Löschung rechtswidriger Hassbeiträge bei Facebook

Verschlechterung von Löschquote und Reaktionszeiten bei User-Meldungen

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 überprüft jugendschutz.net seit 2016 die Effektivität der Beschwerdemechanismen von Facebook im Bereich der Hassinhalte. Der jüngste Test fand Anfang 2017 statt.

Aufbau und Systematik der Tests

Gegenstand der Recherchen

jugendschutz.net überprüfte bei den Tests folgende Aspekte:

- Inhalt der Nutzungsbedingungen und 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
- Reaktion 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 Hassbotschaften, die gegen § 130 StGB (Volksverhetzung, Holocaustleugnung) und § 86a StGB (Verwendung von Kennzeichen verfassungswidriger Organisationen) verstießen (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-Meldung) sowie die Meldemöglichkeit von jugendschutz.net über einen direkten E-Mail-Kontakt.

In einer ersten Phase wurden alle Verstöße über Standard-User-Accounts gemeldet, die jugendschutz.net nicht zugeordnet sind. In einer zweiten Phase meldete jugendschutz.net die jeweils verbliebenen Fälle über eine privilegierte E-Mail-Adresse direkt an den Support. In jeder Phase kontrollierte jugendschutz.net die Aufrufbarkeit der gemeldeten Inhalte nach 24 Stunden, 48 Stunden und einer Woche.

Verstöße wurden u.a. mit zugehöriger URL und einer Beschreibung des Inhalts dokumentiert. Aufgenommen wurden Einzelinhalte (z.B. Kommentare, Fotos, Videos) und übergeordnete Einheiten (z.B. Profile, Seiten). Registriert wurden die Art der Maßnahme, deren Durchführungsdatum, die Reaktion von Facebook sowie die Zeitspanne bis zur Löschung bzw. Sperrung für Deutschland.

In einem Vortest im April/Mai 2016 wurden das Testszenario erprobt und erste Erkenntnisse zu Beschwerdemechanismen und Löschverhalten gewonnen. Im Anschluss optimierte jugendschutz.net den Testaufbau (leichte Verschiebung in der Quotierung, Anpassung der Suchstrategien und Bewertungskriterien). Der erste Haupttest fand mit einer Dauer von 8 Wochen im Juli/August 2016 statt. Die Ergebnisse wurden den Betreibern kommuniziert und Verbesserungen angeregt. Den zweiten Haupttest führte jugendschutz.net über 8 Wochen im Januar/Februar 2017 durch.

Überprüfung von Nutzungsbedingungen und Meldeverfahren

Gemeinschaftsstandards: Sollten erweitert werden

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 sind nicht vollständig abgebildet.

Beschwerdemechanismen: Fehlende Meldeoption für Hassinhalte bei Profilen

Eine Meldefunktion ist für angemeldete User 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. Eine ausdrückliche Meldeoption für Profile mit rechtswidrigen Hassinhalten gibt es nicht (einzige Optionen: "Nacktheit und Pornographie", "Sexuell anzüglich" und "Andere Inhalte").

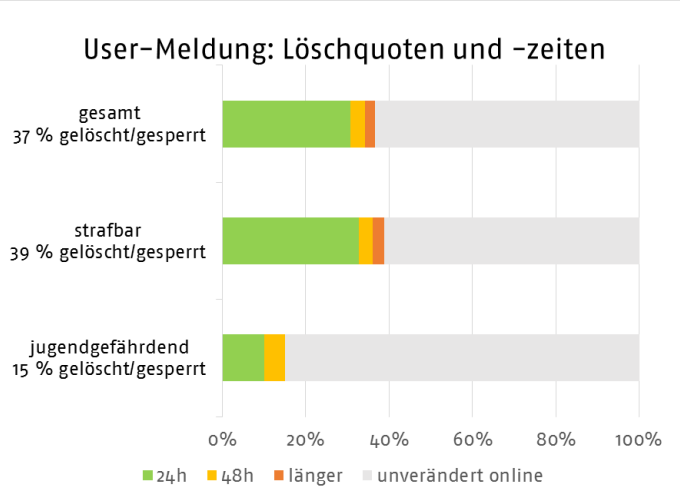
Die gebündelte Weitergabe von Verstoßfällen über einen direkten E-Mail-Kontakt war unkompliziert per Liste möglich. jugendschutz.net erhielt weiterhin nur in Einzelfällen Feedback von Facebook.

Test der Löschpraxis

User-Meldung: Erfolgsquote 37 %

200 Verstöße wurden als User gemeldet. Ergebnis: 37 % wurden gelöscht/gesperrt (minus 8 % im Vergleich zum vorigen Test). Bei 31 % erfolgte die Löschung/Sperrung binnen 24 Stunden (minus 9 %).

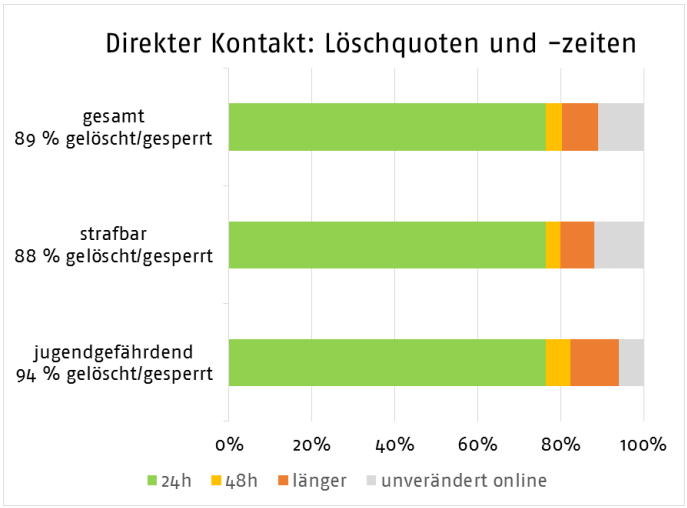
Betrachtet man nur die strafbaren Inhalte (180), liegt die Lösch-/Sperrquote bei 39 % (minus 7 % im Vergleich zum vorigen Test). 33 % wurden binnen 24 Stunden gelöscht/gesperrt (minus 9 %).



Direkter Kontakt: Erfolgsquote 89 %

127 Verstöße, die nach der User-Meldung nicht gelöscht waren, leitete jugendschutz.net nach einer Woche per E-Mail an den Support weiter. Ergebnis: 89 % wurden gelöscht/gesperrt (plus 9 % im Vergleich zum vorigen Test). Bei 76 % erfolgte die Löschung/Sperrung binnen 24 Stunden (plus 30 %).

Betrachtet man nur die strafbaren Inhalte (110), liegt die Lösch-/Sperrquote bei 88 % (plus 4 % im Vergleich zum vorigen Test). 76 % wurden binnen 24 Stunden gelöscht/gesperrt (plus 28 %).



Kumuliertes Ergebnis: Insgesamt 93 % gelöscht

Bei Berücksichtigung aller Maßnahmen, die Facebook nach User-Meldungen und direktem Kontakt ergriffen hat, ergibt sich eine Löschquote von insgesamt 93 % (plus 4 % im Vergleich zum vorigen Test). Betrachtet man nur die strafbaren Inhalte, liegt die Lösch-/Sperrquote bei 93 % (plus 2 %).

Fazit: Geringere Löschquote bei User-Meldungen

Im aktuellen Test haben sich Löschquote und Reaktionszeit von Facebook bei User-Meldungen verschlechtert.

Bei der Nutzung des direkten E-Mail-Kontakts zeigten sich Verbesserungen: Von den übermittelten Fällen wurden insgesamt mehr gelöscht und auch in wesentlich kürzerer Zeit.

Erläuterungen

User-Meldung

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. 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.