Löschung rechtswidriger Hassbeiträge bei Facebook, YouTube und Twitter

Ergebnisse des Monitorings von Beschwerdemechanismen jugendaffiner Dienste

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 die unverzügliche Löschung rechtswidriger Hassbeiträge und die anwenderfreundliche Gestaltung von Meldemöglichkeiten zu.

In einem vom Bundesministerium für Familie, Senioren, Frauen und Jugend (BMFSFJ) und vom BMJV finanzierten Projekt überprüft jugendschutz.net die Reaktionszeiten der Plattformen. Der jüngste Test fand Anfang 2017 statt.

Testaufbau

Für den Test ermittelte jugendschutz.net 540 strafbare Beiträge (§§ 130 und 86a StGB) und meldete sie den Diensten zunächst über einen Standard-User-Account, der jugendschutz.net nicht zugeordnet ist. Inhalte, die binnen einer Woche nicht gelöscht waren, wurden danach über einen akkreditierten Account erneut gemeldet (nur bei YouTube und Twitter möglich). Alle Inhalte, die nach einer weiteren Woche verblieben waren, gab jugendschutz.net abschließend an einen direkten E-Mail-Kontakt weiter.

Die Aufrufbarkeit der gemeldeten Inhalte überprüfte jugendschutz.net jeweils nach 24 Stunden, 48 Stunden und einer Woche. Ein Vortest im April/Mai 2016 diente dazu, das Testszenario zu erproben und für weitere Tests anzupassen. Der erste Haupttest wurde im Juli/August 2016 durchgeführt, der aktuelle Test im Januar/Februar 2017 (jeweils 8 Wochen).

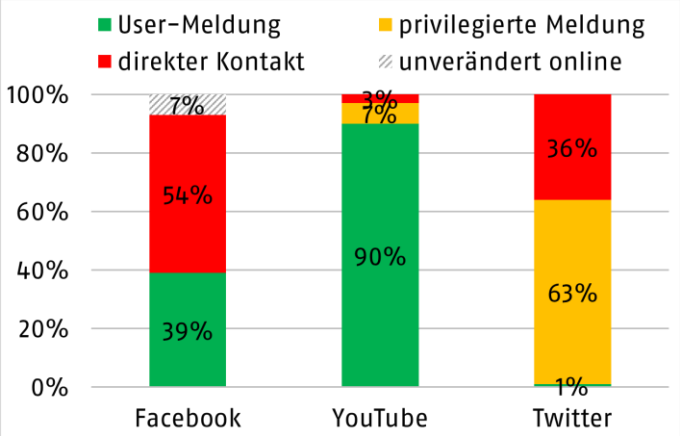
Löschquoten

Von den strafbaren Inhalten, die jugendschutz.net als User meldete, löschten oder sperrten Facebook 39 % (minus 7 % im Vergleich zum letzten Test), YouTube 90 % (plus 80 %) und Twitter weiterhin 1 % (grüne Markierung im Schaubild).

Bei YouTube wurden nach der Meldung als akkreditierter User weitere 7 % gelöscht oder gesperrt, bei Twitter 63 % (gelbe Markierung).

Nach direkten Kontakten per E-Mail löschte Facebook weitere 54 % (Löschquote 88 %), YouTube und Twitter entfernten alle verbliebenen Testfälle (rote Markierung).

Insgesamt löschten somit Facebook 93 % aller gemeldeten strafbaren Inhalte (plus 2 % im Vergleich zum letzten Test), YouTube 100% (plus 2 %) und Twitter 100 % (plus 18 %).



Löschquoten differenziert nach Meldemechanismen. Die Prozentzahlen beziehen sich auf alle gemeldeten Fälle.

Fazit

Die Vereinbarung in der Taskforce, die Mehrzahl der gemeldeten rechtswidrigen Hassbotschaften innerhalb von 24 Stunden zu entfernen, wird derzeit nur von YouTube eingehalten.

Im Vergleich zum letzten Test hat YouTube die Löschquote bei der Meldung strafbarer Inhalte als User enorm verbessern können (von 10 % auf 90 %), Facebook löschte weniger (Rückgang von 46% auf 39 %). Twitter reagierte anhaltend schlecht auf User-Meldungen (weiterhin 1 %).

Bei Berücksichtigung aller Maßnahmen, die die Plattformen nach User-Meldung, Trusted Flagging und direktem Kontakt ergriffen haben, zeigten sich bei allen dreien Verbesserungen und sehr gute Löschquoten (93 % bzw. 100 %).

Facebook, YouTube und Twitter bieten generell gute Meldemöglichkeiten für unzulässige Inhalte. Bei YouTube stehen sie ausschließlich angemeldeten Usern zur Verfügung.

Die Content-Richtlinien aller Dienste müssten optimiert werden. Zwar ist das Verbreiten von Hassinhalten überall ausgeschlossen, deutsche Rechtsverstöße sind jedoch nicht vollständig abgebildet.

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").

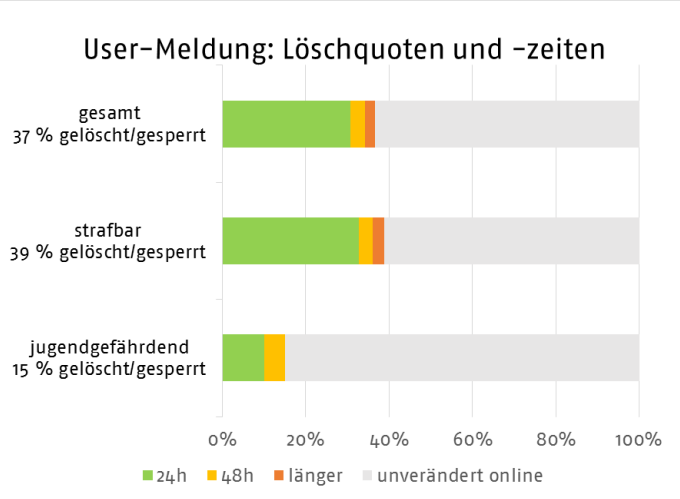
Die gebündelte Weitergabe von Verstoßfällen über einen direkten E-Mail-Kontakt war unkompliziert per Liste möglich. jugendschutz.net erhielt weiterhin nur in Einzelfällen Feedback von Facebook.

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 %).

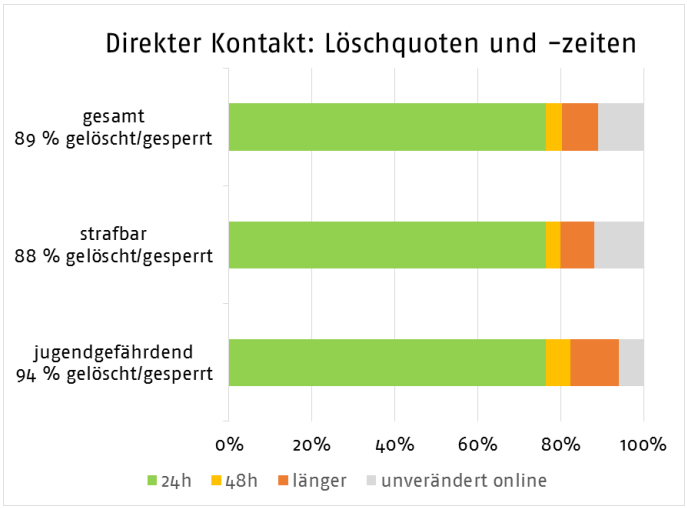
Betrachtet man nur die strafbaren Inhalte (180), liegt die Lösch-/Sperrquote bei 39 % (minus 7 % im Vergleich zum vorigen Test). 33 % wurden binnen 24 Stunden gelöscht/gesperrt (minus 9 %).



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.

Anhang: Detaillierte Übersicht

User-Meldung	Anzahl Fälle	Erfolg bis 24 Stunden	Erfolg bis 48 Stunden	Erfolg später	Erfolg gesamt	unverändert online
§ 130 StGB						
Bild	6	2	0	0	2	4
Kommentar	109	33	2	3	38	71
Post	22	8	2	1	11	11
Video	3	2	0	0	2	1
Gesamt	**140**	**45**	**4**	**4**	**53**	**87**
§ 86a StGB						
Bild	9	7	0	0	7	2
Kommentar	10	2	1	1	4	6
Post	16	5	1	0	6	10
Profil/Seite/Gruppe	4	0	0	0	0	4
Video	1	0	0	0	0	1
Gesamt	**40**	**14**	**2**	**1**	17	**23**
Jugendgefährdung						
Kommentar	20	2	1	0	3	17
Gesamt	**20**	**2**	**1**	**0**	3	**17**
Gesamt	**200**	**61**	**7**	**5**	**73**	**127**

Direkter Kontakt (E-Mail)	Anzahl Fälle	Erfolg bis 24 Stunden	Erfolg bis 48 Stunden	Erfolg später	Erfolg gesamt	unverändert online
§ 130 StGB						
Bild	4	2	0	0	2	2
Kommentar	71	51	4	7	62	9
Post	11	10	0	0	10	1
Video	1	1	0	0	1	0
Gesamt	**87**	**64**	**4**	**7**	75	**12**
§ 86a StGB						
Bild	2	1	0	1	2	0
Kommentar	6	4	0	1	5	1
Post	10	10	0	0	10	0
Profil/Seite/Gruppe	4	4	0	0	4	0
Video	1	1	0	0	1	0
Gesamt	**23**	**20**	**0**	**2**	22	**1**
Jugendgefährdung						
Kommentar	17	13	1	2	16	1
Gesamt	**17**	**13**	**1**	**2**	16	**1**
Gesamt	**127**	**97**	**5**	**11**	**113**	**14**

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.

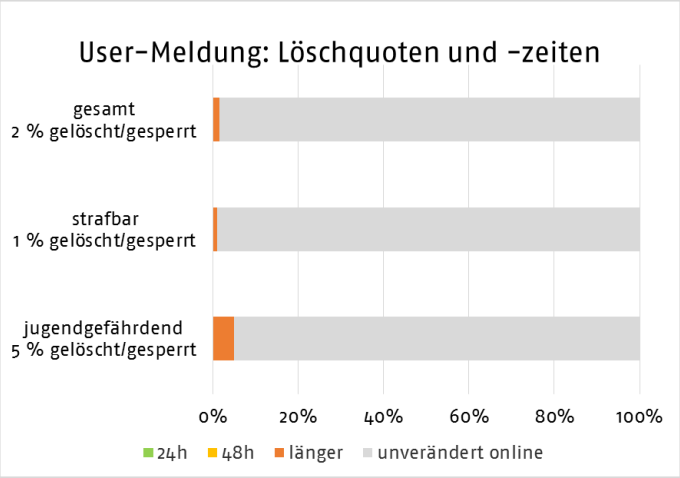
Die Weitergabe von Verstößen per Liste war über einen direkten E-Mail-Kontakt möglich. jugendschutz.net erhielt zeitnah Feedback zum Umgang mit gemeldeten Inhalten.

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.

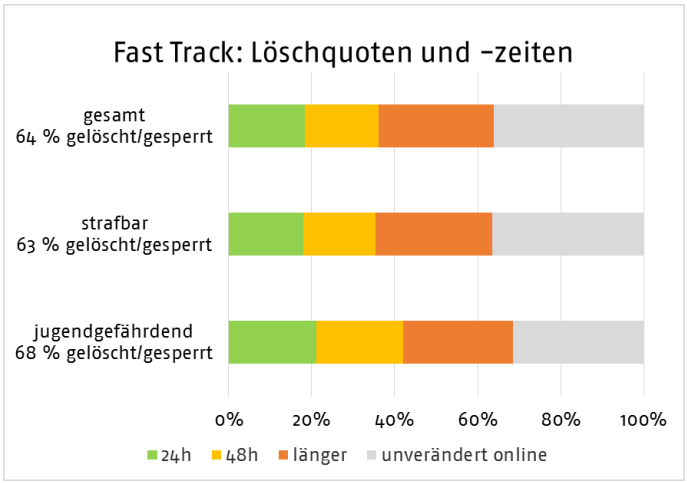
Betrachtet man nur die strafbaren Inhalte (180), liegt die Lösch-/Sperrquote bei 1 % (keine Veränderung zum vorigen Test). In keinem Fall erfolgte die Löschung binnen 24 Stunden.



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 %).

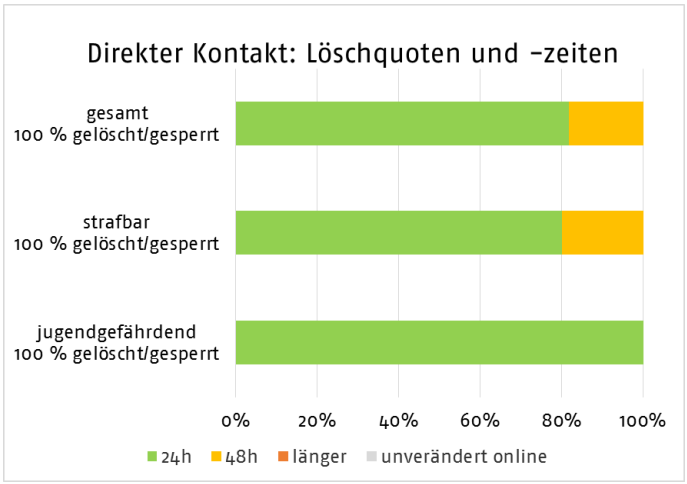
Betrachtet man nur die strafbaren Inhalte (178), liegt die Lösch-/Sperrquote bei 63 % (minus 13 % im Vergleich zum vorigen Test). 18 % wurden binnen 24 Stunden gelöscht/gesperrt (plus 9 %).



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.

Anhang: Detaillierte Übersicht

User-Meldung	Anzahl der Fälle	Erfolg bis 24 Stunden	Erfolg bis 48 Stunden	Erfolg über 48 Stunden	Erfolg gesamt	unverändert online
§ 130 StgB						
Bild	3	0	0	0	0	3
Kommentar	47	0	0	1	1	46
Post	90	0	0	1	1	89
Gesamt	**140**	**0**	**0**	**2**	2	**138**
§ 86a						
Bild	10	0	0	0	0	10
Kommentar	5	0	0	0	0	5
Post	21	0	0	0	0	21
Profil	1	0	0	0	0	1
Video	3	0	0	0	0	3
Gesamt	**40**	**0**	**0**	**0**	**0**	**40**
Jugendgefährdung						
Kommentar	10	0	0	1	1	9
Post	10	0	0	0	0	10
Gesamt	**20**	**0**	**0**	**1**	**1**	**19**
Gesamt	**200**	**0**	**0**	**3**	**3**	**197**

Fast Track (Formular)	Anzahl Fälle	Erfolg bis 24 Stunden	Erfolg bis 48 Stunden	Erfolg über 48 Stunden	Erfolg gesamt	unverändert online
§ 130 StGB						
Bild	3	0	0	1	1	2
Kommentar	46	20	7	12	39	7
Post	89	11	18	20	49	40
Gesamt	**138**	**31**	**25**	**33**	**89**	**49**
§ 86a StGB						
Bild	10	0	0	2	2	8
Kommentar	5	0	3	2	5	0
Post	21	0	3	12	15	6
Profil	1	0	0	0	0	1
Video	3	1	0	1	2	1
Gesamt	**40**	**1**	**6**	**17**	**24**	**16**
Jugendgefährdung						
Kommentar	9	1	2	4	7	2
Post	10	3	2	1	6	4
Gesamt	**19**	**4**	**4**	**5**	13	**6**
Gesamt	**197**	**36**	**35**	**55**	**126**	**71**

Direkter Kontakt (E-Mail)	Anzahl Fälle	Erfolg bis 24 Stunden	Erfolg bis 48 Stunden	Erfolg über 48 Stunden	Erfolg gesamt	unverändert online
§ 130 StGB						
Bild	2	1	1	0	2	0
Kommentar	7	7	0	0	7	0
Post	40	32	8	0	40	0
gesamt	**49**	**40**	**9**	**0**	**49**	**0**
§ 86a StGB						
Bild	8	5	3	0	8	0
Kommentar	0	0	0	0	0	0
Post	6	6	0	0	6	0
Profil	1	0	1	0	1	0
Video	1	1	0	0	1	0
gesamt	**16**	**12**	**4**	**0**	**16**	**0**
Jugendgefährdung						
Kommentar	2	2	0	0	2	0
Post	4	4	0	0	4	0
gesamt	**6**	**6**	**0**	**0**	**6**	**0**
Gesamt	**71**	**58**	**13**	**0**	**71**	**0**

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.

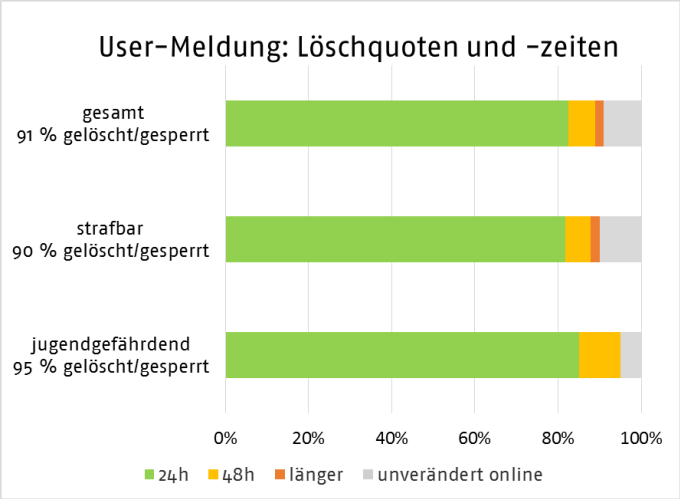
Die gebündelte Weitergabe von deutschen Rechtsverstößen über einen direkten E-Mail-Kontakt war unkompliziert per Liste möglich. jugendschutz.net erhielt in fast allen Fällen binnen 48 Stunden eine Rückmeldung von YouTube zum Umgang mit den gemeldeten Inhalten.

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 %).

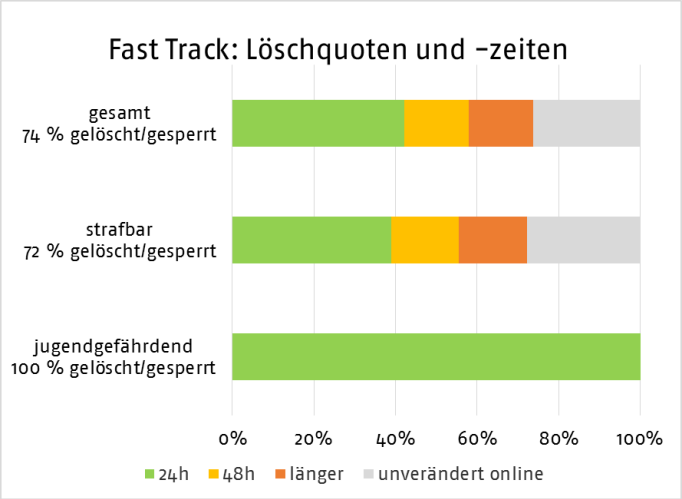
Betrachtet man nur die strafbaren Inhalte (180), liegt die Lösch-/Sperrquote bei 90 % (plus 80 % im Vergleich zum vorigen Test). 82 % wurden binnen 24 Stunden gelöscht/gesperrt (plus 77 %).



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**