Löschung rechtswidriger Hassbeiträge bei Facebook

Test zeigt erheblichen Nachholbedarf beim Umgang mit User-Beschwerden

Die Vielzahl fremdenfeindlicher und rassistischer Hasskommentare im Netz führte 2015 zur Bildung der Task Force "Umgang mit rechtswidrigen Hassbotschaften im Internet" des Bundesministeriums der Justiz und für Verbraucherschutz (BMJV). Die beteiligten Unternehmen (Google, Facebook, Twitter) sicherten unter anderem zu, künftig die Mehrzahl der ihnen gemeldeten, in Deutschland rechtswidrigen Inhalte binnen 24 Stunden zu entfernen.

Im Rahmen eines vom Bundesministerium für Familie, Senioren, Frauen und Jugend (BMFSFJ) und vom BMJV finanzierten Projektes recherchierte jugendschutz.net die Beschwerdemechanismen von Facebook und überprüfte deren Effektivität im Bereich der Hassinhalte.

Recherchegegenstand und Testaufbau

jugendschutz.net überprüfte bei den Tests folgende Aspekte:

- Inhalt der Gemeinschaftsstandards
- Gestaltung von Beschwerdemechanismen für User im Hinblick auf
  - Handhabbarkeit
  - Möglichkeiten, Hassbotschaften zu melden
  - Rückmeldung über Bearbeitungsstand und Bewertung des gemeldeten Inhaltes durch den Support
- Reaktionen und Reaktionszeiten bei
  - User-Meldungen
  - Meldungen über einen direkten Kontakt.

Art der recherchierten Inhalte

Die Verstöße wurden händisch mittels Schlagworten (z.B. "rapefugee", "Heil Hitler") über die Suchfunktionen des Dienstes recherchiert. Zudem erfolgte eine Sichtung des öffentlich einsehbaren Umfelds einschlägiger User (z.B. Freundeslisten, Likes, Gruppenmitgliedschaften). Technische Tools kamen bei der Recherche nicht zum Einsatz.

jugendschutz.net meldete strafbare Verstöße aus dem Bereich der Hassbotschaften (§ 130 StGB Volksverhetzung, Holocaustleugnung; § 86a Verwendung von Kennzeichen verfassungswidriger Organisationen; 90 % der Fälle) sowie Inhalte, die als jugendgefährdend einzustufen wären (10 % der Fälle).

Alle Verstöße wiesen einen deutschen Bezug (deutschsprachiger Inhalt oder User aus Deutschland) auf.

Testaufbau und Kontrolle

jugendschutz.net testete Meldefunktionen, die allen Usern zur Verfügung stehen (User-Flagging) sowie die direkte Meldemöglichkeit per E-Mail-Kontakt.

In einer ersten Phase wurden alle Verstöße über Standard-User-Accounts geflaggt, die jugendschutz.net nicht zugeordnet sind. Die Aufrufbarkeit der gemeldeten Inhalte kontrollierte jugendschutz.net über den Zeitraum von einer Woche täglich. In einer zweiten Phase wurden die jeweils verbliebenen Fälle über eine privilegierte E-Mail-Adresse direkt an den Support von Facebook weitergegeben und die Reaktionen nach der gleichen Systematik überprüft.

Verstöße wurden u.a. mit zugehöriger URL und einer Beschreibung des Inhalts dokumentiert. Aufgenommen wurden alle möglichen Einzelinhalte (z.B. Kommentare, Fotos, Videos) sowie übergreifende Einheiten (z.B. Profile oder Seiten). Registriert wurden die Art der Maßnahme, deren Durchführungsdatum, die Reaktion von Facebook sowie die Zeitspanne bis zur Löschung bzw. Sperrung für Deutschland. Ausgewertet wurden die Löschquoten 24 Stunden, 48 Stunden und eine Woche nach Meldung.

Ein Vortest im April/Mai 2016 diente dazu, das Testszenario zu erproben und erste systematische Erkenntnisse zu Beschwerdemechanismen und Löschverhalten zu gewinnen. Die Ergebnisse wurden den Betreibern kommuniziert und Verbesserungen angeregt. Im Anschluss hat jugendschutz.net das Testszenario optimiert (leichte Verschiebung in der Quotierung, Anpassung der Suchstrategien und Bewertungskriterien). Der Haupttest fand mit einer Dauer von 8 Wochen im Juli/August 2016 statt.

Überprüfung von Nutzungsbedingungen und Meldeverfahren

Gemeinschaftsstandards: greifen zu kurz

Facebook untersagt in seinen Gemeinschaftsstandards "Inhalte, die Personen aufgrund der folgenden Eigenschaften direkt angreifen: Rasse, Ethnizität, Nationale Herkunft, Religiöse Zugehörigkeit, sexuelle Orientierung, Geschlecht bzw. geschlechtliche Identität oder Schwere Behinderungen oder Krankheiten". Zudem sind Organisationen verboten, die an "terroristischen Aktivitäten oder organisierter Kriminalität" beteiligt sind, sowie Inhalte, die solche unterstützen oder Führungspersonen huldigen.

Deutsche Rechtsverstöße wie die Verbreitung von Kennzeichen verfassungswidriger Organisationen (§ 86a StGB) oder holocaustleugnende Inhalte (§ 130 Abs.3 StGB) finden sich nicht explizit.

Beschwerdemechanismen: Fehlende Funktion für Profile

Die Flagging-Option ist für angemeldete User bei Einzelinhalten, Seiten und Profilen unmittelbar erreichbar, die Handhabung einfach und die Nutzung damit ohne große Vorkenntnisse möglich. User können in ihrem "Support-Postfach" nachvollziehen, ob eine Meldung bereits bearbeitet, der Inhalt als Verstoß gegen die Gemeinschaftsstandards bewertet und eine Maßnahme durch den Support ergriffen wurde (z.B. Löschung). Der User hat zudem die Möglichkeit, seine "Nutzererfahrung" zu bewerten und eine Rückmeldung an den Support zu senden. Folgende Probleme zeigten sich beim User-Flagging:

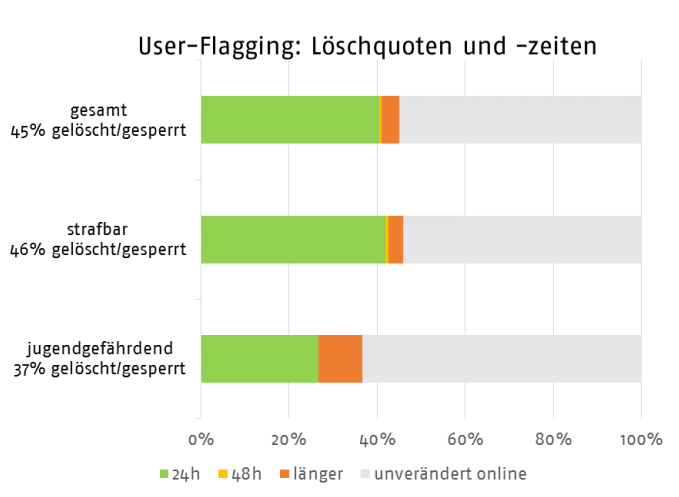
- keine explizite Meldeoption für Verstöße gegen §§ 86a und 130 Abs. 3 StGB
- keine explizite Meldeoption für Profile mit rechtswidrigen Hassinhalten (einzige Optionen: "Nacktheit und Pornographie", "Sexuell anzüglich" und "Andere Inhalte").

Die gebündelte Weitergabe von Verstoßfällen über einen direkten E-Mail-Kontakt war unkompliziert per Liste möglich. jugendschutz.net erhielt jedoch nur in Einzelfällen Feedback von Facebook zum Umgang mit den gemeldeten Inhalten.

Test der Löschpraxis

User-Flagging: Erfolgsquote 45 %

302 Verstöße wurden als User geflaggt. Ergebnis: 45 % wurden gelöscht/gesperrt (Steigerung um 11 % im Vergleich zum Vortest). Bei 40 % erfolgte die Löschung/Sperrung binnen 24 Stunden (Steigerung um 12 %).

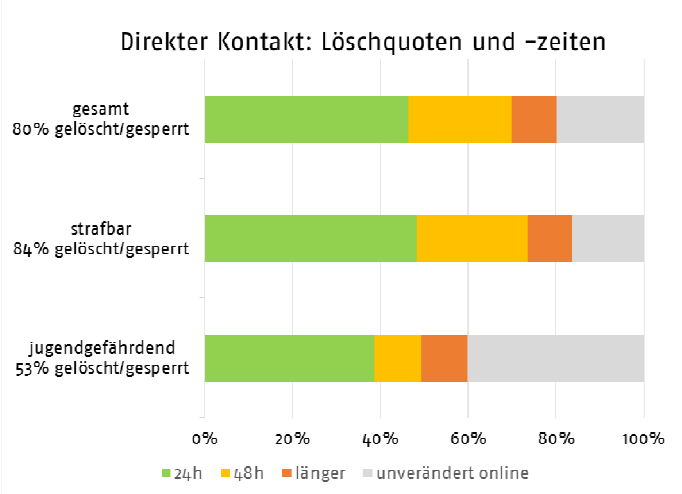


Betrachtet man nur die strafbaren Inhalte (272), liegt die Lösch-/Sperrquote bei 46 % (Steigerung um 8% im Vergleich zum Vortest). 42 % wurden binnen 24 Stunden gelöscht/gesperrt (Steigerung um 10 %).

Direkter Kontakt: Erfolgsquote von 80 %

166 Verstöße, die nach dem User-Flagging nicht gelöscht wurden, leitete jugendschutz.net nach einer Woche per
E-Mail an den Support weiter. Ergebnis: 80 % wurden gelöscht/gesperrt (Steigerung um 10 % im Vergleich zum Vortest). Bei 46 % erfolgte die Löschung/Sperrung binnen 24 Stunden (Steigerung um 30 %).

Betrachtet man nur die strafbaren Inhalte (147), liegt die Lösch-/Sperrquote bei 84 % (Steigerung um 10 % im Vergleich zum Vortest). 48 % wurden binnen 24 Stunden gelöscht/gesperrt (Steigerung um 30 %).



Fazit: Beschwerdemanagement weiter verbessern

Grundsätzlich bietet Facebook mit seinen Meldemöglichkeiten gute Voraussetzungen, um rechtswidrige Hassinhalte schnell und einfach melden zu können. Das Gespräch über die Ergebnisse des Vortests wurde vom Plattformbetreiber genutzt, um das Beschwerdehandling zu optimieren.

Der Haupttest zeigte erste positive Trends: Bei beiden Meldeformen waren Steigerungen bei der Zahl der Löschungen sowie der Geschwindigkeit, in der Maßnahmen ergriffen wurden, zu verzeichnen. Bei Berücksichtigung aller Maßnahmen, die Facebook nach User-Flagging und direkten Kontakten ergriffen hat, ergibt sich eine Löschquote von insgesamt 89 % (Steigerung um 9 % im Vergleich zum Vortest).

Betrachtet man nur die strafbaren Inhalte (272), liegt die Lösch-/Sperrquote bei 91 % (Steigerung um 7% im Vergleich zum Vortest).

Erheblicher Nachholbedarf besteht weiterhin vor allem im Bereich des Flaggings, das jedem User der Plattform zur Verfügung steht: Hier führte noch weniger als die Hälfte der Meldungen dazu, dass strafbare Inhalte gelöscht oder gesperrt werden.

Erläuterungen

User-Flagging

Plattformen bieten Funktionen, mit denen User Inhalte, die gegen Nutzungsrichtlinien oder Rechtsvorschiften verstoßen, melden können. In der Regel ist dies bei Einzelinhalten (z.B. Video, Bild, Kommentar) und übergeordneten Einheiten (z.B. User-Profil, Kanal) direkt während des Nutzungsvorgangs über einen zugeordneten Button möglich. Dieser Meldevorgang wird auch als User-Flagging bezeichnet. Der User hat dabei die Möglichkeit, Angaben zum Verstoß zu machen und seine Beschwerde dann per Mausklick direkt an den Support des Dienstes zu schicken. Der exakte Prozess der Meldung unterscheidet sich von Dienst zu Dienst.

Fast-Track-Mechanismus

Fast Track bezeichnet eine Meldemöglichkeit, über die Organisationen wie jugendschutz.net einfach und schnell Beschwerden unmittelbar an den Support einer Plattform senden können. Die Meldungen werden priorisiert behandelt, da sie aufgrund der inhaltlichen Expertise der Organisationen als besonders verlässlich angesehen werden. Ein Fast Track kann über ein eigens zur Verfügung gestelltes Meldetool (z.B. Trusted Flagging) realisiert werden oder über die Identifizierung beim Meldevorgang (z.B. mittels Account).

Direkter Kontakt

jugendschutz.net hat die Möglichkeit, Verstöße an einen direkten Ansprechpartner per E-Mail zu übermitteln. In den meisten Fällen kann dies in Form einer Liste geschehen, die alle relevanten Informationen (z.B. Fundstelle, Beschreibung des Verstoßes) enthält.

"Löschen" und "Sperren"

Löscht ein Plattformbetreiber einen Inhalt von seinem Server, ist dieser weltweit nicht mehr aufrufbar. Dies geschieht in der Regel dann, wenn ein Inhalt gegen die Nutzungsbedingungen eines Dienstes oder weltweit einheitliches Recht (z.B. Darstellungen des sexuellen Missbrauchs von Kindern) verstößt.

Bei der Sperrung eines Inhalts wird nur der Zugriff eingeschränkt (Geoblocking): Das Abrufen über einen deutschen Internetzugang ist dann nicht mehr möglich, der Inhalt ist in anderen Ländern weiterhin verfügbar. Dies geschieht bei nationalen Rechtsverstößen.