# Löschung rechtswidriger Hassbeiträge bei Twitter

## Test zeigt erheblichen Nachholbedarf beim Umgang mit 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 Twitter und überprüfte deren Effektivität im Bereich der Hassinhalte.**

## Recherchegegenstand und Testaufbau

jugendschutz.net überprüfte bei den Tests folgende Aspekte:

- Inhalt der 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
- 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. Follower, Listen, Likes). 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), die Meldung über einen Fast Track per Formular sowie die Weitergabe von Fällen 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 wurden über den Zeitraum von einer Woche täglich kontrolliert. In einer zweiten (Meldeformular) und dritten (E-Mail) Phase wurden die jeweils verbliebenen Fälle über einen akkreditierten Account von jugendschutz.net an Twitter gemeldet und die Reaktionen 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. Tweets, Fotos, Videos) sowie übergreifende Einheiten (z.B. Profile). Registriert wurden auch Art der Maßnahme, deren Durchführungsdatum, die Reaktion von Twitter 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

### Twitter-Regeln: nicht weitreichend genug

Twitter untersagt in seinen Verhaltensregeln die "Rechtswidrige Nutzung" und verpflichtet "alle internationalen Nutzer (…), alle geltenden nationalen Gesetze und Bestimmungen zu respektieren". Gesondert ausgeschlossen werden Inhalte und Accounts, 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."

Allerdings beziehen sich die Twitter-Regeln auf die Interaktion zwischen Usern – volksverhetzende Aussagen (über Dritte) sind davon nicht per se abgedeckt.

### Beschwerdemechanismen: zu kompliziert und undifferenziert

Twitter bietet allen Usern des Dienstes für Tweets, Bildmaterial und Accounts/Profile leicht zugängliche Flagging-Optionen an. Zudem können Inhalte mittels gesonderten Online-Formularen gemeldet werden. Der User erhält eine automatisierte Eingangsbestätigung an die angegebene E-Mail-Adresse. Folgende Probleme lassen sich beim User-Flagging jedoch feststellen:

- keine explizite Meldeoption für rechtswidrige Hassbotschaften, lediglich über Option "missbräuchlich oder schädigend" (Einzelbeitrag) bzw. "Nimmt an Belästigungen oder Gewalt teil" (Accounts) zu flaggen; Medien wie Bilder und Videos können ausschließlich als "sensibel" gemeldet werden;
- kein explizites Formular für rechtswidrige Hassinhalte, unterschiedliche Zugänge zu Meldemöglichkeiten verwirren.

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

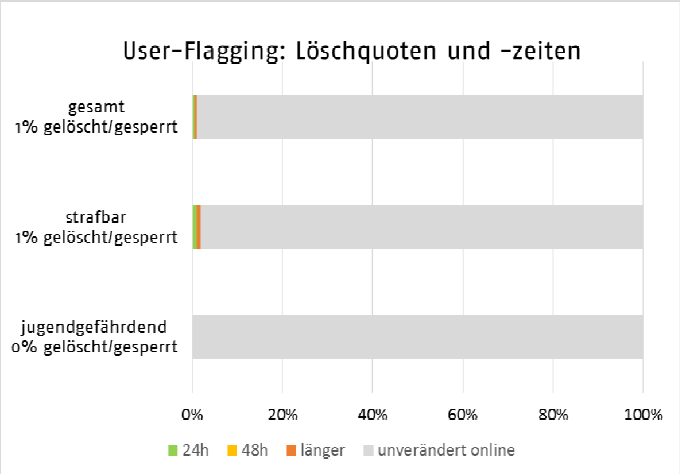
Die Weitergabe von Verstößen per Liste war über einen direkten E-Mail-Kontakt möglich. jugendschutz.net erhielt häufig Feedback zum Umgang mit den gemeldeten Inhalten.

## Test der Löschpraxis

### User-Flagging: Erfolgsquote von 1 %

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

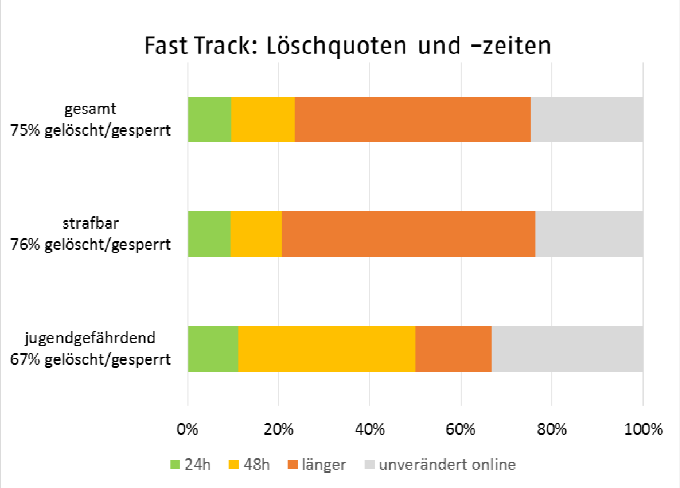
Betrachtet man nur die strafbaren Inhalte (162), liegt die Lösch-/Sperrquote ebenfalls bei 1 %, die Hälfte davon wurde binnen 24 Stunden gelöscht/gesperrt.




### Fast Track: Erfolgsquote von 75 %

178 Verstöße, die nach dem User-Flagging nicht gelöscht wurden, meldete jugendschutz.net nach einer Woche mittels Meldeformular als akkreditierter User. Ergebnis: 75 % wurden gelöscht/gesperrt (Steigerung um 39 % im Vergleich zum Vortest), 10 % binnen 24 Stunden (Reduktion um 4 %).

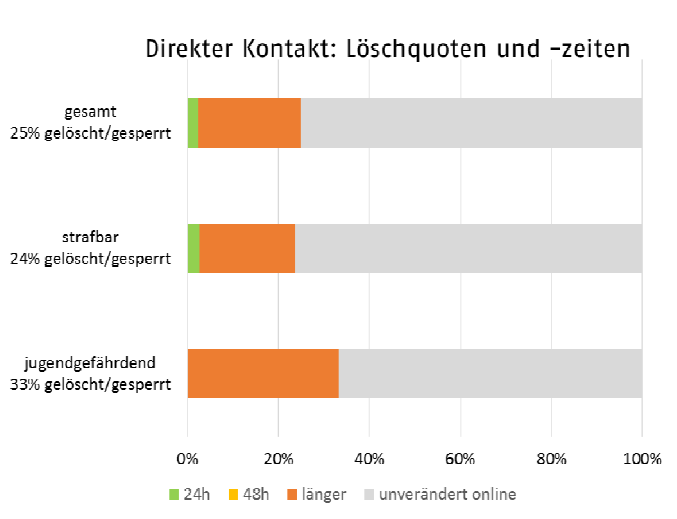
Betrachtet man nur die strafbaren Inhalte (160), liegt die Lösch-/Sperrquote bei 76 % (Steigerung um 34 %). 9 % wurden binnen 24 Stunden gelöscht/gesperrt (Reduktion um 8 %).




### Direkter Kontakt: Erfolgsquote von 25 %

44 Verstöße, die nach der Formularmeldung als akkreditierter User nicht gelöscht wurden, leitete jugendschutz.net nach einer Woche per E-Mail weiter. 25 % wurden gelöscht/gesperrt (Steigerung um 3 % im Vergleich zum Vortest). Bei 2 % erfolgte die Löschung/Sperrung binnen 24 Stunden (Steigerung um 2 %).

Betrachtet man nur die strafbaren Inhalte (38), liegt die Lösch-/Sperrquote bei 24 % (Reduktion um 1 %). 3 % wurden binnen 24 Stunden gelöscht/gesperrt (Steigerung um 3 %).




## Fazit: Beschwerdemanagement muss erheblich verbessert werden

Grundsätzlich bietet Twitter mit seinen Meldemöglichkeiten gute Voraussetzungen, um Hassinhalte schnell und einfach melden zu können. Das Gespräch über die Ergebnisse des Vortests wurde vom Plattformbetreiber genutzt, um das Beschwerdehandlings effektiver zu gestalten.

Der Haupttest zeigte erste positive Trends: Bei Berücksichtigung aller Maßnahmen, die Twitter nach User-Flagging, Formular-Meldung und direkten Kontakten ergriffen hat, ergibt sich eine Löschquote von insgesamt 82 % (Steigerung um 77 % im Vergleich zum Vortest). Diese wurde fast ausschließlich über die akkreditierte Meldung per Formular erreicht.

Betrachtet man nur die strafbaren Inhalte (162), liegt die Lösch-/Sperrquote bei 82% (Steigerung um 75 %) im Vergleich zum Vortest).

Erheblicher Nachbesserungsbedarf besteht weiterhin vor allem im Bereich des Flaggings, das jedem angemeldeten User der Plattform zur Verfügung steht: Hier führte so gut wie keine Meldung dazu, dass strafbare Inhalte gelöscht oder gesperrt werden. Außerdem reagierte Twitter bei allen drei getesteten Meldewegen nur mit großer zeitlicher Verzögerung.

## 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.