Künstliche Medienaufsicht

KI sucht rechtswidrige und jugendgefährdende Inhalte im Internet

Mit einer neuen Software suchen die Landesmedienanstalten im großen Stil nach Rechtsverstößen im Internet. Das System findet derzeit mehr potenziell Illegales, als die Behörden bearbeiten können.

Von Torsten Kleinz

Es gehört zu den Aufgaben der 14 Landesmedienanstalten in Deutschland, Rechtsverstöße in Medien zu bekämpfen. Bis vor wenigen Jahren waren sie fast ausschließlich mit TV- und Radio-Inhalten beschäftigt. Mittlerweile weiten sie ihre Kontrollfunktion auf das Internet aus.

Tobias Schmid, Direktor der nordrhein-westfälischen Landesanstalt für Medien, treibt diese Bemühungen der Medienwächter am aktivsten voran. So hat er YouTube- und Instagram-Influencern blaue Briefe geschickt, wenn sie in ihren Videos Schleichwerbung machten. Schmid war es auch, der gegen das Pornoportal xHamster Netzsperren angeordnet hat, weil es den Jugendschutz nicht mit einem zertifizierten Verfahren sicherstellt.

Bereits 2017 hat die Landesmedienanstalt NRW ihre Initiative „Verfolgen statt nur löschen“ gestartet. Das Projekt hat zum Ziel, dass Inhalte wie Beleidigungen und Verleumdungen von den Betreibern sozialer Netzwerke nicht nur gelöscht oder gesperrt werden, sondern dass auch die Täter strafrechtlich verfolgt werden.

Die Anfänge waren allerdings bescheiden. „Die Idee, mit sieben studentischen Hilfskräften Inhalte im Internet zu regulieren, hat etwas Sozialromantisches“, sagt Schmid heute. Nichtsdestotrotz schlossen sich immer mehr Landesbehörden der Initiative an. Im Frühjahr 2020 beschlossen sie, ein System zu entwickeln, das automatisch auf die Suche nach illegalen Inhalten geht.

KI als Massenscanner

Die Idee, künstliche Intelligenz einzusetzen, lag nahe. Das neue Werkzeug hat den Namen KIVI erhalten, wobei das „VI“ für das lateinische Wort vigilare steht, überwachen. Um das System zusammenzustellen, kombinierte der beauftragte Dienstleister Condat eine Reihe verschiedener Systeme. Für die Bilderkennung kamen zum Beispiel die neuronalen Netze VGG19 und Inception V4 zum Einsatz, die mit eigenen Daten auf ihre neuen Aufgaben trainiert wurden. Texte analysiert das System ebenfalls mit einem neuronalen Netz (DenseNN) sowie mit dem einfacheren Naive-Bayes-Verfahren. Konkret sucht KIVI unter anderem nach Gewaltdarstellungen, Volksverhetzung und der Verwendung verfassungsfeindlicher Kennzeichen.

Für das Training kommen sowohl Positiv- als auch Negativbeispiele zum Einsatz. Als allgemeines Negativmaterial hat Condat zum Beispiel Daten aus Googles Open-Images-Datensatz an das System verfüttert. Die KI werde laut Landesmedienanstalt auch durch tägliches Feedback, ob sich ein gefundener Verdacht bestätigt hat oder nicht, weitertrainiert.

Bild: Landesmedienanstalt NRW

Tobias Schmid, der Direktor der nordrhein-westfälischen Landesmedienanstalt, hat sich die Rechtsdurchsetzung im Netz auf die Fahnen geschrieben.

Im Bereich Pornografie verzichtet Condat auf eine eigene Erkennung und bindet stattdessen den Amazon-Dienst Rekognition ein, der nach Angaben der Medienwächter im Praxisbetrieb eine Erkennungsgenauigkeit von 90 Prozent erreicht. Im Bereich der Verstöße gegen die Menschenwürde und des politischen Extremismus gibt die Landesanstalt eine Erkennungsrate von knapp 40 Prozent an.

Auslese von Hand

Letztlich arbeite KIVI nicht autonom, sondern diene nur der Arbeitserleichterung, betonte die federführende Landesmedienanstalt NRW bei der Vorstellung der ersten Ergebnisse des Projekts im April 2022. Die Entscheidung, ob ein potenzieller Rechtsverstoß an Strafermittlungsbehörden gemeldet wird, treffe immer ein Mensch.

Die Software soll aber einen Großteil der Routinearbeit übernehmen. So verschlagwortet sie jeden Treffer automatisch und überführt ihn in ein Ticketing-System. Die Software sucht auch nach Hinweisen auf den Wohnort des jeweiligen Urhebers, um die Inhalte den jeweiligen Landesbehörden zuzuteilen. Wo dies nicht möglich ist, landen die Inhalte in einem gemeinsamen Pool. Auf einem übersichtlichen Team-Dashboard ist jederzeit der aktuelle Bearbeitungsstatus der eingelesenen Inhalte einsehbar und erlaubt den direkten Absprung in die zu prüfenden Quellen.

Studentische Hilfskräfte sortieren die Treffer weiter vor. Erst nachdem Juristen die Inhalte überprüft haben, werden schließlich weitere Schritte gegen potenzielle Straftäterinnen oder Straftäter ergriffen.

KIVI soll die Arbeit nicht nur effektiver machen, sondern auch stressfreier. Statt Mitarbeiter ständig unvermittelt mit Inhalten wie Erschießungsvideos oder harter Pornografie zu konfrontieren, zeigt

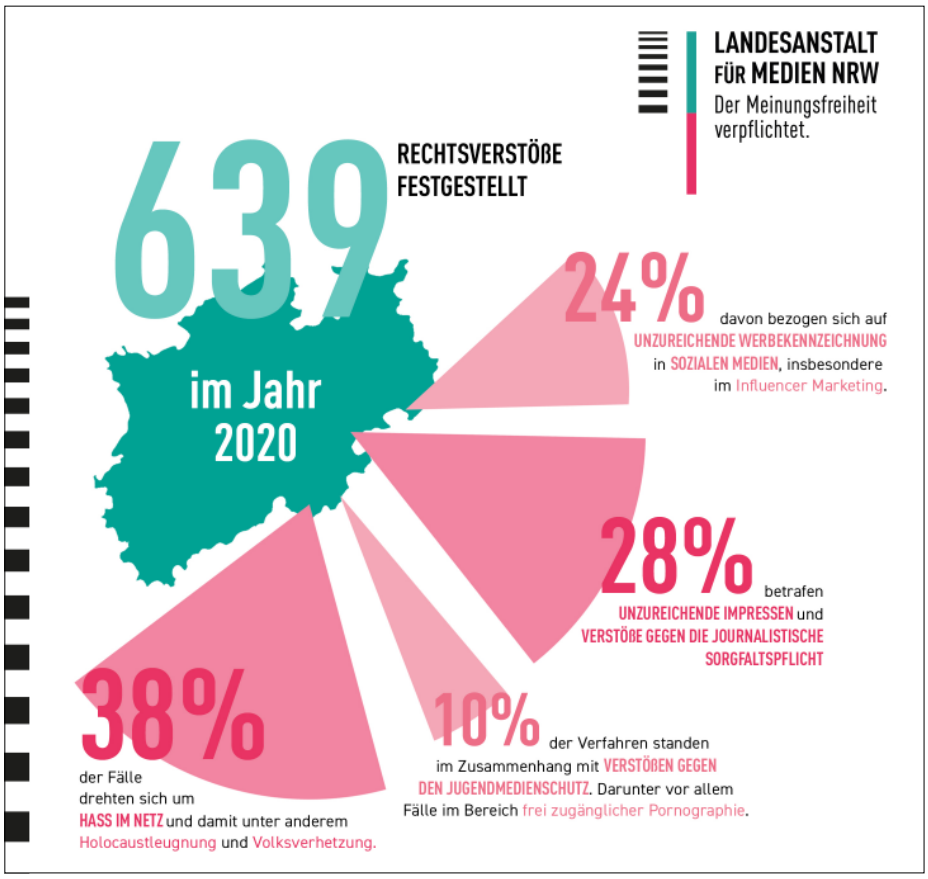

Bild: Landesmedienanstalt NRW

Im Jahr 2020 half noch keine KI bei der Suche nach Verstößen. KIVI soll die Arbeit der Medienwächter deutlich effektiver machen.

die Weboberfläche von KIVI Screenshots zuerst nur verschwommen an. Auf diese Weise können sich die Sachbearbeiter auf die Inhalte einstellen.

Verdoppelte Fallzahlen

Laut der FAQ auf der Homepage der nordrhein-westfälischen Medienaufsicht durchsucht KIVI die verschiedensten Plattformen, von Twitter und YouTube bis zu Telegram und der russischen Plattform VK. Es könne „täglich mehr als 10.000 Seiten automatisch durchsuchen“. Die Medienwächter arbeiten auch daran, weitere relevante Plattformen in KIVI zu integrieren, etwa Reddit. Facebook und Instagram kann der KI-Medienwächter derzeit ebenfalls noch nicht scannen.

Die Bilanz in Düsseldorf kann sich sehen lassen: Innerhalb eines Jahres hat KIVI 20.685 Funde gemeldet, von denen 14.907 geprüft wurden. In 6766 Fällen stellten die Medienwächter einen Verstoß gegen deutsche Gesetze fest. Davon betrafen 692 Verstöße den politischen Extremismus und 67 Delikte den Bereich Gewalt und Menschenwürdeverstöße. Pro Monat resultiere das in zirka 30 Strafanzeigen, die sich meist auf mehrere Verstöße beziehen. Das entspreche einer Verdoppelung im Vergleich zu der Zeit vor dem Einsatz von KIVI. Wo die Polizei nichts erreicht, leiten die Medienwächter die Löschung der Inhalte ein.

Der Flaschenhals ist aber nach wie vor die menschliche Arbeit. So wird KIVI täglich nur für wenige Stunden aktiviert, weil die Anstalten sonst die Flut an möglichen Verstößen nicht bewältigen könnten. Diese Zahl wollen die Medienwächter noch deutlich nach oben schrauben. So sind nun alle Medienanstalten Deutschlands an das System angeschlossen, sodass sich Mehrfachprüfungen vermeiden lassen und sich einzelne Behörden auf bestimmte Einsatzbereiche spezialisieren können.

Zudem hoffen die Medienwächter darauf, dass sich KIVI über die deutschen Grenzen hinaus verbreitet und so noch mehr Beschwerden bearbeitet werden können. Anfragen aus Frankreich, Spanien, Österreich, Belgien und Luxemburg lägen bereits vor, erklärte Schmid Anfang April.

(jo@ct.de) c't

Infos zum Projekt: ***ct.de/yvk9***