erschienen in der FIfF-Kommunikation,
herausgegeben von FIfF e.V. - ISSN 0938-3476
www.fiff.de

Ingo Dachwitz, Markus Reuter

Warum Künstliche Intelligenz Facebooks Moderationsprobleme nicht lösen kann, ohne neue zu schaffen

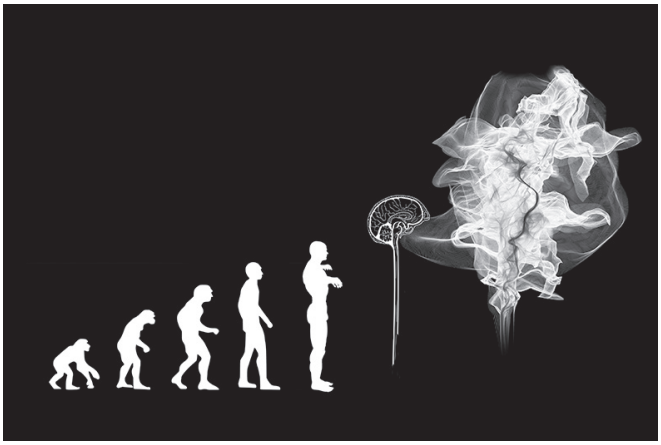
Der Datenkonzern Facebook setzt bei der Moderation von Inhalten zunehmend auf Automatisierung. Eine Quelle erklärt uns erstmals, wie sich die Maschinen auf die Moderationsarbeit auswirken. Auch wenn am Ende heute immer noch Menschen entscheiden: Die automatisierte Inhaltserkennung verändert die digitale Öffentlichkeit grundlegend.

„Auf einmal waren da diese merkwürdigen Tickets." Mika* arbeitet in Essen beim Dienstleister CCC und moderiert dort im Auftrag von Facebook Posts, Videos und Bilder[1], die gegen die Regeln des Konzerns verstoßen könnten. Schon in normalen Wochen bekommen die ModeratorInnen viele Meldungen vorgelegt, die bei ihnen Stirnrunzeln auslösen. An diesem Tag wirkten die Posts besonders wahllos zusammengewürfelt: Oft war an ihnen überhaupt nichts auszusetzen, außer dass ein bestimmtes Wort doppeldeutig war. *Proactive Queue* heißt Ticket-Warteschlage, in der diese merkwürdigen Inhalte zur Moderation vorgelegt wurden.

Es dauerte eine Weile, bis Mika und den KollegInnen dämmerte, womit sie es zu tun hatten: Proaktiv – das heißt, dass Facebooks Software selbst Inhalte sucht und zur Löschung vorschlägt, die sie für verdächtig hält. Heute setzt der Weltkonzern im großen Stil auf diese automatisierte Erkennung unerwünschter Inhalte. Es ist ein qualitativer Sprung: Lange wurden die Beiträge, über die die ModeratorInnen zu entscheiden hatten, nur von Menschen gemeldet. Sie markieren etwa ein Bild als anstößig, gewalttätig oder obszön, so dass es als Ticket in einem Kanal landet und auf dem Bildschirm von Content-ModeratorInnen wie Mika aufschlägt.

Künstliche Intelligenz soll es richten

Wann immer Facebook-Chef Mark Zuckerberg bei den Anhörungen im US-Senat im vergangenen Jahr auf Probleme mit Hetze und anderen unerwünschten Inhalten auf seiner Plattform angesprochen wurde, hatte er für die Politik eine einfache



Evolution wohin?

Antwort parat: „AI will fix this", Künstliche Intelligenz wird es richten. Doch immer mehr ExpertInnen melden Zweifel daran an[2], dass Automatisierung Facebooks Probleme tatsächlich lösen kann. Gleichzeitig können wir beobachten, wie die automatische Vormoderation die digitale Öffentlichkeit schon heute verändert.

Mehr als zwei Milliarden Menschen nutzen laut Unternehmensangaben die von Facebook bereitgestellten Dienste für ihre Kommunikation. Sie diskutieren, streiten, lieben, hassen auf den Plattform des Konzerns. Man findet auf Facebook alles, was das Menschsein ausmacht. Auch Tod und Gewalt in allen Variationen. „Es gibt nichts, was nicht geteilt wird", erzählt Mika lakonisch. Damit Videos von Vergewaltigungen und Enthauptungen

nicht online bleiben, beschäftigt Facebook über Drittfirmen ein Heer von ModeratorInnen – und ein Heer von Maschinen.

Nur für den Dienstgebrauch: Content-Moderation als Staatsgeheimnis

Wie genau das System funktioniert, soll die Öffentlichkeit nicht erfahren. Der Konzern zieht seine Maßnahmen für die Content-Moderation auf wie ein Staatsgeheimnis, ModeratorInnen müssen Geheimhaltungsverträge unterschreiben. Und gerade beim Thema KI lässt sich Mark Zuckerberg ungern in die Karten schauen.

Bekannt ist, dass der Konzern neben digitalen Fingerabdrücken zur automatischen Wiedererkennung bereits gesperrter Inhalte[3] auf maschinelles Lernen setzt. Ein algorithmisches System erkennt hierbei Muster in Trainingsdaten und trifft auf Basis der daraus abgeleiteten Regeln Prognosen zur Bewertung neuer Fälle. Sehr vereinfacht gesagt heißt das: Wenn Post X und Post Y gegen die Gemeinschaftsstandards verstoßen haben, dann tut es Post Z, der ähnliche Eigenschaften aufweist wie X und Y, mit hoher Wahrscheinlichkeit auch.

Durch die schiere Menge zur Verfügung stehender Daten und die gestiegenen Rechenkapazitäten hat diese Form der *Künstlichen Intelligenz* in den vergangenen Jahren enorme Fortschritte gemacht. Sie hilft bei der Diagnose von Krebs, ermöglicht Autos, die fast autonom fahren und Sprachassistenten, die uns zehn Minuten eher wecken, wenn auf dem Weg zur Arbeit Stau herrscht. Doch bei menschlichen Sprache und der komplexen Abwägung, welche auf einer Plattform legitim sind und welche nicht, stößt die Technologie an ihre Grenzen.

Kultur ist nicht maschinenlesbar

Die merkwürdigen Meldungen, über die Mika und KollegInnen sich wunderten, kamen dadurch zustande, dass das System sich auf bestimmte Schlagworte gestürzt hat, die häufig problematisch waren. In einem anderen Kontext, etwa eine in einen Scherz oder Satire eingebette Beleidigung, war ihre Verwendung jedoch vollkommen unproblematisch. „Am Anfang kam da ziemlich viel Schwachsinn", sagt Mika über die Proactive Queue. Mit der Zeit sei das System dann besser geworden. „Wir trainieren die KI, indem wir ihre Vorschläge als richtig oder falsch bewerten." Anfangs häufig wiederkehrende Fehlalarme seien nach einer Weile nicht mehr passiert. Dafür seien neue Fehler aufgetaucht.

Das Problem ist, dass die Unterscheidung dessen, was erlaubt und was verboten ist, in liberalen Gesellschaften eine komplexe Angelegenheit ist. Oft entscheidet der Kontext – und der ist für Maschinen schwer zu erfassen. Meinungsfreiheit ist ein relationales und fluides soziales Konstrukt, dass sich nicht in Formeln übersetzen lässt. Aus diesem Grund bekommen menschliche ModeratorInnen bei der Bearbeitung von gemeldeten Chatnachrichten Mika zufolge nicht nur die eine Nachricht, sondern auch einen Ausschnitt des Nachrichtenverlaufs zu sehen.

Bei Bildern funktioniert die Automatisierung besser: „Pornographie erkennt die Software inzwischen ziemlich zuverlässig", erzählt Mika. Etwa 96 Prozent der wegen Nacktheit wegmoderierten Bilder[4] sind laut Facebook-Angaben durch „Erkennungstechnologie" entdeckt worden. Wie fehleranfällig KI aber auch in diesem Bereich ist, verdeutlicht der Versuch der Blogplattform Tumblr, pornographische Inhalte automatisiert löschen zu lassen. Neben harmlosen Comics[5] sperrte das System auch ein Foto des ehemaligen US-Vizepräsidenten Joe Biden[6]. Viel hängt davon ab, wie gut die KI trainiert ist – dass sie jemals zuverlässig legitime und illegtime Inhalte auseinanderhalten kann, darf bezweifelt werden.

Facebook selbst pflegt neben der Geschichte von KI als Rettung deshalb ein zweites Narrativ, das inzwischen noch häufiger betont wird: Am Ende würden alle relevanten Entscheidungen von Menschen getroffen. Die Software mache zwar Vorschläge zur Moderation, aber die Entscheidungshoheit liege bei den ModeratorInnen. Niemand soll den Eindruck bekommen, Maschinen würden über das hohe Gut der Meinungsfreiheit entscheiden.

Technische Lösungen für soziale Probleme

Doch selbst, wenn jede Moderationsentscheidung am Ende durch einen Menschen geprüft wird, hat die Ausbreitung der Maschinen in der Content Moderation einen profunden Einfluss auf die digitale Öffentlichkeit. Er bedeutet nicht weniger als die Umkehr eines der Grundgesetze des Internets: Bisher durften auch unerwünschte Inhalte auf Plattformen so lange online bleiben, bis sie von irgendwem beim Betreiber einer Seite gemeldet werden. Auch für illegale Inhalte tragen Hoster erst dann eine Verantwortung, wenn sie darauf hingewiesen wurden und den Post trotzdem nicht löschen.

Dieses *Notice and Takedown*[7] genannte Prinzip ist in der EU in der E-Commerce-Richtlinie verankert und konstituierend für ein freies Internet. Bei Plattformbetreibern wie Facebook, Youtube und Twitter hat es lange Zeit zu Verantwortungslosigkeit im Umgang mit verletzenden Inhalten geführt. Aber es hat auch einen Teil der anarchischen Freiheit früher Internettage in die Welt der Plattformmonopole gerettet: Selbst in den Gruppen und Chats von Facebook war Raum für Inhalte, die gegen Regeln wie das rigide Nacktheitsverbot[8] verstoßen.

Mit dem großflächigen Einsatz von Erkennungssoftware kommt Facebook der EU-Kommission zuvor, die derzeit darauf drängt, automatische Erkennung auszuweiten. Die hochumstrittene neue Urheberrechtsrichtlinie[9] wird zur Folge haben, dass Plattformen Inhalte proaktiv und vor deren Veröffentlichung auf Urheberrechtsverletzungen untersuchen. In Anbetracht der täglich von NutzerInnen veröffentlichten Inhalte ist dies nur durch automatische Systeme möglich. Auch für „terroristische Inhalte"[10] will die EU ähnliche Regeln. Was als verboten eingestuft wird, soll nicht nur erkannt werden, sondern gar nicht mehr gepostet werden dürfen. Trotz der bekannten Probleme bei der automatisierten Erkennung des kulturellen Kontextes von Inhalten setzt die EU auf technische Lösungen für soziale Probleme[11].

Wird diese Logik jedoch ausgeweitet, werden die unerwünschten Nebeneffekte[12] zunehmen. Immer wieder werden Vorwürfe laut, das System schieße über das Ziel hinaus. Die Nichtregierungsorganisation *Reporter Ohne Grenzen* etwa machte 2016

auf den Fall des französischen Journalisten und Terrorexperten David Thomson aufmerksam. Sein Account wurde gesperrt, weil auf einem mehrere Jahre alten Bild die Flagge der Terrorgruppe Islamischer Staat/Daesh zu sehen war[13]. Dass ein Mensch sich an dem von Thomson eingeordneten und damals noch nicht verbotenen Symbol störte, ist eher unwahrscheinlich. Stattdessen dürfte der Post in den Schleppnetzen von Facebooks Algorithmus gelandet sein.

Welche Öffentlichkeit wollen wir?

Tatsächlich könnte die automatische Inhalteerkennung sogar noch ausgeweitet werden: In einer vielbeachteten Petition fordert die Kampagnenorganisation Avaaz Mark Zuckerberg auf, automatische Filter auch in den bisher verschlüsselten WhatsApp-Chats zu installieren[14]. Vor dem Hintergrund des von Falschnachrichten geprägten brasilianischen Präsidentschaftswahlkampfes soll dies gegen Desinformation helfen.

Kann Facebook also sein Moderationsprobleme mit Künstlicher Intelligenz lösen? Nur zu einem hohen Preis. Wenn wir über die Zukunft der digitalen Öffentlichkeit nachdenken, sollten wir deshalb gut überlegen, welche Bereiche wir tatsächlich an Maschinen auslagern wollen. Mark Zuckerberg sagt, seine Systeme seien in fünf bis zehn Jahren soweit[15], jegliche Inhalte sauber zu moderieren. Auch Mika selbst rechnet fest damit, auf Kurz oder Lang von der Software ersetzt zu werden: „Irgendwann sind wir überflüssig."

Über diese Recherche und die Quellen:

Unser Wissen über die Organisation des Löschzentrums in Essen beruht auf einem mehrstündigen Gespräch von drei Redakteuren von netzpolitik.org mit einer Quelle bei *Competence Call Center*, die wir im Text geschlechterneutral Mika nennen. Wir können und wollen die Quelle, die wir für glaubwürdig halten, aus Gründen des Informantenschutzes nicht näher beschreiben. Wir sind uns der Probleme und des Risikos bewusst, dass wir uns in Teilen dieser Recherche nur auf eine Quelle stützen können. Deswegen haben wir weite Teile des Artikels durch andere Quellen, auch von anderen Facebook-Dienstleistern verifizieren und bestätigen lassen. Durch diese Quellen können wir heute sagen, dass bei allen Dienstleistern sehr ähnliche oder gar gleiche Systeme eingesetzt werden. Weite Teile dieser Recherche hat außerdem Facebook uns gegenüber bestätigt, die Dienstleister selbst gaben kein Statement ab.

Anmerkungen

1 *https://netzpolitik.org/2019/warum-kuenstliche-intelligenz-facebooks-moderationsprobleme-nicht-loesen-kann-ohne-neue-zu-schaffen/srt*
2 *https://www.theverge.com/2019/2/27/18242724/facebook-moderation-ai-artificial-intelligence-platforms*
3 *https://netzpolitik.org/2017/facebook-twitter-co-upload-filter-gegen-terrorismus-und-extremismus-gestartet/*
4 *https://transparency.facebook.com/community-standards-enforcement#adult-nudity-and-sexual-activity*
5 *https://www.wired.com/story/tumblr-porn-ai-adult-content/*
6 *https://twitter.com/search?f=tweets&q=%23toosexyfortumblr%20joe%20biden&src=typd*
7 *https://netzpolitik.org/2019/plattformen-die-zukunft-von-notice-takedown-in-europa/*
8 *https://netzpolitik.org/2018/facebook-verbannt-mit-neuer-regel-allen-sex-von-seiner-plattform/*
9 *https://netzpolitik.org/tag/eu-urheberrechtsreform/*
10 *https://netzpolitik.org/tag/terreg/*
11 *https://netzpolitik.org/2019/christchurch-es-gibt-keine-technische-loesung-fuer-rechten-terrorismus/*
12 *https://netzpolitik.org/2019/uploadfilter-eine-geschichte-voller-fails/*
13 *https://www.reporter-ohne-grenzen.de/pressemitteilungen/meldung/facebook-muss-sperrung-von-konten-erklaeren/*
14 *https://secure.avaaz.org/campaign/en/wa_destroys_democracy_30/*
15 *https://www.washingtonpost.com/news/the-switch/wp/2018/04/11/ai-will-solve-facebooks-most-vexing-problems-mark-zuckerberg-says-just-dont-ask-when-or-how/*
16 *https://www.leuphana.de/college/bachelor/digital-media.html*
17 *https://pgp.mit.edu/pks/lookup?op=get&search=0x05550760A5E4E814*
18 *http://pgp.mit.edu/pks/lookup?op=get&search=0xB8020A34EF5B7E17*

CC BY NC SA

Ingo Dachwitz und Markus Reuter

Ingo Dachwitz ist Medien- und Kommunikationswissenschaftler, Redakteur bei *netzpolitik.org* und Mitglied beim Verein *Digitale Gesellschaft*. Er schreibt und spricht über Datenkapitalismus, Datenschutz und den digitalen Strukturwandel der Öffentlichkeit. Ingo gibt Workshops für junge und ältere Menschen in digitaler Selbstverteidigung und lehrt im internationalen Studiengang „*Digital Media*"[16] zur politischen Ökonomie digitaler Medien. Gelegentlich moderiert er Veranstaltungen und Diskussionen, etwa auf der *re:publica* oder beim *Netzpolitischen Abend* in Berlin. Ingo ist Mitglied der sozialethischen Kammer der EKD und versucht, auch die Evangelische Kirche mit dem digitalen Zeitalter vertraut zu machen. Kontakt: Ingo ist per Mail an *ingo | ett | netzpolitik.org* (PGP-Key[17]) erreichbar und als *@roofjoke* auf Twitter unterwegs.

Markus Reuter beschäftigt sich mit den Themen Digital Rights, Hate Speech & Zensur, Fake News & Social Bots, Videoüberwachung, Grund- und Bürgerrechte sowie soziale Bewegungen. Bei *netzpolitik.org* seit März 2016 als Redakteur dabei. Er ist erreichbar unter *markus.reuter | ett | netzpolitik.org* (OpenPGP[18]) und auf Twitter unter *@markusreuter_*.