Pressemitteilung der FRA
Wien, 29. November 2023

SPERRFRIST: 29. November 2023 um 06.00 Uhr MEZ

Hass im Internet: Wir müssen die Moderation von Inhalten verbessern, um wirksam gegen Hetze vorzugehen

Beleidigende Kommentare, Mobbing und Belästigung sowie Aufforderungen zur Gewalt werden durch Tools für die Moderation von Inhalten auf Online-Plattformen häufig nicht erkannt, so ein neuer Bericht der Agentur der Europäischen Union für Grundrechte (FRA). Der Bericht zeigt, dass der Großteil der Hetze im Internet gegen Frauen gerichtet ist, jedoch sind auch Menschen afrikanischer Abstammung, Roma und jüdische Menschen davon betroffen. Der mangelnde Zugang zu den Daten der Plattformen und das fehlende Verständnis dafür, was als Hetze einzustufen ist, behindern die Bemühungen zur Bekämpfung von Hass im Internet. Die FRA fordert mehr Transparenz und Leitlinien, um einen sichereren digitalen Raum für alle zu gewährleisten.

Der Bericht der FRA über die Moderation von Online-Inhalten befasst sich mit den Herausforderungen bei der Erkennung und Entfernung von Hetze aus den sozialen Medien.

Er unterstreicht, dass es keine allgemein anerkannte Definition von Hetze im Internet gibt. Auch die Systeme zur Moderation von Online-Inhalten können von den Forschern nicht genau überprüft werden. Dies macht es schwierig, sich ein vollständiges Bild vom Ausmaß der Hetze im Internet zu machen, und behindert die Bemühungen, dagegen vorzugehen.

Laut der von der FRA durchgeführten Analyse der Beiträge (Posts) und Kommentare, die zwischen Januar und Juni 2022 auf Plattformen sozialer Medien veröffentlicht wurden, sind folgende Phänomene zu beobachten:

- **Weitverbreitete Hetze im Internet:** Von 1 500 Posts, die zuvor bereits von Tools für die Moderation von Inhalten bewertet wurden, werden mehr als die Hälfte (53 %) von Codierfachleuten dennoch als Hetze eingestuft.
- **Frauenfeindlichkeit:** Frauen sind auf allen untersuchten Plattformen und in allen Ländern die Hauptziele von Hass im Internet. Hetze gegenüber Frauen beinhaltet zumeist missbräuchliche Sprache, Mobbing und Belästigung sowie Aufstachelung zu sexueller Gewalt.
- **Negative Stereotypisierung:** Menschen afrikanischer Abstammung, Roma und jüdische Menschen sind am häufigsten Ziel negativer Stereotypisierung.
- **Mobbing und Belästigung:** Bei fast der Hälfte (47 %) aller Hasspostings handelt es sich um direktes Mobbing bzw. um direkte Belästigung.

Um Hass im Internet zu bekämpfen, sollten sich die EU und die Online-Plattformen um Folgendes bemühen:

- **Schaffung eines sichereren digitalen Raums für alle:** Um Hass im Internet vorzubeugen, sollten Plattformen bei der Moderation und Überwachung ihrer Inhalte besonders auf geschützte Merkmale wie Geschlecht und ethnische Zugehörigkeit achten. Sehr große Online-Plattformen wie X (ehemals Twitter) oder YouTube sollten Frauenfeindlichkeit gemäß dem Gesetz über digitale Dienste in ihre Risikobewertungs- und Risikominderungsmaßnahmen einbeziehen. Alle EU-Mitgliedstaaten sollten auch das Übereinkommen von Istanbul ratifizieren, um Frauen im Internet besser zu schützen.
- **Bereitstellung von mehr Leitlinien:** Nicht in jedem Fall ist klar, was als Hetze gilt und was im Rahmen der Redefreiheit geschützt ist. Die Regulierungsbehörden auf EU- und auf nationaler Ebene sollten mehr Leitlinien für die Ermittlung rechtsverletzender Hetze im Internet bereitstellen.

- **Erfassung aller Formen von Hass im Internet:** Um sicherzustellen, dass die verschiedenen Arten von Hetze im Internet erkannt werden, sollten die Europäische Kommission und die nationalen Regierungen ein Netzwerk vertrauenswürdiger Hinweisgeber unter Einbeziehung der Zivilgesellschaft einrichten und finanzieren. Polizei, Moderatoren und Hinweisgeber sollten entsprechend geschult werden, um sicherzustellen, dass die Plattformen keine Inhalte übersehen oder diese übermäßig entfernen.
- **Testen der Technologie auf Verzerrungen:** Anbieter und Nutzer automatisierter Tools für die Moderation von Inhalten sollten ihre Technologie auf Verzerrungen testen, um Menschen vor Diskriminierung zu schützen, wie dies auch in dem Bericht über Verzerrungen in KI der FRA aus dem Jahr 2022 hervorgehoben wurde.
- **Gewährleistung des Zugangs zu Daten für unabhängige Forschungsarbeiten:** Die Europäische Kommission sollte sicherstellen, dass die von den Plattformen selbst vorgenommen Risikobewertungen gemäß dem Gesetz über digitale Dienste durch unabhängige Forschungsarbeiten ergänzt werden. Nur durch eine Vielzahl von Ansätzen und Tests lässt sich ein vollständiges Bild davon gewinnen, welche Art von Hass nicht richtig erkannt und daher nicht entfernt wird und welche Auswirkungen dies auf die Grundrechte der Menschen hat.

Der Bericht untersucht vier Social-Media-Plattformen (Reddit, Telegram, X und YouTube) in Bulgarien, Deutschland, Italien und Schweden. Bei dieser Forschungsarbeit konnte die FRA allerdings nicht auf Daten von Facebook und Instagram zugreifen.

Zwischen Januar und Juni 2022 sammelte die FRA fast 350 000 Posts und Kommentare auf der Grundlage bestimmter Schlüsselwörter. Codierfachleute bewerteten etwa 400 zufällig ausgewählte Posts aus jedem Land, um festzustellen, ob sie Hetze enthielten. 40 zufällig ausgewählte Posts wurden dann von Codierern und Rechtsexperten einer eingehenderen Prüfung unterzogen. Dieser Bericht zeigt die verschiedenen Arten von Hetze, die in den erfassten Ländern, Zielgruppen und Plattformen zu finden sind.

Hierzu FRA-Direktor Michael O'Flaherty:

„Das schiere Ausmaß von Hass, das wir in den sozialen Medien festgestellt haben, zeigt deutlich, dass die EU, ihre Mitgliedstaaten und Online-Plattformen ihre Anstrengungen verstärken können, um einen sichereren digitalen Raum für alle zu schaffen, und zwar unter Achtung der Menschenrechte, einschließlich des Rechts auf freie Meinungsäußerung. Es ist nicht hinnehmbar, Menschen im Internet allein aufgrund ihres Geschlechts, ihrer Hautfarbe oder ihrer Religion anzugreifen."

Weitere Auskünfte erhalten Sie unter: media@fra.europa.eu / Tel.: +43 1 580 30 653