte erzeugen. Die Mediane auf X für Politik mit 16,4 %, Beleidigungen mit 11,3 %, Kultur mit 4,1 % und Kriminalität mit 4,0 % legen nahe, dass diese Themen über die Zeit ein beständiges Grundniveau aufweisen. Die Variation ist ebenfalls bei Politik und Beleidigungen am größten. Die Standardabweichungen sind in diesen Kategorien auf allen Plattformen am höchsten, insbesondere auf X mit $SD = 12,3$ für Politik und $SD = 9,4$ für Beleidigungen. Dies weist auf erhebliche Schwankungen im Zeitverlauf hin.

Insgesamt deuten diese Ergebnisse auf deutliche Unterschiede in der Zusammensetzung von Feindseligkeit im Internet zwischen den Plattformen hin. In unserem Datensatz ist X sowohl durch höhere Anteile als auch durch stärkere Schwankungen zwischen den Kategorien gekennzeichnet, insbesondere bei Beleidigungen und politikbezogener Feindseligkeit. Instagram zeigt dagegen relativ niedrigere Anteile, jedoch einen höheren Anteil geschlechtsbezogener Feindseligkeit, sowie eine größere Beständigkeit im Zeitverlauf. TikTok weist bei den meisten von uns erfassten Kategorien feindseliger Inhalte die niedrigsten Anteile auf.

Muster nach Zielpersonen und Verbreitenden

Der folgende Abschnitt berichtet über allgemeine Muster bei den Zielpersonen und Verbreitenden von Feindseligkeit im Internet. Zunächst konzentrieren wir uns darauf, wie stark die Aufmerksamkeit auf bestimmte Zielpersonen gebündelt ist, siehe Abbildung 3.

Abbildung 3. Konzentration der Feindseligkeit im Internet bei Zielpersonen, Eingangsgrad, und Verbreitenden, Ausgangsgrad



Diese Abbildung zeigt, wie sich die Aufmerksamkeit auf Zielpersonen und Verbreitende von Feindseligkeit im Internet verteilt. Ein höherer Gini Koeffizient weist auf eine stärkere Ungleichheit zwischen Zielpersonen oder Verbreitenden hin. Dies bedeutet, dass eine vergleichsweise kleine Gruppe von Konten einen unverhältnismäßig großen Anteil der Beiträge erhält oder versendet. Die zentrale Erkenntnis ist, dass Phasen auftreten, in denen sich feindselige Aufmerksamkeit stärker konzentriert und eine geringe Zahl von Zielpersonen einen unverhältnismäßig großen Anteil aller feindseligen Inhalte in unserem Datensatz erhält. Dies deutet darauf hin, dass bestimmte Ereignisse eine stärkere Fokussierung auf spezifische Zielpersonen erzeugen.

Abbildung 3 zeigt außerdem, wie stark Feindseligkeit im Internet bei den Verbreitenden konzentriert ist. Ein höherer Gini Koeffizient bedeutet hier, dass ein kleinerer Anteil der verbreitenden Konten für einen relativ großen Anteil der in unserem Datensatz identifizierten feindseligen Inhalte verantwortlich ist. Die zentrale Erkenntnis ist der insgesamt geringe Konzentrationsgrad. Dies weist darauf hin, dass Angriffe von vielen unterschiedlichen Konten ausgehen und nicht von einer kleinen Anzahl von Konten, die den Großteil der feindseligen Inhalte in unserem Datensatz erzeugen.

Ein überraschendes Ergebnis besteht darin, dass die Konzentration sowohl bei den Zielpersonen als auch bei den Verbreitenden zu ähnlichen Zeitpunkten zunimmt. Darüber hinaus zeigen sich Ähnlichkeiten in der allmählichen Zunahme der Konzentration bei

Verbreitenden und Zielpersonen, wobei dieser Trend bei den Verbreitenden weniger ausgeprägt ist. Wir vermuten, dass Angriffe auf eine Person konzentrierter erfolgen, wenn diese in der Medienberichterstattung stärker in den Mittelpunkt rückt. Gleichzeitig erzeugt dann ein vergleichsweise kleinerer Anteil von Konten einen größeren Anteil der Feindseligkeit im Internet. Dies werten wir als weiteren Hinweis darauf, dass feindselige Inhalte mit aktuellen Ereignissen, insbesondere Nachrichtenereignissen, zusammenhängen.

Verhaltensmuster

Um die Verbreitenden von Feindseligkeit im Internet weiter zu differenzieren, untersuchen wir ihr Verhalten hinsichtlich fokussierter Adressierung und Beständigkeit. Durch den Vergleich der Verbreitenden anhand von „Adressierung“, also der Anzahl von Zielpersonen, an die jemand feindselige Beiträge sendet, wobei drei oder weniger Zielpersonen als Maß für Fokussierung gelten, und „Beständigkeit“, also der Anzahl von Tagen, an denen jemand feindselige Beiträge gegen dieselbe Zielperson richtet, wobei drei oder mehr Tage innerhalb eines Jahres als Maß für Beständigkeit gelten, unterscheiden wir vier Verhaltensformen, siehe Tabelle 5.

Tabelle 5. Verhaltenscluster unter Verbreitenden: Fokussierung und Beständigkeit

Cluster	Typ	Durchschnittlicher Zielfokus	Durchschnittliche Persistenz	n
1	Generalisten	11.8	1.05	863
2	Fixierte	1	9.33	3
3	Wiederholungstäter	2.98	2.21	174
4	Opportunisten	1.61	1.00	14669

Ein erstes Cluster besteht aus einer großen Gruppe von „Opportunisten“. Dabei handelt es sich um Personen mit geringer Aktivität, die 94 % der Konten ausmachen, welche in unserem Datensatz problematische Inhalte verbreiten. Diese Gruppe umfasst Konten mit kaum ausgeprägter Adressierung und ohne Beständigkeit: Sie veröffentlichen lediglich einzelne feindselige Beiträge, die sich gegen eine oder zwei Zielpersonen richten.

Ein zweites Cluster ist eine mittelgroße Gruppe von Generalisten mit 863 Konten. Dabei handelt es sich um aggressive Konten, die viele unterschiedliche Zielpersonen adressieren,

jedoch nur sehr geringe Beständigkeit aufweisen. Es sind einmalige Ausfälle, die sich gleichzeitig gegen viele verschiedene Personen richten.

Ein drittes Cluster wiederholt auffällige Personen ist klein und umfasst 174 Konten. Es weist eine mässige Adressierung und eine mässige Beständigkeit auf. Diese Personen befassen sich über längere Zeiträume mit unterschiedlichen Themen und verwenden dabei wiederholt problematische Sprache.

Ein letztes Cluster besteht aus fixierten Personen. Dabei handelt es sich um extreme Ausreißer, die sich über mehr als eine Woche hinweg auf eine einzige Zielperson konzentrieren. Entgegen unseren Erwartungen sind solche Konten in unserem Datensatz sehr selten. Eine mögliche Erklärung ist eine wirksame Plattformmoderation: Besonders problematische Konten könnten gemeldet und gesperrt werden, wodurch sie von unserer Datenerhebung ausgeschlossen würden.

Muster nach Kategorien

Beschreibung der Kategorien

Die folgenden Kategorien sind das Ergebnis mehrerer Durchgänge induktiver Kodierung. Es ist zu beachten, dass diese Kategorien nicht erschöpfend sind. Auch wenn sie die für die Annotierung aus unserem Datensatz gezogenen Stichproben abdecken, ist davon auszugehen, dass Daten aus anderen Ländern, von anderen Plattformen oder aus anderen Zeiträumen eine Erweiterung dieser Liste erfordern werden. Darüber hinaus bleibt ein gewisser Interpretationsspielraum bestehen. Die Kategorien können Sprache enthalten, die von unseren annotierenden Personen als feindselig identifiziert wurde, während andere Lesende sie möglicherweise nicht als feindselig bewerten würden. Zudem schließen sich diese Kategorien nicht gegenseitig aus: Ein Beitrag kann sowohl als „vulgäre Sprache“ als auch als „Beleidigung“ kategorisiert werden.

Die Gesamtmenge der Feindseligkeit im Internet kann wie folgt weiter nach Kategorien differenziert werden, siehe Tabelle 6.

Tabelle 6. Deskriptive Statistiken nach Kategorie, alle Plattformen und Länder

Kategorie	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

Deskriptive Statistiken der Prozentwerte (Anzahl der Beiträge in einer Kategorie als Anteil an der Gesamtzahl der täglich als hasserfüllt identifizierten Beiträge).

Erstens sind mehrere Kategorien primär ereignisbezogen. Diese Kategorien weisen eine hohe Schwankung auf: „Rasse“ mit 0,9 % und $SD = 0,968$, „sexuelle Einschüchterung“ mit 0,3 % und $SD = 0,291$ sowie „gewaltvolle Sprache“ mit 0,3 % und $SD = 0,307$. Solche Formen der Feindseligkeit lassen sich möglicherweise nur schwer vorhersehen und könnten von schnellen Reaktionen profitieren, da sie häufig stark auf wenige Zielpersonen konzentriert sind und überwiegend von nur vorübergehend aktiven Nutzenden verbreitet werden.

Zweitens verbinden die Kategorien „Politik“, „Beleidigungen“, „Kriminalität“ und „Kultur“ ein beständiges Grundniveau mit Schwankungen. Diese Kategorien machen in unserem Datensatz den größten Anteil der Feindseligkeit im Internet aus. Auf Politik sowie Beleidigungen und persönliche Angriffe entfallen täglich jeweils 8,3 % beziehungsweise 5,1 % aller Formen von Feindseligkeit im Internet. Kriminalität mit 2,7 % und $SD = 1,100$ weist eine bemerkenswert hohe Variabilität auf. Dies legt nahe, dass diese Kategorie primär ereignisbezogen ist. Bei Kultur mit 1,7 % und $SD = 0,970$ ist die Variabilität vergleichsweise niedrig. Dies deutet darauf hin, dass solche Inhalte täglich beständig auftreten. Da Median und Modalwert bei Politik und Beleidigungen dem Mittelwert ähneln, weisen beide Kategorien ein stärkeres tägliches Grundniveau auf. Zugleich deuten die Standardabweichungen auf eine relativ hohe

Variabilität bei Aktivitätsspitzen hin. Dies legt nahe, dass feindselige Inhalte in diesen Kategorien sowohl durch ein stabiles Grundniveau an Aktivität als auch durch episodische Aktivitätsspitzen geprägt sind.

Drittens sind Kategorien mit stabilem Grundniveau weniger stark durch Ereignisse geprägt und treten über die Zeit beständiger auf. Dazu gehören „Geschlecht“ mit 1,4 % und $SD = 0,823$, „Religion“ mit 0,7 % und $SD = 0,509$, „sexuelle Orientierung“ mit 0,8 % und $SD = 0,803$ sowie „vulgäre Sprache“ mit 0,8 % und $SD = 0,540$. Die Standardabweichungen weisen hier auf eine vergleichsweise geringe Variabilität hin. Dies legt nahe, dass solche Inhalte täglich beständig auftreten, statt stark ereignisbezogen zu sein. Diese Kategorien bilden ein strukturelles tägliches Grundniveau feindseliger Inhalte. Sie könnten von kontinuierlicher Beobachtung wiederkehrender Zielpersonen profitieren, anstatt von reaktiven Interventionen oder Maßnahmen, die sich gegen wiederholt auffällige Personen richten.

Zwei wichtige Einschränkungen müssen hervorgehoben werden. Erstens haben wir, wie im Methodikteil erläutert, Daten aus sozialen Medien gesammelt, die zum Zeitpunkt der Datensammlung noch öffentlich verfügbar waren. Beiträge waren somit bereits einer Plattformmoderation unterzogen worden, bevor sie in unseren Datensatz aufgenommen wurden. Es ist plausibel, dass die Kategorien der ersten Gruppe am stärksten moderiert werden und deshalb in unserer Analyse vergleichsweise seltener und stärker schwankend auftreten. Zweitens betonen wir, dass Prävalenz nicht mit Wirkung gleichzusetzen ist. Beiträge in diesen Kategorien enthalten die schwerwiegendsten Formen von Feindseligkeit im Internet in unserem Datensatz. Es kann daher plausibel angenommen werden, dass diese Kategorien unverhältnismäßig starke Auswirkungen haben, auch wenn sie seltener auftreten.

Insgesamt machen Politik und persönliche Beleidigungen den größten Anteil der von unseren Modellen als problematisch markierten Beiträge aus. Dies legt nahe, dass Feindseligkeit im Internet über Länder und Plattformen hinweg nicht zufällig auftritt, sondern strukturelle Merkmale aufweist: Einige Formen sind stark ereignisbezogen, andere weisen ein beständigeres Grundniveau auf. Dies kann dabei helfen zu entscheiden, in welche Interventionsformen gegen Feindseligkeit im Internet Ressourcen investiert werden sollten.

Im Folgenden untersuchen wir für jede Kategorie die Zielpersonen sowie das Verhalten der Verbreitenden.

Kriminalität

Diese Kategorie umfasst Sprache über kriminelles Verhalten, illegale Aktivitäten und die Polizei.

- **Spanisches Beispiel:** „quien venga a delinquir fuera“.
- **Deutsches Beispiel:** „Das sind kriminell die wo andere Mitmenschen gefährden die müssen alle in Therapie und weggesperrt werden !!!“.
- **Niederländisches Beispiel:** „Die gore crimineel zou harder gestraft moeten worden“.

Inhalte mit Bezug zu Kriminalität treten in unserem Datensatz mäßig häufig auf. Auf Instagram zeigt sich ein allmählicher Anstieg. Auf TikTok sind die Werte variabler und nach frühen Spitzen insgesamt leicht rückläufig. Auf X ist die Kategorie schwankend, ohne einen anhaltenden Trend. Im Ländervergleich handelt es sich in Deutschland, verglichen mit den anderen Kategorien, um eine relativ hervortretende Form von Inhalten, in Spanien um eine mäßig hervortretende und in den Niederlanden um eine der weniger hervortretenden Formen.

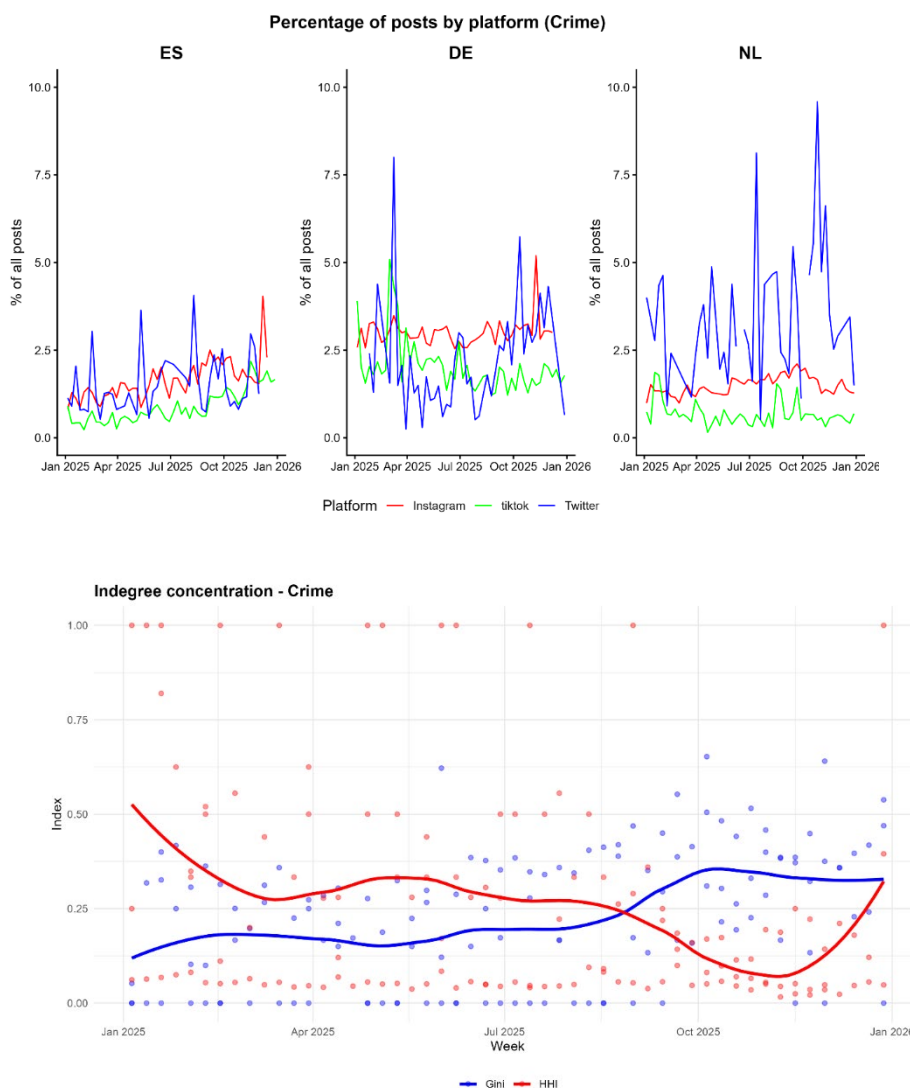
Hinsichtlich der Zielpersonen stellen wir eine moderate Konzentration in Verbindung mit einer relativ stabilen Gruppe von Zielpersonen fest, mit einem mittleren Gini Koeffizienten von 0,23 und $HHI = 0,25$. Die Konzentration bleibt im Zeitverlauf beständig, ohne deutliche systematische Zu oder Abnahme. Gegen Ende des Jahres sinkt der HHI parallel zu einem Anstieg der Aktivität. Dies weist auf schwankende Aufmerksamkeit über die Zeit hin. Die Zielpersonen dieser Kategorie bleiben jedoch über die Zeit vergleichsweise beständig konzentriert. Dies deutet auf ein Muster wechselnder Aufmerksamkeit innerhalb einer sich verändernden Gruppe von Zielpersonen hin, nicht auf eine anhaltende Fokussierung auf bestimmte Konten.

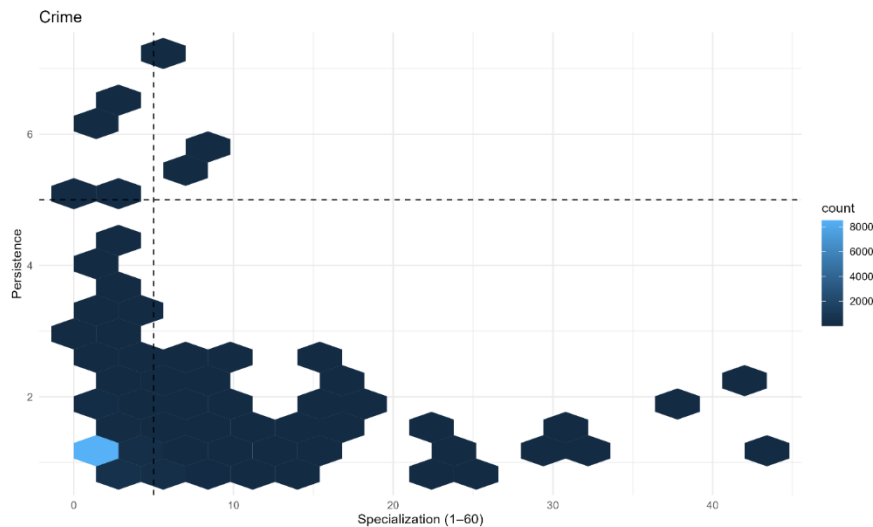
Verbreitende von Inhalten mit Bezug zu Kriminalität sind im Vergleich zu anderen Kategorien in unserem Datensatz ähnlich stark konzentriert. 95 % zeigen fokussierte Adressierung bei geringer Beständigkeit. Zugleich weist die Kategorie mit 4,9 % einen der höchsten Anteile stärker gestreuter Adressierung bei geringer Beständigkeit auf. Der Anteil fokussierter Adressierung bei hoher Beständigkeit ist mit 0,1 % vernachlässigbar. Dies deutet darauf hin,

dass Feindseligkeit im Internet im Zusammenhang mit Kriminalität nur selten über längere Zeit aufrechterhalten wird und mit einer relativ breiten Adressierung einhergeht.

Hinsichtlich möglicher Interventionen legen diese Ergebnisse nahe, dass Feindseligkeit im Internet im Zusammenhang mit Kriminalität relativ häufig und über die Zeit beständig auftritt. Sie richtet sich gegen eine relativ große, aber stabile Gruppe von Zielpersonen und geht von einer vorübergehend aktiven und wenig fokussierten Gruppe von Verbreitenden aus. Dies macht sie zu einer der stärker gestreuten Kategorien, bei denen Interventionen schwieriger sein könnten, da es weniger Hinweise auf vorhersehbare Ereignisse gibt.

Abbildung 4. Vergleich kriminalitätsbezogener Feindseligkeit: Anteile, Zielpersonen und Verbreitende





Kultur

Diese Kategorie umfasst Sprache über spezifische Kulturen und kulturelle Stereotype.

- Spanisches Beispiel: „Menos clase que un gitano recién llegado de Rumanía“.
- Deutsches Beispiel: „Sofort auch raus mit den Migranten !!!! Keiner will euch !!!!“.
- Niederländisches Beispiel: „smerige halve Marokkaan“.

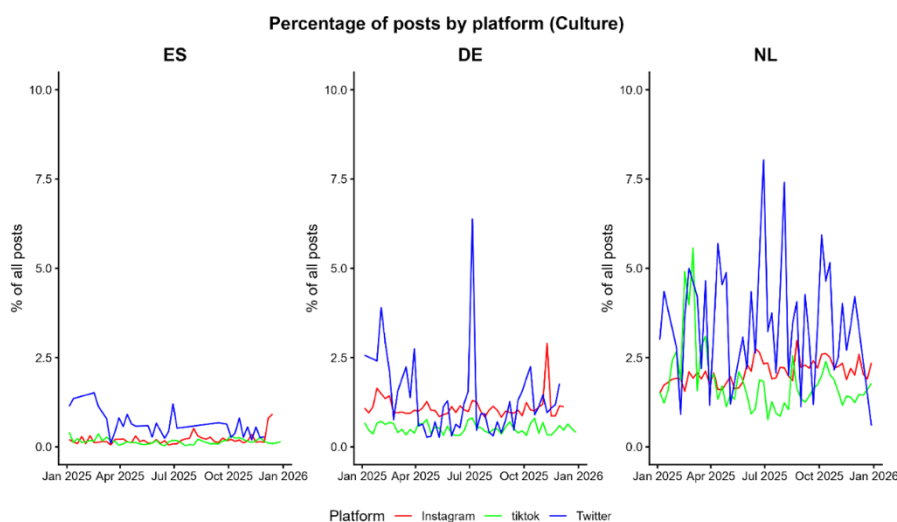
Inhalte mit Bezug zu Kultur treten in unserem niederländischsprachigen Datensatz sehr häufig auf, jedoch deutlich seltener in unseren spanischsprachigen und deutschsprachigen Datensätzen. In den Niederlanden kommen sie auf allen drei Plattformen beständig vor und sind, verglichen mit anderen Kategorien, sowohl auf Instagram als auch auf TikTok relativ deutlich ausgeprägt. Auf X tritt Feindseligkeit im Internet im Zusammenhang mit Kultur häufig und in Aktivitätsspitzen auf.

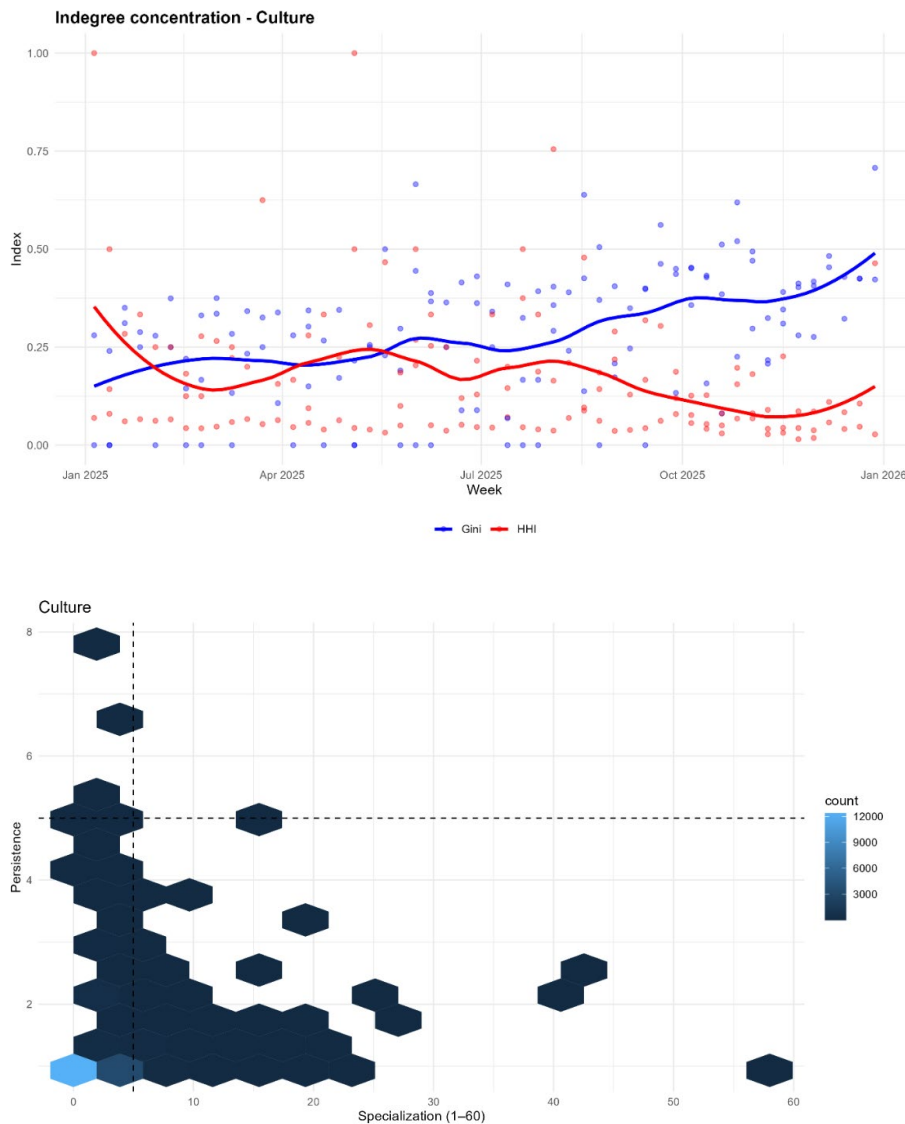
Hinsichtlich der Zielpersonen zählt Feindseligkeit im Internet im Zusammenhang mit Kultur zu den Kategorien mit der stärksten Ungleichverteilung, bei zugleich relativ geringer Dominanz einzelner Zielpersonen. Der mittlere Gini Koeffizient beträgt 0,28 und $HHI = 0,17$. Dies weist auf eine konzentrierte Gruppe von Zielpersonen hin, die sich relativ häufig verändert. Beide Modelle bleiben über die Zeit beständig. Dies deutet darauf hin, dass sich diese Form der Feindseligkeit ungleich auf Zielpersonen verteilt, ohne dass dauerhaft zentrale Zielpersonen bestehen.

Die Verbreitenden kulturbezogener Feindseligkeit im Internet gehören im Vergleich zu anderen Kategorien in unserem Datensatz zu den am stärksten fokussierten und nur vorübergehend aktiven Gruppen. 97,4 % zeigen fokussierte Adressierung bei geringer Beständigkeit. Nur 2,6 % der Nutzenden richten sich an unterschiedliche Zielpersonen, ohne dass sich Hinweise auf Beständigkeit im Zeitverlauf zeigen. Dies legt nahe, dass diese Verbreitenden stark fragmentiert sind und sich lediglich mit einer einzelnen Botschaft befassen.

Hinsichtlich möglicher Interventionen legen diese Ergebnisse nahe, dass Feindseligkeit im Internet im Zusammenhang mit Kultur im Niederländischen relativ häufig ist und auf allen Plattformen beständig auftritt, im Spanischen und Deutschen jedoch deutlich seltener vorkommt. Länderübergreifend ist die Aufmerksamkeit in dieser Kategorie ungleich verteilt, weist jedoch eine größere Streuung zwischen den Zielpersonen auf. Sie geht von einer sehr vorübergehend aktiven, aber vergleichsweise fokussierten Gruppe von Verbreitenden aus. Damit gehört Kultur zu den stärker gestreuten Kategorien, bei denen Interventionen schwieriger sein könnten, da es weniger Hinweise auf vorhersehbare Ereignisse gibt.

Abbildung 5. Vergleich kulturbezogener Feindseligkeit: Anteile, Zielpersonen und Verbreitende





Geschlecht

Diese Kategorie umfasst intolerante Sprache, die sich gegen spezifische Aspekte geschlechtlicher Identität und geschlechtliche Stereotype richtet. Sie beinhaltet Geschlechternormen und Frauenfeindlichkeit.

- **Spanisches Beispiel:** „sin maquillaje,es una bruja“.
- **Deutsches Beispiel:** „Ich kann diese dumme Hexe nicht mehr ertragen.“.
- **Niederländisches Beispiel:** „Alle vrouwen zijn dom“.

Geschlechtsbezogene Feindseligkeit im Internet tritt in unserem Datensatz auf allen Plattformen und in allen Ländern selten auf. Im Niederländischen ist sie auffallend selten, während sie im Spanischen deutlich häufiger vorkommt. Verglichen mit anderen Kategorien

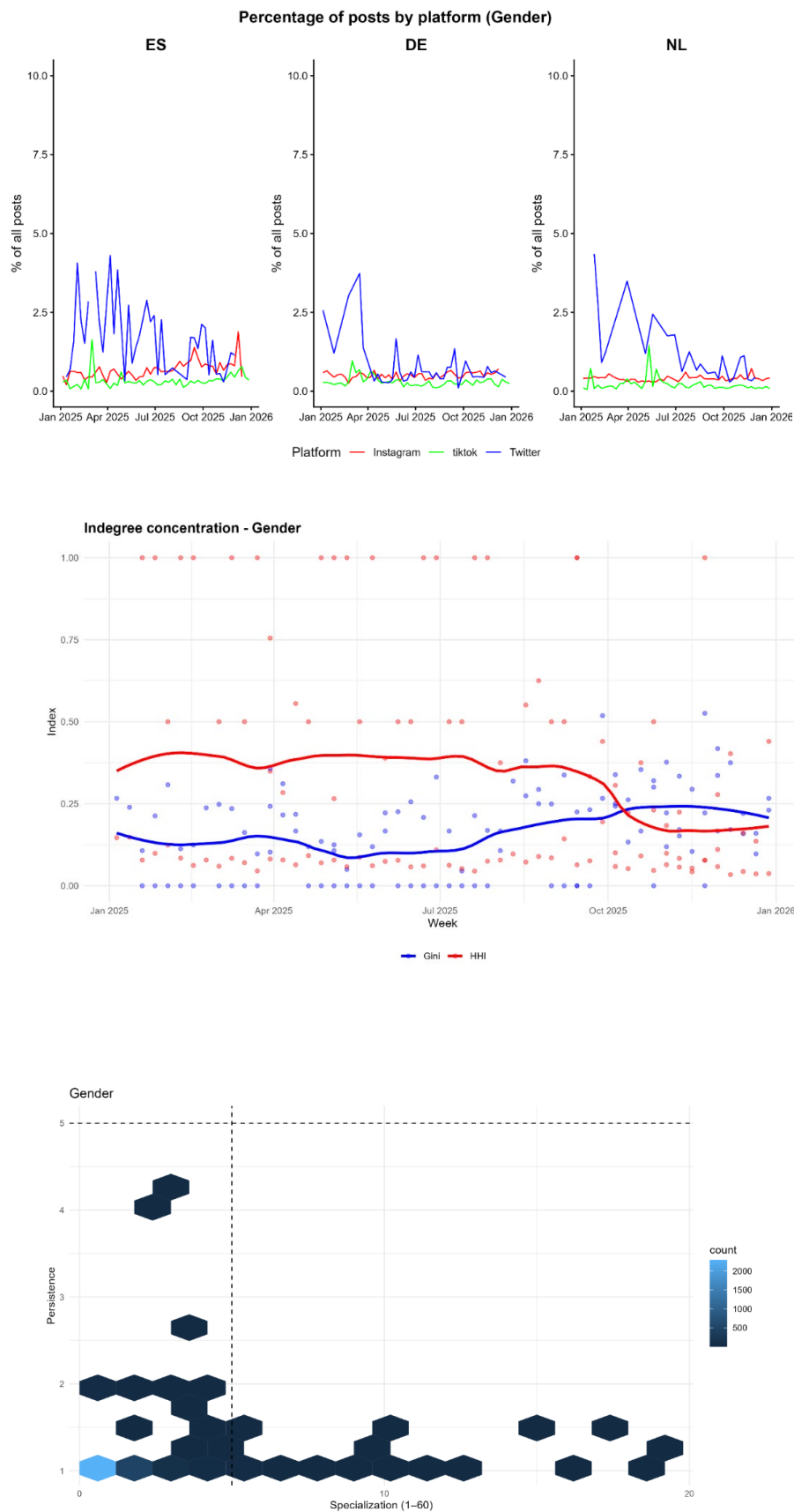
sind die Schwankungen auf X ebenfalls deutlich geringer. Dieser relativ stabile Trend bleibt über die Zeit bestehen.

Hinsichtlich der Zielpersonen ist geschlechtsbezogene Feindseligkeit im Internet relativ breit verteilt, bei zugleich relativ hoher Dominanz. Der mittlere Gini Koeffizient beträgt 0,16 und $HHI = 0,32$. Dies legt nahe, dass die Ungleichverteilung begrenzt ist, sich die Aufmerksamkeit jedoch auf einige zentrale Zielpersonen konzentriert. Dieser Trend bleibt über die Zeit beständig.

Hinsichtlich der Verbreitenden ist Geschlecht in unserem Datensatz insgesamt die am stärksten konzentrierte Kategorie. 98,3 % zeigen fokussierte Adressierung bei geringer Beständigkeit, bei minimaler Diversifizierung von 1,7 % und ohne beständig aktive Nutzende. Dies legt nahe, dass diese Form der Feindseligkeit stark auf bestimmte Zielpersonen ausgerichtet ist und überwiegend aus einzelnen Interaktionen mit nahezu keiner Beständigkeit besteht.

Hinsichtlich möglicher Interventionen ist Geschlecht eine Kategorie mit einem beständigen Grundniveau an Aktivität, die im Spanischen mehr Aufmerksamkeit erfordert als in den anderen Sprachen. Länderübergreifend ist die Aufmerksamkeit in dieser Kategorie relativ breit verteilt, weist jedoch einen vergleichsweise stabilen Kern von Zielpersonen auf. Sie geht von einer vorübergehend aktiven und fokussierten Gruppe von Verbreitenden aus. Interventionen in dieser Kategorie können sich weniger an ereignisbezogener Aufmerksamkeit orientieren, sondern eher an hervortretenden Konten, die beständig solche Feindseligkeit auf sich ziehen.

Abbildung 6. Vergleich geschlechtsbezogener Feindseligkeit: Anteile, Zielpersonen und Verbreitende



Beleidigungen und persönliche Angriffe

Diese Kategorie umfasst ein breites Spektrum an Sprache, die darauf abzielt, eine Person zu beleidigen oder persönlich anzugreifen, ohne ausdrücklich auf Gruppenidentitäten Bezug zu nehmen. Es handelt sich um eine weit gefasste Kategorie, die nicht nur Feindseligkeit im Internet, sondern auch allgemeiner Respektlosigkeit und Unhöflichkeit umfasst, einschließlich grober, beleidigender und herabsetzender Sprache.

- **Spanisches Beispiel:** „Eres amiga de este no hay mas preguntas miserable“.
- **Deutsches Beispiel:** „Du bist so erbärmlich.“.
- **Niederländisches Beispiel:** „Wat ben jij een ontzettende eikel“.

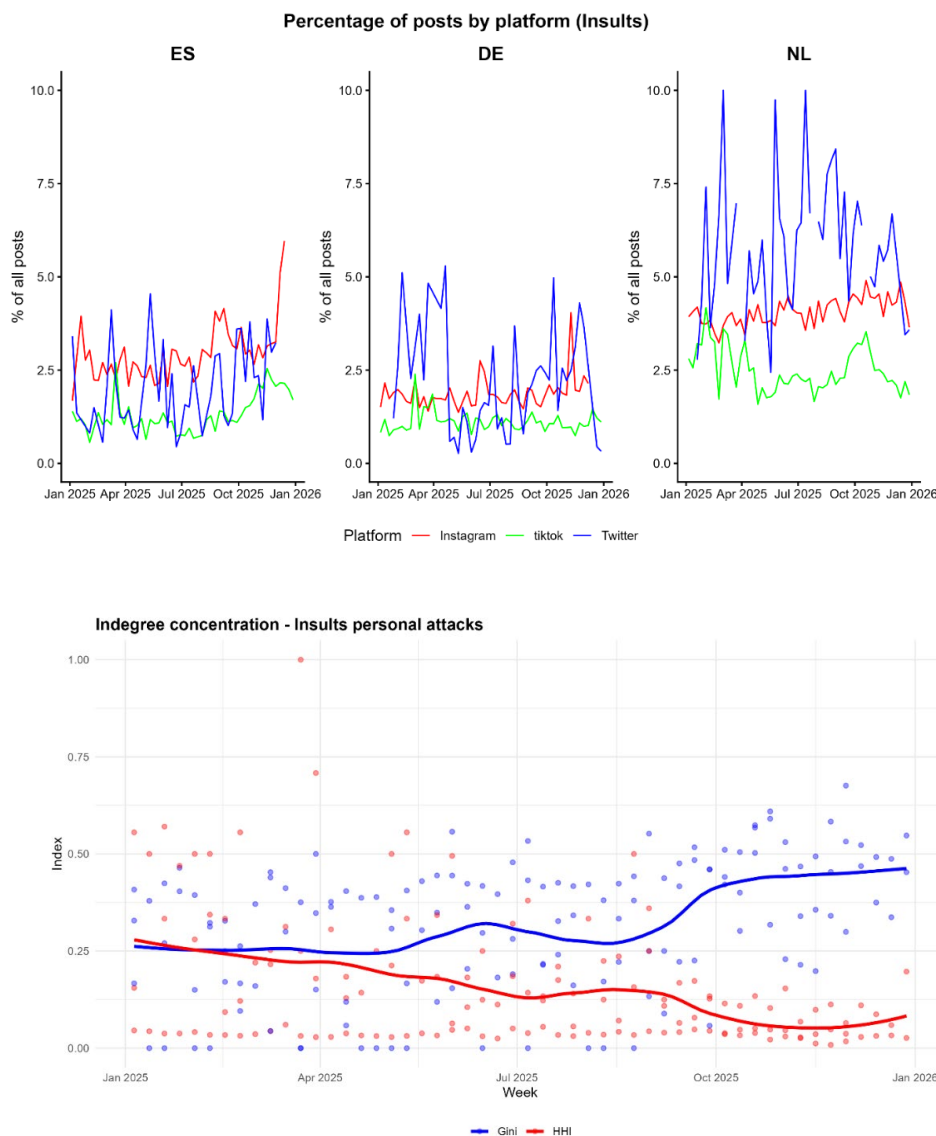
Feindseligkeit im Internet in Form von Beleidigungen oder persönlichen Angriffen ist in unserem Datensatz über alle Plattformen und Länder hinweg die zweithäufigste Kategorie und im Niederländischen besonders deutlich ausgeprägt. Im Vergleich zu anderen Kategorien tritt sie auf Instagram auffällig häufig und auf TikTok etwas seltener auf. Auf X zeigt sie beständige Schwankungen. Instagram weist konstant hohe Werte mit moderater Variation im Zeitverlauf auf, während TikTok eine ähnliche Stabilität mit gelegentlichen Spitzen zeigt.

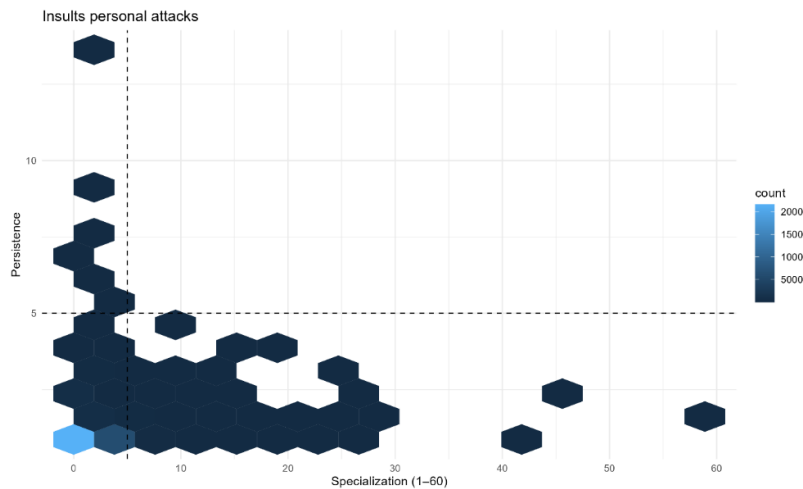
Hinsichtlich der Zielpersonen ist dies die Kategorie mit der zweithöchsten Ungleichverteilung in unserem Datensatz, bei zugleich relativ geringer Dominanz einzelner Zielpersonen. Der mittlere Gini Koeffizient beträgt 0,32 und $HHI = 0,15$. Dies legt nahe, dass Beleidigungen stark konzentriert auftreten, ohne dass dauerhaft zentrale Zielpersonen bestehen. Obwohl wir in unserem Datensatz eine zunehmende Auseinanderentwicklung zwischen stärkerer Konzentration und geringerer Dominanz beobachten, erscheint der Trend über die Zeit relativ stabil. Insgesamt deuten diese Modelle auf eine breit gestreute, aber ungleich verteilte Adressierung hin.

Bei den Verbreitenden von Beleidigungen stellen wir im Vergleich zu anderen Kategorien mehr Variation fest. Obwohl die Verbreitung weiterhin stark konzentriert und nur wenig beständig ist, mit 97,2 % fokussierter Adressierung bei geringer Beständigkeit, gibt es einen größeren Anteil von Konten, die eine breitere Gruppe von Zielpersonen beleidigen, mit 2,7 % gestreuter Adressierung. Auch der Anteil beständig angreifender Konten ist mit 0,1 % vergleichsweise höher.

Hinsichtlich möglicher Interventionen ist die Aufmerksamkeit in dieser Kategorie in allen Sprachen stark ungleich verteilt und weist eine hohe Streuung zwischen den Zielpersonen auf. Zwar ist es schwierig vorherzusehen, welche Ereignisse solche Inhalte auslösen könnten, und die Aufmerksamkeit verteilt sich weitgehend auf viele unterschiedliche Zielpersonen. Diese Modelle deuten jedoch auch darauf hin, dass Kerngruppen von Verbreitenden bestehen, die bestimmte Zielpersonen beständig beleidigen.

Abbildung 7. Beleidigungen und persönliche Angriffe: Vergleich von Anteilen, Zielpersonen und Verbreitenden





Politik

Diese Kategorie umfasst ein breites Spektrum an Sprache zu Politik, Politikerinnen und Politikern, politischen Parteien, konkreten politischen Maßnahmen, Kriegen und Konflikten, Ideologie sowie der Polarisierung zwischen politischer Linker und Rechter.

- Spanisches Beispiel: „Basura woke“.
- Deutsches Beispiel: „Selten so gelacht Hochverschuldung haben wir durch das Russen Drecksvolk die Freude von dem AFD Gesindel“.
- Niederländisches Beispiel: „Geert Wilders is een engnek“.

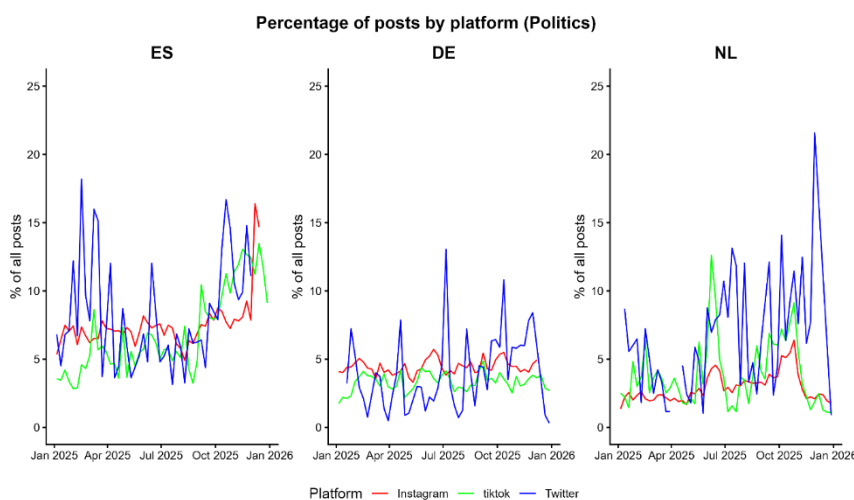
Politikbezogene Onlinefeindseligkeit ist in unserem Datensatz über alle Plattformen und Länder hinweg die häufigste Kategorie. Sie ist auf allen Plattformen relativ volatil und weist durchgängig hohe zeitliche Schwankungen auf, die auf X besonders ausgeprägt sind. Dies legt nahe, dass solche Inhalte durch Ereignisse und Nachrichtenzyklen beeinflusst werden könnten.

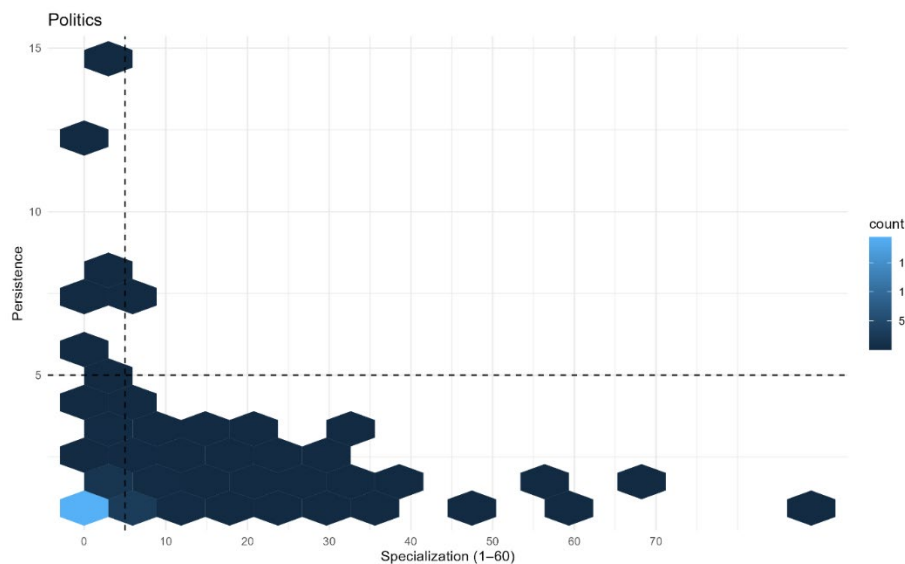
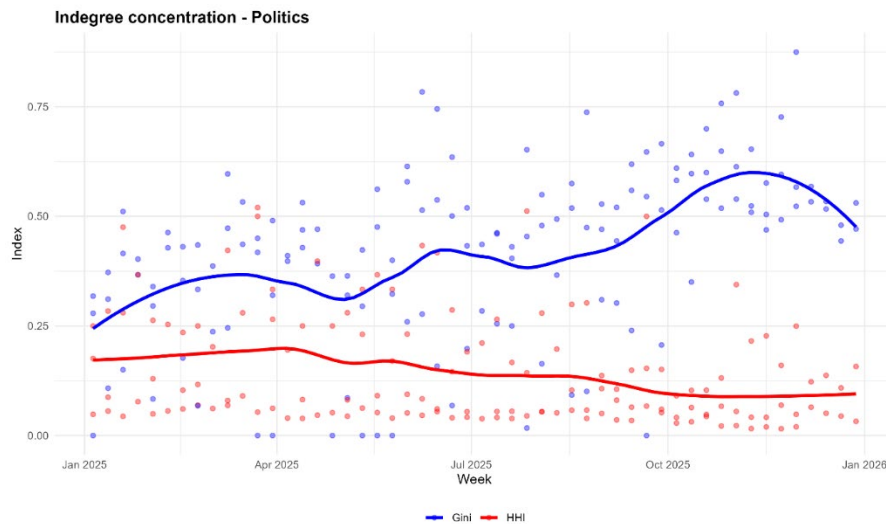
Hinsichtlich der Zielgruppen ist dies die ungleichste Kategorie in unserem Datensatz, verbunden mit einer relativ geringen Dominanz einzelner Zielgruppen (mittlerer Gini Koeffizient = 0,42, HHI = 0,14). Dies deutet darauf hin, dass politische Feindseligkeit zwar stark konzentriert ist, sich jedoch kaum auf wenige zentrale Zielgruppen verdichtet. Obwohl unser Datensatz eine zunehmende Divergenz zwischen stärkerer Konzentration und geringerer Dominanz erkennen lässt, erscheint der Trend über die Zeit relativ stabil. Insgesamt weisen diese Muster auf eine diffuse Zielauswahl ohne beständigen Fokus hin.

Bei den Verbreitern politikbezogener Feindseligkeit finden wir in unserem Datensatz mehr Variation als in den meisten anderen Kategorien. Obwohl die Verbreitung weiterhin stark konzentriert ist und nur geringe Persistenz aufweist (94,1 % fokussierte Zielauswahl bei geringer Persistenz), hat diese Kategorie den größten Anteil an Accounts, die eine breitere Gruppe von Zielpersonen beleidigen (5,8 % diffuse Zielauswahl), sowie einen vergleichsweise höheren Anteil persistenter Angreiferinnen und Angreifer (0,1 %). Dies legt nahe, dass politische Feindseligkeit häufiger gegen mehrere Zielpersonen gerichtet wird und von einigen Nutzergruppen dauerhaft ausgeht.

Hinsichtlich der Interventionen ist die Aufmerksamkeit in dieser Kategorie stark ungleich verteilt, mit großen Unterschieden zwischen den Zielgruppen in allen Sprachräumen. Zwar ist es schwierig vorherzusehen, welche Ereignisse diese Inhalte auslösen könnten, und die Aufmerksamkeit verteilt sich diffus auf viele unterschiedliche Zielgruppen. Diese Muster werden jedoch wahrscheinlich durch Nachrichtenzyklen beeinflusst und deuten darauf hin, dass es Kerngruppen von Verbreitern gibt, die bestimmte Zielpersonen fortlaufend beleidigen.

Abbildung 8. Vergleich politikbezogener Feindseligkeit: Anteile, Zielpersonen und Verbreitende





Rasse

Diese Kategorie umfasst intolerante Sprache, die sich gegen spezifische Aspekte der Identität im Zusammenhang mit Rasse und ethnischer Zugehörigkeit sowie gegen rassistische Stereotype richtet.

- **Spanisches Beispiel:** „Qué les pasa a los hombres n*? Por qué son la mayoría así?“.
- **Deutsches Beispiel:** „Was hat der N* mit deutscher Jugend zu tun?“
- **Niederländisches Beispiel:** „Terug naar je eigen LAND VIEZE AAP“.

Feindseligkeit im Internet im Zusammenhang mit Rasse tritt in unserem Datensatz nicht häufig auf. In den spanischen und deutschen Daten sowie auf niederländischem Instagram und

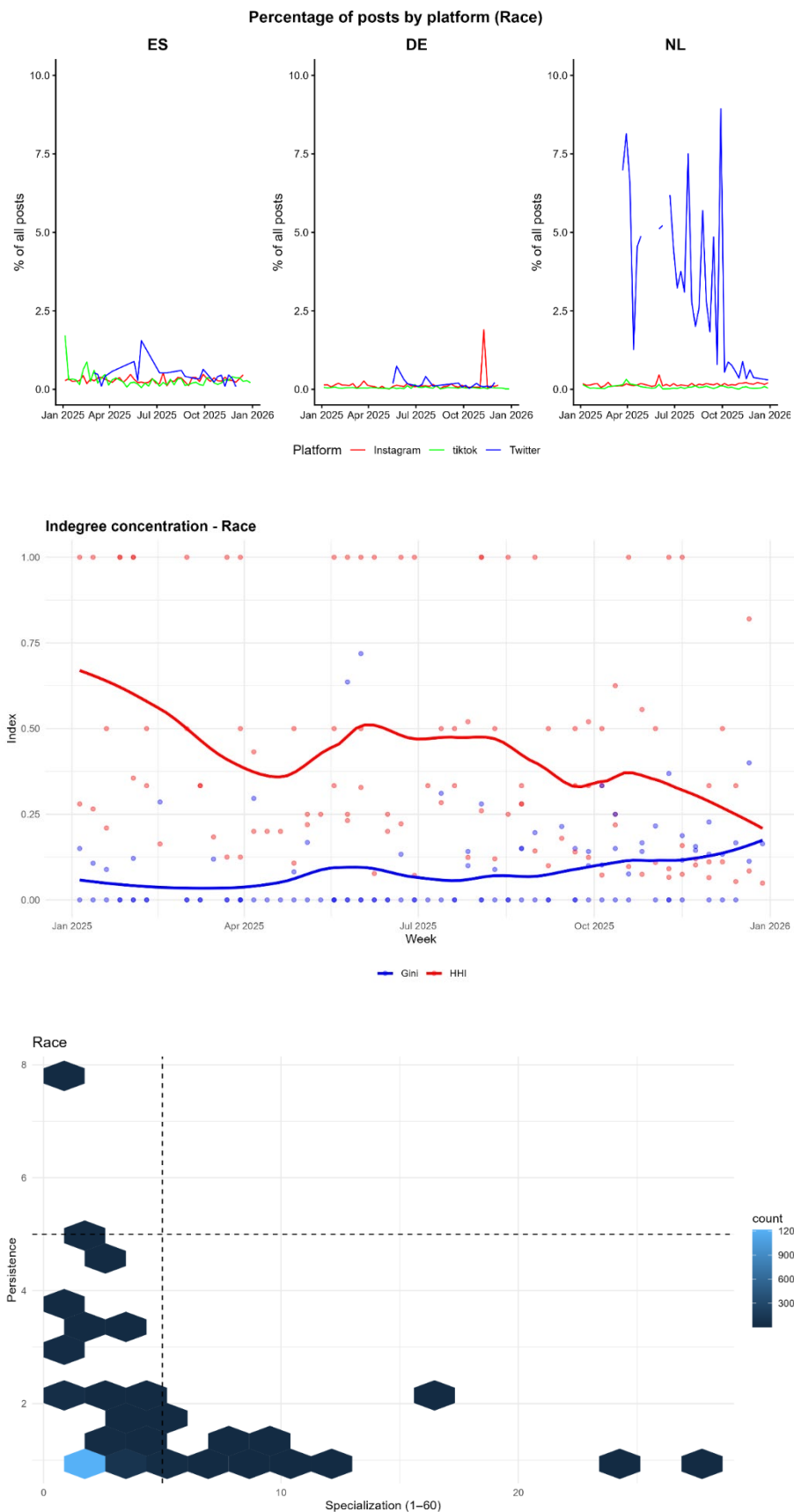
TikTok finden wir nur wenige verbleibende Hinweise darauf. Im niederländischsprachigen Teil des Datensatzes tritt sie jedoch auf X mit erheblichen Schwankungen auf. Wir erwarteten Unterschiede zwischen den Plattformen, da X im Beobachtungszeitraum angeblich die strenge Moderation eingestellt hat. Die Modellleistung ist akzeptabel. In Verbindung mit der starken Übereinstimmung zwischen den Kodierenden legt dies nahe, dass diese Sprache zuverlässig erkannt werden kann. Zusammengenommen erkennen wir an, dass die geringe Anzahl an Beobachtungen in dieser Kategorie in unserem Datensatz auf die Methodik zurückzuführen sein könnte und nicht zwangsläufig das tatsächliche Fehlen solcher Inhalte belegt.

Hinsichtlich der Aufmerksamkeit und der Zielpersonen weist Rassismus in unserem Datensatz einen sehr niedrigen mittleren Gini Koeffizienten von 0,08, aber einen hohen HHI von $HHI = 0,42$ auf. Dies legt nahe, dass die Ungleichverteilung zwar gering ist, die Aufmerksamkeit jedoch stark auf eine kleine Gruppe von Zielpersonen konzentriert ist. Der Trend ist über die Zeit beständig. Aufgrund der begrenzten Anzahl an Beobachtungen sind die Schwankungen jedoch relativ hoch. Zusammen mit den Aktivitätsspitzen deutet dies auf episodische Adressierung hin, bei der sich gebündelte Aufmerksamkeit auf wenige Zielpersonen richtet. Dies könnte mit Nachrichtenereignissen zusammenhängen, durch die bestimmte Personen zu besonders hervortretenden Zielpersonen werden.

Hinsichtlich der Verbreitenden stellen wir ein stark konzentriertes Muster fest: 96,9 % zeigen fokussierte Adressierung bei geringer Beständigkeit. Verglichen mit anderen Kategorien in unserem Datensatz weist Rassismus einen relativ großen Anteil von Nutzenden mit breiterer Adressierung auf, nämlich 2,9 % mit gestreuter Adressierung bei geringer Beständigkeit. Wir finden einen kleinen Anteil beständig aktiver Nutzender von 0,1 %.

Hinsichtlich möglicher Interventionen legen diese Ergebnisse nahe, dass Rassismus möglicherweise bereits streng moderiert wird. Nach den in unserem Datensatz verbleibenden Inhalten erscheint Rassismus weitgehend episodisch, mit wenig anhaltender oder breiter Adressierung. Es besteht jedoch ein Kern beständig angreifender Konten, besonders deutlich auf X.

Abbildung 9. Vergleich rassismusbezogener Feindseligkeit: Anteile, Zielpersonen und Verbreitende



Religion

Diese Kategorie umfasst intolerante Sprache, die sich gegen spezifische Aspekte der Identität im Zusammenhang mit Religion sowie gegen religiöse Stereotype richtet.

- **Spanisches Beispiel:** „Yo voy a entrar en una Mezquita y me voy a cargar 😂😂“.
- **Deutsches Beispiel:** „Ein verlogener Jude, das ist ja nichts Neues“.
- **Niederländisches Beispiel:** „Weer een dag dat Soennieten zich niet normaal kunnen gedragen“.

Feindseligkeit im Internet im Zusammenhang mit Religion tritt in unserem Datensatz im Deutschen und Niederländischen über alle Plattformen hinweg nicht sehr häufig auf. Im Spanischen ist sie auf allen Plattformen in unserem Datensatz auffallend selten. Dort, wo religionsbezogene Feindseligkeit in unserem Datensatz vorkommt, tritt sie auf Instagram und TikTok über die Zeit relativ beständig auf, während sie auf X starke Schwankungen zeigt. Diese Trends bleiben über die Zeit bestehen und legen eine Kombination aus einem Grundniveau religionsbezogener Hassrede und ereignisbedingten Aktivitätsspitzen nahe.

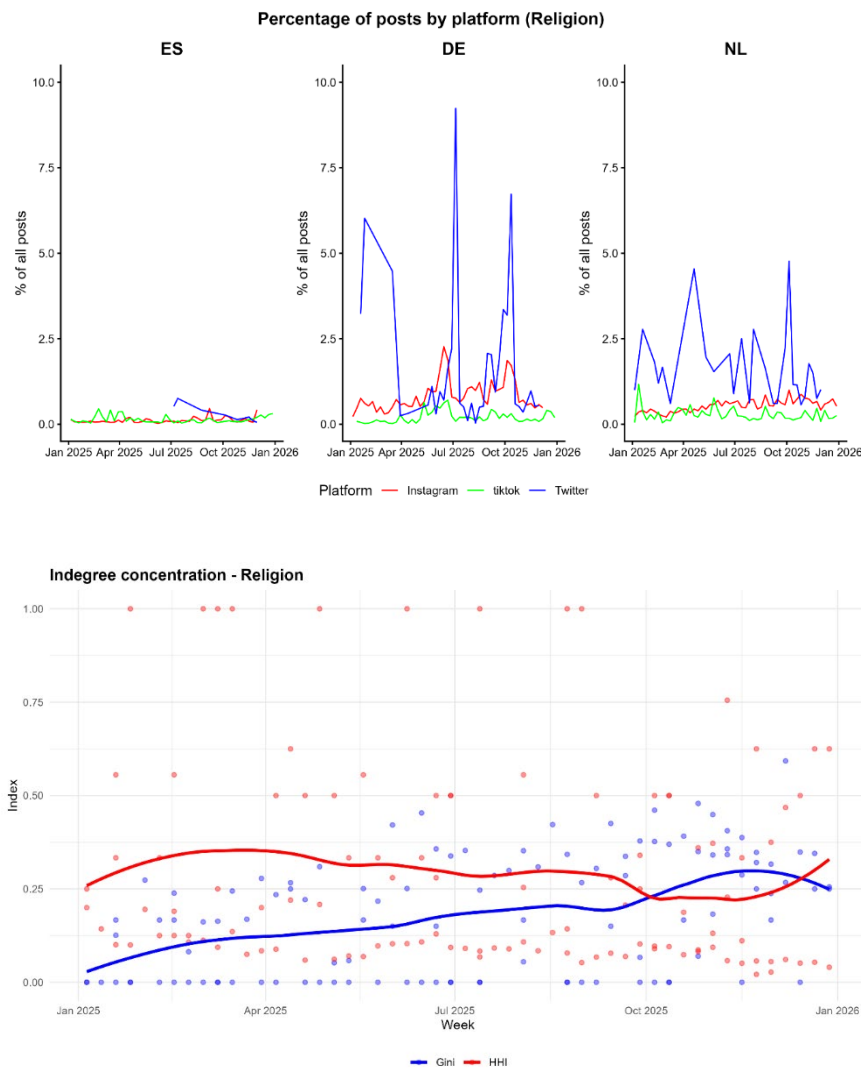
Hinsichtlich der Zielpersonen stellen wir im Vergleich zu anderen Kategorien eine moderate Konzentration bei zugleich moderater Dominanz fest. Der mittlere Gini Koeffizient beträgt 0,18 und $HHI = 0,29$. Dieser Trend erscheint über die Zeit relativ stabil, mit einer leichten Zunahme der Konzentration. Diese Muster deuten auf eine relativ breit gestreute Gruppe von Zielpersonen hin.

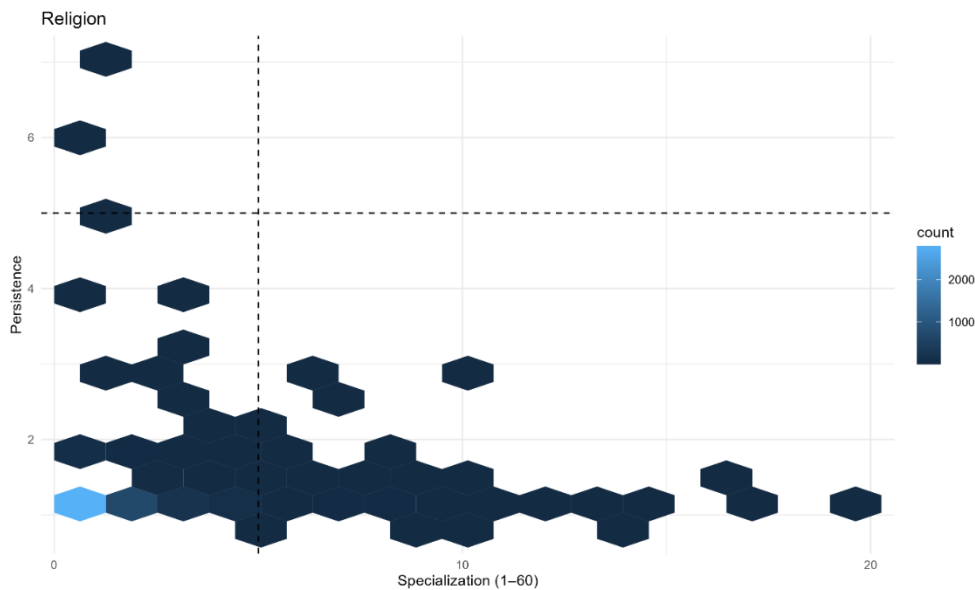
Bei den Verbreitenden religionsbezogener Feindseligkeit handelt es sich in unserem Datensatz um eine relativ homogene Kategorie. Die meisten Konten richten sich mit geringer Beständigkeit gegen eine einzige Zielperson. 97,1 % zeigen fokussierte Adressierung bei geringer Beständigkeit. Der Anteil gestreuter Adressierung ist mit 2,8 % vergleichsweise groß. Beständigkeit tritt nur sehr selten auf. Dies deutet auf eine stark episodische, aber vergleichsweise breit gestreute Adressierung hin.

Hinsichtlich möglicher Interventionen ist die Aufmerksamkeit in dieser Kategorie relativ breit verteilt, weist jedoch eine gewisse Beständigkeit bei den Zielpersonen auf. Wir stellen auffällige Unterschiede zwischen den Sprachen fest, die wahrscheinlich mit methodischen Einschränkungen zusammenhängen. Bei dieser Art von Inhalten werden Interventionen schwierig sein, da es kaum Hinweise darauf gibt, welche Ereignisse solche Inhalte auslösen

könnten, es keine klar erkennbaren wiederholt auffälligen Akteure gibt und die Beständigkeit unter den Verbreitenden über die Zeit sehr gering ist. Abhängig vom jeweiligen Länderkontext kann es jedoch beständige Zielpersonen geben.

Abbildung 10. Vergleich religionsbezogener Feindseligkeit: Anteile, Zielpersonen und Verbreitende





Sexuelle Einschüchterung

Diese Kategorie umfasst Sprache, die sexuell anzüglich oder herabsetzend ist und sich gegen eine Einzelperson richtet, ohne Bezug auf eine Gruppe oder Geschlechterrolle zu nehmen.

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

Sexuelle Einschüchterung tritt in unserem Datensatz nicht häufig auf. In den spanischen und deutschen Daten sowie auf niederländischem Instagram und TikTok finden wir nur wenige verbleibende Hinweise darauf. Im niederländischsprachigen Teil des Datensatzes tritt sie jedoch auf X mit erheblichen Schwankungen auf. Wir erwarteten Unterschiede zwischen den Plattformen, da X im Beobachtungszeitraum angeblich die strenge Moderation eingestellt hat. Zudem ist die Modellleistung akzeptabel. In Verbindung mit der starken Übereinstimmung zwischen den Kodierenden legt dies nahe, dass diese Sprache zuverlässig erkannt werden kann. Zusammengefasst erkennen wir, ähnlich wie bei Rassismus, an, dass die geringe Zahl an Beobachtungen in dieser Kategorie in unserem Datensatz auf die Methodik zurückzuführen sein könnte und nicht zwangsläufig das tatsächliche Fehlen solcher Inhalte belegt.

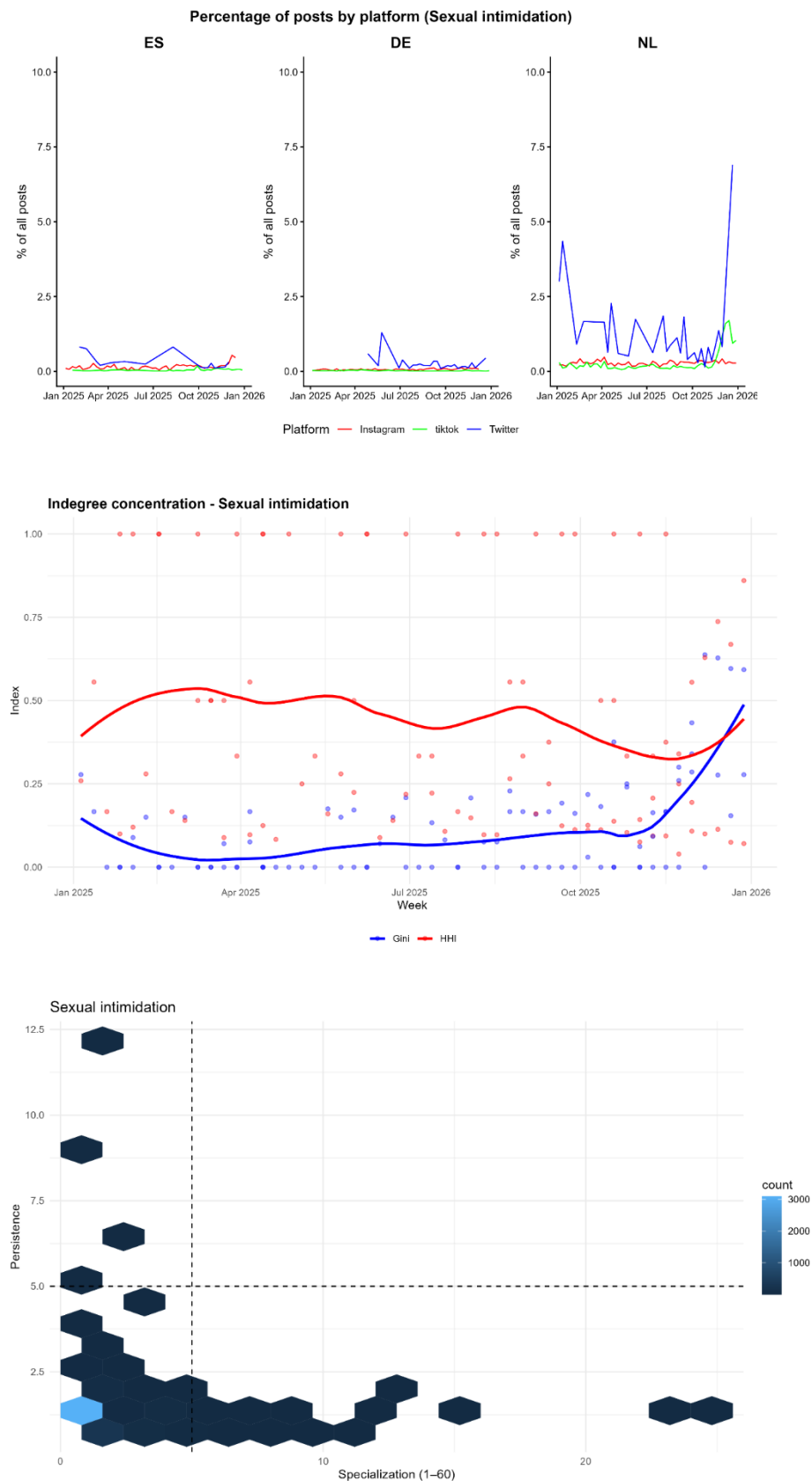
Die Muster, die wir feststellen, bleiben über die Zeit stabil: Auf X zeigen sich Schwankungen, während Instagram und TikTok in allen drei Ländern ein beständiges Grundniveau sexueller Einschüchterung aufweisen.

Hinsichtlich der Aufmerksamkeit und der Zielpersonen weist sexuelle Einschüchterung in unserem Datensatz einen sehr niedrigen mittleren Gini Koeffizienten von 0,11, aber einen hohen HHI von $HHI = 0,44$ auf. Dies legt nahe, dass die Ungleichverteilung zwar insgesamt gering ist, die Aufmerksamkeit jedoch stark auf eine kleine Gruppe von Zielpersonen konzentriert ist. Der Trend ist über die Zeit beständig. Aufgrund der begrenzten Anzahl an Beobachtungen sind die Schwankungen jedoch relativ hoch. Zusammen mit den Aktivitätsspitzen auf X und, in geringerem Umfang, auf Instagram deutet dies auf episodische Adressierung hin, bei der sich gebündelte Aufmerksamkeit auf wenige Zielpersonen richtet. Ähnlich wie bei Rassismus könnte dies mit Nachrichtenereignissen zusammenhängen, durch die bestimmte Personen zu besonders hervortretenden Zielpersonen werden.

Hinsichtlich der Verbreitenden stellen wir ein stark konzentriertes Muster fest: 97,7 % zeigen fokussierte Adressierung bei geringer Beständigkeit. Ähnlich wie bei Rassismus und im Unterschied zu anderen Kategorien in unserem Datensatz weist sexuelle Einschüchterung einen vergleichsweise großen Anteil von Nutzenden mit breiterer Adressierung auf, nämlich 2,3 % mit gestreuter Adressierung bei geringer Beständigkeit. Wir finden in unserem Datensatz keine bedeutsame Beständigkeit.

Hinsichtlich möglicher Interventionen erscheint sexuelle Einschüchterung auf Grundlage der in unserem Datensatz verbleibenden Hinweise weitgehend episodisch, mit teilweise beständiger Adressierung, aber begrenzter Beständigkeit und ohne eine klar erkennbare Kerngruppe von Verursachenden, an die sich Maßnahmen richten könnten. Diese Ergebnisse legen nahe, dass sexuelle Einschüchterung auf Instagram und TikTok möglicherweise bereits streng moderiert wird, gegebenenfalls einschließlich Meldungen und anschließender Sperrungen von Konten, und auf X weniger stark moderiert wird.

Abbildung 11. Vergleich von Feindseligkeit im Zusammenhang mit sexueller Einschüchterung: Anteile, Zielpersonen und Verbreitende



Sexuelle Orientierung

Diese Kategorie umfasst intolerante Sprache, die sich gegen spezifische Aspekte der Identität im Zusammenhang mit sexueller Orientierung sowie gegen damit verbundene Stereotype richtet.

- **Spanisches Beispiel:** „la persona gay en muuuy promiscua y guarra en el aspecto sexual. Tu eres gay? Pues disfrutalo“.
- **Deutsches Beispiel:** „Dreckige Schwuchtel“.
- **Niederländisches Beispiel:** „Donderop man vieze flikker“.

Feindseligkeit im Internet im Zusammenhang mit sexueller Orientierung ist keine besonders hervortretende Kategorie. In unserem Datensatz tritt sie auf allen Plattformen im Deutschen am deutlichsten auf, im Niederländischen vor allem auf X und im Spanischen nur selten. Auf X zeigt sie starke Schwankungen, während sie auf TikTok und Instagram relativ stabil auftritt. Dies deutet auf ein Grundniveau solcher Inhalte hin, das durch ereignisbezogene Aktivitätsschübe und Nachrichtenzyklen ergänzt wird.

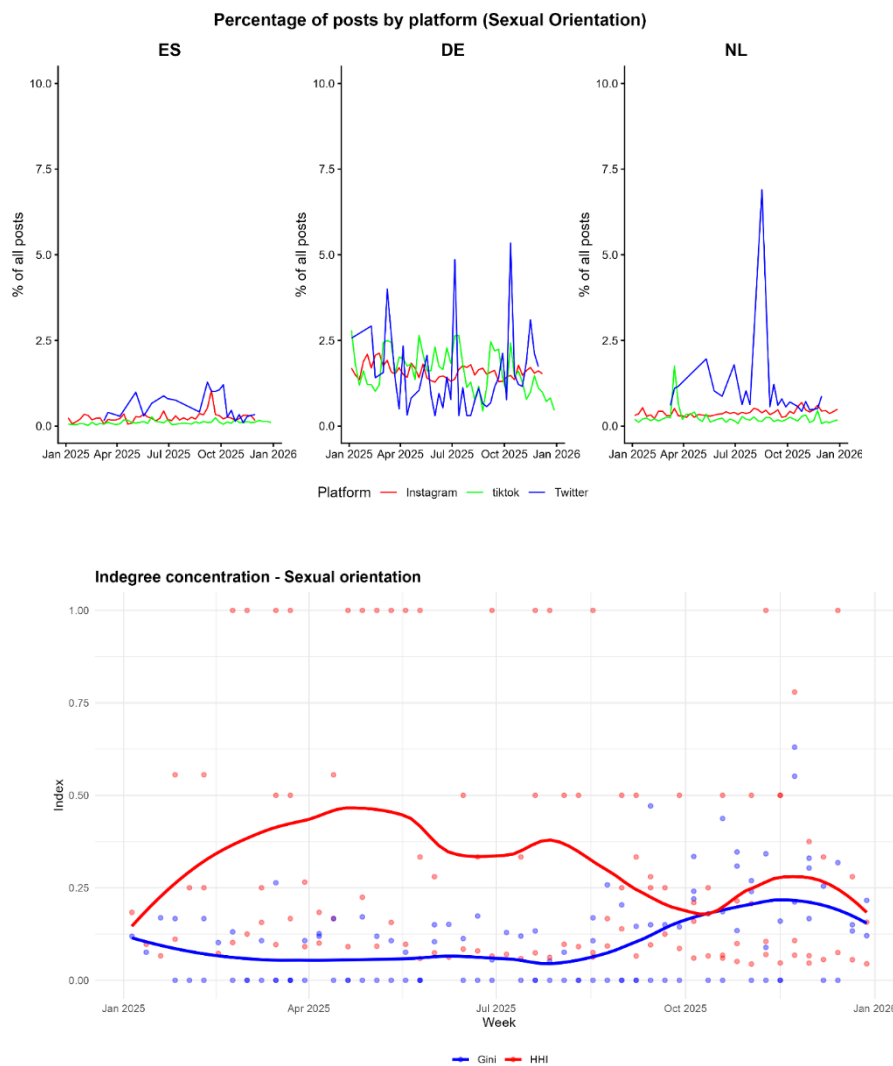
Hinsichtlich der Zielpersonen handelt es sich um eine der am wenigsten konzentrierten Kategorien in unserem Datensatz, bei zugleich moderater Dominanz einzelner Zielpersonen. Der mittlere Gini Koeffizient beträgt 0,11 und $HHI = 0,32$. Dies legt nahe, dass Feindseligkeit im Zusammenhang mit sexueller Orientierung relativ breit verteilt ist, sich jedoch in gewissem Maße auf zentrale Zielpersonen konzentriert. Obwohl wir in unserem Datensatz eine Annäherung an stärkere Konzentration und geringere Dominanz beobachten, erscheint der Trend über die Zeit relativ stabil. Insgesamt deuten diese Muster auf eine im Allgemeinen breit gestreute Konzentration mit moderat beständigen Zielpersonen hin.

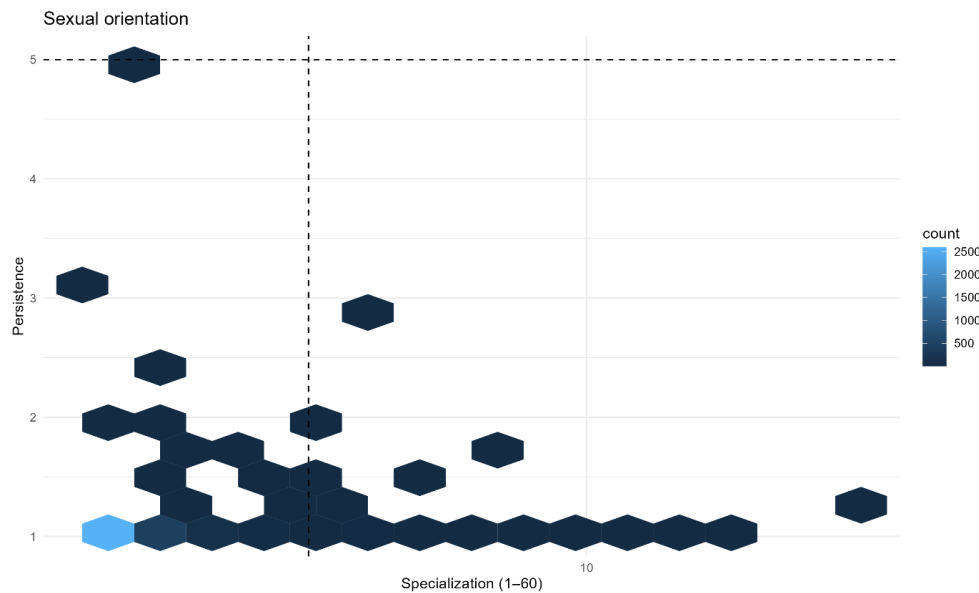
Bei den Verbreitenden von Feindseligkeit im Zusammenhang mit sexueller Orientierung finden wir einen relativ großen Anteil vorübergehend angreifender Konten. 98,4 % zeigen fokussierte Adressierung bei geringer Beständigkeit. Gestreute Adressierung und Beständigkeit sind vernachlässigbar. Dies legt nahe, dass Feindseligkeit aufgrund sexueller Orientierung von einzelnen Angriffen dominiert wird.

Hinsichtlich möglicher Interventionen legen diese Ergebnisse nahe, dass Feindseligkeit aufgrund sexueller Orientierung möglicherweise bereits wirksam moderiert wird. Nach den in unserem Datensatz verbleibenden Inhalten erscheint Feindseligkeit im Zusammenhang mit

sexueller Orientierung überwiegend vorübergehend, mit sehr wenig anhaltender oder breiter Adressierung.

Abbildung 12. Vergleich von Feindseligkeit im Zusammenhang mit sexueller Orientierung: Anteile, Zielpersonen und Verbreitende





Gewaltvolle Sprache

Diese Kategorie umfasst Sprache, die zu Gewalt aufruft oder Gewalt verherrlicht, einschließlich gezielter Todesdrohungen.

- **Spanisches Beispiel:** „Esa soberbia con la que aplaude y niega es merecedora de una guillotina“.
- **Deutsches Beispiel:** „Der gehört an die Wand“.
- **Niederländisches Beispiel:** „Tering jong, zal jou doodtrappen.“.

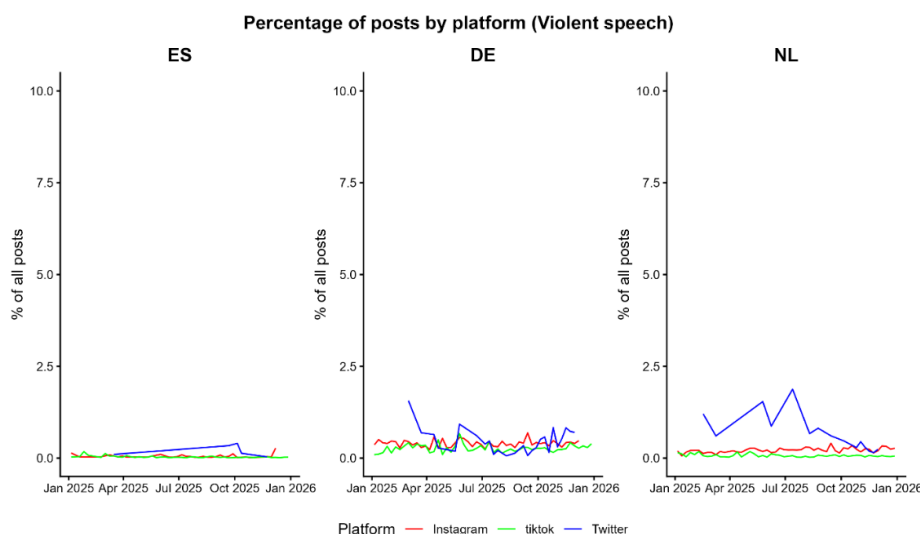
Gewaltvolle Sprache tritt in unserem Datensatz nicht häufig auf. In den spanischen und deutschen Daten sowie auf niederländischem Instagram und TikTok finden wir nur wenige verbleibende Hinweise darauf. Im niederländischsprachigen Teil des Datensatzes tritt sie jedoch auf X mit gewissen Schwankungen auf. Wir erwarteten Unterschiede zwischen den Plattformen, da X im Beobachtungszeitraum angeblich die strenge Moderation eingestellt hat. Die Modellleistung ist akzeptabel. In Verbindung mit der starken Übereinstimmung zwischen den Kodierenden legt dies nahe, dass diese Sprache zuverlässig erkannt werden kann. Zusammengefasst erkennen wir, ähnlich wie bei Rasse und sexueller Einschüchterung, an, dass die geringe Zahl an Beobachtungen in dieser Kategorie in unserem Datensatz auf die Methodik zurückzuführen sein könnte und nicht zwangsläufig das tatsächliche Fehlen solcher Inhalte belegt.

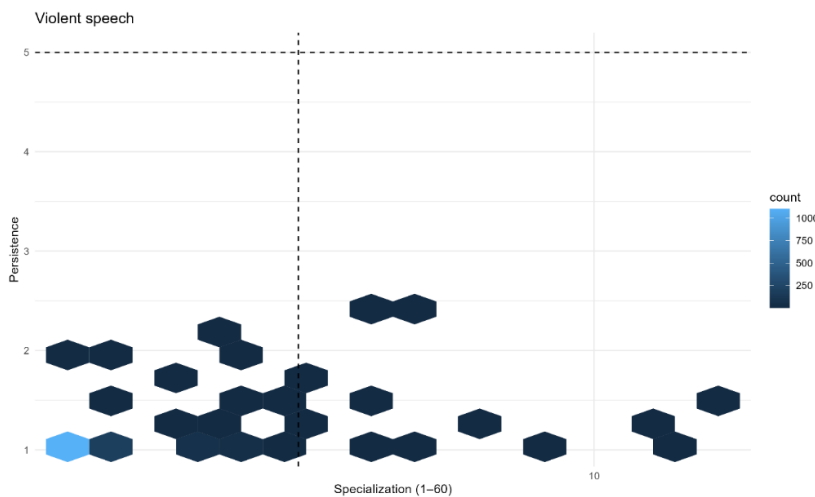
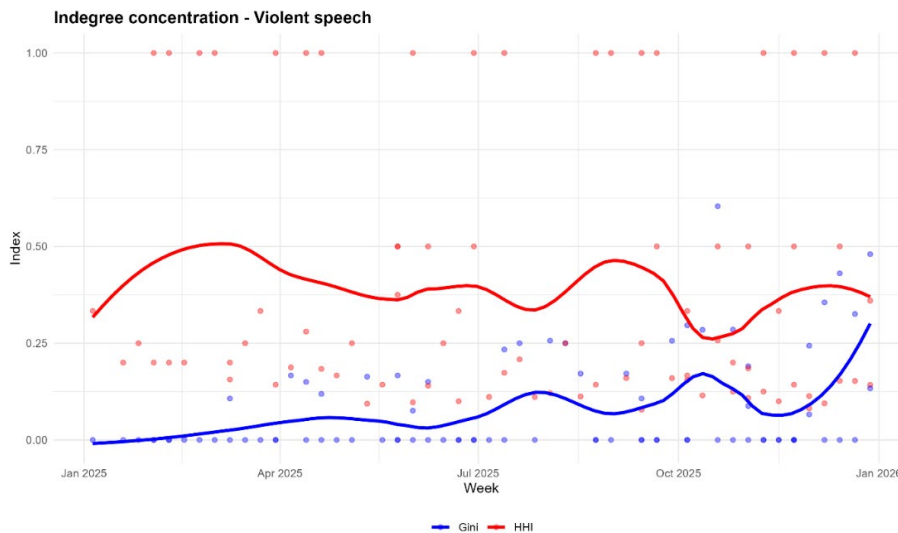
Hinsichtlich der Aufmerksamkeit und der Zielpersonen weist gewaltvolle Sprache in unserem Datensatz einen sehr niedrigen mittleren Gini Koeffizienten von 0,08, aber einen hohen HHI von $HHI = 0,40$ auf. Dies legt nahe, dass die Ungleichverteilung zwar gering ist, die Aufmerksamkeit jedoch stark auf eine kleine Gruppe von Zielpersonen konzentriert ist. Der Trend ist über die Zeit beständig, wobei die begrenzte Anzahl an Beobachtungen berücksichtigt werden muss. In Verbindung mit lediglich moderaten Aktivitätsspitzen deutet dies eher auf ein Grundniveau gewaltvoller Sprache als auf episodische Schübe gebündelter Aufmerksamkeit hin. Die Kategorie scheint somit weniger stark von Nachrichtenzyklen abhängig zu sein.

Hinsichtlich der Verbreitenden stellen wir ein stark konzentriertes Muster fest: 97,7 % zeigen fokussierte Adressierung bei geringer Beständigkeit. Verglichen mit anderen Kategorien in unserem Datensatz weist gewaltvolle Sprache einen relativ großen Anteil von Nutzenden mit breiterer Adressierung auf, nämlich 2,3 % mit gestreuter Adressierung bei geringer Beständigkeit. In unserem Datensatz finden sich nur sehr wenige Hinweise auf beständig angreifende Konten, die sich über die Zeit auf dieselben Zielpersonen konzentrieren.

Hinsichtlich möglicher Interventionen legen die in unserem Datensatz verbleibenden Hinweise darauf nahe, dass gewaltvolle Sprache möglicherweise bereits wirksam moderiert wird. Nach den verbleibenden Inhalten erscheint gewaltvolle Sprache überwiegend vorübergehend, mit sehr wenig anhaltender oder breiter Adressierung.

Abbildung 13. Vergleich gewalttätiger Äußerungen: Anteile, Zielpersonen und Verbreitende





Vulgäre Sprache

Diese Kategorie umfasst Sprache, die als unhöflich interpretiert werden kann und beleidigend sein mag, sich jedoch nicht gegen eine bestimmte Gruppe oder Einzelperson richtet. Dazu gehören Fluchen, Schimpfen und Kraftausdrücke.

- **Spanisches Beispiel:** „Coño señora entonces que hace usted y todo su equipo de gobierno“.
- **Deutsches Beispiel:** „drecks Politik!“.
- **Niederländisches Beispiel:** „God verdomme vuile moordenaars.“.

Vulgäre Sprache ist in unserem Datensatz keine besonders hervortretende Kategorie. In unseren Daten tritt sie auf allen Plattformen im Spanischen relativ häufiger auf, außerdem auf

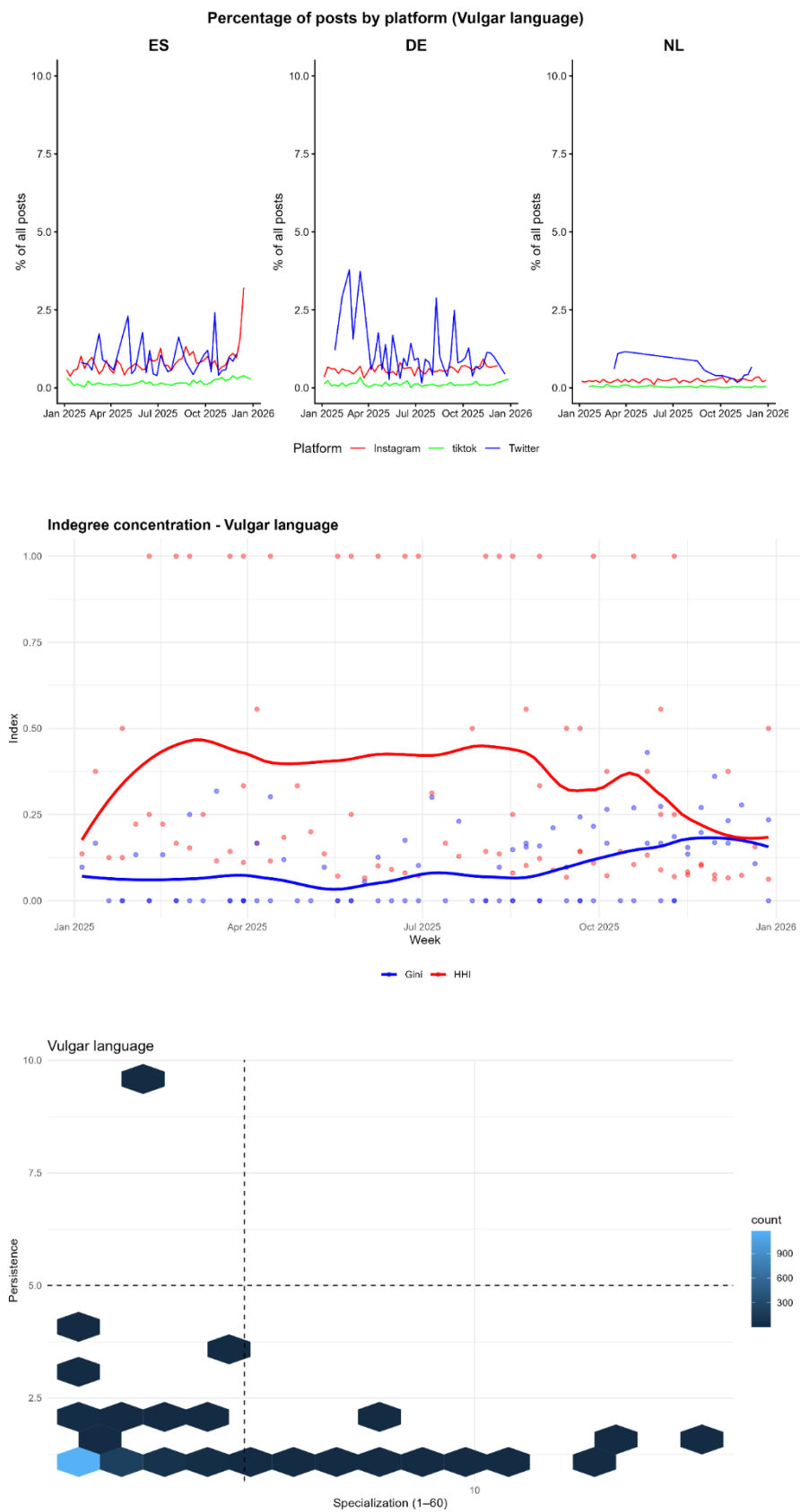
dem deutschsprachigen X, während sie im Niederländischen selten vorkommt. Sie tritt auf allen Plattformen relativ beständig auf, auch auf X im Vergleich zu anderen Kategorien. Dies deutet eher auf ein Grundniveau an Aktivität als auf ereignisbezogene Spitzen hin.

Hinsichtlich der Aufmerksamkeit und der Zielpersonen weist vulgäre Sprache in unserem Datensatz einen sehr niedrigen mittleren Gini Koeffizienten von 0,11 und einen hohen HHI von $HHI = 0,36$ auf. Dies legt nahe, dass die Ungleichverteilung zwar gering ist, die Aufmerksamkeit jedoch stark auf eine kleine Gruppe von Zielpersonen konzentriert ist. Der Trend ist über die Zeit beständig, wobei die begrenzte Anzahl an Beobachtungen berücksichtigt werden muss. Zusammen mit lediglich moderaten Aktivitätsspitzen deutet dies auf ein Grundniveau vulgärer Sprache und nicht auf episodische Schübe gebündelter Aufmerksamkeit hin. Die Kategorie scheint somit weniger stark von Nachrichtenzyklen abhängig zu sein.

Hinsichtlich der Verbreitenden stellen wir ein stark konzentriertes Modell fest: 98,5 % zeigen fokussierte Adressierung bei geringer Beständigkeit. Verglichen mit anderen Kategorien in unserem Datensatz weist vulgäre Sprache einen moderaten Anteil von Nutzenden mit breiterer Adressierung auf, nämlich 1,4 % mit gestreuter Adressierung bei geringer Beständigkeit. Wir finden einen kleinen Anteil beständig aktiver Nutzender von 0,1 %.

Hinsichtlich möglicher Interventionen legen diese Ergebnisse nahe, dass vulgäre Sprache auf Instagram und TikTok möglicherweise bereits weitgehend moderiert wird. Nach den in unserem Datensatz verbleibenden Inhalten erscheint vulgäre Sprache weitgehend episodisch, mit teilweise breiter Adressierung und einem Kern beständig angreifender Konten, besonders deutlich auf X.

Abbildung 14. Vergleich vulgärer Feindseligkeit: Anteile, Zielpersonen und Verbreitende



Schlussfolgerungen

Insgesamt legen unsere Ergebnisse nahe, dass Feindseligkeit im Internet weit verbreitet ist und überwiegend aus einzelnen feindseligen Kommentaren besteht. Hinsichtlich möglicher Interventionen bedeutet dies, dass die zentrale Herausforderung nicht darin liegt, einer kleinen Zahl wiederholt auffälliger Personen entgegenzuwirken, sondern einer großen Zahl einzelner feindseliger Kommentare, die von vielen unterschiedlichen Konten ausgehen.

Wir betonen, dass die geringe Zahl an Beobachtungen von Kategorien wie „Rasse“ oder „gewaltvolle Sprache“ in unserem Datensatz auf Verzerrungen bei der Datenerhebung und in der Methodik zurückzuführen sein könnte und nicht zwangsläufig das tatsächliche Fehlen solcher Inhalte belegt. Zudem weisen wir auf die begrenzte Genauigkeit und Sensitivität der hier verwendeten Sprachmodelle zur Erkennung unterschiedlicher Formen von Feindseligkeit hin. Aus dem Vergleich der Kategorien, Länder und Plattformen leiten wir die folgenden Erkenntnisse ab.

Im Allgemeinen stellen wir fest, dass Verbreitende von Feindseligkeit breiter gestreut sind als erwartet. Feindseligkeit im Internet geht in unserem Datensatz überwiegend von vielen unterschiedlichen Konten aus, statt sich auf wenige Konten zu konzentrieren. Zwischen den Kategorien finden wir bei den Verbreitenden nur geringe Überschneidungen. Dies legt nahe, dass die meisten Konten nur eine Art von Feindseligkeit äußern, anstatt ihr Repertoire zu diversifizieren. Unter den Verbreitenden von Feindseligkeit im Internet finden wir sehr wenige Hinweise auf fokussierte und beständige Angriffe. Über alle Kategorien hinweg ist das vorherrschende Model fokussierte Adressierung bei geringer Beständigkeit. Dies legt nahe, dass sich die meisten Nutzenden unabhängig vom Thema nur kurzzeitig und gegen wenige Zielpersonen äußern. Wir schließen daraus, dass der Großteil der Feindseligkeit im Internet in unserem Datensatz aus einzelnen Beiträgen besteht und es nur sehr wenige wiederholt auffällige Personen gibt.

Zugleich stellen wir fest, dass Zielpersonen von Feindseligkeit im Internet stark konzentriert sind. Dasselbe Model zeigt sich bei einer Differenzierung nach Art der Feindseligkeit. Über alle Kategorien hinweg finden wir eine hohe Konzentration von Zielpersonen, wobei sich der Großteil der Aufmerksamkeit auf relativ wenige Zielpersonen richtet. Zugleich variiert, wie beständig dieselben Konten adressiert werden. Bei einer weiteren Differenzierung nach

Adressierung, feindseligem Veröffentlichungsverhalten und Schwankungen zeigen sich deutliche Unterschiede zwischen den Kategorien. Diese Unterschiede sind für unterschiedliche Interventionsstrategien relevant.

Empfohlene Interventionsstrategien

Hinsichtlich möglicher Interventionen schlagen wir drei umsetzbare Erkenntnisse auf Grundlage unserer Ergebnisse vor, die in Tabelle 7 zusammengefasst sind.

Tabelle 7. Empfohlene Interventionsstrategie nach Kategorietypen

Kategorie	Zielkonzentration	Auffälliger Verbreitertyp	Volatilität	Interventionsfokus
Race	Konzentriert + stabil	Wiederholungstäter (Diffus + persistent)	episodisch	Spezifische Ziele + spezifische Verbreiter + Ereignisse
Sexual intimidation	Diffus + stabil	Generalisten (Diffus + vorübergehend)	Konsistent + episodisch	Spezifische Ziele + Ereignisse
Religion	Diffus + instabil	Generalisten (Diffus + vorübergehend)	Konsistent + episodisch	Spezifische Ziele + Ereignisse
Sexual orientation	Diffus + stabil	Generalisten (Diffus + vorübergehend)	Konsistent + episodisch	Spezifische Ziele + Ereignisse
Vulgar language	Konzentriert + stabil	Wiederholungstäter (Diffus + persistent)	Konsistent	Spezifische Ziele + spezifische Verbreiter
Crime	Diffus + stabil	Generalisten (Diffus + vorübergehend)	Konsistent + episodisch	Ereignisse
Culture	Konzentriert + instabil	Opportunisten (Fokussiert + vorübergehend)	Konsistent + episodisch	Ereignisse

Politics	Konzentriert + instabil	Wiederholungstäter (Diffus + persistent)	Episodisch	Ereignisse
Gender	Diffus + stabil	Opportunisten (Fokussiert + vorübergehend)	Konsistent	Spezifische Ziele
Violent language	Konzentriert + stabil	Generalisten (Diffus + vorübergehend)	Konsistent	Spezifische Ziele
Insults	Konzentriert + instabil	Wiederholungstäter (Diffus + persistent))	Konsistent	Spezifische Verbreiter

Eine erste Gruppe von Kategorien ist durch episodische Aktivität gekennzeichnet: Rasse, sexuelle Einschüchterung, sexuelle Orientierung, Religion, gewaltvolle Sprache, Kriminalität, Kultur und Politik. Im Vergleich zu anderen Kategorien weisen diese relativ stärkere Schwankungen auf und sind wahrscheinlich stärker durch Ereignisse geprägt. Interventionsstrategien für diese Kategorien könnten darauf ausgerichtet sein, Ereignisse, die Feindseligkeit im Internet erzeugen, vorherzusehen und darauf zu reagieren.

Eine zweite Gruppe von Kategorien ist durch die Konzentration auf Zielpersonen gekennzeichnet: Rasse, sexuelle Einschüchterung, Religion, sexuelle Orientierung, Geschlecht, vulgäre Sprache und gewaltvolle Sprache. Im Vergleich zu anderen Kategorien sind die Zielpersonen dieser Formen von Feindseligkeit relativ konzentriert und oder stabil. Interventionsstrategien für diese Kategorien könnten sich auf die Beobachtung von Konten richten, die wahrscheinlich Zielpersonen von Feindseligkeit im Internet werden.

Eine dritte Gruppe von Kategorien ist durch Beständigkeit unter den Verbreitenden gekennzeichnet: Rasse, vulgäre Sprache und Beleidigungen. Im Vergleich zu anderen Kategorien sind Verbreitende dieser Formen von Feindseligkeit relativ beständig aktiv. Interventionsstrategien für diese Kategorien könnten sich auf wiederholt auffällige Personen richten, die Feindseligkeit im Internet erzeugen.

Wir betonen, dass die geringe Sensitivität und Präzision der Sprachmodelle eine zentrale Einschränkung darstellen. In der Praxis werden die hier eingesetzten Modelle viele Beiträge mit Feindseligkeit im Internet übersehen und viele Fälle fälschlicherweise als problematisch markieren. Mit Blick auf unser Ziel, Moderation zu unterstützen, halten wir Fehlalarme jedoch

für vorzuziehen gegenüber dem Übersehen zahlreicher problematischer Beiträge. Die Genauigkeit wird sich voraussichtlich mit der Zeit sowie bei der Anwendung auf anderen Plattformen weiter verschlechtern. Wir raten davon ab, diese Modelle für automatisierte Entscheidungen einzusetzen. Ihr praktischer Nutzen liegt darin, Inhalte zur Überprüfung durch Fachpersonen zu markieren.

Methodische Einschränkungen

Die Datensammlung und Methodik, auf denen dieser Bericht beruht, gehen mit wichtigen Einschränkungen einher.

Erstens weisen wir darauf hin, dass die am weitesten verbreiteten Formen von Feindseligkeit aufgrund unserer Strategie der Datensammlung in unserem Datensatz unterrepräsentiert sein könnten. Unsere Datensätze sind zwar umfangreich, aber nur teilweise repräsentativ. Die Beschränkung auf öffentlich verfügbare Daten, die aus Gründen des Datenschutzes und angesichts praktischer Einschränkungen bei Daten aus sozialen Medien erforderlich war, bedeutet, dass die am strengsten moderierten Inhalte gelöscht wurden, bevor wir sie erheben konnten. Daher ist es wahrscheinlich, dass Rassismus und gewaltvolle Sprache im Allgemeinen sowie geschlechtsbezogene Gewalt im Besonderen in der Praxis stärker ausgeprägt sind, als unsere Analyse nahelegt. Auch wenn Moderation solche Inhalte nach einiger Zeit löscht, kann der Schaden bereits entstanden sein.

Zweitens bestehen ebenfalls im Zusammenhang mit der Datensammlung Einschränkungen, die sich aus der Priorisierung von Breite und Vielfalt ergeben. Unser anfänglicher Fokus auf die Identifizierung prominenter Personen bedeutet, dass weniger prominente oder sichtbare Akteure unterrepräsentiert sein könnten. Dies haben wir durch eine gezielte Stichprobenziehung nach Identitätsmerkmalen zu berücksichtigen versucht. Merkmale wie Rasse, ethnische Zugehörigkeit, Geschlechtsidentität, sexuelle Orientierung, Religion, Migrationshintergrund, Sprache und politische Ideologie wurden berücksichtigt, um sicherzustellen, dass der Datensatz Gruppen einschließt, die unverhältnismäßig häufig von Feindseligkeit im Internet betroffen sind, aber in der öffentlichen Debatte möglicherweise weniger sichtbar sind. Wir versuchten, Minderheiten zu erfassen, die für den jeweiligen nationalen Kontext spezifisch sind, etwa sprachliche Minderheiten in Spanien und Personen,

die sich in Deutschland an Debatten über historisch sensible Themen beteiligen. Eine damit verbundene Einschränkung unseres Ansatzes besteht in Überschneidungen zwischen Bereichen, etwa Politik und Aktivismus. Diese erschwerten die Bemühungen, die Kategorien ausgewogen abzubilden. Letztlich führte dies dazu, dass einige Kategorien, etwa Politik, stark vertreten sind, während andere, etwa Sport, weniger berücksichtigt werden. Wir gehen davon aus, dass unser Ansatz eine umfangreiche Abdeckung sozialer Bereiche und Identitätsgruppen gewährleistet und zugleich kontextuelle sowie kulturelle Unterschiede im Ausdruck von Hass im Internet berücksichtigt.

Drittens erkennen wir an, dass Beiträge in sozialen Medien weder neutral noch natürlich entstanden sind und dass Konten in sozialen Medien nicht mit natürlichen Personen gleichgesetzt werden können. Unser Datensatz enthält wahrscheinlich unterschiedliche Formen von Automatisierung, Spam und andere Formen böswilliger Absichten. Die Identifizierung von Bots und Formen der Koordinierung liegt außerhalb des Rahmens dieses Projekts. Wir empfehlen jedoch nachdrücklich, das Verhältnis zwischen koordinierten Formen der Manipulation und den verschiedenen in diesem Bericht identifizierten Formen von Feindseligkeit weiter zu untersuchen.

Viertens erkennen wir an, dass der Vergleich von Daten unterschiedlicher Plattformen sozialer Medien, also Instagram, X und TikTok, sowie unterschiedlicher nationaler Kontexte, also Deutschland, Spanien und die Niederlande, erheblich eingeschränkt ist. Jede Plattform sozialer Medien weist unterschiedliche Nutzergruppen, spezifische Sprachgebräuche und besondere Nutzungsmöglichkeiten auf, die eng mit dem Sprachgebrauch auf diesen Plattformen verbunden sind. Darüber hinaus erkennen wir an, dass die Bedeutung von Sprache sehr stark vom Kontext abhängt. Vergleichende Analysen der Bedeutung von Sprache und damit verbundener Unterschiede zwischen Ländern und Plattformen erfordern qualitative Forschung. Wir halten die in diesem Bericht vorgenommenen, stärker quantitativ orientierten Vergleiche jedoch durch unser Ziel gerechtfertigt, Muster in der Verbreitung von Sprache zwischen Ländern und Plattformen zu identifizieren. Wir verstehen diese Vergleiche als Hinweise auf Trends und nicht als präzise Abbildungen.

Fünftens ist die automatisierte Sprachklassifikation hinsichtlich ihrer Leistung grundsätzlich begrenzt. Wir erkennen an, dass die geringe Sensitivität und Präzision der Modelle zur Erkennung von Hassrede und, in geringerem Maße, der Klassifikationsmodelle eine zentrale

Einschränkung der hier beschriebenen Methode darstellen. Dies gilt insbesondere für Kategorien, die in unserem Datensatz weniger häufig vorkommen und für die daher weniger Beobachtungen zum Training der Modelle vorliegen. Praktisch bedeutet dies, dass die Modelle viele falsch positive Ergebnisse erzeugen werden. Mit Blick auf unser Ziel, Moderation zu unterstützen, halten wir Fehlalarme für vorzuziehen gegenüber dem Übersehen zahlreicher problematischer Beiträge.

Die drei bereits bestehenden Modelle zur Erkennung von Hassrede wurden nach einer vorläufigen Leistungsanalyse in der Phase des Forschungsdesigns ausgewählt. Zum Zeitpunkt der Entwicklung des Forschungsdesigns wurde auf Large Language Models als leistungstärkere Alternative zu regelbasierten Lexika verzichtet, und zwar aus Gründen der praktischen Umsetzbarkeit sowie aufgrund von Einschränkungen im Datenmanagement und Datenschutz. Die Feinabstimmung von BERT basierten Modellen mit unseren annotierten Daten bleibt ein Ziel für zukünftige Arbeiten. Für die Zwecke unserer Analyse haben wir diese Einschränkung durch ein teilüberwachtes Verfahren adressiert, das automatisierte Erkennung mit menschlicher Annotierung und Klassifikation verbindet. Darüber hinaus versuchten wir, diese Einschränkung durch fortlaufende kritische Reflexion, die Bewertung der Übereinstimmung zwischen Kodierenden, Kalibrierung, manuelle Überprüfung und Stichprobenkontrollen zu berücksichtigen, um die Validität unserer Ergebnisse sicherzustellen.

Dennoch erkennen wir an, dass die Klassifikationsmodelle, auf denen unsere Analysen beruhen, hinsichtlich Genauigkeit, Präzision und Sensitivität fehlerhaft sind. Da sich Sprache und ihre Verwendung weiterentwickeln und Feindseligkeit im Internet auf andere Wörter und Bedeutungen verlagert, wird die Genauigkeit dieser Modelle mit der Zeit wahrscheinlich weiter abnehmen. Wir raten davon ab, diese Modelle für automatisierte Entscheidungen einzusetzen. Ihr praktischer Nutzen liegt darin, Inhalte zur Überprüfung durch Fachpersonen zu markieren.

Sechstens weisen wir darauf hin, dass wir nur begrenzt Aussagen über die Auswirkungen von Feindseligkeit im Internet treffen können. Unsere Analysen geben Einblicke in die Modelle unterschiedlicher Formen von Feindseligkeit, können jedoch nicht zwischen den Auswirkungen auf das Leben der Personen unterscheiden, die diese Inhalte erhalten. Wir erkennen an, dass einige Formen von Feindseligkeit stärker wirken können als andere, auch

wenn Kategorien wie Rassismus in unserer Analyse nicht am deutlichsten hervortreten. Zudem kann das Zusammenwirken verschiedener Kategorien, etwa Rasse und Geschlecht, schädliche Folgen weiter verstärken.

Insgesamt bietet diese Methodik einen umfassenden und anpassungsfähigen Rahmen für länderübergreifende Datensammlungen. Durch die systematische Identifizierung von Konten, die im öffentlichen Raum agieren und verschiedenen Formen von Angriffen im Internet ausgesetzt sind, deckt der resultierende Datensatz ein breites Spektrum von Feindseligkeit im Internet ab. Die für diese Analyse erstellten Lexika sind frei zugänglich verfügbar, siehe van Haperen et al. 2026. Bei angemessener Berücksichtigung lokaler Kontexte ist dieser Ansatz replizierbar und kann die Moderation sowohl offener als auch subtilerer Formen von Feindseligkeit im Internet unterstützen.

Danksagungen

Vielen Dank an alle, die zum Verfahren und zur Analyse in diesem Bericht beigetragen haben: 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. Vielen Dank für Ihre Beiträge und wertvollen Einblicke.

Rechtlicher Hinweis

Ziel dieses Berichts ist es, Trends und Muster zu identifizieren. Er stellt keine rechtliche Bewertung von Meinungen oder Aussagen einzelner Personen dar. Soweit möglich, werden personenbezogene Daten vor der Analyse gemäß der DSGVO und den einschlägigen europäischen Vorschriften aggregiert, pseudonymisiert oder anonymisiert. Die Analysen in diesem Bericht beruhen auf der automatisierten Erkennung und Interpretation öffentlich

zugänglicher Online Daten und unterliegen methodischen sowie kontextuellen Einschränkungen.

Datenerklärung und Lizenz

Die im Rahmen dieses Projekts erstellten Lexika, die der vorliegenden Analyse zugrunde liegen, sind als Open Source im Repository der Universität Amsterdam verfügbar⁴.

Datensatzname: Violence, Intolerance and Impoliteness thematic Corpus v1.0

Entwickler des Datensatzes: 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.

Verfasser der Datenerklärung: Sander van Haperen (University of Amsterdam)

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

Sie dürfen:

- **Teilen:** das Material in jedwedem Format oder Medium vervielfältigen und weiterverbreiten.
- **Bearbeiten:** das Material neu zusammenstellen, verändern und darauf aufbauen. Der Lizenzgeber kann Ihnen diese Freiheiten nicht entziehen, solange Sie die Lizenzbedingungen einhalten.

Unter folgenden Bedingungen:

- **Namensnennung:** Sie müssen eine angemessene Urheber und Rechteangabe machen, einen Link zur Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden. Sie dürfen dies in jeder angemessenen Weise tun, jedoch nicht in einer Weise, die den Eindruck erweckt, der Lizenzgeber unterstütze Sie oder Ihre Nutzung.
- **Nicht kommerziell:** Sie dürfen das Material nicht für kommerzielle Zwecke nutzen.

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

- **Weitergabe unter gleichen Bedingungen:** Wenn Sie das Material neu zusammenstellen, verändern oder darauf aufbauen, dürfen Sie Ihre Beiträge nur unter derselben Lizenz wie das Original verbreiten.
- **Keine zusätzlichen Einschränkungen:** Sie dürfen keine zusätzlichen Klauseln oder technischen Verfahren einsetzen, die anderen rechtlich etwas untersagen, was die Lizenz erlaubt.

Hinweise: Sie müssen die Lizenzbedingungen nicht einhalten, soweit Teile des Materials gemeinfrei sind oder Ihre Nutzung durch eine anwendbare Ausnahme oder Beschränkung erlaubt ist. Es werden keine Gewährleistungen gegeben. Die Lizenz gewährt Ihnen möglicherweise nicht alle Berechtigungen, die für Ihre beabsichtigte Nutzung erforderlich sind. Beispielsweise können andere Rechte, etwa Persönlichkeitsrechte, Datenschutzrechte oder Urheberpersönlichkeitsrechte, die Nutzung des Materials einschränken.

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

Literaturverzeichnis

- 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., Quattrociocchi, 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://eooH.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.