Löschung rechtswidriger Hassbeiträge bei Twitter

Weiterhin sehr schlechte Löschquoten 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 Twitter 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 Allgemeinen Geschäftsbedingungen und Twitter-Regeln
- 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 Fast-Track-Mechanismen und 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. Follower, Listen, Likes). 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 Meldemöglichkeiten von jugendschutz.net über einen Fast Track per Formular und 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 (Meldeformular) und dritten (E-Mail) Phase meldete jugendschutz.net die jeweils verbliebenen Fälle über akkreditierte Accounts. 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. Tweets) sowie übergeordnete Einheiten (z.B. Profile). Registriert wurden die Art der Maßnahme, deren Durchführungsdatum, die Reaktion von Twitter 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 dem Betreiber kommuniziert und Verbesserungen angeregt. Den zweiten Haupttest führte jugendschutz.net über 8 Wochen im Januar/Februar 2017 durch.

Überprüfung von Nutzungsbedingungen und Meldeverfahren

Twitter-Regeln: Sollten erweitert werden

Twitter schließt in seinen Nutzungsregeln Inhalte und Accounts aus, die "Gewalt gegen andere Personen fördern, sie direkt angreifen oder ihnen drohen, wenn diese Äußerungen aufgrund von Abstammung, ethnischer Zugehörigkeit, nationaler Herkunft, sexueller Orientierung, Geschlecht,

Geschlechtsidentität, religiöser Zugehörigkeit, Alter, Behinderung oder Krankheit erfolgen."

Die Twitter-Regeln wurden seit dem ersten Haupttest angepasst. Bezog sich der o.g. Ausschluss bestimmter Inhalte bislang nur auf Interaktionen, also z.B. Beleidigungen anderer User, untersagt Twitter nun auch "Accounts, deren Hauptziel darin besteht, basierend auf diesen Kategorien, Schaden gegen andere anzustiften". Deutsche Rechtsverstöße sind nicht vollständig abgebildet.

Beschwerdemechanismen: Neue Meldeoptionen für Hassbotschaften

Eine Meldefunktion ist für angemeldete User von Twitter unmittelbar erreichbar, die Handhabung einfach und die Nutzung damit ohne große Vorkenntnisse möglich. Zudem können Inhalte mittels gesonderter Formulare gemeldet werden. Der User erhält danach eine automatisierte Eingangsbestätigung an die angegebene E-Mail-Adresse.

Neu geschaffen wurden Meldeoptionen für rechtswidrige Hassbotschaften bei Einzelinhalten und Profilen. Zusätzlich hat Twitter das Meldeformular für missbräuchliches Verhalten um eine Option für "Hassäußerungen" erweitert.

Der Fast-Track-Mechanismus bei Twitter beschränkt sich auf die Meldung per Formular mit Angabe eines akkreditierten Accounts. Die Nutzung ist kompliziert und zeitaufwändig. jugendschutz.net erhielt in den meisten Fällen ein Feedback über die ergriffenen Maßnahmen.

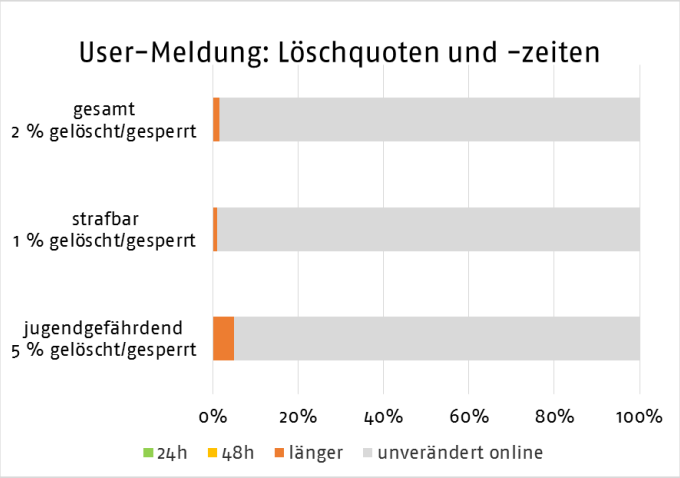
Die Weitergabe von Verstößen per Liste war über einen direkten E-Mail-Kontakt möglich. jugendschutz.net erhielt zeitnah Feedback zum Umgang mit gemeldeten Inhalten.

Test der Löschpraxis

User-Meldung: Erfolgsquote 2 %

200 Verstöße wurden als User gemeldet. Ergebnis: 2 % wurden gelöscht/gesperrt (plus 1 % im Vergleich zum vorigen Test). In keinem Fall erfolgte die Löschung/Sperrung binnen 24 Stunden.

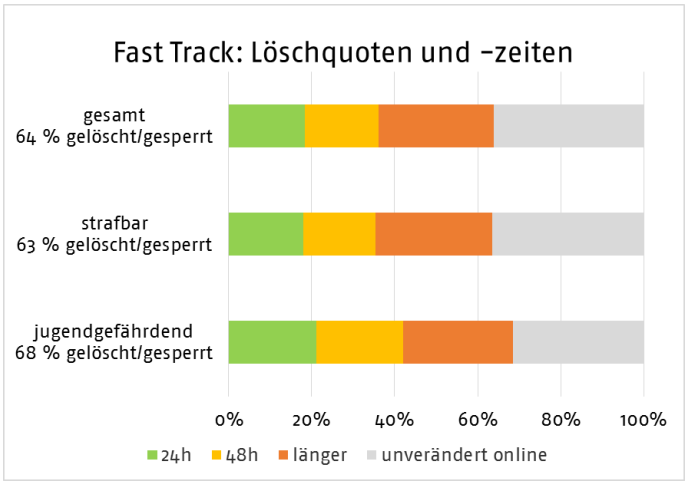
Betrachtet man nur die strafbaren Inhalte (180), liegt die Lösch-/Sperrquote bei 1 % (keine Veränderung zum vorigen Test). In keinem Fall erfolgte die Löschung binnen 24 Stunden.



Fast Track: Erfolgsquote 64 %

197 Verstöße, die nach der User-Meldung nicht gelöscht wurden, meldete jugendschutz.net nach einer Woche mittels Meldeformular als akkreditierter User. Ergebnis: 64 % wurden gelöscht/gesperrt (minus 11 % im Vergleich zum vorigen Test), 18 % binnen 24 Stunden (plus 8 %).

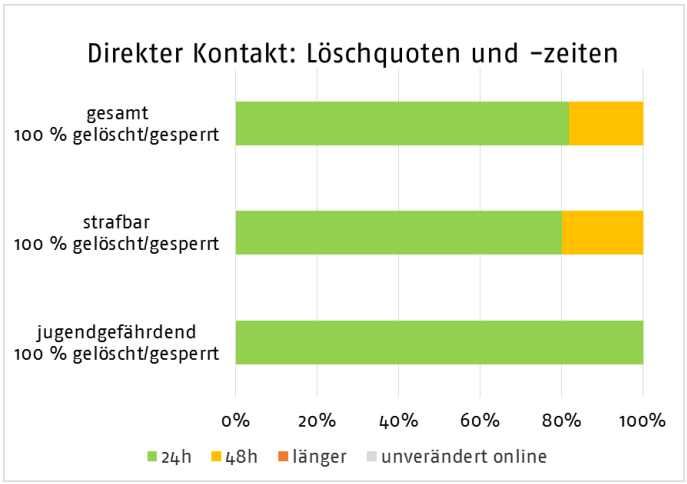
Betrachtet man nur die strafbaren Inhalte (178), liegt die Lösch-/Sperrquote bei 63 % (minus 13 % im Vergleich zum vorigen Test). 18 % wurden binnen 24 Stunden gelöscht/gesperrt (plus 9 %).



Direkter Kontakt: Erfolgsquote 100 %

71 Verstöße, die nach der Formularmeldung als akkreditierter User nicht gelöscht wurden, leitete jugendschutz.net nach einer Woche per E-Mail weiter. 100 % wurden gelöscht/gesperrt (plus 75 % im Vergleich zum vorigen Test).

Bei 82 % erfolgte die Löschung/Sperrung binnen 24 Stunden (plus 80 %). Betrachtet man hier nur die strafbaren Inhalte (65), liegt die Quote bei 80 % (plus 77 %).



Kumuliertes Ergebnis: Insgesamt 100 % gelöscht

Bei Berücksichtigung aller Maßnahmen, die Twitter nach User-Meldung, Fast Track und direktem E-Mail-Kontakt ergriffen hat, ergibt sich eine Löschquote von insgesamt 100 % (plus 18 % im Vergleich zum vorigen Test; plus 13 % bei den strafbaren Fällen).

Fazit: Keine Verbesserung der Reaktion bei User-Meldungen

Trotz neuer Meldeoptionen für Hassbotschaften hat Twitter beim aktuellen Test der Reaktion auf User-Meldungen kaum einen strafbaren Inhalt gelöscht.

Bei der Weiterleitung über einen direkten E-Mail-Kontakt zeigten sich dagegen Verbesserungen: Twitter löschte alle Fälle und die Löschungen erfolgten 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.