Löschung rechtswidriger Hassbeiträge bei YouTube

Enorme Verbesserung von Löschquote und Reaktionszeiten bei 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 überprüft jugendschutz.net seit 2016 die Effektivität der Beschwerdemechanismen von YouTube 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 Community-Richtlinien
- 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 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. Playlists, Related Videos). 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), Fast-Track-Mechanismen als bevorzugte Meldeoption für privilegierte Organisationen (Trusted Flagging) sowie die direkte Meldemöglichkeit von jugendschutz.net über einen E-Mail-Kontakt.

In einer ersten Phase wurden alle Verstöße über Standard-User-Accounts gemeldet (Flagging-Funktion), die jugendschutz.net nicht zugeordnet sind. In einer zweiten (Trusted Flagging) 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. Kommentare, Fotos, Videos) und übergeordnete Einheiten (z.B. Profile, Kanäle). Registriert wurden die Art der Maßnahme, deren Durchführungsdatum, die Reaktion von YouTube 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

Community-Richtlinien: Sollten erweitert werden

Hasserfüllte Inhalte 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 sind nicht vollständig abgebildet.

Beschwerdemechanismen: Neues Formular für deutsche Rechtsverstöße

Eine Meldefunktion ist für angemeldete User unmittelbar 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 den ergriffenen Maßnahmen erhalten User jedoch nicht. Werden Kommentare gemeldet, erfolgt kein Feedback. Der Test zeigte weiterhin keine Meldefunktion für User, die nicht angemeldet sind, obwohl unzulässige Beiträge allen Nutzerinnen und Nutzern der Plattform zugänglich sind.

Im Nachgang des ersten Haupttests stellte YouTube für deutsche User ein zusätzliches Formular zur Meldung von Volksverhetzung/Hassrede bereit (rechtliche Beschwerde). Der Link zu diesem wird standardmäßig angezeigt, nachdem ein Inhalt als Hassbotschaft geflaggt wurde.

Akkreditierte User können Verstöße gegen die Community-Richtlinien über den Trusted-Flagging-Mechanismus einfach und schnell melden. Für Trusted Flagger, deren Einschätzung vom YouTube-Support als besonders vertrauenswürdig eingestuft wird, ist der Status der Bearbeitung von Beschwerden jederzeit detailliert im Meldecenter einsehbar. Der Mechanismus war ursprünglich nur für das Melden von Videos vorgesehen, nach dem ersten Haupttest erweiterte YouTube die Trusted-Flagging-Möglichkeit für jugendschutz.net auch auf Kommentare.

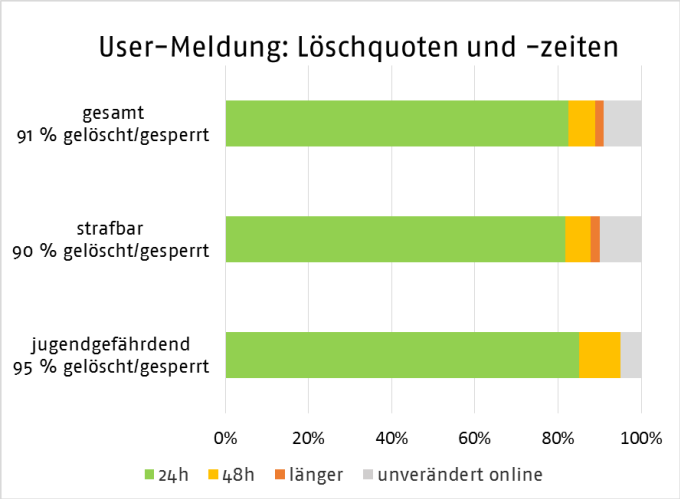
Die gebündelte Weitergabe von deutschen Rechtsverstößen über einen direkten E-Mail-Kontakt war unkompliziert per Liste möglich. jugendschutz.net erhielt in fast allen Fällen binnen 48 Stunden eine Rückmeldung von YouTube zum Umgang mit den gemeldeten Inhalten.

Test der Löschpraxis

User-Meldung: Erfolgsquote 91 %

200 Verstöße wurden als User gemeldet (Flagging-Funktion). Ergebnis: 91 % wurden gelöscht/gesperrt (plus 82 % im Vergleich zum vorigen Test). Bei 82 % erfolgte die Sperrung/Löschung binnen 24 Stunden (plus 78 %).

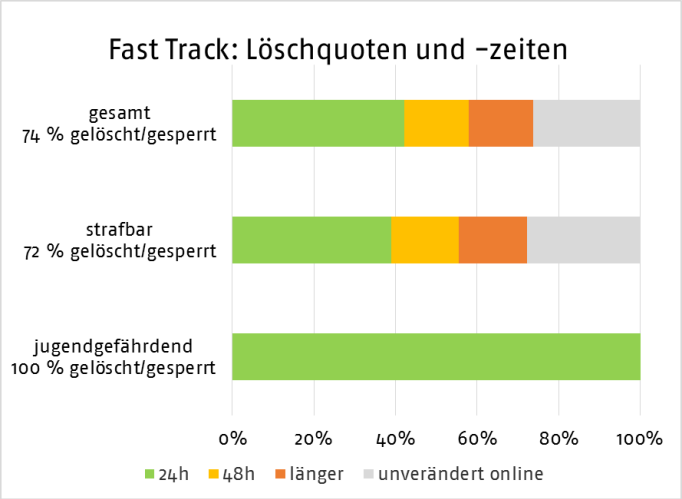
Betrachtet man nur die strafbaren Inhalte (180), liegt die Lösch-/Sperrquote bei 90 % (plus 80 % im Vergleich zum vorigen Test). 82 % wurden binnen 24 Stunden gelöscht/gesperrt (plus 77 %).



Fast Track: Erfolgsquote 74 %

19 Verstöße (davon 18 strafbare Inhalte), die nach der User-Meldung nicht gelöscht wurden, meldete jugendschutz.net nach einer Woche über den Trusted-Flagging-Account. Ergebnis: 74 % wurden gelöscht/gesperrt (plus 39 % im Vergleich zum vorigen Test). Bei 42 % erfolgte die Lösch-ung/Sperrung binnen 24 Stunden (plus 13 %).

Betrachtet man nur die strafbaren Inhalte (18), liegt die Lösch-/Sperrquote bei 72 % (plus 33 % im Vergleich zum vorigen Test). 39 % wurden binnen 24 Stunden gelöscht/gesperrt (plus 7 %).



Direkter Kontakt: Erfolgsquote 100 %

Die restlichen 5 Beiträge (alle strafbar), die nach dem Trusted Flagging nicht gelöscht wurden, leitete jugendschutz.net nach einer Woche per E-Mail weiter. Alle wurden daraufhin gelöscht/gesperrt – 4 binnen 24 Stunden, die restlichen binnen 48 Stunden.

Kumuliertes Ergebnis: Insgesamt 100 % gelöscht

Bei Berücksichtigung aller Maßnahmen, die YouTube nach User-Meldung, Fast Track und direktem Kontakt ergriffen hat, ergibt sich eine Löschquote von 100 % (plus 5 % im Vergleich zum vorigen Test; plus 2 % bei den strafbaren Fällen).

Fazit: Sehr starke Optimierung bei User-Meldungen

Im aktuellen Test der Reaktion auf User-Meldungen hat YouTube die Löschquoten sehr stark verbessert: Der Support entfernte neun von zehn strafbaren Inhalten.

Die hohen Löschquoten und die verkürzten Reaktionszeiten im kompletten Testverlauf zeigen, dass YouTube sein Beschwerdemanagement grundsätzlich verbessert hat.

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. 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.

Anhang: Detaillierte Übersicht

User-Meldung (Flagging)	Anzahl Fälle	Erfolg bis 24 Stunden	Erfolg bis 48 Stunden	Erfolg über 48 Stunden	Erfolg gesamt	unverändert online
§ 130 StGB						
Kommentar	41	27	5	2	34	7
Video	99	84	6	2	92	7
Gesamt	**140**	**111**	**11**	**4**	126	**14**
§ 86a StGB						
Bild	1	0	0	0	0	1
Kommentar	10	9	0	2	9	1
Video	29	27	0	0	27	2
Gesamt	**40**	**36**	**0**	**2**	**36**	**4**
Jugendgefährdung						
Kommentar	7	4	2	0	6	**1**
Video	13	13	0	0	13	**0**
Gesamt	**20**	**17**	**2**	**0**	**19**	**1**
Gesamt	**200**	**164**	**13**	**4**	**181**	**19**

Fast Track (Trusted Flagging)	Anzahl Fälle	Erfolg bis 24 Stunden	Erfolg bis 48 Stunden	Erfolg über 48 Stunden	Erfolg gesamt	unverändert online
§ 130 StGB						
Kommentar	7	3	0	2	5	2
Video	7	2	3	1	6	1
Gesamt	**14**	**5**	**3**	**3**	**11**	**3**
§ 86a StGB						
Bild	1	0	0	0	0	1
Kommentar	1	0	0	0	0	1
Video	2	2	0	0	2	0
Gesamt	**4**	**2**	**0**	**0**	**2**	**2**
Jugendgefährdung						
Kommentar	1	1	0	0	1	**0**
Video	0	0	0	0	0	**0**
Gesamt	**1**	**1**	**0**	**0**	**1**	**0**
Gesamt	**19**	**8**	**3**	**3**	**14**	**5**

Direkter Kontakt (E-Mail)	Anzahl Fälle	Erfolg bis 24 Stunden	Erfolg bis 48 Stunden	Erfolg über 48 Stunden	Erfolg gesamt	unverändert online
§ 130 StGB						
Kommentar	2	1	1	0	2	0
Video	1	0	1	0	1	0
Gesamt	**3**	**1**	**2**	**0**	3	**0**
§ 86a StGB						
Bild	1	0	1	0	1	0
Kommentar	1	1	0	0	1	0
Video	0	0	0	0	0	0
Gesamt	**2**	**1**	**1**	**0**	**2**	**0**
Jugendgefährdung						
Kommentar	0	0	0	0	0	**0**
Video	0	0	0	0	0	**0**
Gesamt	**0**	**0**	**0**	**0**	**0**	**0**
Gesamt	**5**	**2**	**3**	**0**	**5**	**0**