Platforms, Experts, Tools: Specialised Cyber-Activists Network

Monitoring Report

2018 – 2019

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

Über das Projekt

Das von der EU geförderte project **sCAN** – *Platforms, Experts, Tools: Specialised Cyber-Activists Network* (2018-2020), koordiniert von Licra (International League Against Racism and Antisemitism), zielt 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-Hassreden zu entwickeln. Das Projekt stützt sich auf die Ergebnisse bereits realisierter, erfolgreicher europäischer Projekte, darunter "*Research, Report, Remove: Countering Cyber-Hate phenomena*" und "*Facing Facts",* und ist bestrebt, die von der Zivilgesellschaft entwickelten Initiativen zur Bekämpfung von Hassrede fortzusetzen, zu präzisieren und zu stärken.

Durch die europaweite Kooperation werden die Projektpartner ihre fruchtbare Zusammenarbeit (weiter) verstärken und intensivieren. Die **sCAN**-Projektpartner werden zur Auswahl und Bereitstellung relevanter, automatischer Überwachungsinstrumente beitragen, um die Erkennung hasserfüllter Inhalte zu verbessern. Ein weiterer wichtiger Aspekt von **sCAN** ist die Stärkung der, von der Europäischen Kommission eingerichteten, Monitoring-Maßnahmen (z.B. der Monitoring-Übungen). Zudem werden die Projektpartner gemeinsam Wissen und Erkenntnisse sammeln, um Trends des Cyberhasses auf länderübergreifender Ebene besser zu identifizieren, zu erklären und zu verstehen. Darüber hinaus zielt das Projekt darauf ab, europaweit Fähigkeiten von Cyber-Aktivisten, Moderatoren und Tutoren zu entwickeln, indem E-Learning-Kurse über die Facing Facts Online Plattform angeboten werden.

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

Haftungsausschluss

Dieser Monitoring Report wurde durch das Programm der Europäischen Union für Rechte, Gleichstellung und Unionsbürgerschaft (2014-2020) finanziert.

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

Inhalt

Einleitung

Im ersten Jahr des Projekts nahmen die sCAN-Partnerorganisationen an zwei gemeinsamen Monitoringübungen mit der Europäischen Kommission und dem International Network Against Cyber Hate (INACH) teil. Ziel der Monitoringübungen war es, die Einhaltung des von der Europäischen Kommission in 2016 entwickelten Verhaltenskodexes der IT-Unternehmen Facebook, Twitter, YouTube und Instagram zur Bekämpfung illegaler *Hatespeech (*Hassrede) im Internet zu bewerten. Zwischen 2016 und 2018 gab es vier Monitoringzeiträume zur Überprüfung des Verhaltenskodex der Europäischen Kommission. Die meisten sCAN-Partner haben bereits an den vorangegangenen Monitoringübungen 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 dieEntfernung illegaler Hassrede in weniger als 24 Stunden zu prüfen"[1] und den Zugang zu Inhalten, die gegen ihre Community-Richtlinien und/oder das nationale Recht verstoßen, zu entfernen oder einzuschränken. 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 diesen Meldungen gegeben hat.

Das erste Monitoring während der Laufzeit des sCAN-Projekts wurde von der Europäischen Kommission organisiert und zwischen dem 05. November 2018 und dem 14. Dezember 2018 durchgeführt. Während dieser ersten Monitoringübung meldeten die sCAN-Partner 762 Fälle illegaler online *Hatespeech* an die IT-Unternehmen Facebook, Twitter, YouTube, Instagram, Google+, Dailymotion und Jeuxvidéo. Das zweite Monitoring wurde gemeinsam von den sCAN-Partnern und dem International Network Against Cyber Hate (INACH) organisiert. Es wurde vom 6. Mai 2019 bis zum 21. Juni 2019 durchgeführt. Während dieses Monitorings meldeten die Partner 432 Fälle an die IT-Unternehmen Facebook, Twitter, YouTube und Instagram.

Neun sCAN-Partner nahmen an der Monitoringübung 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 INACH-Partnerorganisationen Greek Helsinki Monitor (Griechenland) und Never Again Association (Polen) am zweiten Monitoring teil.

Die Ergebnisse dieser Monitoringübung sollten nicht als eine umfassende Studie über die 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 überprüften Plattformen aufgefunden haben. Einige teilnehmende Organisationen konzentrierten ihre Arbeit zudem hauptsächlich auf bestimmte Arten von online *Hatespeech*. Dies kann sich auf die während der Überprüfung gemeldeten Fälle auswirken und wird im Folgenden näher erläutert. Des Weiteren lag der Schwerpunkt der Monitoringübung auf der Reaktion der IT-Unternehmen und nicht auf den konkreten Inhalten der identifizierten illegalen Hassrede.

[1] European Commission (2016). Code of Conduct on countering illegal hate speech online. Verfügbar unter https://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=54300 (Letzter Abruf: 22.07.2019).

Methodologie

Die Methodik der Monitoringübungen folgte dem Monitoringprozess, den die Europäische Kommission in den vorangegangenen Zeiträumen 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 Nutzerschaft zu testen, wurde 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 Nutzer*innen 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 Berichtswege meldeten, die nur denjenigen Organisationen zur Verfügung stehen, die von den IT-Unternehmen als "trusted flaggers" (vertrauenswürdige Melder) anerkannt sind. Nach dieser zweiten Berichterstattung 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 geoblockiert, so dass sie für Benutzer, 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.

Um eine gemeinsame Untersuchung und einen Vergleich der Ergebnisse zu ermöglichen, haben die teilnehmenden Organisationen ihre Fälle in länderübergreifenden Datenbanken erfasst. Für das erste Monitoring wurde die Datenerhebung über eine von der Europäischen Kommission entwickelte und verwaltete Online-Vorlage durchgeführt. Für das zweite Monitoring vereinbarten die Partner, die INACH-Datenbank zu Hassrede zu nutzen. Die INACH-Datenbank wurde eingerichtet, um ein internationales Instrument zur Dokumentation und Analyse von Cyberhass-Fällen bereitzustellen und als zentrale Anlaufstelle zur Meldung von Cyberhass-Fällen für Nutzer*innen zu fungieren.

[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 (Letzter Abruf: 22.07.2019).

Kennzahlen

Die Ergebnisse dieses Monitorings sollten nicht als eine umfassende Studie über die 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 Hatespeech. Dies kann sich auf die während der Überwachung gemeldeten Fälle auswirken und wird im Folgenden näher erläutert. Zudem lag der Schwerpunkt der Überwachungsübung auf der Reaktion der IT-Unternehmen und nicht auf dem konkreten Inhalt der identifizierten illegalen Hassrede.

Erstes Monitoring: 5. November 2018 – 14. Dezember 2018

Das erste Monitoring wurde vom 05.11.2018 bis zum 14.12.2018 durchgeführt. Die sCAN-Partner meldeten 762 Fälle illegaler online *Hatespeech* an die IT-Unternehmen Facebook (311 Fälle), Twitter (190), YouTube (142), Instagram (86), Google+ (23), Dailymotion (8) und Jeuxvidéo (2). Um die Reaktion von IT-Unternehmen auf Meldungen ihrer allgemeinen Nutzerschaft zu testen, wurden 755 Benachrichtigungen anonym über öffentlich zugängliche Kanäle gesendet. In einem zweiten Schritt wurden 165 Fälle, die nach Benachrichtigung durch allgemeine Benutzer nicht entfernt wurden, erneut über Berichtskanäle gemeldet, die nur für vertrauenswürdige *Flagger* verfügbar sind. Sieben Fälle wurden direkt über die Kanäle der vertrauenswürdige *Flagger* der Partner gemeldet. Insgesamt wurden 172 Benachrichtigungen an die IT-Unternehmen über die Kanäle der vertrauenswürdige *Flagger* gesendet.

Analyse des Hassarten:

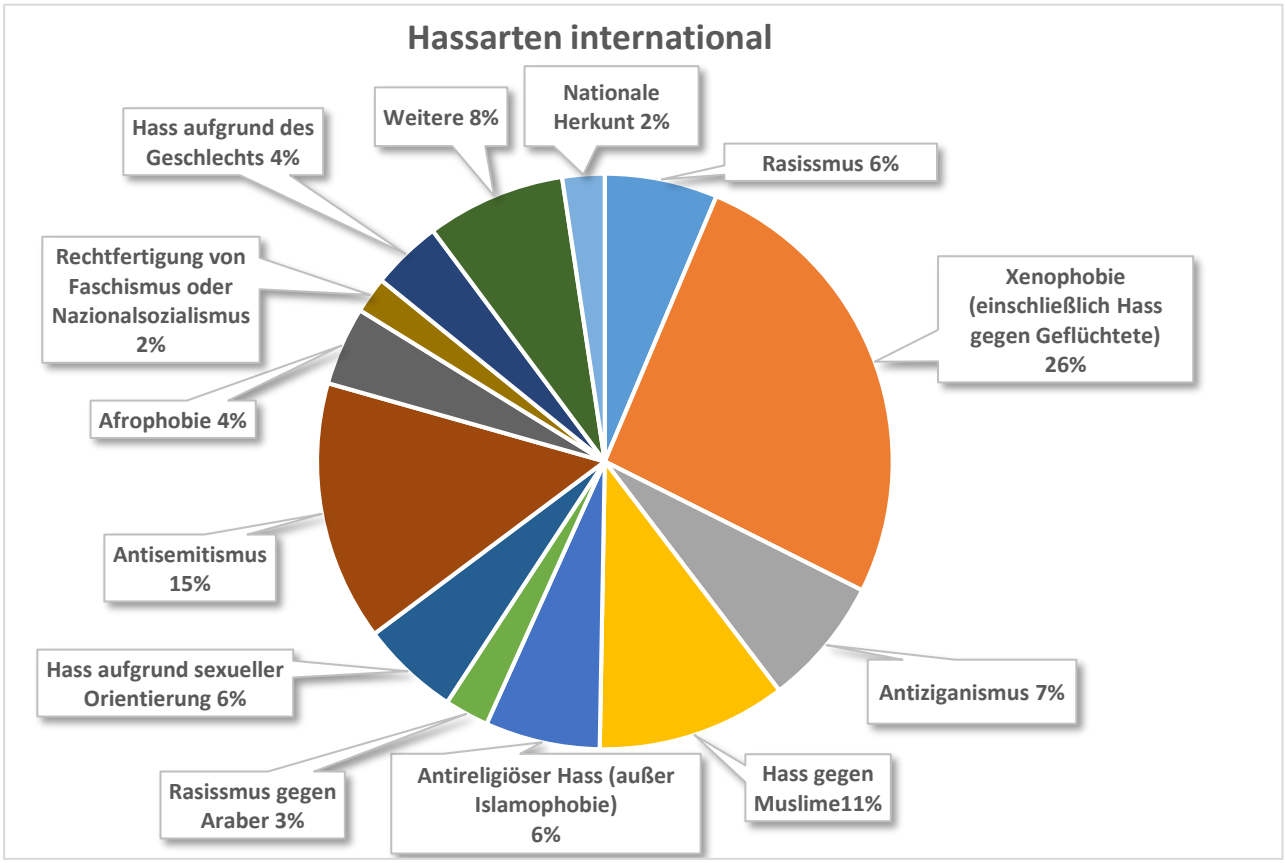

Abbildung 1: Hassarten international; Quelle: sCAN-Monitoring

Für das erste Monitoring wurden Informationen über Hassarten innerhalb der von der Europäischen Kommission für alle am Monitoring beteiligten Organisationen festgelegten Kategorien erfasst. Die häufigsten Hassarten in der Stichprobe der sCAN-Projektpartner waren Fremdenfeindlichkeit (einschließlich Hassrede gegen Geflüchtete) (30%), Antisemitismus (17%) und antimuslimische Hassrede (12%).

Entfernungsraten:

Insgesamt haben die überwachten Unternehmen in 73% der Fälle Maßnahmen ergriffen, indem sie den Inhalt entweder entfernten (67%) oder geo-blockierten (6%). Die Entfernungsraten variierten je nach Berichtsweg, über den die Inhalte gemeldet wurden. Insgesamt reagierten die IT-Unternehmen auf 62% der über allgemeine Nutzerkanäle gemeldet Inhalte (58% Entfernung, 4% Geo-Blocking) sowie auf 60% der über vertrauenswürdige *Flagger*-Kanäle gemeldeten Inhalte (42% Entfernung, 18% Geo-Blocking). Die meisten IT-Unternehmen reagierten häufiger auf Meldungen von vertrauenswürdigen *Flaggern,* als auf solche, die den allgemeinen Nutzer*innen der Plattformen zur Verfügung stehen.

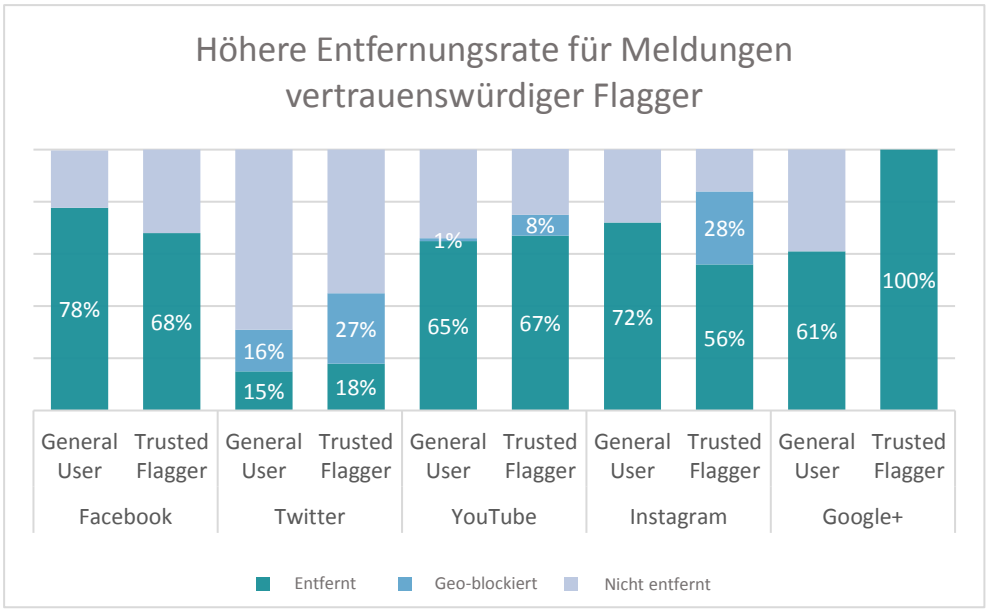

Figure 1: Entfernungsraten allgemeine Nutzerschaft/ vertrauenswürdige Flagger nach Plattform; Quelle: sCAN-Monitoring

Die Entfernungsraten waren nicht nur über die Berichtskanälen, über die die Benachrichtigungen versendet wurden hinweg unterschiedlich, die sCAN-Partner beobachteten auch länderspezifische Unterschiede in der Reaktion auf Meldungen vertrauenswürdige *Flagger*. Facebook hat 100% der von sCAN-Partnern aus Lettland, Deutschland, Frankreich und der Tschechischen Republik über vertrauenswürdige *Flagger* gemeldeten Fälle entfernt, wohingegen nur 50% der vom österreichischen Partner, über vertrauenswürdige *Flagger*-Kanäle gemeldeten Fälle entfernt wurden[3]. Twitter und Instagram wendeten *Geo-Blocking* nur auf Meldungen vertrauenswürdiger *Flagger* des französischen Partners an.

[3] Die vom österreichischen Partner als vertrauenswürdige *Flagger* gemeldeten Fälle, die nicht von Facebook entfernt wurden, wurden von Rechtsexperten überwiegend als strafrechtlich relevant bewertet, weisen aber ein hohes Maß an Komplexität auf. Die Beurteilung dieser Fälle als illegal und als Verstoß gegen die Community-Standards von Facebook erfordert ein gründliches Verständnis des jeweiligen Kontextes. Vermutlich konnte Facebook die (rechtliche) Komplexität der gemeldeten Fälle aufgrund des mangelnden Verständnisses des nationalen Kontextes, der Kenntnis aktueller politischer Ereignisse und einer differenzierten Betrachtung einer Reihe deutscher Dialekte nicht erfassen. Facebook hat die Gelegenheit nicht genutzt, den vertrauenswürdigen Flagger zu konsultieren.

Entfernungszeiten:

Da der Zeitpunkt der Überprüfung eines Berichts für externe Organisationen nicht abschätzbar ist, haben die sCAN-Partner den Zeitpunkt erfasst, zu dem das gemeldete Unternehmen Maßnahmen ergriffen hat. Zwei der überprüften IT-Unternehmen haben den Großteil der Inhalte in weniger als 24 Stunden nach Erhalt einer Meldung über allgemeine Nutzerkanäle entfernt: Facebook (76%) und YouTube (58%). Instagram entfernte 47% dieser Inhalte in weniger als 24 Stunden und Google+ 35%. Twitter hat 12% der Inhalte innerhalb von 24 Stunden entfernt und 13% geogeblockt. Wenn an YouTube über vertrauenswürdige Flagger-Kanäle gemeldet wurde, entfernte das Unternehmen den Inhalt in 67% der Fälle und geoblockierte weitere 8% in weniger als 24 Stunden; Instagram entfernte 50% und geoblockierte 28% der Fälle, Twitter entfernte 17% und geoblockierte 27% der gemeldeten Fälle, während Facebook 32% der Inhalte in diesem Zeitraum entfernte. Google+ hat keinen der von vertrauenswürdigen *Flaggern* gemeldeten Inhalte in weniger als 24 Stunden entfernt.

Feedback:

Insgesamt gaben die IT-Unternehmen Feedback zu 48% der Markierungen über die allgemeinen Nutzerkanäle (46% in weniger als 24 Stunden) und zu 55% der Benachrichtigungen über die vertrauenswürdigen Berichtskanäle (45% in weniger als 24 Stunden). Facebook war das einzige IT-Unternehmen, das systematisch Feedback an alle seine Nutzer*innen gab, während Twitter und YouTube eher Feedback an vertrauenswürdige *Flagger* gaben als an allgemeine Nutzer*innen. Instagram lieferte nur Feedback an vertrauenswürdige *Flagger*. Google+ hat während des Monitoringzeitraums kein einziges Feedback gegeben. Die Bereitstellung von Feedback zu Benutzerbenachrichtigungen ist jedoch unerlässlich, um die Nutzer*innen aktiv und motiviert zu halten, illegale Inhalte an die Unternehmen zu melden.

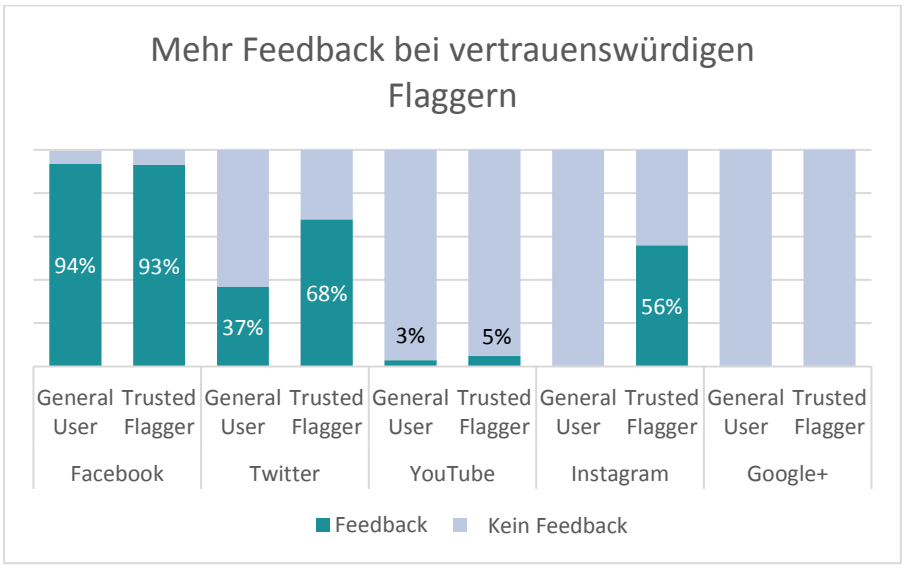

Figure 2: Feedback allgemeine Nutzerschaft/ vertrauenswürdige Flagger nach Plattform; Quelle: sCAN-Monitoring

Erfahrungen und Beobachtungen

Während des Beobachtungszeitraums erhielten Facebook, Twitter und YouTube die höchste Anzahl von Markierungen jeder Partnerorganisation. Instagram wurde zum ersten Mal in die Monitoringübung einbezogen, weshalb einige Partner nur begrenzte Erfahrung mit der Überprüfung dieser Plattform hatten. Die Partner berichteten jedoch, dass sie, obwohl es für sie schwieriger war, relevante Inhalte auf Instagram zu finden, dennoch eine wichtige Anzahl illegaler *Hatespeech* fanden. Google+ wurde im April 2019 geschlossen, weshalb die Partner nur wenige Meldungen an diese Plattform tätigen konnten. Dailymotion und Jeuxvidéo erhielten nur eine geringe Anzahl von Meldungen, aufgrund ihrer geringen Reichweite und Relevanz nur in Kroatien und Frankreich.

Das Wohlergehen der mit der Durchführung des Monitorings beauftragten Forscher war den sCAN-Partnern ein wichtiges Anliegen, weshalb sie dafür sorgten, dass ihre Mitarbeiter während des gesamten Prozesses gut geschult und unterstützt wurden. Um die Privatsphäre und Sicherheit ihrer Mitarbeiter zu gewährleisten, aber dennoch die Erfahrungen der allgemeinen Nutzerschaft beim Melden über öffentlich zugängliche Meldekanäle widerzuspiegeln, richteten die Partner anonymisierte E-Mail-Adressen und Fake-Profile auf den von ihnen überwachten Plattformen ein.

Mehrere Partner beobachteten, dass Twitter auf einige Berichte von vertrauenswürdigen *Flaggern* mit einer Anfrage nach detaillierteren Informationen reagierte, einschließlich Informationen über den Mitarbeiter, der die Meldung getätigt hatte. Den Partnern war nicht klar, warum Twitter nach diesen Informationen fragte. Sie erhielten nach der Bereitstellung der angeforderten Informationen weder weiteres Feedback, noch eine Bewertung und der Inhalt blieb online.

Zweites Monitoring: 6. Mai 2019 – 21. Juni 2019

Die zweite Monitoringübung wurde vom 6. Mai bis zum 21. Juni 2019 durchgeführt. Die sCAN-Partner meldeten 432 Fälle an vier verschiedene Social Media-Plattformen. Diese waren: Facebook (200 Fälle), Twitter (107 Fälle), YouTube (90 Fälle) und schließlich Instagram (35 Fälle). Die IT-Unternehmen wurden durch öffentlich zugängliche Kanäle über alle diese Fälle informiert. Nach dieser ersten Runde meldeten einige sCAN-Partner 90 nicht entfernte Fälle über Kanäle, die nur für vertrauenswürdige *Flagger* zugänglich sind.

Die Partner beobachteten während der Übung eine Vielzahl verschiedener Hassarten. Einige von ihnen waren stärker verbreitet als andere.

Das untenstehende Kreisdiagramm spiegelt eine Momentaufnahme der Trends in online *Hatespeech* wider. Dem Monitoring durch unsere Partner zufolge, waren Hass gegen Flüchtlinge, Homophobie, Hass gegen Muslime und Antisemitismus die häufigsten Hassarten innerhalb des sechswöchigen Beobachtungszeitraums.

Diese Ergebnisse können nur ein kurzlebiges Bild der Inhalte vermitteln, die die teilnehmenden Organisationen in diesem spezifischen Zeitraum von sechs Wochen gefunden haben. Einige teilnehmende Organisationen konzentrierten ihre Arbeit hauptsächlich auf bestimmte Arten von Online-Hassreden. Um die während des Monitoringszeitraums gefundenen Hassarten besser bewerten zu können, gaben die Partner detaillierte Informationen über die von ihnen markierten Hassarten an.

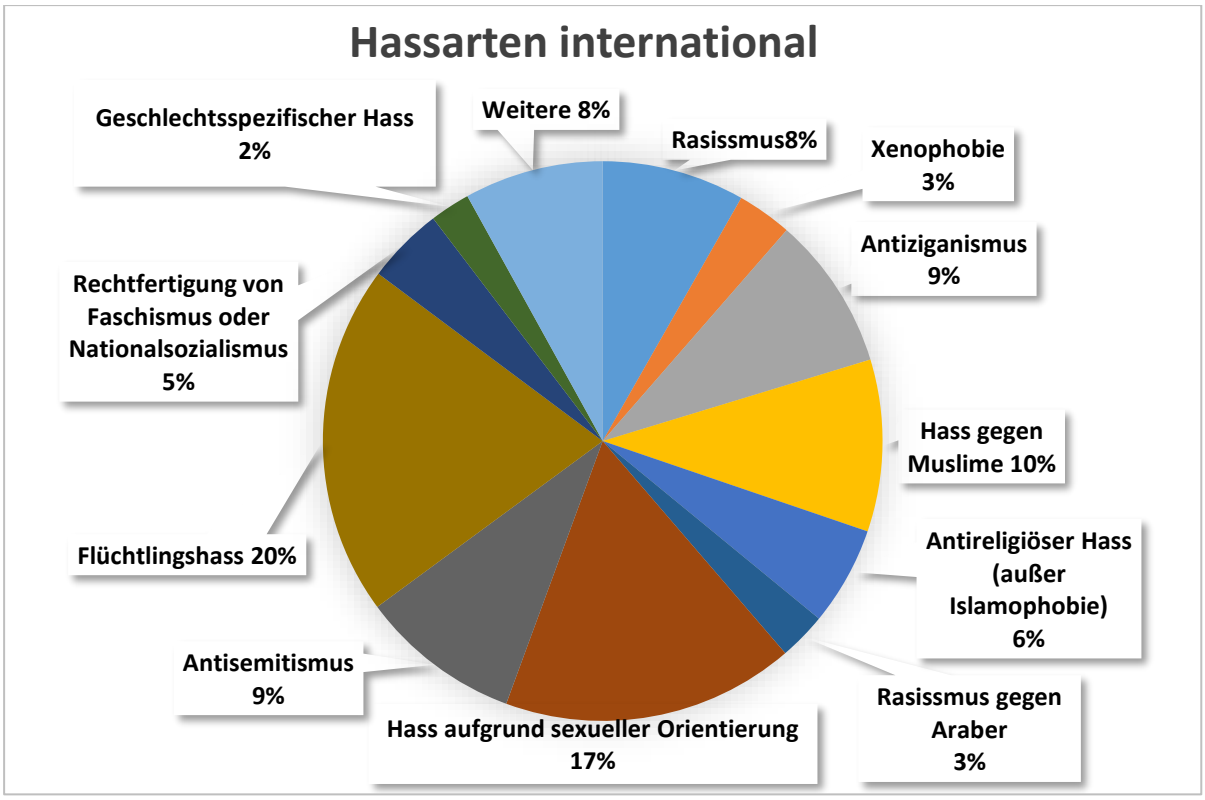

Abbildung 2: Hassarten international; Quelle: sCAN-Monitoring

Im Allgemeinen scheinen die gefundenen Hassarten das generelle Bild der *Hatespeech* in den jeweiligen Ländern widerzuspiegeln. In Österreich und Slowenien ist die Hassrede gegen Flüchtlinge den Erfahrungen der Partnerorganisationen zu folge eine der am weitesten verbreiteten. In der Tschechischen Republik sind Sinti und Roma die Minderheit, die am dauerhaftesten Ziel von Hassrede ist. Seit 2015 ist zudem auch die Hassrede gegen Muslime, Araber, Flüchtlinge und People of Colour weit verbreitet.

In anderen Fällen spiegelte die während des Monitorings identifizierte Hassrede aktuelle Diskurse und Entwicklungen innerhalb der überprüften Länder wider. In Italien gibt es eine deutliche Aufforderung zum Hass, die von der Obrigkeit selbst ausgeht. Die meisten der vom italienischen Partner gemeldeten Fälle von *Hatespeech* waren Reaktionen auf Beiträge und Inhalte hochrangiger Politiker oder politischer Parteien. In Frankreich ist seit Anfang des Jahres eine Häufung antisemitischer Inhalte zu beobachten. In Kroatien wurden vor allem die während des Monitoringzeitraums stattfindenden Pride Paraden von homophober Hassrede im Internet ins Visier genommen.

In Lettland wurde antisemitische *Hatespeech* durch einen Gesetzentwurf zur Entschädigung der jüdischen Gemeinschaft für verloren gegangenes Gemeinschaftseigentum während des Holocausts sowie durch die Tatsache, dass der neu gewählte lettische Präsident lettischer-jüdischer Herkunft ist, losgetreten. Homophobe Hassrede wurden durch den Bericht über Hassdelikte gegen LGBT in Lettland, den vom Parlament abgelehnten Entwurf eines Lebensgemeinschaftsgesetzes sowie Angriffe auf Homosexuelle in Tschetschenien getriggert. Fremdenfeindliche *Hatespeech* wurden durch Diskussionen über Gesetzesänderungen, die es ausländischen Studenten erlauben, Vollzeit in Lettland zu arbeiten, angeregt.

Der deutsche Partner jugendschutz.net überwacht Rechtsextremismus und islamistischen Extremismus kontinuierlich. Islamistische online Propaganda, die vom Deutschen Partner gemeldet wurde, richteten sich meist an all diejenigen, die der islamistischen Ideologie nicht folgen. Die meisten rechtsextremen Fälle, die markiert wurden, enthielten Verherrlichungen des Nationalsozialismus. Es ist zu-

dem wichtig, zur Kenntnis zu nehmen, dass in Deutschland die Verwendung und Verbreitung von Symbolen verfassungswidriger Organisationen verboten ist. Die in Deutschland gemeldeten Fälle beinhalteten daher häufig Symbole, die mit islamistischen Terrororganisationen in Verbindung stehen (z.B. die Flagge des sogenannten IS) oder Symbole nationalsozialistischer Organisationen (z.B. Hakenkreuze oder der SS-Totenkopf).

Entfernungsraten

Insgesamt haben die Social Media Unternehmen 67% der während des Monitorings gemeldeten Inhalte entfernt und 4% eingeschränkt. Aufgrund von Meldungen über die für allgemeine Nutzer verfügbaren Kanäle, haben die IT-Unternehmen 59% der Inhalte entfernt und 3% eingeschränkt. Andere Inhalte wurden erst entfernt, nachdem sie ein zweites Mal über die Kanäle vertrauenswürdiger F*lagger* der Partner gemeldet wurden. Die Unternehmen reagierten auf 43% der Meldungen über Kanäle vertrauenswürdigen *Flagger*, indem sie 40% der gemeldeten Inhalte entfernten und 3% einschränkten.

Überraschenderweise wurden, im Gegensatz zu früheren Monitoringübungen, Fälle, die über die Kanäle vertrauenswürdige *Flagger* gemeldet wurden, nicht in höherem Maße auf den Plattformen entfernt. Twitter war die einzige Plattform, bei der das vertrauenswürdige Markieren einen signifikanten und positiven Unterschied bedeutete.

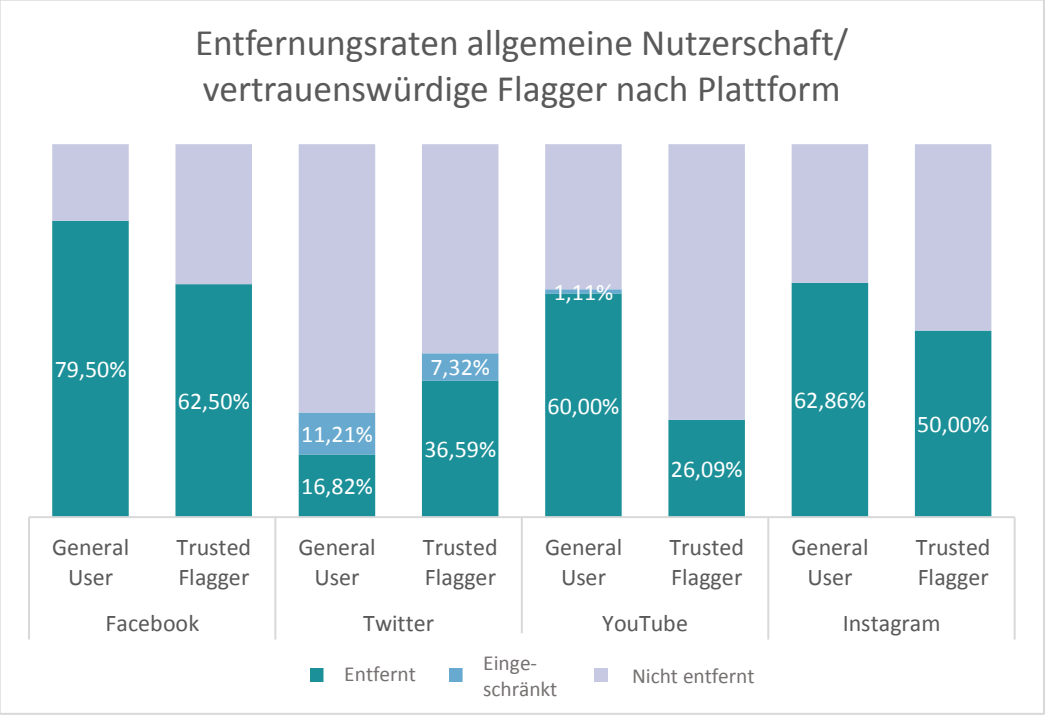

Figure 3: Entfernungsraten allgemeine Nutzerschaft/ vertrauenswürdige Flagger; Quelle: sCAN-Monitoring

Dennoch ist zu beachten, dass die Entfernungsraten für Fälle, die über die Kanäle vertrauenswürdige *Flagger* gemeldet wurden, ziemlich hoch waren und das, obwohl dieselben Fälle bereits abgelehnt worden waren, als sie von normalen Benutzern markiert wurden. Es kann daher immer noch festgehalten werden, dass vertrauenswürdiges Markieren von Hassrede effektiver ist, als jenes durch öffentliche Kanäle, die allen Benutzern auf den überwachten Plattformen zur Verfügung stehen.

Entfernungszeiten

Bezüglich Entfernungsraten und -zeiten sind die Zahlen sehr unterschiedlich. Sie schwanken sowohl von Land zu Land als auch von Plattform zu Plattform. Allgemein kann jedoch festgehalten werden, dass Facebook bei weitem am effektivsten in der Entfernung und Beseitigung von Cyberhass auf Grundlage des Verhaltenskodex ist.

Von den Fällen, die über die den allgemeinen Nutzern zur Verfügung stehenden Kanäle gemeldet wurden, entfernte Facebook 64% innerhalb von 24 Stunden. Instagram hat 43% dieser Fälle innerhalb von 24 Stunden entfernt oder eingeschränkt, YouTube 23%.

Twitter hat innerhalb dieses Zeitraums nur 17% der gemeldeten Inhalte entfernt oder eingeschränkt. Darüber hinaus gab es Länder, in denen das Unternehmen keine der markierten Hassrede entfernt hat: Romea in Tschechien und UL-FDV in Slowenien waren in ihren Versuchen, Inhalte von der Plattform zu entfernen, nicht erfolgreich. Twitter hat in Slowenien nichts entfernt, selbst dann nicht, wenn UL-FDV über Kanäle vertrauenswürdiger *Flagger* an das Unternehmen meldete. In Italien schränkte Twitter nur einen der von CESIE gemeldeten Fälle ein.

Beim Melden über Kanäle vertrauenswürdiger *Flagger* sieht es ähnlich aus: Facebook ist mit Abstand das reaktionsschnellste Unternehmen, wenn es darum geht, Inhalte innerhalb von 24 Stunden zu entfernen (44%). Twitter ist jedoch reaktionsschneller, Inhalte zu entfernen, auch innerhalb von 24 Stunden (37%), wenn sie über die Kanäle vertrauenswürdiger *Flagger* gemeldet werden. Sowohl Instagram als auch YouTube hatten eine Entfernungsrate von 30% innerhalb von 24 Stunden.

Feedback

Das Erhalten von Feedback zu gemeldeter *Hatespeech* in den Social Media ist sowohl für die Nutzer als auch für vertrauenswürdige *Flagger* in den Social Media äußerst wichtig. Darüber hinaus fordert es der Verhaltenskodex der Unternehmen, rechtzeitig substanzielles Feedback zu geben. Der Unterschied zwischen den Unternehmen ist hinsichtlich der Bereitstellung von aussagekräftigem und zeitnahem Feedback groß.

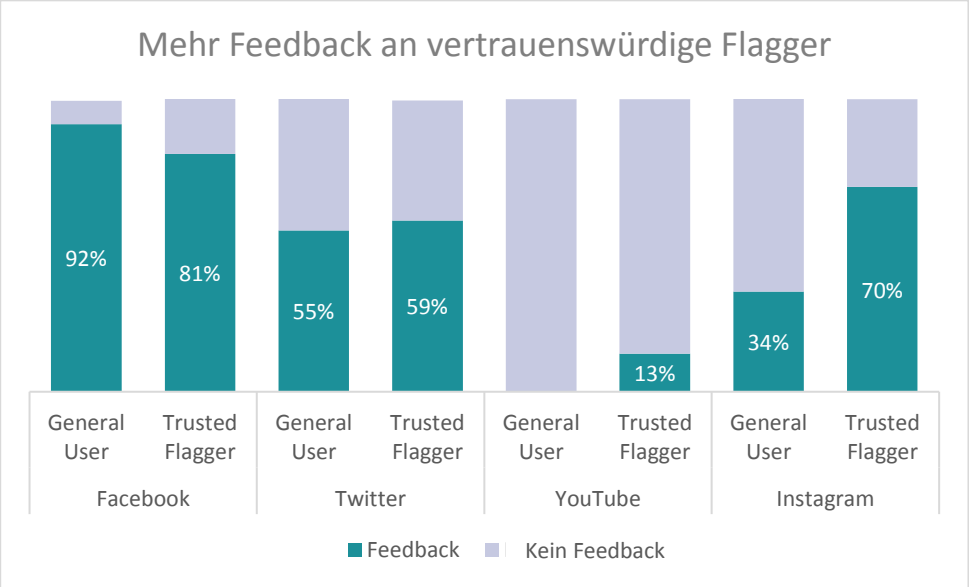

Figure 4: Feedback an die allgemeine Nutzerschaft/ vertrauenswürdige Flagger nach Plattform; Quelle: sCAN-Monitoring

Facebook ist bei weitem der effizienteste Bereitsteller von Feedback. Das Unternehmen reagierte auf 92% der Benachrichtigungen durch die allgemeine Nutzerschaft (74% innerhalb von 24 Stunden). Twitter reagierte auf 55 % dieser Meldungen (45 % innerhalb von 24 Stunden), und Instagram auf 34 % der

Benachrichtigungen von allgemeinen Nutzern innerhalb des Untersuchungszeitraums. YouTube hat kein Feedback zu den von den sCAN-Partnern über allgemeine Berichtswege gemeldeten Fällen gesendet.

Wie aus der obigen Grafik ersichtlich wird, haben die meisten Plattformen mehr Feedback an vertrauenswürdige *Flagger* gesendet als auf allgemeine Benutzerberichte reagiert. Die Ausnahme ist Facebook, das häufiger auf allgemeine Benutzerberichte als auf vertrauenswürdige *Flagger* reagiert hat.

Erfahrungen und Beobachtungen

Um einen besseren Einblick in die während des Monitorings gemachten Erfahrungen der Partnerorganisationen zu erhalten, wurde nach Abschluss des Monitorings ein Bewertungsbogen verteilt.

Die Partner berichteten, dass nur wenige Fälle von den IT-Unternehmen eingeschränkt, statt beseitigt wurden. Von diesen wurde die überwiegende Mehrheit geoblockiert. Hingegen berichtete der französische Partner, dass während des zweiten Monitorings alle Fälle homophober Hassrede, die sie an Twitter gemeldet hatten, eingeschränkt und nicht entfernt wurden. Da die Inhalte somit online bleiben und Methoden zur Umgehung der Restriktionen in der Online-Community weit verbreitet sind, halten die sCAN-Partner diesen Ansatz für nur bedingt effektiv. Es gab keine Hinweise darauf, warum Twitter diese Methode nur speziell auf französische homophobe Inhalte anwendete.

Der tschechische Partner stellte fest, dass Tschechien nicht als Standort im Bereich "Trends" auf Twitter gewählt werden kann. Dies erschwert den Monitoringprozess, da es schwieriger wird, relevante nationale Debatten auf hasserfüllte Inhalte zu überwachen. Ein weiteres Problem während des Monitorings war, dass direkte Links zu gemeldeten Kommentaren auf Facebook nicht zuverlässig funktionierten. In langen Konversationen mit einer Vielzahl von Kommentaren ist es daher sehr schwierig zu überprüfen, ob der gemeldete Kommentar entfernt wurde.

Feedback wurde von den IT-Unternehmen meist in Form automatischer Antworten gegeben, die jedoch keine Details über den konkreten Fall oder die Gründe für ihre Entscheidung enthielten. Darüber hinaus haben manche IT-Unternehmen einige Partner anders behandelt, als andere. Während Facebook die höchste Feedbackquote hatte und sowohl den allgemeinen Nutzern als auch den vertrauenswürdigen *Flaggern* Feedback gab, stellten einige Partner fest, dass das Feedback nicht sofort dann gesendet wurde, wenn das Unternehmen Maßnahmen ergriff, sondern nur einige Tage später.

Der italienische Partner erhielt individuelles Feedback für alle Inhalte, die er an Facebook meldete, während der slowenische Partner nur automatisierte Antworten erhielt, wenn er als allgemeiner Nutzer meldete. Für Markierungen als vertrauenswürdiger *Flagger* erhielten sie Feedback per E-Mail. Diese E-Mails verwiesen jedoch nicht immer direkt auf den gemeldeten Inhalt, was die Nachbereitung der Fälle sehr kompliziert machte.

Fazit

Die Ergebnisse der ersten und zweiten Monitorings weisen ähnliche Entfernungsraten auf. Da Facebook jedoch die meisten Meldungen von den teilnehmenden Organisationen erhalten hat, spiegelt das Gesamtergebnis nicht die Leistung aller überwachten IT-Unternehmen wider. Betrachtet man die Performance der Unternehmen einzeln, ist festzustellen, dass die Plattformen, mit Ausnahme von Facebook, im zweiten Monitoring nicht so gut abschnitten. Hinsichtlich der Inhalte, die über Kanäle vertrauenswürdige *Flagger* gemeldet wurden, ist festzustellen, dass während dieses Monitorings weniger Fälle von den IT-Unternehmen entfernt oder eingeschränkt wurden als bei früheren von der Europäischen Kommission organisierten Monitoringmaßnahmen.

Mit dem Verhaltenskodex haben sich die Unternehmen bereit erklärt, Inhalte, die gegen nationales Recht oder ihre Nutzungsbedingungen verstoßen, innerhalb von 24 Stunden zu bewerten und zu entfernen. Doch nur Facebook und Instagram schafften es innerhalb dieses Zeitraums, eine vertretbare

Quote des Entfernens gemeldeter *Hatespeech* zu erreichen, während Twitter und YouTube nicht einmal eine Rate von 50% erreichten.

Des Weiteren verpflichteten sich die Unternehmen, den Meldern illegaler Inhalte substanzielles und zeitnahes Feedback zu geben. Die Beseitigung von Hassrede ist wichtig, aber das Feedback zu Meldungen ist mindestens ebenso wichtig, wenn nicht sogar noch wichtiger: Die Bereitstellung von Feedback, auch als automatisierte Antwort, wird als entscheidend angesehen, um Transparenz über die Handlungen gegenüber den Nutzer*innen zu wahren und diese zu ermutigen, die Social Media bei der Bekämpfung von online *Hatespeech* zu unterstützen.

Die Partner beobachteten einen Rückgang des Feedbacks im Vergleich zu früheren Monitoringübungen. Facebook ist das einzige Unternehmen, das während beider Überwachungsübungen systematisch Feedback sowohl zu Markierungen allgemeiner Nutzer als auch zu Meldungen vertrauenswürdiger *Flagger* gab. Es war zudem das einzige Unternehmen, das den Großteil des Feedbacks innerhalb von 24 Stunden nach Erhalten der Markierung bereitstellte. YouTube wurde ausdrücklich dafür kritisiert, dass es kaum Feedback gab – weder an normale Nutzer noch an vertrauenswürdige *Flagger*.

Um den Verhaltenskodex gänzlich zu verwirklichen und illegale *Hatespeech* im Internet wirksam zu bekämpfen, ist es von entscheidender Bedeutung, dass Social-Media-Plattformen auf alle Nutzermeldungen zeitnah reagieren, unabhängig davon, wer berichtet oder in welchen Berichtszeitraum. Kontinuierliche Bemühungen um Monitoring und Gegenmaßnahmen sind entscheidend für die Gewährleistung eines sicheren und respektvollen *Online Space* in der gesamten EU und über sie hinaus.