Platforms, Experts, Tools: Specialised Cyber-Activists Network

Monitoring Report

2019 – 2020

Project funded by the European Union's Rights, Equality and Citizenship Programme (2014-2020)

Über das Projekt

Das von der EU geförderte Projekt **sCAN** – *Platforms, Experts, Tools: Specialised Cyber-Activists Network* (2018-2020), koordiniert von Licra (International League Against Racism and Antisemitism), zielte darauf ab, Fachwissen, Tools, Methodik und Wissen über Cyberhass im Internet zu sammeln und länderübergreifende, umfassende Praktiken zur Identifizierung, Analyse, Berichterstattung und Bekämpfung von Online-Hassrede zu entwickeln. Das Projekt stützte sich auf die Ergebnisse bereits realisierter, erfolgreicher europäischer Projekte, darunter "*Research, Report, Remove project: Countering Cyber-Hate phenomena*" und *"Facing Facts",* und war bestrebt, die von der Zivilgesellschaft entwickelten Initiativen zur Bekämpfung von Hassrede fortzusetzen, zu präzisieren und zu stärken.

Durch die europaweite Kooperation haben die Projektpartner ihre fruchtbare Zusammenarbeit (weiter) verstärkt und intensiviert. Die **sCAN**-Projektpartner haben zur Auswahl und Bereitstellung relevanter, automatischer Überwachungsinstrumente beigetragen, um die Erkennung hasserfüllter Inhalte zu verbessern. Ein weiterer wichtiger Aspekt von **sCAN** war die Stärkung der von der Europäischen Kommission eingerichteten Monitoring-Maßnahmen. Zudem haben die Projektpartner gemeinsam Wissen und Erkenntnisse gesammelt, um Trends des Cyberhasses auf länderübergreifender Ebene besser zu identifizieren, zu erklären und zu verstehen. Darüber hinaus zielte das Projekt darauf ab, europaweit Fähigkeiten von Cyber-Aktivist*innen, Moderator*innen und Tutor*innen zu entwickeln, indem E-Learning-Kurse über die Facing Facts! Online-Plattform angeboten wurden.

sCAN wurde von zehn verschiedenen europäischen Partnern umgesetzt: ZARA - Zivilcourage und Anti-Rassismus-Arbeit aus Österreich, CEJI – A Jewish contribution to an inclusive Europe aus Belgien, Human Rights House Zagreb aus Kroatien, Romea aus Tschechien, Licra – International League Against Racism and Antisemitism sowie Respect Zone aus Frankreich, jugendschutz.net aus Deutschland, CESIE aus Italien, Latvian Centre For Human Rights aus Lettland und die Universität Ljubljana, Fakultät für Sozialwissenschaften aus Slowenien.

Das **sCAN**-Projekt wurde von der Generaldirektion Justiz und Verbraucher der Europäischen Kommission im Rahmen des Programms für Rechte, Gleichstellung und Unionsbürgerschaft (REC) der Europäischen Union finanziert.

Project funded by the European Union's Rights, Equality and Citizenship Programme
(2014-2020)

Inhalt

Einleitung

Im zweien Projektjahr nahmen die sCAN-Partnerorganisationen an zwei Monitoring-Runden teil, eine davon mit der Europäischen Kommission und die zweite mit dem International Network Against Cyber Hate (INACH) sowie dem Projekt „Open Code for Hate-Free Communication“ (OpCode). Ziel des Monitorings war es, die Einhaltung des von der Europäischen Kommission 2016 entwickelten Verhaltenskodexes für die Bekämpfung illegaler Hassreden im Internet durch die IT-Unternehmen Facebook, Twitter, YouTube und Instagram zu überprüfen. Die sCAN-Partner haben bereits an den vorangegangenen Monitoring-Runden teilgenommen, die von der Europäischen Kommission und INACH organisiert wurden.

Im Verhaltenskodex verpflichten sich die IT-Unternehmen, „die Mehrheit der gültigen Meldungen in Bezug auf die Entfernung illegaler Hassreden in weniger als 24 Stunden zu prüfen“ und den Zugang zu solchen Inhalten, die gegen ihre Gemeinschaftsrichtlinien und/oder das nationale Recht verstoßen, zu entfernen oder einzuschränken[1]. Da der Zeitpunkt der Überprüfung von Meldungen für externe Organisationen nicht abschätzbar ist, haben die sCAN-Partner den Zeitpunkt erfasst, zu dem das Unternehmen, an das gemeldet wurde, Maßnahmen ergriffen oder Feedback zu den betreffenden Meldungen gegeben hat.

Zwischen dem 4. November und dem 13. Dezember 2019 nahmen die sCAN-Partner an der fünften Monitoring-Runde teil, die von der Europäischen Kommission seit 2016 organisiert wurde. Während des Monitoringzeitraums meldeten die Partner 635 Fälle von illegaler Hassrede online an die IT-Unternehmen Facebook, Twitter, YouTube, Instagram, Dailymotion und Jeuxvidéo.

Vom 20. Januar bis 28. Februar 2020 arbeitete das sCAN-Projekt bei der Organisation eines unangekündigten Monitorings mit INACH und dem Projekt OpCode zusammen. Der Zeitpunkt dieser Überwachung wurde so gewählt, dass er der sCAN-Projektdauer bis Ende April 2020 entsprach. Während des Monitorings meldeten die sCAN-Partner 484 Fälle illegaler Hassrede online an die IT-Unternehmen Facebook, Twitter, YouTube und Instagram.

Neun sCAN-Partner nahmen an dem Monitoring teil:

- ZARA (Österreich)
- CEJI (Belgien)
- Human Rights House Zagreb (Kroatien)
- Romea (Tschechische Republik)
- Licra (Frankreich)
- jugendschutz.net (Deutschland)
- CESIE (Italien)
- Latvian Center for Human Rights (Lettland)
- Universität Ljubljana, Fakultät Sozialwissenschaften (UL-FDV) (Slowenien)

Neben den sCAN Organisationen, nahmen zudem das INACH Sekretariat und die Partnerorganisationen des Projekts OpCode – ActiveWatch (Rumänien), DigiQ (Slowakei), Estonian Human Rights Centre (Estland), Movimiento contra la Intolerancia (Spanien) und Never Again (Polen) – an der Monitorübung teil. Aus Gründen der Vergleichbarkeit enthält der sCAN-Monitoring-Bericht nur die von den sCAN-Projektpartnern gemeldeten Fälle.

Die Ergebnisse dieser Monitoring-Runden sollten nicht als eine umfassende Studie zur Verbreitung von Hassrede in den Social Media interpretiert werden. Sie können nur eine Momentaufnahme der Inhalte darstellen, die die teilnehmenden Organisationen in einem spezifischen Zeitraum von sechs Wochen auf den von ihnen überwachten Plattformen aufgefunden haben. Einige teilnehmende Organisationen

[1] Europäische Kommission (2016). *Verhaltenskodex für die Bekämpfung illegaler Hassreden im Internet*. Verfügbar unter ec.europa.eu/justice/fundamental-rights/files/hate_speech_code_of_conduct_en.pdf (Zuletzt abgerufen am: 27.04.2020)

konzentrierten ihre Arbeit zudem hauptsächlich auf bestimmte Arten von online Hassrede. Dies kann sich auf die während des Monitoringzeitraums gemeldeten Fälle auswirken und wird im Folgenden näher erläutert. Des Weiteren lag der Schwerpunkt des Monitorings auf der Reaktion der IT-Unternehmen und nicht auf den konkreten Inhalten der identifizierten illegalen Hassrede.

Methodologie

Wie bereits in den vorherigen Monitoring-Runden folgte die Methodologie dem Prozess, den die Europäische Kommission in den vorangegangenen Monitoring-Runden eingeführt hat. In einem ersten Schritt sammelten die teilnehmenden Organisationen Fälle illegaler Hassrede auf den in das Monitoring einbezogenen Social-Media-Plattformen. Die Rechtmäßigkeit der Inhalte wurde auf der Grundlage der nationalen Rechtsvorschriften zur Umsetzung des Rahmenbeschlusses 2008/913/JI zur strafrechtlichen Bekämpfung bestimmter Formen und Ausdrucksformen von Rassismus und Fremdenfeindlichkeit bewertet[2].

Um die Reaktion der IT-Unternehmen auf Benachrichtigungen aus ihrer allgemeinen Nutzerbasis zu testen, wurden die Inhalte zunächst über die öffentlichen Berichtswege der jeweiligen Unternehmen gemeldet. Im Anschluss an diese Meldungen erfassten die Partnerorganisationen, ob die IT-Unternehmen auf die Meldung reagierten, indem sie den Inhalt innerhalb gemeinsam vereinbarter Zeiträume (24h, 48h, 1 Woche) entweder entfernten oder eingeschränkten (Geo-Blocking, Eingeschränkte Funktionen usw.). Darüber hinaus erfassten die Partner, ob und wann sie von den IT-Unternehmen Feedback zu ihrer Meldung erhielten. Die Bereitstellung von Feedback zu Benutzerbenachrichtigungen ist unerlässlich, um diese aktiv und motiviert zu halten, illegale Inhalte an die Unternehmen zu melden.

Einige Partnerorganisationen nahmen an einem zusätzlichen Monitoring-Schritt teil, indem sie solche Inhalte, die nicht innerhalb einer Woche nach der ersten Meldung entfernt wurden, über Kanäle meldeten, die nur denjenigen Organisationen zur Verfügung stehen, die von den IT-Unternehmen als „trusted flaggers“ (vertrauenswürdige Melder) anerkannt sind. Nach dieser zweiten Meldung durchliefen die Partnerorganisationen erneut den Prozess des Monitorings und erfassten die Reaktion und das Feedback der IT-Unternehmen.

Die sCAN-Organisationen einigten sich darauf, zwischen Inhalten, die von der Plattform entfernt wurden, und solchen, die von den IT-Unternehmen eingeschränkt, aber nicht entfernt wurden zu unterscheiden. Fast alle (99%) eingeschränkten Inhalte wurden geo-blockiert, so dass sie für Benutzer*innen, die sich aus dem Land einloggten, aus dem der Inhalt ursprünglich gemeldet wurde, nicht mehr verfügbar waren. Andere Formen der Einschränkung umfassen die Beschränkung bestimmter Merkmale des Inhalts (z.B. Kommentarfunktion) oder die Kennzeichnung als sensibler Inhalt. Die sCAN-Partner betrachten die Einschränkung von Inhalten nur als teilweise effektiv, da die Inhalte online bleiben und Methoden zur Umgehung der Einschränkungen in der Online-Community allgemein bekannt sind.

Für die erste Monitoring-Runde, die im vorliegenden Bericht behandelt wird, wurde die Datenerhebung mit Hilfe einer Online-Vorlage durchgeführt, die von der Europäischen Kommission entworfen und verwaltet wurde. Die Fälle wurden zusätzlich zur internen Analyse durch die sCAN-Partner in Excel-Dateien erfasst. Für das zweite Monitoring einigten sich die Partner auf die Verwendung einer standardisierten Excel-Vorlage, die auf Vorschlägen der sCAN-Teilnehmer*innen basierte und vom INACH-Sekretariat erstellt wurde.

[2] European Union (2008). *COUNCIL FRAMEWORK DECISION 2008/913/JHA on combating certain forms and expressions of racism and xenophobia by means of criminal law*. Verfügbar unter https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32008F0913&from=EN (Zuletzt abgerufen am: 26.03.2020).

Kennzahlen

Die Ergebnisse dieser Monitoring-Runden sollten nicht als eine umfassende Studie zur Verbreitung von Hassrede in den Social Media interpretiert werden. Sie können nur eine Momentaufnahme der Inhalte darstellen, die die teilnehmenden Organisationen in einem spezifischen Zeitraum von sechs Wochen auf den von ihnen überwachten Plattformen aufgefunden haben. Einige teilnehmende Organisationen konzentrierten ihre Arbeit hauptsächlich auf bestimmte Arten von online Hassrede. Dies kann sich auf die während des Monitorings gemeldeten Fälle auswirken und wird im Folgenden näher erläutert. Zudem lag der Schwerpunkt des Monitorings auf der Reaktion der IT-Unternehmen und nicht auf dem konkreten Inhalt der identifizierten illegalen Hassrede.

Drittes Monitoring: 4. November bis 13. Dezember 2019

Das dritte sCAN-Monitoring wurde während der von der Europäischen Kommission vom 4. November bis 13. Dezember 2019 organisierten Monitoring-Runde durchgeführt. In diesem Zeitraum von sechs Wochen meldeten die sCAN-Partner 635 Fälle illegaler Hassrede an die IT-Unternehmen Facebook, Instagram, Twitter, YouTube, Dailymotion und Jeuxvideo. Facebook erhielt von den sCAN-Partnern die meisten Meldungen (280 Fälle), gefolgt von Twitter mit 198 Fällen. YouTube erhielt 102 Meldungen über illegale Hassrede und Instagram erhielt 37 solcher Meldungen von den sCAN-Partnern.

84 Fälle wurden über Kanäle gemeldet, die nur den Trusted Flaggern der IT-Unternehmen zur Verfügung standen, nachdem sie nicht innerhalb einer Woche nach der ersten Meldung über die Meldewege der allgemeinen Benutzer*innen entfernt worden waren. Twitter erhielt 59 Meldungen von Trusted Flaggern, Facebook und Instagram jeweils 10 und YouTube erhielt 5 Meldungen über Kanäle Trusted Flagger. Es wurden keine Fälle an Dailymotion und Jeuxvideo gemeldet.

Analyse der Hassarten

Die Partner klassifizierten die verschieden Arten von Hass, die die Grundlage für illegale Hassrede bilden, auf Basis von der Europäischen Kommission erstellter Kategorien. Das in der Stichprobe am weitesten verbreitete Hassphänomen war Fremdenfeindlichkeit, einschließlich des Hasses gegen Flüchtlinge (31% der Fälle). Nach der Erfahrung von INACH[3] und von Projekten, die zuvor von den sCAN-Partnern durchgeführt wurden, entwickelte sich der Hass gegen Flüchtlinge im Jahr 2015 mit dem Beginn der sogenannten „Flüchtlingskrise“ [4] zu einem weit verbreiteten Hassphänomen. Wir raten deshalb dringend, in weiteren Analysen die Zahlen für Fremdenfeindlichkeit, die nicht mit dem (vermeintlichen) Flüchtlingsstatus der Zielperson zusammenhängen, von dem eindeutig flüchtlingsfeindlichen Hass zu trennen.

Die zweithäufigste Hassart war der antimuslimische Hass (15 %). Antimuslimischer Hass wird oft mit flüchtlingsfeindlichem Hass in Verbindung gebracht, da Hassredner dazu neigen, alle Flüchtlinge als Muslime und alle Muslime als Flüchtlinge zu betrachten. [5] 11 % der von den sCAN-Partnern gemeldeten Fälle von Hassrede beruhten auf Rassismus.

---

[3] INACH (2016). *"Kick them back into the sea" – Online hate speech against refugees*. Verfügbar unter: https://www.inach.net/kick-them-back-into-the-sea/ (Zuletzt abgerufen am: 26.03.2020).

[4] Der Begriff ist negativ konnotiert, da er andeutet, Geflüchtete selbst seien problematisch und die Aufnahme von Geflüchteten sei per se kritisch. Nach dem Verständnis der Autoren zeigt der Begriff jedoch die stark kontroverse öffentliche Debatte zu dem Thema, die zunehmende Skandalisierung von Migration und die hasserfüllte Atmosphäre gegenüber Geflüchteten.

[5] Weitere Informationen zu intersektionaler Hassrede auf Grund der Religion und der angenommenen ethnischen finden Sie hier: sCAN project (2020). *Intersektionale Hassrede Online*. Verfügbar unter: http://scan-project.eu/wp-content/uploads/sCAN_Intersektionale-Hassrede_DEU_final.pdf (Zuletzt abgerufen am: 27.04.2020).

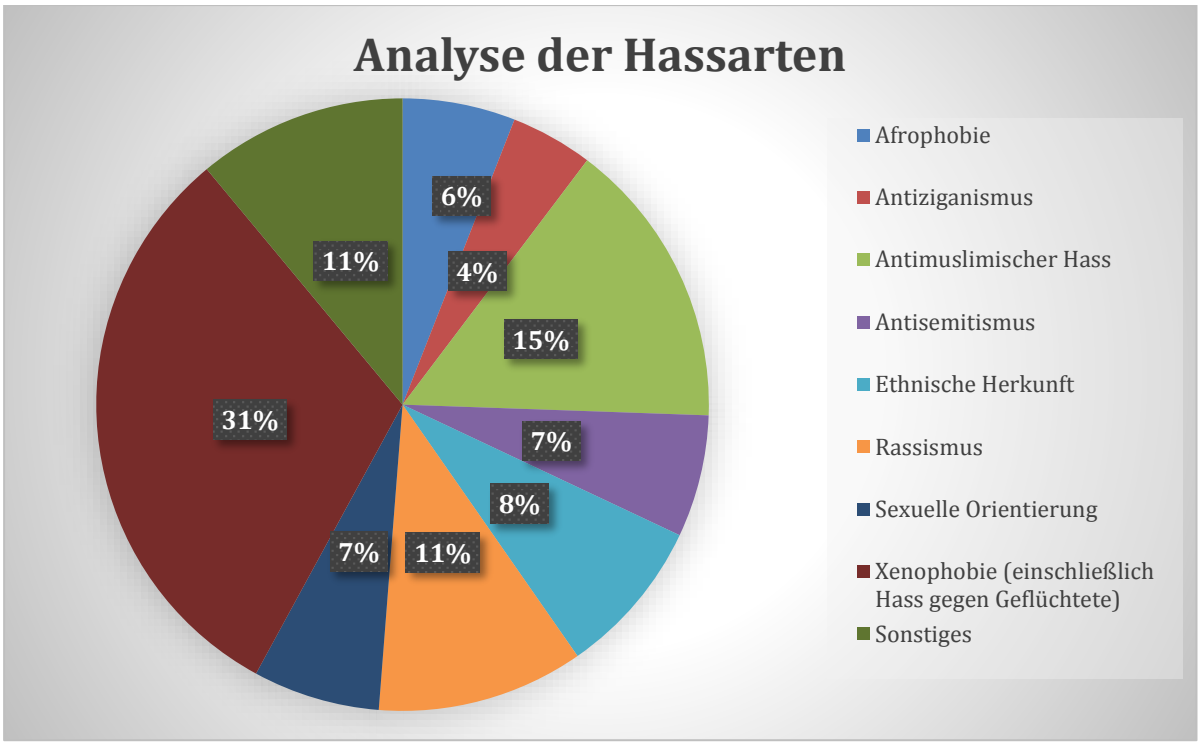

Abbildung 1: Analyse der Hassarten; sCAN Monitoring 4. November – 13. Dezember 2019

Löschquote

Insgesamt waren 67,56 % der Inhalte am Ende des Monitoringzeitraums in dem Land, aus dem sie gemeldet wurde, nicht mehr verfügbar (64,25 % entfernt, 3,31 % eingeschränkt). Diese Zahl steht im Einklang mit den Ergebnissen früherer Monitorings, die von den sCAN-Partnern durchgeführt wurden. Die IT-Unternehmen handelten in 58,74 % der Fälle direkt nach dem ersten Melden über die normalen Benutzerkanäle (57,80 % entfernt, 0,94 % eingeschränkt). Einige Partner meldeten Inhalte, die nicht innerhalb einer Woche nach der Erstmeldung entfernt wurden, erneut über die Kanäle, die *Trusted Flaggern* zur Verfügung stehen. Die Unternehmen reagierten auf 66,67% der Meldungen von *Trusted Flaggern* (48,81% entfernt, 17,86% eingeschränkt).

Dailymotion und Jeuxvideo waren dem Verhaltenskodex kurz vor der Monitoring-Runde beigetreten. Da ihnen nur sehr wenige Fälle gemeldet wurden (6 bzw. 12) und sie nicht in die vierte Monitoring-Runde einbezogen wurden, werden ihre Zahlen aus Gründen der Vergleichbarkeit getrennt dargestellt**. Jeuxvideo entfernte 100% der Fälle, die ihnen über Kanäle der allgemeinen Benutzer*innen gemeldet wurden innerhalb von 24 Stunden.** Dailymotion entfernte 33% der ihnen gemeldeten Fälle. Wenn sie Fälle entfernten, taten sie dies innerhalb von 24 Stunden. Weder an Dailymotion noch Jeuxvideo wurden Fälle über Trusted Flagger Kanäle gemeldet.

Von den anderen überwachten Plattformen erzielte Facebook die höchste Löschquote (83,21%) bei Fällen, die über Kanäle der allgemeinen Benutzer*innen gemeldet wurden. YouTube entfernte 76% dieser Fälle, Instagram 46% und Twitter ergriff nur in 16% der Fälle Maßnahmen, indem es 13% entfernte und weitere 3% einschränkte (Geo-blocking).

Alle Plattformen schnitten bei Meldungen durch Trusted Flagger Kanäle erheblich besser ab. YouTube entfernte 100 % der Inhalte, die über Trusted Flagger Kanäle gemeldet wurden. Facebook ergriff in 90 % der Fälle Maßnahmen, indem es 70 % einschränkte und 20% entfernte. Den Projektpartnern ist nicht bekannt, warum sich die Plattform dafür entschieden hat, einen so großen Prozentsatz der Fälle nur zu beschränken, anstatt sie zu entfernen. Instagram entfernte 60% der von Trusted Flaggern gemeldeten Fälle. Der signifikanteste Anstieg der Löschquote wurde bei Twitter verzeichnet. Das Unternehmen ergriff in 61 % der Fälle Maßnahmen (47 % entfernt, 14 % eingeschränkt), das ist **fast viermal so viel**

wie bei Fällen, die über Kanäle gemeldet wurden, die ihrer allgemeinen Nutzerbasis zur Verfügung stehen.

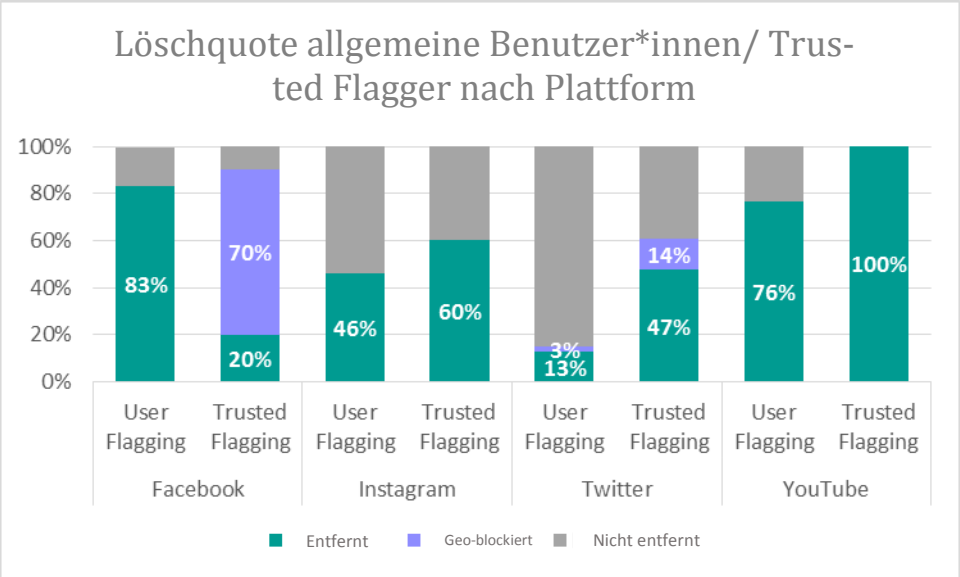

Abbildung 2: Löschquote nach Plattform; sCAN Monitoring vom 4.November – 13. Dezember 2019

Bei den Löschzeiten zeigte Facebook erneut die beste Leistung. Es war die einzige Plattform, die bei der Mehrzahl der über allgemeine Nutzerkanäle gemeldeten Inhalte innerhalb von 24 Stunden (76,43 %) aktiv wurde. YouTube entfernte 47,06 % und Instagram 40,54 % innerhalb von 24 Stunden. Twitter wurde innerhalb dieses Zeitrahmens nur bei 8,59 % der über öffentlich zugängliche Kanäle gemeldeten Inhalte aktiv.

Die meisten Unternehmen reagierten schneller auf Inhalte, die über Trusted Flagger Kanäle gemeldet wurden. Insbesondere Twitter steigerte seine Leistung und entfernte 47,46 % der über diesen Weg gemeldeten Inhalte innerhalb von 24 Stunden. Facebook ergriff bei 70 % dieser Inhalte innerhalb von 24 Stunden Maßnahmen. Instagram entfernte 60 % und YouTube 40 % der Inhalte, die von Trusted Flaggern innerhalb dieses Zeitraums gemeldet wurden.

Feedback

Das Erhalten von Feedback zu gemeldeter Hassrede in sozialen Medien ist sowohl für die Nutzer*innen als auch für Trusted Flagger von entscheidender Bedeutung, um Vertrauen aufzubauen und den Nutzer*innen ein besseres Verständnis dafür zu vermitteln, wie die Plattformen die Inhalte moderieren. Auch der 2016 unterzeichnete Verhaltenskodex fordert von den Unternehmen, zeitnah Feedback zu geben. Dennoch besteht ein nennenswerter Unterschied zwischen den am Verhaltenskodex beteiligten Unternehmen, wenn es um die Bereitstellung von Feedback geht. Eines der Hauptziele der fortgesetzten Organisation von Monitoring-Runden und des vorliegenden Berichtes ist es, die Öffentlichkeit über das Feedback zu gemeldeter Hassrede in sozialen Medien zu informieren.

Insgesamt haben die IT-Unternehmen zu 36,9 % der Meldungen über die für allgemeine Nutzer*innen verfügbaren Kanäle und 56,7 % der Meldungen über die Trusted Flagger Kanäle Feedback gegeben. In Übereinstimmung mit früheren Monitoring-Runden ist die Feedbackrate für Trusted Flagger im Vergleich zur Feedbackrate für normale Benutzer*innen höher – insbesondere in weniger als 24 Stunden (fast 20 Prozentpunkte Unterschied). Es ist wichtig zu beachten, dass die Situation zwischen den IT-Unternehmen hinsichtlich der Feedbackrate sehr unterschiedlich ist.

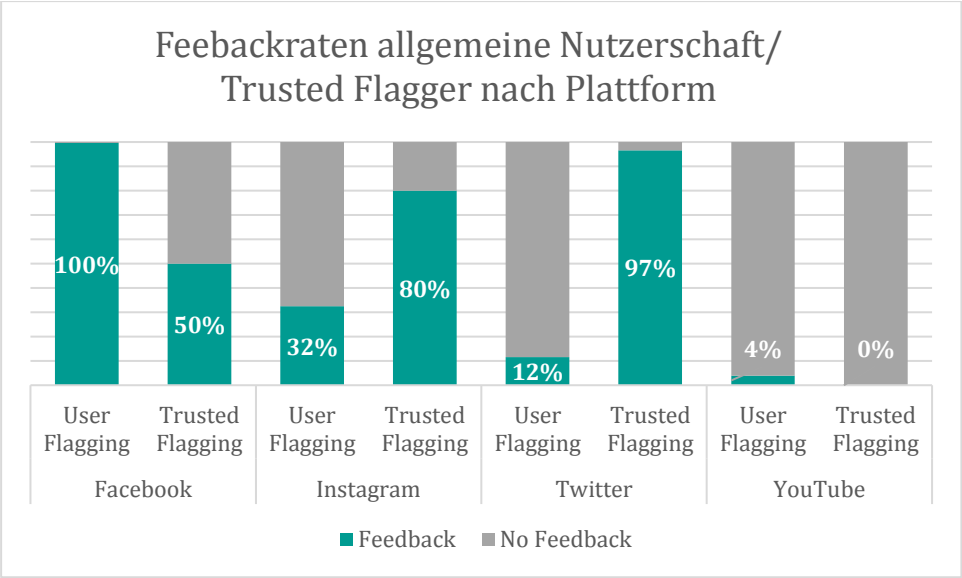

Abbildung 3: Feedbackraten nach Plattform; sCAN Monitoring vom 4.November – 13. Dezember 2019

Facebook kann in diesem Zusammenhang als Einzelfall betrachtet werden: Normale Nutzer*innen erhielten nach dem Melden von Inhalten fast immer eine Rückmeldung. Es war damit zum wiederholten Male das einzige IT-Unternehmen, das allen seinen Nutzer*innen systematisch Rückmeldung gab. Dennoch erhielten Trusted Flagger nur in 50% der Fälle eine spezifische Rückmeldung.

Bei Instagram und Twitter wurde den Meldewegen für Trusted Flagger hinsichtlich der Feedbackrate (80 % und 96,6 %) Priorität eingeräumt. Es scheint, dass deren Richtlinien dazu tendieren, Trusted Flaggern mehr Aufmerksamkeit zu schenken. Auf der anderen Seite hat YouTube nur selten, wenn überhaupt, Rückmeldungen an normale Benutzer*innen und Trusted Flagger gegeben: Für keinen der Fälle, die von den sCAN-Teilnehmern über Trusted Flagger Kanäle gemeldet wurden, haben diese Feedback erhalten.

Wenn IT-Firmen Rückmeldung gaben, so fast immer innerhalb von 24 Stunden. Insgesamt erfolgte das Feedback in 91,2% der über allgemeine Meldewege gemeldeten Fälle in weniger als 24 Stunden. Diese Zeitspanne konnte sich für Trusted Flagger hingegen ausdehnen: nur 75 % des Feedbacks wurden in weniger als 24 Stunden übermittelt.

Viertes Monitoring: 20. Januar bis 28. Februar 2020

Das vierte sCAN-Monitoring fand zwischen dem 20. Januar und dem 28. Februar 2020 statt. Es handelte sich um ein unangekündigtes „stilles“ Monitoring in Zusammenarbeit mit dem INACH-Sekretariat und dem OpCode-Projekt. Die sCAN-Partner meldeten 484 Fälle illegaler Online-Hassrede an die IT-Unternehmen Facebook (242 Fälle), Twitter (127), YouTube (66) und Instagram (49). Um die Reaktion der IT-Firmen auf Meldungen ihrer allgemeinen Nutzerbasis zu testen, wurden die Meldungen zunächst anonym über öffentlich zugängliche Kanäle versandt. In einem zweiten Schritt wurden 94 Fälle, die nach der Benachrichtigung als allgemeine Nutzer*innen nicht entfernt worden waren, erneut über Kanäle gemeldet, die nur für Trusted Flagger zur Verfügung stehen.

Analyse der Hassarten

Um ein ebenso umfassendes und tiefgehendes Bild des Cyberhasses innerhalb der Europäischen Union zu liefern, haben das sCAN-Projekt und INACH die Fälle von Online-Hassrede während dieser Monitoring-Runde in fünfzehn verschiedene Kategorien eingeteilt. Es wurde ebenfalls eine zusätzliche Kategorie „Sonstiges“ hinzugefügt, um diejenigen Fälle erfassen zu können, die sonst durch die Maschen

fallen würden. So berichtete beispielsweise der französische Partner Licra über Fälle anti-asiatischer Hassrede im Zusammenhang mit dem Ausbruch von Covid-19, der bereits im Februar 2020 einige Aufmerksamkeit erregte.

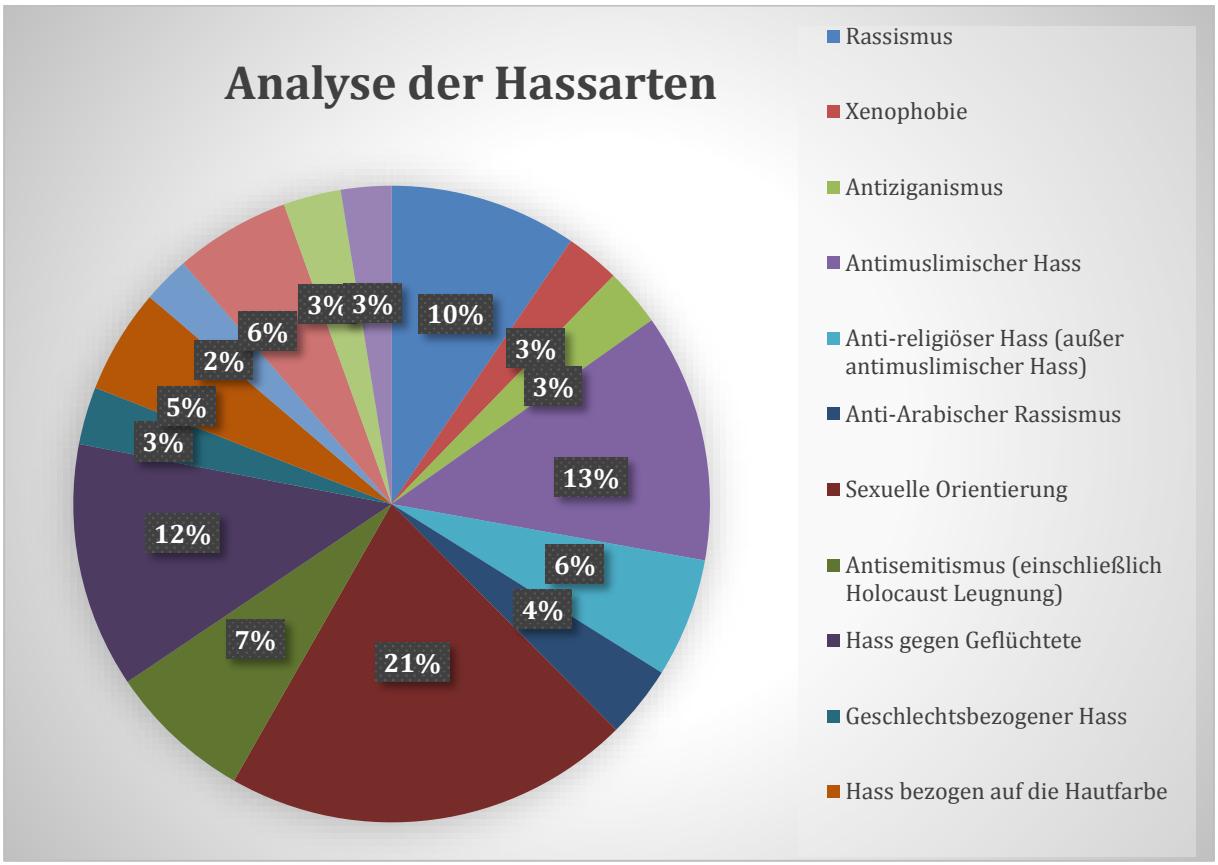

Abbildung 4: Analyse der Hassarten; sCAN Monitoring 20. Januar – 28. Februar 2020

Hass im Zusammenhang mit der sexuellen Orientierung war das am häufigsten beobachtete Hassphänomen in der Stichprobe von Fällen, die im Rahmen dieser Monitoring-Runde gesammelt wurden. Dies kann mit spezifischen, lokalen Ursachen und Ereignissen in Verbindung gebracht werden, die die öffentliche Diskussion in einigen Ländern während des Monitoringzeitraums geprägt haben. Die meisten Fälle von Homophobie wurden in Kroatien und Lettland registriert. Das Human Rights House Zagreb (HRHZ) sammelte insgesamt 49 Fälle, von denen 48 homophob waren, und das Latvian Center for Human Rights (LCHR) sammelte fast 50 homophober Fälle, und meldete insgesamt fast doppelt so viele Fälle wie jede andere am Monitoring beteiligte Organisation. Sowohl das HRHZ als auch das LCHR berichteten, dass es während des Monitoringzeitraumes mehrere Nachrichten gab, die LGBTQ+-Themen in den Mittelpunkt der öffentlichen Debatte rückten. So gab es in Kroatien eine anhaltende öffentliche Debatte darüber, ob schwule Paare Pflegeeltern sein dürfen. In Lettland hingegen stand ein lettischer Politiker im Zentrum der Aufmerksamkeit, der seinen gleichgeschlechtlichen Partner in Deutschland heiratete, und es wurde über eine lettische Basketballspielerin berichtet, die ein Kind mit ihrer gleichgeschlechtlichen Partnerin hat. Diese Ereignisse wirkten als Triebkraft für Hassrede gegenüber der LGBTQ+-Gemeinschaft in den betroffenen Ländern.

Abgesehen von dieser Besonderheit stimmen die in dieser Stichprobe repräsentierten Hassarten mit den Erkenntnissen früherer Monitoring-Runden überein. Antimuslimischer Hass (12,61%) und Hass gegen Geflüchtete (12,43%) führen nach wie vor die Liste an. Ein nicht überraschendes Ergebnis, da die beiden Typen eng miteinander verbunden sind. Es folgen Rassismus (9,51%) und Antisemitismus (7,31%). Antisemitische Hassrede, einschließlich der Leugnung des Holocaust, war besonders um den Holocaust-Gedenktag am 27. Januar weit verbreitet.

Löschquoten

Insgesamt waren nur 58 % der gemeldeten Fälle am Ende des Monitorings nicht mehr verfügbar. Dies ist ein erheblicher Rückgang im Vergleich zur 3. Monitoring-Runde, die nur einen Monat zuvor durchgeführt wurde. Das Ergebnis unterstreicht die Bedeutung einer konsistenten Fallbearbeitung durch die Plattformen, unabhängig von den offiziellen, von der Europäischen Kommission organisierten Monitoring-Runden.

51 % der Fälle wurden bereits nach der ersten Meldung als allgemeine Nutzer*innen entfernt (allgemeine Meldewege). Instagram erreichte die höchste Löschquote mit 75,51 % der Fälle, die nach dem Melden über diese Kanäle entfernt wurden. Facebook entfernte 71,49 % der Fälle nach der Erstmeldung. YouTube und Twitter schnitten deutlich schlechter ab. YouTube entfernte 25,76 % der Fälle nach Meldung als allgemeine Nutzer, während Twitter nur 9,45 % der Fälle entfernte und 4,72 % der Fälle einschränkte.

94 Fälle wurden über Trusted Flagger Kanäle erneut gemeldet, nachdem sie von den Unternehmen bei Meldung über allgemeine Benutzerbenachrichtigungskanäle nicht entfernt worden waren. Von diesen wurden 39 % von den IT-Unternehmen entfernt. Instagram entfernte alle Fälle, die ihnen ein zweites Mal über Trusted Flagger Kanäle gemeldet wurden. Facebook entfernte 68,75% der Fälle, die über Trusted Flagger Kanäle gemeldet wurden. Twitter entfernte einen wesentlich höheren Anteil der Fälle, wenn sie über Trusted Flagger Kanäle gemeldet wurden (41,86 %), und schränkte weitere 4,65% ein, während YouTube weniger Fälle (6,45 %) entfernte, als wenn sie von allgemeinen Nutzer*innen gemeldet wurden.

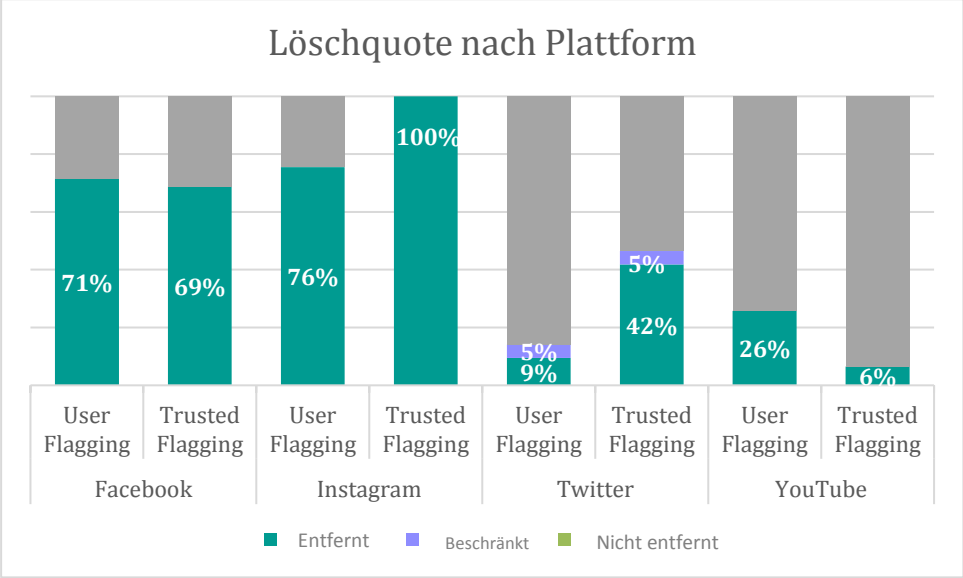

Abbildung 5: Löschquote nach Plattform; sCAN Monitoring 20.Januar – 28. Februar 2020

Bearbeitung und Löschzeitenzeiten

Im Verhaltenskodex verpflichten sich die Plattformen, die Mehrzahl der ihnen gemeldeten illegalen Inhalte in weniger als 24 Stunden zu bearbeiten und gegebenenfalls zu löschen. Die sCAN-Partner erfassten die Löschung von Inhalten und/oder die Abgabe von Feedback zur Bewertung. Bei dieser Monitoring-Runde erreichten die Plattformen das gesetzte Ziel für 44,21 % der Fälle, die ihnen von normale Nutzer*innen gemeldet wurden. 4,13 % der Fälle wurden nach 48 Stunden und 4,96 % innerhalb einer Woche bearbeitet. In 46,69 % der Fälle gab es eine Woche nach dem Erstbericht keinen Hinweis auf eine Bearbeitung.

Nur zwei Plattformen entfernten innerhalb von 24 Stunden den Großteil der über allgemeine Medlewege gemeldeten Inhalte. Instagram entfernte 71,43 % der gemeldeten Inhalte innerhalb von 24 Stunden, sein Mutterunternehmen Facebook entfernte im selben Zeitraum 59 % der gemeldeten Inhalte.

Twitter (2,36 %) und YouTube (1,52 %) entfernten innerhalb von 24 Stunden kaum Inhalte, die als allgemeine Nutzer*innen gemeldet wurden. Wenn die Partner Inhalte über ihre Trusted Flagger Kanäle meldeten, wurden innerhalb von 24 Stunden mehr Inhalte entfernt. Der größte Unterschied in den Löschzeiten zwischen der Meldung normaler Nutzer und dem Melden über Trusted Flagger Kanäle wurde bei Twitter beobachtet, das 37,21 % der von Trusted Flaggern gemeldeten Inhalte innerhalb von 24 Stunden entfernte. Instagram (75 %) und Facebook (62,5 %) verbesserten ihre Leistung weiter, wenn Inhalte von Trusted Flaggern gemeldet wurden. YouTube entfernte innerhalb von 24 Stunden 6,45 % dieser Inhalte.

Fast drei Viertel der Fälle, die die Unternehmen innerhalb einer Woche nach der Meldung untersuchten, wurden anschließend gelöscht, einhergehend mit einer Rückmeldung an den betreffenden Partner. 10 % der untersuchten Fälle wurden gelöscht, aber die meldende Organisation erhielt kein Feedback vom Unternehmen. In 17 % der untersuchten Fälle gab das Unternehmen eine Rückmeldung, in der es die meldende Organisation darüber informierte, dass der Inhalt als nicht gegen die Gemeinschaftsstandards verstoßend erachtet und daher nicht entfernt wurde.

In den Fällen, in denen nach einer Woche kein Hinweis auf eine Bewertung vorlag, überprüften die Partner am Ende des Monitorings, ob die Fälle noch online waren. Die überwiegende Mehrheit (86,36 %) dieser Fälle war nach dem Ende des Monitorings immer noch online, und die Partner erhielten keine Rückmeldung zu ihrer MEldung. In 4,09 % der Fälle erhielten die Partner mehr als eine Woche nach ihrer Meldung eine Rückmeldung, um sie darüber zu informieren, dass der Inhalt entfernt worden war. In 1,36 % der Fälle wurde der Inhalt nicht entfernt, aber die Partner erhielten mehr als eine Woche nach ihrer Meldung an die IT-Unternehmen eine Rückmeldung, die sie über diese Entscheidung informierte. 8,18 % der Fälle wurden irgendwann zwischen einer Woche nach der Meldung und dem Ende des Monitorings entfernt, wobei die IT-Unternehmen die meldende Organisationen nicht über die Entfernung informierten. Es ist daher unmöglich zu sagen, ob die Fälle als Ergebnis des Monitorings oder aus anderen Gründen entfernt wurden.

Die Mehrzahl der erneut gemeldeten Fälle wurde innerhalb von 24 Stunden nach der Meldung bearbeitet (52 %). 2 % wurden innerhalb von 48 Stunden und 5 % innerhalb einer Woche bearbeitet. In 41 % der erneut gemeldeten Fälle gab es keinen Hinweis auf eine Bearbeitung. Die IT-Unternehmen entfernten diese Fälle nicht und gaben den meldenden Organisationen kein Feedback, obwohl diese Organisationen als Trusted Flagger registriert sind.

Die Plattformen schnitten bei der Bearbeitung der Fälle, die ihnen von Trusted Flaggern gemeldet wurden, unterschiedlich ab. Twitter (83,72 %), Instagram (75 %) und Facebook (62,5 %) bearbeiteten die Mehrzahl dieser Fälle in weniger als 24 Stunden. Bei den Fällen, die als Trusted Flagger an YouTube gemeldet wurden, gab es jedoch keinen Hinweis auf eine Bearbeitung (weder Rückmeldung noch Entfernung). Der Erhalt von Feedback zu gemeldeter Hassrede (auch wenn sie von der Plattform nicht entfernt werden) ist entscheidend für eine fruchtbare Zusammenarbeit zwischen den Social-Media-Plattformen und ihren Trusted Flaggern. Es wäre daher sehr zu begrüßen, wenn YouTube mehr Rückmeldungen über die gemeldete Fälle schicken würde, um die Zusammenarbeit zu verbessern.

Feedback

Die IT-Unternehmen gaben Feedback zu 51,45 % der Meldungen über die für normale Benutzer*innen verfügbaren Kanäle (42,56 % in weniger als 24 Stunden) und 60 % der Meldungen über die Trusted Flagger Kanäle (52,63 % in weniger als 24 Stunden). Aus früheren Analysen von Monitoring-Runden geht hervor, dass die Feedbackrate für Trusted Flagger im Vergleich zur Feedbackrate für normale Nutzer*innen höher ist – insbesondere im Rahmen der ersten 24 Stunden nach Erstmeldung: fast 10 Prozentpunkte Unterschied für ein Feedback in weniger als 24 Stunden. Im Vergleich zur letzten Monitoring-Runde hat sich dieser Unterschied jedoch um die Hälfte verringert.

Facebook gab den Nutzer*innen im Vergleich zum letzten Monitoring weniger Feedback - etwa 88%. Bei dieser Monitoring-Runde stieg jedoch die Feedbackrate für Trusted Flagger auf etwa 70,6 %.

Die Feedbackraten von Instagram haben sich erheblich verbessert: Jene an normale Nutzer*innen haben sich fast verdoppelt (von 32 % im dritten Monitoring auf 61 % beim vierten), und in 100 % der Fälle erhielt der Trusted Flagger ein Feedback.

Die Rücklaufquoten von Twitter und YouTube blieben niedrig. Die Feedbackrate von Twitter an allgemeine Nutzer*innen hat sich verschlechtert und lag in der vierten Monitoring-Runde bei 11,6 %. Die Feedbackrate für Trusted Flagger der Plattform lag bei 96,6 %.

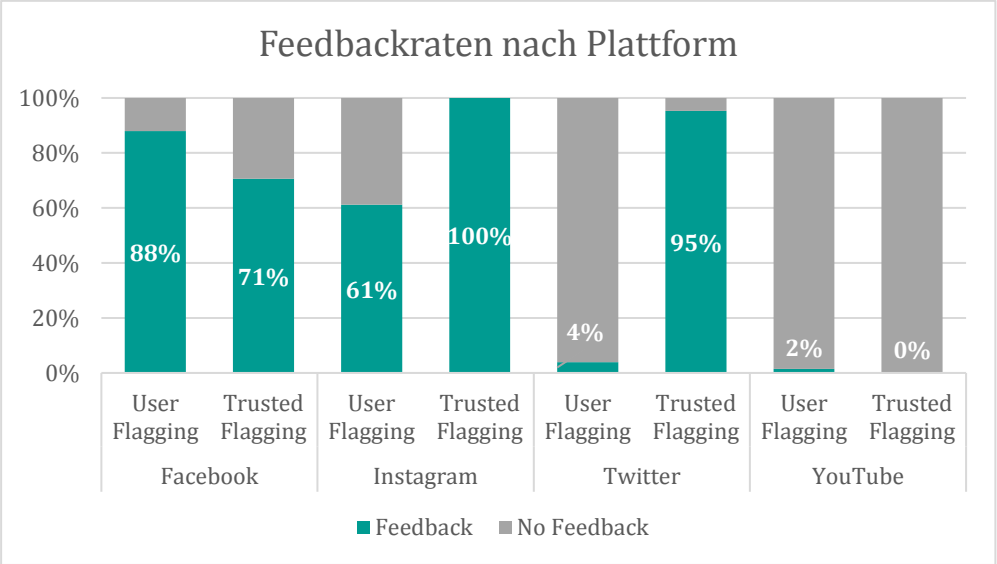

Abbildung 6: Feedbackraten nach Plattform; sCAN Monitoring 20.Januar-28.Februar 2020

Bei YouTube ist die Situation für Meldungen über normale Benutzerkanäle und über Trusted Flagger Kanäle, wie bereits in der dritten Monitoring-Runde, im Grunde dieselbe: YouTube schickte kein Feedback zu Fällen, die von den sCAN-Partnern über Trusted Flagger Kanäle gemeldet wurden.
Die Rückmeldungen von Facebook und Instagram nahmen in dieser vierten Monitoring-Runde etwas mehr Zeit in Anspruch. Facebook lieferte für 71,9 % der Fälle, die von allgemeinen Benutzer*innen gemeldet wurden, in weniger als 24 Stunden Feedback. Bei Instagram lag die Rate bei 61,2 %.

Hervorzuheben ist auch, dass sich die Feedbackzeiten bei Twitter deutlich verlängert haben: Während unserer dritten Monitoring-Runde wurde zu 82 % der Fälle innerhalb von 24 Stunden ein Feedback an normale Benutzer*innen gegeben. Bei unserer vierten Monitoring-Runde traf dies jedoch nur auf 1,57 % der Fälle zu.

Fast die Hälfte der Plattformen gab Trusted Flaggern mehr Feedback als allgemeinen Benutzer*innen. Bei Facebook ist, wie bereits in früheren Zeiträumen, die Feedbackrate bei allgemeinen Nutzermeldungen höher als beim Melden als Trusted Flagger. Was YouTube betrifft, so sind, wie bereits erwähnt, beide Raten verschwindend gering.

Erfahrungen und Beobachtungen

Um von den Erfahrungen und Beobachtungen der Partner für künftige Monitoring-Runden zu profitieren, füllten die sCAN-Partner am Ende des Monitorings einen Evaluationsbogen aus. Einige wichtige Beobachtungen werden in diesem Abschnitt erörtert.

Die Partner berichteten, dass das für die Berichterstattung verwendete Gerät offenbar einen Einfluss darauf hatte, ob sie von Instagram Feedback erhielten oder nicht. Während Partner, die über die mobile App Inhalte meldeten, berichteten, dass sie Feedback von der Plattform erhielten, erreichte Partner, die an Instagram mit einem Desktop-Computer meldeten, kaum Feedback.

Während des Monitoringzeitraums bemerkten die Partner mehrere Accounts, auf denen große Mengen illegaler Hassrede-Kommentare und -Beiträge veröffentlicht wurden. Einige dieser Seiten oder Accounts haben täglich eine beträchtliche Anzahl rassistischer, frauenfeindlicher und extrem gewalttätiger Kommentare veröffentlicht. Daher empfehlen wir, dass die IT-Unternehmen diese genauer überwachen und entschieden gegen jeden Fall illegaler Hassrede auf diesen Accounts vorgehen.

Bei der Beantwortung der Frage, wie sich das Monitoring auf diejenigen auswirkte, die es durchführten, betonten die Partner die zeitintensive Auswirkung auf ihre Arbeit. Einige lösten dieses Problem, indem sie die Arbeitsbelastung auf verschiedene Mitarbeiter*innen aufteilten. Ein weiteres Problem war, dass im regulären Arbeitsfluss der Organisation eine erhöhte Anzahl von gemeldeten Fällen zu bewältigen war und daher die Kapazität zum erneuten Melden von Fällen über Trusted Flagger Kanäle im Rahmen des Monitorings abnahm.

Für einige Organisationen mit umfangreicher Erfahrung in der Durchführung von Monitorings war dies nur eine weitere Monitoring-Runde, und es gab keine zusätzlichen Auswirkungen auf die Expert*innen, Assistent*innen oder Fallbearbeiter*innen.

Darüber hinaus wiesen einige Partner auf eine Reihe psychologischer Faktoren im Zusammenhang mit der Monitoringarbeit hin. So berichtete bspw. die Mitarbeiterin von ROMEA, dass die Monitoringarbeit bei ihr ein Gefühl „des Ekels und der Hilflosigkeit“, aber auch „der Entschlossenheit zum Weitermachen“ hervorrief.

Es ist von besonderem Interesse, dass einige Organisationen eine Form der Trauma- oder Monitoring-bezogenen Stressberatung bereitstellten. Im Fall von CESIE geschah dies, indem im Team diskutiert wurde, wie die Stimmung der Teilnehmer*innen durch die Inhalte, denen sie ausgesetzt waren, beeinflusst wurde. Sie berichteten auch, dass eine ihrer Forscherinnen auch außerhalb ihrer Arbeitszeit und über ihr Arbeitspensums hinaus übermäßig aktiv im Melden von Inhalten wurde. Im Fall von jugendschutz.net haben die Mitarbeiter*innen regelmäßig Zugang zu einer Entlastungsberatung.

Zusammenfassend können wir sagen, dass Monitoring psychologischen und emotionalen Stress bei den Teilnehmenden hervorrufen, sie zeitaufwendig sind, vorzugsweise innerhalb der Organisation von mehreren Mitarbeiter*innen geteilt werden sollte und dass die Verfügbarkeit einer Form von Entlastungsberatung vorteilhaft ist. Die Abstimmung des Monitoringzeitraums mit dem Rhythmus der Organisation kann ebenfalls eine Rolle für den Erfolg des Monitorings spielen.

Fazit

Die Ergebnisse dieser Monitoring-Runden unterstreichen die Notwendigkeit eines konsequenteren Einsatzes von IT-Unternehmen bei der Beseitigung illegaler Hassreden im Internet. Die Gesamtlöschquote von 58% im vierten Monitoring ist fast 10 Prozentpunkte niedriger als die Gesamtlöschquote bei den vorherigen Monitoring-Runden. Dies schließt auch die dritte sCAN Monitoring-Runde im November und Dezember 2019 ein, nur einen Monat zuvor. Die IT-Unternehmen müssen zu jeder Zeit sicherstellen, dass sie zeitnah reagieren und illegale Online-Hassrede entfernen.

Die meisten Unternehmen geben mehr Feedback an Trusted Flagger als an ihre allgemeine Nutzerbasis. Dies kann problematisch sein, da Organisationen der Zivilgesellschaft, die als Trusted Flagger anerkannt sind, nicht in der Lage sind, die Gesamtheit aller illegalen Hassrede selbst zu überwachen und zu melden.

Die Einbeziehung aller Plattformnutzer*innen in das Melden von illegalen Inhalten ist entscheidend, um Hassrede online wirksam zu bekämpfen. Feedback ist ein wichtiger Aspekt, um die Nutzer*innen engagiert und motiviert zu halten und ihnen ein besseres Verständnis dafür zu vermitteln, wie die Plattformen die Inhalte moderieren und ihre Gemeinschaftsrichtlinien durchsetzen.

Quellen

European Union (2008). *COUNCIL FRAMEWORK DECISION 2008/913/JHA on combating certain forms and expressions of racism and xenophobia by means of criminal law*. Verfügbar unter: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32008F0913&from=EN (Zuletzt abgerufen am: 26.03.2020).

Europäische Kommission (2016). Verhaltenskodex für die Bekämpfung illegaler Hassreden im Internet. Verfügbar unter ec.europa.eu/justice/fundamental-rights/files/hate_speech_code_of_conduct_en.pdf (Zuletzt abgerufen am: 27.04.2020)

INACH (2016). *"Kick them back into the sea" – Online hate speech against refugees*. Verfügbar unter: https://www.inach.net/kick-them-back-into-the-sea/ (Zuletzt abgerufen am: 26.03.2020).

sCAN project (2020). *Intersectional Hate Speech Online*. Verfügbar unter: http://scan-project.eu/wp-content/uploads/sCAN_intersectional_hate_final.pdf (Zuletzt abgerufen am: 26.03.2020).