Das Projekt wird im Rahmen der Ausschreibung »Zukunftsdiskurse« vom Niedersächsischen Ministerium für Wissenschaft und Kultur gefördert.

Impulspapier Zukunftsdiskurs

Das Phänomen Hate Speech und seine Erkennung durch KI: interdisziplinär – international – erklärbar?

HASeKI

Hate Speech oder Hate Speech online hat sich zu einem sehr kontrovers diskutierten und interdisziplinär erforschten Thema entwickelt.

Dieses Impulspapier fasst die wesentlichen Positionen zusammen, wie sie im Projekt HASEKI erarbeitet wurden. Dieser Text hat das Ziel, die Ergebnisse des Projektes allgemeinverständlich zu formulieren. Literaturhinweise wurden zur Sicherung der wissenschaftlichen Transparenz eingefügt, jedoch kann das Papier ohne diese verstanden werden.

In sozialen Medien finden sich mittlerweile häufig problematische Äußerungen, die dem Phänomen der Hassrede (= Hate Speech) zuzuschreiben sind. Millionen Nutzer*innen weltweit erstellen dort ununterbrochen eine Vielzahl von Nachrichten (so berichtet zum Beispiel Krikorian (2013) von täglich 500 Millionen auf Twitter verfassten Nachrichten). Betreiber von Plattformen sind in der Verantwortung, da sie den nutzergenerierten Inhalt für jedermann zugänglich machen; diese Verantwortlichkeit besteht auch nach der Erstveröffentlichung fort (Gillespie, 2018). Geeignete Methoden müssen eingesetzt werden, um den richtigen Grad einer freien Kommunikation nach den Möglichkeiten des Internets zu ermöglichen, gleichzeitig jedoch mit den gesellschaftlichen – zum Teil auch gesetzlichen – Vorschriften und Regeln konform zu bleiben.

Hate Speech als Forschungsgegenstand der Linguistik

Hate Speech als genuin interdisziplinärer Forschungsgegenstand hat in den letzten Jahren in den verschiedensten wissenschaftlichen Bereichen gesteigerte Aufmerksamkeit erfahren. Laut Meibauer (2013: 1) kann Hate Speech oder Hassrede verstanden werden als „der sprachliche Ausdruck von Hass gegen Personen oder Gruppen […], insbesondere durch die Verwendung von Ausdrücken, die der Herabsetzung und Verunglimpfung von Bevölkerungsgruppen dienen."

Das Interesse der Linguistik ist es, die kommunikativen und dabei schwerpunktmäßig die sprachlichen Mechanismen von Hate Speech offenzulegen. Diese grundlegende Beschreibung von Hassrede im Netz kann eine Grundlage für Forschung zur automatischen Erkennung bieten. Eine grundlegende Verständigung über das Inventar der verwendeten sprachlichen Muster ist auch dafür notwendig, anhand konkreter sprachlicher Kriterien eindeutiger über Rechtswidrigkeiten entscheiden zu können (Ruzaite, 2018: 94). Tatsächlich wissen wir bislang jedoch weniger über solche Muster, als das starke Forschungsinteresse in diesem Bereich suggerieren mag. Grund hierfür ist, dass es bestimmte Medien, Sprachen, Diskurse und sprachliche Charakteristika gibt, die bislang besonders viel Aufmerksamkeit auf sich gezogen haben, während andere noch ein weitgehendes Desiderat darstellen. Auch genuin sprachwissenschaftliche Forschungsprojekte bestehen weniger häufig als erwartet (zum Beispiel der EU-geförderte Action Grant C.O.N.T.A.C.T, vgl. Assimakopoulos et al., 2017, oder das Anschlussprojekt XPEROHS, vgl. Baumgarten et al., 2019). In folgenden Abschnitt soll die sprachwissenschaftliche Forschung zum Thema Hate Speech und verwandten kommunikativen Phänomenen überblickshaft dargestellt werden.

Forschungsüberblick Linguistik

Generell lassen sich die Arbeiten zum Thema nach ihrer **methodischen Herangehensweise** klassifizieren. Hier unterscheidet man allgemein zwischen qualitativen und quantitativen Ansätzen. Qualitative Forschung fokussiert die Beschreibung von Einzelbeispielen ohne Berücksichtigung statistischer Aspekte. Hierzu gehören zum Beispiel diskurslinguistische Ansätze wie von Musolff (2015). Bei quantitativer Forschung dagegen bemüht man sich jedoch um die Auswertung großer Datensätze. Hierbei gibt es wiederum eine ganze Reihe konkreter Methoden, die bei der Hate-Speech-Forschung eingesetzt werden, zum Beispiel Keywordanalysen, Kollokations- und Konkordanzanalysen oder Sentimentanalysen (z.B. Brindle, 2016; Jaki und De Smedt, 2019; Jaki et al., 2019; Ruzaite, 2018). Interessant sind auch Arbeiten, die quantitative mit qualitativer Analyse kombinieren und somit auch

Grenzen beider Herangehensweisen aufzeigen können (z.B. Hardaker und McGlashan, 2016; Jaki und De Smedt, 2019; Lewandowska-Tomaszczyk, 2017).

Im Bereich der linguistischen Hate-Speech-Forschung gibt es einige **Diskurse und Targets** (d.h. Menschengruppen, gegen die sich Hasskommentare richten), die bereits besser erforscht sind als andere. Dies trifft insbesondere auf Kommentare zu, die sich gegen Migrant*innen allgemein oder Geflüchtete im Speziellen richten, oder auf Texte, in denen es um Migration geht (z.B. Agnetta, 2018; Kreis, 2017 und Musolff, 2015). Zu rechtsextremistischer Kommunikation forschen beispielsweise Baumgarten (2017) oder Jaki und De Smedt (2019). Daneben erntet auch der Bereich des Sexismus und Antifemismus zunehmend mehr Aufmerksamkeit (zum Beispiel Hardaker und McGlashan, 2016; Jaki et al., 2019) – in Bezug auf den deutschen Raum könnte man mutmaßen, dass dies u.a. auch mit dem gesteigerten Interesse für Genderfragen zu tun hat, die durch die Debatten um gendergerechte Sprache noch einmal verstärkt angestoßen werden.

Unter den **untersuchten Medien** dominieren immer noch die Plattformen Facebook und Twitter (z.B. Dynel, 2021; Greule et al., 2020; Hardaker und McGlashan, 2016; Kreis, 2017; Marx, 2018 oder Opiłowski, 2020), denn gerade die sozialen Medien können als Beförderer von Polarisierung und Hate Speech gesehen werden. Allerdings gilt, dass „[v]om Virus der Hasssprache heute fast alle Domänen der Internetkommunikation betroffen [sind]: Blogs, Chats, Soziale Medien oder Leserforen“ (Smułczyński, 2019: 227). Gleichermaßen dominiert das Englische als **untersuchte Sprache** (z.B. Baumgarten, 2017; Brindle, 2016; Hardaker und McGlashan, 2016; Jaki et al., 2019; Lewandowska-Tomaszczyk, 2017; Musolff, 2015). Jedoch beschäftigen sich auch zahlreiche Autor*innen mit der deutschen Sprache, so u.a. Agnetta (2018), Baumgarten et al. (2019), Marx (2018), Smułczyński (2019) oder Stojić und Brala-Vukanović (2017).

Die meisten Arbeiten befassen sich mit lexikalischen **Charakteristika**, also Eigenschaften auf der Wortebene, da diese auf der sprachlichen Oberfläche besonders hervorstechen. Hierzu gehört pejorative Lexik, zum Beispiel abwertende Personenbezeichnungen, die in sozialen Netzwerken unter anderem als Hashtags zu finden sind (*#Gutmensch, #Covidiot, #Bahnhofsklatscher*). Solche Bezeichnungen weisen bisweilen auch abwertende Endungen wie wie dt. *–ler* oder engl. *–tard* auf. Typisch sind insbesondere negative Bezeichnungen für fremde Ethnien (sog. Ethnophaulismen) wie *Kanaken*, *Nafris* oder *Polacken*. Charakteristisch für Hate Speech sind überdies Entmenschlichungsmetaphern (z.B. Musolff, 2015). Hierbei wird Gruppen oder Individuen der Status eines Menschen abgesprochen – sie werden als Tiere (*Schwein*, *Affe*, *Ratte*), Ungeziefer (*Zecken*, *Parasiten*) oder völlig Unbelebtes (*Müll*, *Abfall*) konzeptualisiert. Bekannt ist, dass Entmenschlichung

zur Verringerung von Empathie mit Menschen führen und schlimmsten Fall sogar die Hemmschwelle für Gewalt reduzieren kann (vgl. Cassese, 2020: 108). Auf der sprachlichen Oberfläche zeigt sich Hate Speech jedoch auch außerhalb der Wortebene, zum Beispiel in Konstruktionen wie *Ich bin kein Rassist*, *aber* oder *Ich habe nichts gegen X, aber* und *die ach so* (Baumgarten et al., 2019). Zunehmend wird nun auch die pragmatische Sprachebene untersucht. So unterscheidet Opiłowski (2020) beispielweise Facebook-Kommentare mittels verschiedener Sprachhandlungen (BELEHREN, HERABWÜRDIGEN, BEDROHEN, BELEIDIGEN und KRITISIEREN AM VERHÖHNEN) und Jaki und De Smedt (2019: 20f) geben einen exemplarischen Einblick in Sprechakte in rechtsextremistischer Hate Speech und kommen zu dem Schluss, dass die expressiven Sprechakte den höchsten Grad an Aggressivität aufweisen. In diesen Bereichen ist in Zukunft verstärkte Forschungsaktivität zu erwarten.

Hate Speech abseits der sprachlichen Oberfläche

Allerdings genügt eine Analyse auf der sprachlichen Oberfläche nicht, um die kommunikativen Mechanismen von Hate Speech zu beschreiben, und hier lassen sich weitere Trends für künftige Forschung erkennen: Zum einen gibt es zahlreiche Formen indirekter Hate Speech (vgl. z.B. Ruzaite, 2018), so zum Beispiel Ironie. Andererseits können scheinbar beleidigende Äußerungen auch scherzhaft gemeint sein (vgl. z.B. Dynel, 2021) – unter Freunden ist dies ein häufiges Phänomen. Da es sich bei Hassrede immer um vollständige kommunikative Handlungen handelt, ist der Kontext für die Interpretation stets von Belang. Gerade solche indirekten Formen entziehen sich häufig einer korrekten Klassifizierung bei der automatischen Erkennung von Hate Speech.

Auch besteht Hate Speech im Internet und besonders in sozialen Medien häufig nicht nur aus sprachlichem Material, sondern schließt weitere Komponenten wie Verlinkungen, Emojis oder eingebettete Bilder mit ein. Während Emojis in Hate Speech primäre intensivierend wirken, auf Emotionen schließen lassen und als Interpretationshinweise dienen, sind eingebettete Bilder häufig für die Konstitution von Hate Speech ausschlaggebend – so kann ein Post-Text beispielsweise neutral wirken, während das Bild jedoch klar diskriminierend ist, oder die Abwertung entsteht erst aus dem Zusammenspiel von Sprache und Bild. Das heißt, dass die Linguistik trotz ihres Hauptaugenmerks Sprache künftig verstärkt weitere Bedeutungsressourcen berücksichtigen muss, insbesondere auf Bildebene.

- *Nicht alle sprachliche Muster von Hate Speech sind bereits ausreichend untersucht*
- *Ein häufiges Phänomen sind Entmenschlichungsmetaphern*
- *Qualitative und quantitative Forschung in der Linguistik gehen grundsätzlich unterschiedlich vor*
- *Die Analyse erfolgt oft auf Wortebene, jedoch muss auch der Kontext durch vorhergehende Nachrichten und vor allem Bilder berücksichtigt werden*

Automatische Erkennung von Hate Speech

Eine manuelle Sichtung vor Veröffentlichung ist unmöglich zu realisieren, was eine automatische Regulierung des Inhalts notwendig macht. Dies kann als eine Aufgabe der Computerlinguistik formuliert werden, welche Methoden zur vorrangig automatischen Verarbeitung von Sprach- und Textdaten erforscht. Hier ist nun speziell die Analyse und Erkennung von Nachrichten, die Äußerungen des Phänomens Hate Speech sind, gefragt. Solche Methoden werden eingesetzt in Anwendungen für Betreiber, Autoren und Konsumenten von sozialen Medien, zum Schutz der Bevölkerung vor illegalem Inhalt und zur Verhinderung von dessen Weiterverbreitung, also zum Blockieren von inakzeptablem Inhalt, worunter auch Hate Speech fällt.

Im Fokus in der computerlinguistischen Forschung zu Hate Speech sind daher Methoden zu deren automatischer Erkennung. Einen generellen Überblick zu Ansätzen geben Schmidt & Wiegand (2017). Dabei ist die Aufgabe für ein System, für jede gegebene Nachricht zu entscheiden, ob sie als Hate Speech kategorisiert werden sollte, oder ob sie komplett aus unkritischem Inhalt besteht. Vorab muss dazu definiert werden, nach welchen Richtlinien diese Entscheidung zu treffen ist. Auch ist üblicherweise ein von Menschen gesichteter, vorannotierter Datensatz nötig, der Tausende Beispiele und Gegenbeispiele des Phänomens enthält.

Die größten Sammlungen computerlinguistischer Forschungen zum Thema Hate Speech finden sich in sogenannten Shared Tasks (zum Beispiel Bosco et al., 2018; Wiegand et al., 2018; Zampieri et al., 2019; Basile et al., 2019; Struß et al., 2019; Mandl et al., 2020). Hierbei wird ein vorannotierter Datensatz für gemeinsame Untersuchungen bereitgestellt. Als Grundaufgabe wird üblicherweise die Erkennung von Hate Speech in Kurznachrichten als eine binäre Klassifikation formuliert. Dabei werden jedoch häufig

mehrere Vereinfachungen getroffen. Beiträge werden separat analysiert, also ohne den Kontext auf der Plattform, in dem sie geschrieben und später auch dargestellt werden. Außerdem ist es oft aufgrund von Datenschutzrechten nicht möglich, Nachrichten zusammen mit den zugehörigen Metadaten zu Verfügung zu stellen. Auch werden Anonymisierungsverfahren angewandt, bei denen zum Beispiel Nutzernamen und andere personenbezogene Daten gelöscht werden (Townsend & Wallace, 2017). Lösungsansätze können daher in der Forschung oft nur auf der Basis eines Teilausschnitts der Realität entwickelt werden.

Automatische Erkennungsmethoden

Lösungsansätze zur Hate Speech-Erkennung lassen sich in drei Kategorien einteilen: Lexikon-basierte Erkennung, Methoden auf der Basis von erklärbaren maschinellen Lernsystemen und Methoden auf der Basis von neuronalen Netzwerken.

Bei Lexikon-basierten Methoden wird ein vorab erstelltes Lexikon verwendet, welches Wörter und Wortverbindungen enthält, die als Merkmale für die Erkennung von Hate Speech genutzt werden können. Beim Anwenden dieser Methodik bleibt immer nachvollziehbar, anhand welcher Merkmale (aus dem Lexikon) das System eine Entscheidung getroffen hat. Jedoch bringt sie auch einen hohen Aufwand bei der Erstellung und fortlaufenden Aktualisierung der Lexikoneinträge mit sich, da diese an Veränderungen des Phänomens Hate Speech angepasst werden müssen. Daher haben Lexikon-basierte Systeme häufig zwei Schwachstellen: Zum einen liefern sie Erkennungsfehler bei neuen, unbekannten, oder gar nur leicht veränderten Formen des Phänomens. Zum anderen haben sie es schwer, komplexe Merkmalskombinationen, wie beispielsweise bei impliziter Hate Speech, zu erkennen. Nichtsdestotrotz eignen sich Lexikon-basierte Methoden als automatisches Werkzeug zur Vorauswahl für eine manuelle Weiterverarbeitung.

Maschinelle Lernsysteme hingegen lernen aus Merkmalen deren Gewichtungen und Kombinationen, die mitunter sehr komplex sein können, für eine Vorhersage. Dazu werden Trainingsdaten genutzt, mit denen das System lernen kann, wie häufig und wie prominent gewisse Muster in Beispielen von Hate Speech-Beiträgen im Vergleich zu Gegenbeispielen vorkommen. Hierbei kann man erklärbare Ansätze von Ansätzen auf der Basis von neuronalen Netzwerken unterscheiden. Als erklärbare Ansätze, also solche, die verständlich ausgeben können, wie sie zu Entscheidungen gekommen sind, werden Merkmale vordefiniert, anhand denen gelernt werden soll, das Phänomen zu erkennen. Bei neuronalen Netzwerken hingegen ist das Vorgehen üblicherweise so, dass die Merkmalsauswahl aus Rohdaten komplett vom System selbst übernommen wird. Bei der Erkennung von Hate Speech ergibt

sich bei der Verwendung von erklärbaren maschinellen Lernansätzen eine Schwierigkeit, ähnlich wie beim Einsatz von Lexika: nämlich zu entscheiden, welches sinnvolle Merkmale für das komplexe Phänomen sind. Dadurch, dass das System aber selbst mittels Trainingsdaten lernt, die Merkmale zu gewichten, ist es eher anpassungsfähig an Veränderungen des Phänomens und kann auch lernen, welche Merkmale sinnvoller sind als andere.

In den letzten Jahren ist es in der Computerlinguistik sehr populär geworden, neuronale Netzwerke für komplexe Problem einzusetzen, da diese die Fähigkeit besitzen, sowohl die Merkmalsextraktion, als auch die Gewichtung und Kombination von Merkmalen in einem einzigen Modell durchzuführen. Steuerungs- und Anpassungsmöglichkeiten sind hierbei unter anderem durch die Wahl der Netzwerkstruktur gegeben. Zwar sind durch neuronale Netzwerke in vielen Evaluierungen die besten Leistungen erzielt worden, aber auch hier gilt, dass komplexe Strukturen eine große Menge an Trainingsdaten benötigen, um statistisch ausreichend Gewissheit für gewählte Gewichtungen zu erzielen. Auch ist die Entscheidungsfindung des Netzwerks am Ende in der Regel schwer nachvollziehbar, was bei folgenschweren Klassifikationen problematisch sein kann. Erklären kann man die Entscheidung von neuronalen Netzwerken dennoch mittels der Trainingsdaten: daher möchten wir an dieser Stelle die Wichtigkeit von sorgfältig ausgewählten Trainingsdaten für diesen Ansatz hervorheben. Ist die manuelle Auswahl der Trainingsdaten einseitig, so wird auch das trainierte System nicht alle Schattierungen des Phänomens erkennen können. Neuronale Netzwerke werden meist nur auf einzelne Trainingsdatensätze optimiert, was mittlerweile allerdings sehr gut gelingt.

Ausblick Automatische Erkennungsmethoden

Zusammenfassend lassen sich Ansätze zur automatischen Erkennung von Hate Speech so darstellen, dass sie manuell vorannotierte Trainingsdaten verwenden, um Merkmale und deren Kombinationen für das Phänomen zu lernen. Die Forschung fokussiert dabei auf eine binäre Klassifikation von Kurznachrichten, mit der relevante Beiträge von unbedenklichen getrennt werden sollen. Hate Speech ist dabei jedoch ein komplexes und diverses sprachliches Verhalten, das nur zum Teil in solchen Beiträgen zur Äußerung kommt. Eine kontextualisierte Interpretation scheint notwendig, ist aber aktuell nicht realisierbar. Daher scheint es angebracht, eher präziser eingegrenzte Unterkategorien zu erforschen (siehe zum Beispiel Schäfer & Boguslu, 2021) und statt ganzen Kurztexten eher Teile von Äußerungen zu klassifizieren, was, selbst beim Einsatz von modernen Methoden auf der Basis von neuronalen Netzwerken, zu nachvollziehbaren Klassifikationsentscheidungen führen könnte.

- *Automatische Erkennung ist alternativlos*
- *Die Erkennung wird von Systemen anhand von Trainingsdaten erlernt*
- *Hauptsächlich werden drei Methoden verwendet: Lexikon-basiert, erklärbare maschinelle Lernsysteme und neuronale Netzwerke*
- *Erkennung im Kontext ist noch kaum möglich*

Trainingsdaten für die Automatische Erkennung

Das maschinelle Lernen konnte in den letzten Jahren erhebliche Fortschritte erzielen. Algorithmen suchen nicht nach einzelnen Wörtern oder anhand von nachvollziehbaren Regeln nach sprachlichen Mustern, sondern bauen aus vielen Beispielen Klassifikationsverfahren auf, die letztlich nach vergleichbaren Posts suchen

Diese Trainingsdaten bestehen aus Beispielen für unangemessene und akzeptable Inhalte. Da die Themen innerhalb der hasserfüllten Inhalte äußerst heterogen sein können, sollten sie möglichst breit durch Trainingsdaten abgedeckt sein. Die für das Training verwendeten Texte und die Entscheidungen dazu, sind für die Entwicklung von KI-Verfahren entscheidend. Ihre Zusammenstellung hat den größten Einfluss bei allen Design-Entscheidungen bei der Implementierung von Detektionssystemen für Hassrede.

Die Erstellung von Datensammlungen bildet ein zentrales Instrument für die Forschung zur Erkennung und Bekämpfung von Hassrede. Hierzu werden echte Tweets oder Posts aus sozialen Netzwerken systematisch gesammelt und zunächst von Menschen in zwei oder mehrere Klassen kategorisiert. Der Aufwand für die Erstellung solcher Daten ist hoch und er kann nicht von allen Forscher*innen geleistet werden. Dementsprechend hat sich für diese Forschung das Prinzip der offenen Forschungsdaten etabliert. Wenige Forschungsgruppen entwickeln Daten und stellen diese zur Verfügung. Das führt zudem zu dem positiven Effekt, dass die verschiedensten Algorithmen von mehreren Forschungsgruppen anhand der gleichen Daten verglichen werden. Die Ergebnisse bei der Klassifikation sind direkt vergleichbar.

Neben dem Einordnen als Hassrede sollen die problematischen Tweets in die Klassen implizit und explizit sortiert werden. Gemessen wird die Klassifika¬tionsgenauigkeit der Systeme mit dem F1-Maß, welches Recall und Precision zusammenfasst. Das beste System erreichte für den binären Task (Inhalt problematisch oder nicht) ein F1-Maß von 0,76 (Struß et al. 2019).

Eine der großen Herausforderungen bei der Annotation von Hassrede ist der fehlende Kontext. In einem sozialen Netzwerk steht jede Äußerung in einem kommunikativen Zusammenhang und wird vom Leser unter Einbezug möglicher vorhergehender Äußerungen interpretiert. Aufgrund der Kürze von Texten auf Online Plattformen werden gerade dort Bezüge nicht explizit genannt, sondern der Sender vertraut auf den gegebenen Kontext. Betrachtet man nun jede Äußerung für sich, können völlig unterschiedliche Interpretationen entstehen.

Bei der Annotation mit und ohne Kontext zeigte sich, dass fast 50% der als Hassbotschaften ausgezeichneten Nachrichten mit Kontext nicht mehr als solche betrachtet wurden. Der umgekehrte Effekt war geringer (Menini et al., 2021). Diese Problematik des Kontexts greift auch der HASOC Contextual Subtask 2021 auf. Dabei sollte auf eine weitgehend einheitliche Größe des Kontexts Werte gelegt werden und auch alle Kontext-Tweets sollten einheitlich annotiert werden. Zudem war ein weiteres Ziel, den Anteil der Hassrede relativ hoch zu halten, um ihn für maschinelles Lernen angemessen zu gestalten (Satapara et al., 2021).

Zwar scheinen Algorithmen Entscheidungen objektiv zu treffen, gleichwohl zahlreiche Entscheidungen. Somit ist der Aufbau von Trainingsmengen eine soziale Konstruktion, die in einem bestimmten Kontext unter Rahmenbedingungen und Zwängen erfolgt. Die Entwickler von Daten treffen bei der Gestaltung der Trainingsdaten bewusst oder unbewusst Entscheidungen, die sich auf die Daten auswirken und somit auch die Wirksamkeit der KI Verfahren beeinflussen.

Die Trainingsmenge sollte repräsentativ für die bekannten Formen der problematischen Inhalte sein, wobei aber aufgrund der Vielfältigkeit sprachlicher Ausdrucksformen unklar ist, wie diese Repräsentativität erreicht oder erkannt werden kann. Somit kann lediglich der Prozess der Erstellung von Hate Speech Datenmengen betrachtet und bewertet werden. Er besteht üblicherweise aus den folgenden Schritten (Vidgen & Derczynski 2020):

- Erstellung einer Strategie zur Vorauswahl von Inhalten aus sozialen Netzwerken
- Umsetzung der Strategie mit Werkzeugen und Extraktion von Posts aus großen Mengen aus sozialen Netzwerken

- Annotation einer Vorauswahl durch Menschen

Die Strategie besteht zum einen oft im Auswählen von Begriffen, die für Hate Speech typisch sein könnten (GermEval) oder auch im Erstellen eines Vorab-Klassifizierers (HASOC 2020, Mandl et al. 2020). In beiden Fällen können bestimmte Inhalte präferiert werden. So spielen bei der Auswahl von Begriffen notwendigerweise Vorkenntnisse bzw. Annahmen über Hate Speech eine erhebliche Rolle. Ganze Komplexe problematischer Inhalte könnten übersehen werden, wenn lediglich bereits bekannte Themen selektiert werden.

Die Auswahl von bestimmten Hate Speech Beispielen durch manuelles Suchen kann dazu führen, dass diese Beispiele immer von einigen Autor*innen stammen, während die neutralen Äußerungen von anderen Autor*innen stammen. Das kann sogar darin münden, dass ein Klassifikationssystem letztlich eine Autorenschaft-Erkennung durchführt. Eine solche Stil-Erkennung erkennt dann evtl. ganz andere Merkmale und ist nicht in der Lage, in einer realen Umgebung eine gute Erkennungsqualität für Hassbotschaften zu liefern (Arango et al. 2020). Deswegen sollten von jedem Profil in sozialen Netzwerken immer mehrere Posts gesammelt werden, um pro Autor*in Beispiele für problema¬tische und unproblematische Inhalte einzubauen.

Alle Trainingsmengen weisen jedoch deutlich höhere Anteile auf, da es schwierig ist, mit nur einem geringen Anteil von Beispielen eine Klassifikation zu trainieren. Zudem liefern die Algorithmen bessere Ergebnisse, wenn die Klassen vergleichbar oft vorkommen. Aus dieser Perspektive spiegeln die Trainingsmengen in keiner Weise die Realität wider.

Messung von Verzerrungen

Die Zuverlässigkeit von Datensets kann mit einigen Methoden überprüft werden. Die Sprachmodelle der Trainingsmenge sowie gegebenenfalls auch der Testmenge können untereinander und mit dem allgemeinen Sprachmodell im Korpus verglichen werden. Dazu kann z.B. das Maß Mutual Information eingesetzt werden. Es wurde schon beobachtet, dass bestimmte Begriffe in den Hate Korpora häufiger vorkommen als allgemein. Dies führt teils zu guten Klassifikationsergebnissen bei der Entwicklung, die aber unter realen Einsatzbedingungen nicht erreicht werden.

Eine Analyse von Ross et al. (2016) konnte sogar zeigen, dass selbst schriftliche Richtlinien keinen hohen Einfluss auf die Übereinstimmung zwischen Annotierenden haben.

Grenzfälle weisen eine deutlich höhere Abweichung auf (Salminen et al. 2019). Dies bedeutet, dass ein gutes Interrater Agreement auch bedeuten kann, dass lediglich sehr klare Fälle in der Menge vorkommen. Die Häufigkeit von Grenzfällen in den gesammelten Daten ist vorab ja nicht bekannt und lässt sich auch nicht steuern. Wenige oder unsicher bewertete Grenzfälle können es den Algorithmen allerdings zusätzlich erschweren, die Grenze deutlich zu ziehen. Somit muss das Interrater Agreement als Qualitätsmerkmal auch hinterfragt werden.

Die Genauigkeit der Vorhersage von Hate Speech variiert meist stark je nach Datenmenge. Dies weist auf die Bedeutung der Trainingsdaten hin. Für einen realen Einsatz ist natürlich viel wichtiger, wie gut ein System bei völlig anderen Daten unter echten Bedingungen funktioniert. Dann stellt sich die Frage, ob die Systeme aus der Forschung robust genug für einen Dauerbetrieb wären. Dies lässt sich transparent und mit Daten, die außerhalb von kommerziellen Plattformen zur Verfügung stehen, nur schwer überprüfen. Als gängigste Methode wird hierfür mit den Trainingsdaten eines Benchmarks ein Modell trainiert und dann mit den Testdaten anderer Benchmarks getestet. Wenn ein Klassifikationssystem mit einer Menge trainiert wird und dann auch eine andere angewendet wird, zeigt sich zu einem gewissen Maß, ob diese das gleiche Konzept abbilden bzw. Verzerrungen in den Daten vorliegen. Ergebnisse bei solchen Cross-Validitäts-Studien zeigen, dass teils deutlich niedrigere Trefferquoten erzielt werden. Experimente von Fortuna et al. (2021) zeigen, dass die Performanz für einen Datensatz um über 30% Genauigkeit schwanken kann, je nachdem mit welchem anderen Datensatz trainiert wurde.

- *Die Qualität der Trainingsdaten ist entscheidend für die Automatische Erkennung*
- *Noch gibt es jedoch kaum Methoden für die Messung der Qualität.*
- *Ziel ist eine Erkennung, die nicht nur in Trainingsdaten gut funktioniert, sondern auch auf den realen Einsatz übertragbar ist*
- *Verzerrungen können sehr leicht durch das Vorgehen beim Sammeln, Auswählen oder Annotieren entstehen*
- *Vergleichende Forschung zu Methoden der Erstellung ist notwendig*

Die Regulierung von Internetinhalten am Beispiel Hassrede

Digitale Kommunikation und politische Regulierung stehen in einem grundlegenden Spannungsverhältnis. Das Internet ist dezentral, transnational und die erforderlichen Infrastrukturen weitgehend in privater Hand. Hinzu kommt die relative Anonymität der Netzwerkkommunikation. Dies alles fordert hergebrachte Strukturen und Routinen der Regulierung, insbesondere durch Staaten, heraus. Neue und flexible Formen des Regierens und der Koordination – in der Wissenschaft wird auch von „Governance“ gesprochen – sind also gefragt. Dies gilt gerade auch, wenn es um die Regulierung von Internetinhalten geht. In früheren Phasen der Internetentwicklung überwogen optimistische Erwartungen im Hinblick auf die politische Teilhabe, die demokratische Diskursqualität und die grenzüberschreitende Kommunikation. Seit einigen Jahren wird jedoch verstärkt über die Schattenseiten digitaler Kommunikation debattiert. Dabei nehmen Sorgen vor der Verbreitung von Hassrede, Hetze und extremistischen Inhalten sowie ihren schädlichen Wirkungen auf das gesellschaftliche Leben und den demokratischen Diskurs einen prominenten Platz ein. Diese Sorgen sind in den vergangenen Jahren zunehmend auch von politischen Entscheidungsträger*innen aufgegriffen worden und haben zu verstärkten Regulierungsbemühungen auf verschiedenen Ebenen geführt.

Der beschriebene Trend hat dabei auch demokratische Systeme erfasst. Für liberale Demokratien ist das eingangs genannte Spannungsverhältnis besonders groß, weil jeder Eingriff in die Online-Kommunikation gegen das für den demokratischen Regimetyp konstitutive Recht auf freie Meinungsäußerung und Information sorgsam abgewogen werden muss (Schejter & Han 2011: 248). Die für demokratische Medienpolitik deshalb so typische, normativ begründete Zurückhaltung, wenn es um Eingriffe in die Medienkommunikation geht, ist im Zuge einer gründlichen Liberalisierung auch der klassischen Medienmärkte in der sog. ‚westlichen Welt‘ eher größer geworden. Vor diesem Hintergrund stellt sich die Frage nach einer allgemeinen Rechtfertigung von Eingriffen in die Redefreiheit durch liberaldemokratische Regime heute in neuer Dringlichkeit. Das im deutschen Staatsrecht bekannte Prinzip der „wehrhaften Demokratie“ bietet eine Antwortmöglichkeit, allerdings sind die Spielräume für staatliche Eingriffe auch hier eng beschränkt (Sirsch 2013; Sunstein 1995).

Haben Expert*innen und Beobachter*innen aufgrund der besonderen Regulierungsarmut in der Internet-Governance über lange Zeit von einer Art ‚Internet-Exzeptionalismus‘ gesprochen, ist heute immer häufiger von dessen Ende die Rede. Vollzieht die digitalpolitische Regulierung also derzeit einen Paradigmenweschel zugunsten einer politischen Institutionalisierung, wie wir sie bereits im Hinblick auf andere Medientypen, also Presse, Rundfunk und Fernsehen kennen? Ist dies eine gleichsam

nachholende Entwicklung? Oder bleibt die Governance digitaler Medien doch eine Sphäre mit eigenen Gesetzlichkeiten und spezifischen Governance-Formaten?

Medien- und Kommunikationswissenschaftler*innen haben in diesem Zusammenhang vielfach festgestellt, dass Massenmedien seit jeher politischer Institutionalisierung unterliegen (Jarren 2007) und dass technologischer Wandel im Mediensektor stets neue Regulierungsdebatten und -ansätze hervorgerufen hat (Schejter & Han 2011, S. 245), die mit zeitlicher Verzögerung ("zweistufiger Institutionalisierungsprozess" als medienhistorische Regel, Stöber 2007, S. 107) auch das neue Feld einer angepassten Regulierung unterwerfen und dabei medienpolitischen Traditionslinien folgen (2011: 245). Allerdings steht die empirische Überprüfung dieser historisch-institutionalistischen Hypothese bislang für die aktuelle Entwicklung der digitalpolitischen Regulierung noch aus.

Online-Hassrede und ihre Bekämpfung spielen in den gesellschaftlichen Debatten aktuell eine wichtige Rolle. Doch was ist das Besondere an Online-Hass? In der öffentlichen und wissenschaftlichen Debatte werden immer wieder die Besonderheiten und genuin neuen Herausforderungen durch digitale Hassrede betont, bspw. die (geografische) Distanz zwischen Täter*innen und Opfern. Diese führe zu einer abnehmenden Wahrscheinlichkeit, dass sich wie bei persönlicher Interaktion Mitgefühl oder Sympathie einstellen könnten (Brown 2015, 134f.). In diesem Kontext werden in der demokratietheoretischen Literatur seit langer Zeit die schädlichen Wirkungen der relativen Anonymität von Kommunikationspartner*innen im Netz diskutiert, wobei ein vergleichsweise offensives, polarisierendes und extremes Diskursverhalten erwartet wird (Barber 2001, S. 216). Dadurch ständen gerade die Chancen für politische Deliberation als zentrales demokratisches Versprechen der Internetentwicklung schlecht. Dies wird immer wieder zur Rechtfertigung regulatorischer Maßnahmen angeführt.

Passend zu den wahrgenommenen Besonderheiten besteht die Erwartung, dass dem Phänomen nicht mit etablierten Maßnahmen begegnet werden könne. Diese Wahrnehmung erscheint insofern berechtigt, als die Digitalisierung zu drastisch gewandelten Kommunikationsumgebungen mit veränderten Veröffentlichungsmöglichkeiten und (international) begrenzten Regulierungsspielräumen geführt hat. Medien- und Kommunikationswissenschaften haben die veränderten Rahmenbedingungen für die Produktion und Verbreitung von Medieninhalten auf unterschiedlichen Stufen der Internetentwicklung, insbesondere im Hinblick auf soziale Medien, theoretisch und empirisch gründlich bearbeitet (stellvertretend für viele s. Benkler 2006; Shirky 2011). Die Regulierungsspielräume für staatliche Autoritäten sind mit technischen Begründungen als gering eingestuft oder aus normativ-ethischen Gründen zurückgewiesen worden (Mueller 2017; Johnson & Post 1996; dazu kritisch: Goldsmith & Wu 2006; Shearing & Wood 2003).

Die meisten politikwissenschaftlichen Studien, die sich explizit mit der Online-Inhalteregulierung befassen, sind bislang auf dem Feld der Vergleichenden Politikforschung entstanden. Dabei hat vor allem die Regimedifferenz die Annahmen zur Variation von Politikansätzen beeinflusst. So befasst sich ein dominanter Strang mit Online-Kontrolle und Zensurmaßnahmen ausschließlich in autokratischen Regimen (Rød & Keremoğlu & Weidmann 2020; Hellmeier 2016). Ergebnisse internationaler Vergleichsstudien zur Regulierungspraxis, die gesteigerte Aktivitäten auch für Demokratien dokumentieren (Freedom House 2018, 2019, 2020), zeigen aber, dass auch liberale Demokratien nicht mehr vor der Regulation von Internetinhalten zurückschrecken. Der Trend zum Ende staatlicher Zurückhaltung in liberalen Demokratien erfolgt dabei zumeist unter Bezugnahme auf die Verbreitung von Hassrede oder Desinformation.

Busch et al. haben diese Befunde aufgegriffen und eine auf Demokratien bezogene Entwicklungsannahme formuliert. Demnach stelle sich die Frage, ob die gesteigerten Regulationsaktivitäten demokratischer Regime auf einen Lerneffekt zurückzuführen sind, wonach Demokratien von Autokratien lernen ("Learning from Autocracies", Busch et al. 2018; Busch 2017; aber auch Gomez 2004). Diese Konvergenzhypothese hat Vorteile darin, dass sie ein Erklärungsangebot für jüngere empirische Befunde in groß angelegten Vergleichsstudien bietet. Die tatsächlichen Überprüfungen der Annahme leiden indes ihrerseits an der Beschränkung auf liberale Demokratien. Damit besteht bis heute ein augenfälliger Mangel an regimetypübergreifenden Arbeiten zur Internetkontrolle im Allgemeinen und zur Online-Content-Regulierung im Besonderen (Ausnahmen bilden Stier 2017; Timofeeva 2006). Jenseits der Regimetypendifferenz wären dabei alternative Erklärungsfaktoren zu prüfen, etwa die historische Pfadabhängigkeit. Die medienwissenschaftliche Forschungstradition könnte hier in zweierlei Hinsicht aushelfen. Zum einen bietet sie einen differenzierten Blick auf die medienpolitische Regulierung über verschiedene Medientypen und technologische Entwicklungsstufen hinweg auch für Demokratien (Humphreys 1996), zum anderen hält sie Typologien von Mediensystemen bereit, die womöglich geeigneter sind, Variationen in aktuell relevanten digitalpolitischen Regulierungsbereichen zu erklären (Hallin & Mancini 2004).

Gerade Deutschland kann im Hinblick auf die Regulierung von Hassrede als Beispiel dienen. Denn hierzulande führten Bedenken über die Folgen von Hassrede und Extremismus im Netz – bspw. die begünstigende Wirkung auf politische Gewalt (Müller & Schwarz 2020) – zu neuen regulatorischen Maßnahmen auf nationaler Ebene. Dadurch wurden die Betreiber sozialer Medien 2017 zur Löschung strafbarer Inhalte nach dem sog. ‚notice-and-takedown'-Prinzip verpflichtet (s. Netzwerkdurchsetzungsgesetz). Das bedeutet, dass große Plattformen illegale Inhalte umgehend löschen müssen,

sobald sie auf diese hingewiesen worden sind. In diesem Jahr sind die Vorgaben noch einmal verschärft worden und schließen nun auch eine Meldung bei den Strafverfolgungsbehörden mit ein (Gesetz zur Bekämpfung des Rechtsextremismus und der Hasskriminalität). Das NetzDG ist international viel beachtet, auch viel kritisiert, worden. An seinem Vorbild haben sich andere Länder in ihrer Rechtsetzung orientiert. Auch der im Dezember 2020 von der EU präsentierte und mit großen Erwartungen begleitete Entwurf zu einem Digital Services Act, also einer grundlegenden europäischen Rechtsetzung im Hinblick auf digitale Dienste, macht Anleihen am Beispiel des NetzDG.

Abschließend kann festgehalten werden, dass die verbreitete These eines weitgehend unregulierten oder schwach regulierten digitalen Kommunikationsraumes – dessen Staatsferne mitunter aus normativer Perspektive begrüßt worden ist – heute klar infragegestellt werden muss. Vielmehr beobachten Wissenschaftler*innen eine gesteigerte Regulationsaktivität in und durch liberale Demokratien. Diese könnte mit Blick auf die Geschichte der Medienregulierung als eine nachholende Entwicklung verstanden werden. In jedem Fall wirken die Sorgen vor der Verbreitung von Hassrede als Treiber dieser Entwicklung.

- *Die negativen Auswirkungen von Hassrede für den politischen Diskurs werden weitgehend akzeptiert*
- *Eingriffe in Internet-Inhalte sind auch ein Eingriff in die Meinungsfreiheit und somit müssen somit sorgfältig abgewogen werden*
- *Es ist noch unklar, ob sich die Digitale Governance als Sphäre mit eigenen Gesetzlichkeiten etabliert*
- *Die Eingriffe nehmen auch in liberalen Demokratien zu, wobei umstritten bleibt, ob diese von Autokratien gelernt haben*
- *Deutschland mit dem NetzDG und die EU werden für ihre Regulierung zwar auch kritisiert, bilden jedoch Vorbilder*

Literatur

Gillespie, Tarleton. (2018). Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press.

Krikorian, Raffi. (2013). New Tweets per second record, and how! Friday, 16 August 2013. In: blog.twitter.com, URL: https://blog.twitter.com/engineering/en_us/a/2013/new-tweets-per-second-record-and-how.html

Literatur zur Linguistik

Agnetta, Marco. (2018). Die Entmachtung der Metapher: Zur Dekonstruktion sprachlich vermittelter Feindbilder im europäischen Flüchtlingsdiskurs. In Metaphorik.de, 28, 11-46.

Assimakopoulos, Stavros, Fabienne H. Baider und Sharon Millar, Hg. (2017). Online Hate Speech in the European Union. A Discourse-Analytic Perspective. Cham: Springer.

Baumgarten, Nicole, Eckhard Bick, Klaus Geyer, Ditte Aakær Iversen, Andrea Kleene, Anna V. Lindø, Jana Neitsch, Oliver Niebuhr, Rasmus Nielsen und Esben N. Petersen. (2019). Towards balance and boundaries in public discourse: expressing and perceiving online hate speech (XPEROHS). In RASK: International Journal of Language and Communication, 50, 87-108.

Brindle, Andrew. (2016). The Language of Hate: A Corpus Linguistic Analysis of White Supremacist Language. New York/Abingdon: Routledge.

Cassese, Erin C. (2020). Dehumanization of the Opposition in Political Campaigns. In Social Science Quarterly, 101(1), 107-120. doi.org/10.1111/ssqu.12745

Dynel, Marta. (2021). Desperately seeking intentions: Genuine and jocular insults on social media. In Journal of Pragmatics, 179, 26-36. doi.org/10.1016/j.pragma.2021.04.017

Geyer, Klaus. (2019). Die ‚Grammatik' der Hassrede – am Beispiel des Dänischen. In Jürg Strässler (Hg.), Sprache(n) für Europa. Mehrsprachigkeit als Chance. Berlin u.a.: Peter Lang, 195-207.

Greule, Albrecht, Sandra Reimann und Julia Enzinger. (2020). Abkehr vom Frieden? Eine medien-und politolinguistische Untersuchung von Facebook-Einträgen der Organisation Pegida. In Makowski (Hg.), Hassrede–ein multidimensionales Phänomen im interdisziplinären Vergleich. Lodz: Wydawnictwo Uniwersytetu Łódzkiego, doi.org/10.18778/8142-633-6.11

Hardaker, Claire und Mark McGlashan. (2016). "Real men don't hate women": Twitter rape threats and group identity. In Journal of Pragmatics, 91, 80-93. doi.org/10.1016/j.pragma.2015.11.005

Jaki, Sylvia und Tom De Smedt. (2019). Right-wing German hate speech on Twitter: Analysis and automatic detection. arXiv:1910.07518.

Jaki, Sylvia, Tom De Smedt, Maja Gwóźdź, Rudresh Panchal, Alexander Rossa und Guy De Pauw. (2019). Online hatred of women in the Incels.me forum: Linguistic analysis and automatic detection. In *Journal of Language Aggression and Conflict,* 7(2), 240-268. doi.org/10.1075/jlac.00026.jak

Lewandowska-Tomaszczyk, Barbara. (2017). Incivility and confrontation in online conflict discourses. In Lodz Papers in Pragmatics, 13(2), 347-363. doi.org/ 10.1515/lpp-2017-0017

Marx, Konstanze. (2018). Hate Speech–Ein Thema für die Linguistik. In Albers und Katsivelas (Hg.), Recht & Netz. Baden-Baden: Nomos, 37-57.

Meibauer, Jörg. (2013). Hassrede–von der Sprache zur Politik. In Maibauer (Hg.), Hassrede/Hate Speech. Interdisziplinäre Beiträge zu einer aktuellen Diskussion. Gießen: Gießener Elektronische Bibliothek, 1-17. http://geb.uni-giessen.de/geb/volltexte/2013/9251/pdf/HassredeMeibauer_2013.pdf

Musolff, Andreas. (2015). Dehumanizing metaphors in UK immigrant debates in press and online media. In Journal of Language Aggression and Conflict, 3(1), 41-56. doi.org/ 10.1075/jlac.3.1.02mus

Opiłowski, Roman. (2020). Netzhass in deutschen und polnischen Nutzerkommentaren aus multimodaler Sicht. In Makowski (Hg.), Hassrede–ein multidimensionales Phänomen im interdisziplinären Vergleich. Lodz: Wydawnictwo Uniwersytetu Łódzkiego, 167-185. doi.org/10.18778/8142-633-6.10

Ruzaite, Jurate. (2018). In search of hate speech in Lithuanian public discourse: A corpus-assisted analysis of online comments. In Lodz Papers in Pragmatics, 14(1), 93-116. doi.org/10.1515/lpp-2018-0005

Smułczyński, Michal. (2019). Wo liegen die Grenzen der Hasssprache? Kommentare zum Anschlag in Manchester in sozialen Netzwerken in Deutschland, Dänemark und Polen. In Linguistische Treffen in Wrocław, 15(1), 225-232.

Stojić, Aneta und Marija Brala-Vukanović. (2017). Gewalt der Sprache: Lexikalische Abwertung als (Ab)bild einer Sprachgemeinschaft. In Linguistik Online, 82(3), 65-77. doi.org/10.13092/lo.82.3715

Literatur zur Automatischen Erkennung

Basile, Valerio, Cristina Bosco, Elisabetta Fersini, Nozza Debora, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso und Manuela Sanguinetti. (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In: 13th International Workshop on Semantic Evaluation, 54-63. Association for Computational Linguistics, 2019.

Bosco, Cristina, Dell'Orletta Felice, Fabio Poletto, Manuela Sanguinetti und Tesconi Maurizio. (2018). Overview of the EVALITA 2018 hate speech detection task. In: EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, 2263, 1-9. CEUR.

Thomas Mandl, Sandip Modha, Gautam Kishore Shahi, Amit Kumar Jaiswal, Durgesh Nandini, Daksh Patel, Prasenjit Majumder and Johannes Schäfer. (2020). Overview of the HASOC track at FIRE 2020: Hate Speech and Offensive Content Identification in Indo-European Languages. In: Proceedings of the Forum for Information Retrieval Evaluation (FIRE 2020), December 16–20, 2020, Hyderabad, India.

Schäfer, Johannes und Kübra Boguslu. (2021). Towards annotating illegal hate speech: A computational linguistic approach. Detect Then Act (DTCT) Technical Report 3, ISSN 2736-6391. URL: https://dtct.eu/wp-content/uploads/2021/10/DTCT-TR3-CL.pdf

Schmidt, Anna und Michael Wiegand. (2017). A survey on hate speech detection using natural language processing. In: Proceedings of the Fifth International workshop on natural language processing for social media, 1-10.

Struß, Julia Maria, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand und Manfred Klenner. (2019). Overview of GermEval Task 2, 2019 shared task on the identification of offensive language. In: Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): 354-365. Erlangen, Germany.

Townsend, Leanne und Claire Wallace. (2017). The ethics of using social media data in research: A new framework. In: The ethics of online research. Emerald Publishing Limited.

Wiegand, Michael, Melanie Siegel und Josef Ruppenhofer. (2018). Overview of the GermEval 2018 shared task on the identification of offensive language. In: Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018). Vienna, Austria: Austrian Academy of Sciences.

Zampieri, Marcos, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra und Ritesh Kumar. (2019). SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In: Proceedings of the 13th International Workshop on Semantic Evaluation, 75-86.

Literatur zu Trainingsdaten für die Automatische Erkennung

Arango, A., Pérez, J., and Poblete, B. 2020. Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). *Information Systems*, 101584.

Fortuna, P., Soler-Company, J., and Wanner, L. 2021. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3), 102524. https://doi.org/10.1016/j.ipm.2021.102524

Mandl, T., Modha, S., M, A. K., and Chakravarthi, B. R. 2020. Overview of the HASOC Track at FIRE 2020: Hate Speech and Offensive Language Identification in Tamil, Malayalam, Hindi, English and German. In: *Proceedings of the 12th Annual Meeting of the Forum for Information Retrieval Evaluation (FIRE)*. ACM. https://doi.org/10.1145/3441501.3441517

Menini, S., Aprosio, A. P., and Tonelli, S. 2021. Abuse is Contextual, What about NLP? The Role of Context in Abusive Language Annotation and Detection. *arXiv preprint* arXiv:2103.14916.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., and Wojatzki, M. 2016. Measuring the reliability of hate speech annotations: The case of the European refugee crisis. In: *NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*.

Salminen, J., Almerekhi, H., Kamel, A. M., Jung, S. G., and Jansen, B. J. 2019. Online Hate Ratings Vary by Extremes: A Statistical Analysis. In: *Proceedings Conference on Human Information Interaction and Retrieval*. (CHIIR) ACM. S. 213-217. https://doi.org/10.1145/3295750.3298954

Satapara, S., Modha, S., Mandl, T., Madhu, H., and Majumder, P. 2021. Overview of the HASOC Subtrack at FIRE 2021: Conversational Hate Speech Detection in Code-mixed language. in: *Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation*. CEUR, 2021.

Struß, J. M., Siegel, M., Ruppenhofer, J., Wiegand, M., and Klenner, M. 2019. Overview of GermEval Task 2, 2019 shared task on the identification of offensive language. In: *Proceedings of the 15th Conference on Natural Language Processing* (KONVENS) Nürnberg/Erlangen. https://doi.org/10.5167/uzh-178687

Vidgen, B., and Derczynski, L. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PloS one* 15(12) https://doi.org/10.1371/journal.pone.0243300

Literatur zur Politischen Regulierung

Barber, B. R. (2001). Which Technology for which Democracy? Which Democracy for which Technology? In B. Holznagel, A. Grünwald, & A. Hanssmann (Eds.), Schriftenreihe Information und Recht: Bd. 24. Elektronische Demokratie: Bürgerbeteiligung per Internet zwischen Wissenschaft und Praxis (pp. 209–217). München: Beck.

Benkler, Y. (2006). The wealth of networks. How social production transforms markets and freedom. New Haven, Conn. [u.a.]: Yale Univ. Press.

Brown, A. (2015). Hate Speech Law: A Philosophical Examination: Routledge.

Busch, A. (2017). Netzzensur in liberalen Demokratien. In A. Croissant, S. Kneip, & A. Petring (Eds.), Demokratie, Diktatur, Gerechtigkeit: Festschrift für Wolfgang Merkel (pp. 331–352). Wiesbaden: Springer VS.

Busch, A., Theiner, P., & Breindl, Y. (2018). Internet Censorship in Liberal Democracies: Learning from Autocracies? In J. Schwanholz, T. Graham, & P.-T. Stoll (Eds.), Managing Democracy in the Digital Age: Internet Regulation, Social Media Use, and Online Civic Engagement (pp. 11–28). Cham: Springer Publishing. https://doi.org/10.1007/978-3-319-61708-4_2

Freedom House (2018). Freedom on the Net 2018: The Rise of Digital Authoritarianism. Retrieved from https://freedomhouse.org/sites/default/files/2020-02/10192018_FOTN_2018_Final_Booklet.pdf

Freedom House (2019). Freedom on the Net 2019: The Crisis of Social Media. Retrieved from https://www.freedomonthenet.org/sites/default/files/2019-11/11042019_Report_FH_FOTN_2019_final_Public_Download.pdf

Freedom House (2020). Freedom on the Net 2020: The Pandemic's Digital Shadow. Retrieved from https://freedomhouse.org/sites/default/files/2020-10/10122020_FOTN2020_Complete_Report_FINAL.pdf

Goldsmith, J. L., & Wu, T. (2006). Who controls the Internet?: Illusions of a borderless world. New York: Oxford Univ. Press.

Gomez, J. (2004). Dumbing Down Democracy:Trends in Internet Regulation, Surveillance and Control in Asia. Pacific Journalism Review, 10, 130–150.

Hallin, D. C., & Mancini, P. (2004). Comparing media systems: Three models of media and politics. Communication, society and politics. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511790867

Hellmeier, S. (2016). The Dictator's Digital Toolkit: Explaining Variation in Internet Filtering in Authoritarian Regimes. Politics & Policy, 44(6), 1158–1191. https://doi.org/10.1111/polp.12189

Humphreys, P. (1996). Mass media and media policy in Western Europe. European Policy Research Unit series. Manchester: Manchester Univ. Press.

Jarren, O. (2007). Die Regulierung der öffentlichen Kommunikation. Zeitschrift für Literaturwissenschaft und Linguistik, 37(2), 131–153. https://doi.org/10.1007/BF03379662

Johnson, D. R., & Post, D. (1996). Law and Borders: The Rise of Law in Cyberspace. Stanford Law Review, 48(5), 1367. https://doi.org/10.2307/1229390

Keremoğlu, E., & Weidmann, N. B. (2020). How Dictators Control the Internet: A Review Essay. Comparative Political Studies, 12(5), 001041402091227. https://doi.org/10.1177/0010414020912278

McQuail, Denis (2010). McQuail's mass communication theory. 6. ed. Los Angeles, Calif.: Sage.

Mueller, M. (2017). Will the Internet Fragment?: Surveillance, Cybersecurity and Internet Governance: Polity Press.

Müller, K., & Schwarz, C. (2020). Fanning the Flames of Hate: Social Media and Hate Crime. Journal of the European Economic Association. Advance online publication. https://doi.org/10.1093/jeea/jvaa045

Schejter, A. M., & Han, S. (2011). Regulating the media: Four perspectives. In D. Lévi-Faur (Ed.), Handbook on the politics of regulation (pp. 243–279). Cheltenham: Elgar.

Shearing, C., & Wood, J. (2003). Nodal Governance, Democracy, and the New 'Denizens'. Journal of Law and Society, 30(3), 400–419. https://doi.org/10.1111/1467-6478.00263

Sirsch, J. (2013). Die Regulierung von Hassrede in liberalen Demokratien. In J. Meibauer (Ed.), Linguistische Untersuchungen: Vol. 6. Hassrede/Hate Speech: Interdisziplinäre Beiträge zu einer aktuellen Diskussion (pp. 165–193). Gießen: Gießener elektronische Bibliothek.

Sunstein, C. R. (1995). Democracy and the problem of free speech: With a new afterword (1. Free Press paperback ed.). New York, NY: Free Press.

Stier, S. (2017). Internet und Regimetyp: Netzpolitik und politische Online-Kommunikation in Autokratien und Demokratien. Vergleichende Politikwissenschaft. Wiesbaden: Springer Fachmedien Wiesbaden. Retrieved from http://ebookcentral.proquest.com/lib/gbv/detail.action?docID=4800435

Stöber, R. (2007). Kommunikationsfreiheit und ihre Feinde. Zeitschrift für Literaturwissenschaft und Linguistik, 37(2), 104–119. https://doi.org/10.1007/BF03379660

Timofeeva, Y. (2006). Censorship in cyberspace: New regulatory strategies in the digital age on the example of freedom of expression (1. Aufl.). Schriften zur Governance-Forschung: Vol. 6. Baden-Baden: Nomos.

Autor*innen

Thomas Mandl, Institut für Informationswissenschaft und Sprachtechnologie

Ulrich Heid, Institut für Informationswissenschaft und Sprachtechnologie

Sylvia Jaki, Institut für Übersetzungswissenschaft und Fachkommunikation

Wolf J. Schünemann, Institut für Sozialwissenschaften

Johannes Schäfer, Institut für Informationswissenschaft und Sprachtechnologie

Stefan Steiger, Institut für Sozialwissenschaften