Löschung rechtswidriger Hassbeiträge bei YouTube

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 YouTube und überprüfte deren Effektivität im Bereich der Hassinhalte.

Recherchegegenstand und Testaufbau

jugendschutz.net überprüfte bei den Tests folgende Aspekte:

- Inhalt der Community Guidelines
- 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
  - Hinweisen über Fast-Track-Mechanismen und über direkte Kontakte.

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. Freunde, Playlists). 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), Fast-Track-Mechanismen als bevorzugte Meldeoption für privilegierte Organisationen (Trusted Flagging) sowie die Meldung an den Dienst über einen direkten 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 (Trusted Flagging) und dritten (E-Mail) Phase wurden die jeweils verbliebenen Fälle über einen akkreditierten Account von jugendschutz.net an
YouTube gemeldet und die Reaktionen des Supports 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, Bilder, Videos) sowie übergreifende Einheiten (z.B. Profile oder Kanäle). Registriert wurden die Art der Maßnahme, deren Durchführungsdatum, die Reaktion von YouTube 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

Community Guidelines: Sollten erweitert werden

Hasserfüllte Beiträge werden bei YouTube laut Community-Richtlinien nicht geduldet. Darunter fallen Inhalte, "die Gewalt gegen Einzelpersonen oder Gruppen aufgrund von ethnischer Zugehörigkeit, Religion, Behinderung, Geschlecht, Alter, Nationalität, Veteranenstatus oder sexueller Orientierung/geschlechtlicher Identität fördern bzw. billigen, oder Inhalte, deren Ziel hauptsächlich darin besteht, Hass in Zusammenhang mit diesen Eigenschaften zu animieren."

Deutsche Rechtsverstöße wie die Verbreitung von Kennzeichen verfassungswidriger Organisationen (§ 86a StGB) oder holocaustleugnenden Inhalten (§ 130 Abs.3 StGB) werden in den Nutzungsbedingungen von YouTube nicht ausdrücklich untersagt. Google verbietet zwar allgemein die Verbreitung rechtswidriger Inhalte und sperrt diese Inhalte, wenn sie der Rechtsabteilung gemeldet werden. Dazu ist jedoch ein direkter Kontakt erforderlich, ein einfacher Meldemechanismus für User steht dafür nicht zur Verfügung.

Beschwerdemechanismen: Flagging nur für angemeldete User

Die Flagging-Funktion ist für angemeldete User unmittelbar bei der Nutzung von Videos, Kommentaren und Bildern erreichbar, die Handhabung einfach und die Nutzung damit ohne große Vorkenntnisse möglich. Nach Meldung eines Videos wird eine Zusammenfassung der Angaben angezeigt. Rückmeldung zum Bearbeitungsstatus oder der ergriffenen Maßnahmen erhalten User jedoch nicht. Werden Kommentare gemeldet, erfolgt kein Feedback. Der Test zeigt darüber hinaus folgende Probleme beim User-Flagging:

- keine Flagging-Möglichkeit für User, die nicht angemeldet sind, obwohl unzulässige Beiträge allen Nutzerinnen und Nutzern der Plattform zugänglich sind
- keine explizite Meldeoption für rechtswidrige Hassinhalte bei Profil- und Kanalbildern (keine Differenzierungsmöglichkeit)
- keine explizite Meldeoption für Verstöße gegen §§ 86a und 130 Abs. 3 StGB

Meldungen von Videos durch Mitglieder des Trusted-Flagging-Programms von YouTube werden vom Support als besonders vertrauenswürdig behandelt. Verstöße gegen die Content Richtlinien können einfach und schnell gemeldet werden. Für Trusted Flagger stellt der Dienst den Meldeverlauf detailliert im Meldecenter dar. Dort kann jederzeit der aktuelle Status der Bearbeitung eingesehen werden. Für die Meldung von Kommentaren war das Trusted Flagging laut Aussage von YouTube nicht vorgesehen.

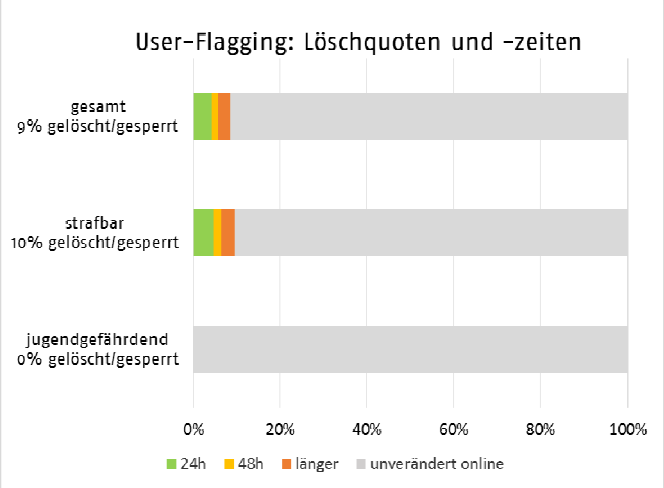
Die gebündelte Weitergabe von deutschen Rechtsverstößen über einen direkten E-Mail-Kontakt war unkompliziert per Liste möglich. jugendschutz.net erhielt bei fast allen Fällen binnen 48 Stunden Rückmeldung von YouTube zum Umgang mit den gemeldeten Inhalten.

Test der Löschpraxis

User-Flagging: Erfolgsquote von 9 %

210 Verstöße wurden als User geflaggt. Ergebnis: 9 % wurden gelöscht/gesperrt (Steigerung um 5 %im Vergleich zum Vortest). Bei 4 % erfolgte die Sperrung/Löschung binnen 24 Stunden (Steigerung um 1 %).

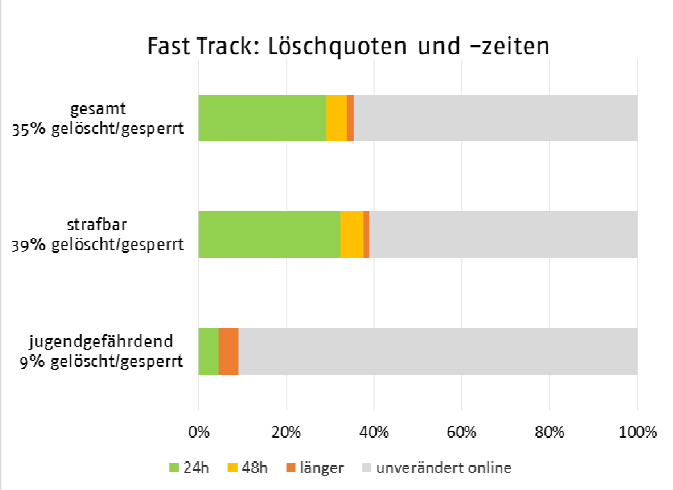
Betrachtet man nur die strafbaren Inhalte (188), liegt die Lösch-/Sperrquote bei 10 % (Steigerung um 6 % im Vergleich zum Vortest). 5 % wurden binnen 24 Stunden gelöscht/gesperrt (Steigerung um 2 %).



Fast Track: Erfolgsquote von 35 %

192 Verstöße, die nach dem User-Flagging nicht gelöscht wurden, meldete jugendschutz.net nach einer Woche über den Trusted Flagging-Account. Ergebnis: 35 % wurden gelöscht/gesperrt (Steigerung um 12% im Vergleich zum Vortest). Bei 29 % erfolgte die Löschung/ Sperrung binnen 24 Stunden (Steigerung um 11 %).

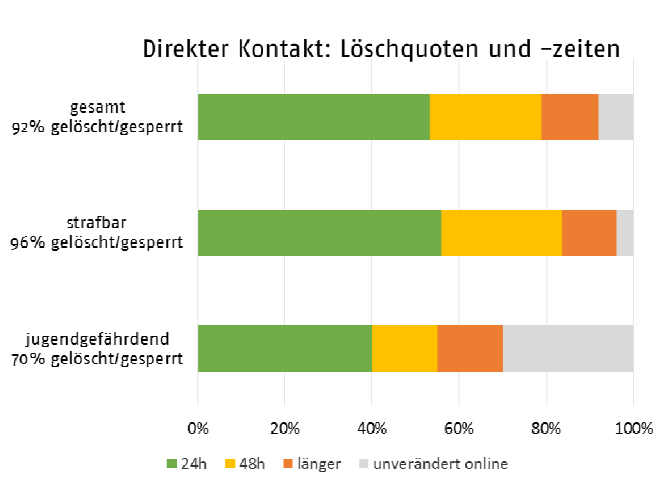
Betrachtet man nur die strafbaren Inhalte (170), liegt die Lösch-/Sperrquote bei 39 % (Steigerung um 13 %). 32 % wurden binnen 24 Stunden gelöscht/gesperrt (Steigerung um 12 %).



Direkter Kontakt: Erfolgsquote von 92 %

124 Verstöße, die nach dem Trusted Flagging nicht gelöscht wurden, leitete jugendschutz.net nach einer Woche per E-Mail weiter. Ergebnis: 92 % wurden gelöscht/gesperrt (Steigerung um 10 % im Vergleich zum Vortest). Bei 53 % erfolgte die Löschung/Sperrung binnen 24 Stunden (Steigerung um 40 %).

Betrachtet man nur die strafbaren Inhalte (104), liegt die Lösch-/Sperrquote bei 96 % (Steigerung um 15 %). 56 % wurden binnen 24 Stunden gelöscht/gesperrt (Steigerung um 41 %).



Fazit: Beschwerdemanagement weiter verbessern

Grundsätzlich bietet YouTube 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 allen Meldeformen waren Steigerungen bei der Zahl der Löschungen sowie der Geschwindigkeit, in der Maßnahmen ergriffen wurden, zu verzeichnen. Bei Berücksichtigung aller Maßnahmen, die YouTube nach User-Flagging, Trusted-Flagging und direkten Kontakten ergriffen hat, ergibt sich eine Löschquote von insgesamt 95 % (Steigerung um 10 % im Vergleich zum Vortest).

Betrachtet man nur die strafbaren Inhalte (188), liegt die Lösch-/Sperrquote bei 98 % (Steigerung um 13 %) im Vergleich zum Vortest).

Erheblicher Nachholbedarf besteht weiterhin vor allem im Bereich des Flaggings, das zudem nur angemeldeten Usern der Plattform zur Verfügung steht: Hier führte nur jede zehnte Meldung 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.