Hate Speech online: Hartknäckiges Phänomen und interdisziplinärer Forschungsgegenstand

Sylvia Jaki und Stefan Steiger

1 Hate Speech und ihre Bekämpfung

Während in der Frühphase des Internets die erwarteten positiven Wirkungen digitaler Kommunikation auf demokratische Diskurse betont wurden, hat sich die Perspektive in den vergangenen Jahren deutlich getrübt. Die gezielte Verbreitung von Desinformationen zur Beeinflussung von demokratischen Wahlen oder während der Coronapandemie stehen ebenso exemplarisch hierfür wie die Verbreitung von Hate Speech. Diese Phänomene bedrohen den demokratischen Diskurs und erschweren bspw. die Konsensfindung bei entscheidenden politischen Fragen.

Hate Speech im digitalen Raum stellt eine wachsende gesellschaftliche Herausforderung dar, das Problem ist aber nicht genuin neu. Schon in der Frühphase der Internetentwicklung wurden potenziell unerwünschte Folgen der Internetkommunikation für demokratische Diskurse debattiert (Buchstein, 1996). Die globale Verbreitung digitalisierter Kommunikation und insbesondere die massenhafte Nutzung von sozialen Medien haben die neue Qualität der Herausforderung aber eindrücklich verdeutlicht. In jüngster Vergangenheit gab es beispielsweise gehäuft Fälle von digitaler Hate Speech im Kontext der Coronapandemie (Lee & Li, 2021; Uyheng & Carley, 2020). Die Auseinandersetzung um die Maßnahmen

S. Jaki (✉)
Institut für Übersetzungswissenschaft und Fachkommunikation, Universität Hildesheim, Hildesheim, Deutschland
E-Mail: jakisy@uni-hildesheim.de

S. Steiger
Universitätsrechenzentrum, Universität Heidelberg, Heidelberg, Deutschland
E-Mail: stefan.steiger@urz.uni-heidelberg.de

S. Jaki und S. Steiger (Hrsg.), *Digitale Hate Speech*,
https://doi.org/10.1007/978-3-662-65964-9_1

zur Eindämmung der Coronapandemie stellen dabei aber nur einen der aktuellen Kristallisationspunkte der Hate Speech im Netz dar. Demokratische Gemeinwesen sind durch die immer intensiver geführten Auseinandersetzungen grundlegend herausgefordert, ist der freie Diskurs und der Austausch verschiedener Positionen doch elementar für eine gelingende demokratische Entscheidungsfindung. Aufgrund der negativen Folgen haben demokratische Regierungen schon vor mehreren Jahren mit Initiativen gegen Hate Speech begonnen.

Bereits 2016 vereinbarten die Betreiber sozialer Medien und weitere Internetunternehmen mit der EU einen Code of Conduct zur Bekämpfung von Hate Speech. Hierin erklärten die Unternehmen ihre Bereitschaft, an einer Verbesserung des Onlinediskursklimas mitzuwirken (EU, 2016). Diese auf Freiwilligkeit basierenden Maßnahmen waren aus Sicht verschiedener Regierungen allerdings unzureichend. In Deutschland wird die Debatte um die Folgen digitaler Hate Speech spätestens seit dem Mord am ehemaligen Kasseler Regierungspräsidenten Walter Lübcke 2019 intensiv geführt. Die negativen Folgen für den demokratischen Diskurs allgemein und politische Gewalt im Besonderen haben in Deutschland aber bereits mit dem Netzwerkdurchsetzungsgesetz (NetzDG) 2017 dazu geführt, dass strafbare Inhalte zeitnah von den Betreibern sozialer Medien gelöscht werden müssen (Bundesgesetzblatt, 2017). Das Gesetz zur Bekämpfung des Rechtsextremismus und der Hasskriminalität verpflichtet die Betreiber seit 2021 sogar dazu, besonders schwerwiegende Fälle wie beispielsweise Verstöße gegen § 130 StGB (Volksverhetzung) direkt an das Bundeskriminalamt zu melden. Diese gesetzgeberische Dynamik ist aber auch mit Kritik bedacht worden: So wurde im NetzDG mit Blick auf die potenziell weitgehenden Löschungen durch die Unternehmen ein unverhältnismäßiger Eingriff in den freien Diskurs durch nicht ausreichend legitimierte Wirtschaftssubjekte gesehen. Weiterhin wurde vor einer Fragmentierung des Kommunikationsraums gewarnt, da das NetzDG Betreiber nur dazu verpflichtet, strafbare Inhalte für die deutschen Nutzer*innen zu entfernen. Sofern die Fälle nicht gegen die plattformeigenen Community Standards verstoßen, bleiben die Nachrichten international weiter sichtbar (Eickelmann et al., 2017).

Diese regulatorischen Bemühungen sind stets mit den verschiedenen Herausforderungen bei der Bekämpfung digitaler Hate Speech konfrontiert, und zu einer erfolgreichen Bewältigung müssen unterschiedliche Fachexpertisen verknüpft werden.

Zu diesen Herausforderungen gehört, dass Hate Speech erstens nicht immer leicht zu erkennen ist. Auf sprachlicher Ebene können Verdachtsfälle von Hate Speech zwar mitunter lexikalisch anhand eines spezifischen Vokabulars identifiziert werden. Letztlich beurteilt werden kann Hate Speech dann aber erst mit

dem entsprechenden Kontext. Auch ist Hate Speech nicht immer direkt auf der sprachlichen Oberfläche erkennbar. Mitunter haben sich in Gruppen spezifische Bezeichnungen zur Verschleierung offener Hate Speech etabliert, die zunächst decodiert werden müssen. Auch zwischen Sprachen können signifikante Unterschiede beispielsweise in der Wahrnehmung verschiedener Arten von Hate Speech bestehen. Aus linguistischer Sicht stellen sich daher Fragen wie: Welche sprachlichen Eigenheiten weist Hate Speech auf? Wie manifestiert sich Hate Speech an der sprachlichen „Oberfläche"? Welche subtileren Formen gibt es und wie lassen sie sich identifizieren? Welche Interaktionen bestehen zwischen verschiedenen Kommunikationsebenen? Inwiefern unterscheidet sich Hate Speech zwischen verschiedenen Sprachräumen? Wie wird Hate Speech (auch über unterschiedliche Quellen) wahrgenommen? Dies sind einige Fragen, denen sich der linguistische Part des Bandes annähert.

Zweitens ist eine systematische Bekämpfung von Hate Speech schwierig, weil es allein die Masse digitaler Kommunikation unmöglich macht, die Aufgabe nach Suche und Identifikation von Hate Speech allein Personen zu überlassen. Es bedarf verlässlicher automatisierter Verfahren zur Erkennung und ggf. Löschung von Hate Speech. Die im linguistischen Teil debattierten Fragen und Probleme stellen sich hier mit neuer Dringlichkeit, wenn Manifestationen von Hate Speech automatisiert klassifiziert werden müssen. Fragen, die in diesem Kontext untersucht werden, beziehen sich beispielsweise darauf, welche Verfahren sich für die Durchsuchung großer Datenmengen besonders eignen. Ferner geht es darum, nicht nur die Performanz der Systeme zu erfassen, sondern auch die Erklärbarkeit der Ergebnisse zu berücksichtigen. In diesem Zusammenhang ist auch die Frage nach der Evaluation derartiger Verfahren von zentraler Bedeutung, möchte man abschätzen, inwiefern diese Hilfsmittel verlässlich dazu beitragen können, dem Problem zu begegnen.

Drittens wird eine systematische Bekämpfung von Hate Speech durch die transnationale Kommunikationsumgebung erschwert. Generell sind liberalen Demokratien bei Eingriffen in die freie Rede enge Grenzen gesetzt. Die Bereitschaft, regulatorisch in Debatten einzugreifen, ist dabei aber zudem international verschieden stark ausgeprägt, so dass unterschiedliche Positionen zur Regulation von Hate Speech aufeinandertreffen. Aus politikwissenschaftlicher Perspektive stellen sich daher Fragen wie: Inwiefern sollten und dürfen demokratische Regierungen in die freie Rede eingreifen? Wie und durch welche Akteure kann Hate Speech sinnvoll reguliert werden? Wie können automatisierte Verfahren reguliert werden, wenn sie möglicherweise in Zukunft große Teile der Erkennung und Löschung von Inhalten übernehmen?

Diese kurzen Überlegungen und Fragen zeigen bereits die Vielschichtigkeit, die mit dem Problem Hate Speech verbunden ist. Die vielen Facetten der Thematik legen eine interdisziplinäre Betrachtung des Phänomens nahe, möchte man die Thematik umfassend ausleuchten und analysieren. Dieser Band versammelt daher Expertisen aus den Sprach-, Informations- und Politikwissenschaften.

Der Band ist eines der Resultate aus einem interdisziplinären Forschungsprojekt an der Universität Hildesheim, in dessen Kontext sich Forschende intensiv mit dem Phänomen digitaler Hate Speech auseinandersetzen. Unter dem *Titel Das Phänomen Hate Speech und seine Erkennung durch KI* (HASeKI) befassten sich Informations-, Sprach- und Politikwissenschaftler*innen mit verschiedenen Aspekten digitaler Hate Speech. Gefördert wurde das Projekt im Rahmen der Ausschreibung „Zukunftsdiskurse" des Niedersächsischen Ministerium für Wissenschaft und Kultur. Der vorliegende Sammelband richtet sich an Wissenschaftler*innen, Studierende und Interessierte, die sich mit den verschiedenen Facetten von Hate Speech näher befassen möchten. Er vereint überblicksartige Darstellungen der Forschungsstände verschiedener Disziplinen mit detaillierten Analysen zur Erkennung, Rezeption und Regulation von Hate Speech.

2 Einschätzungen zum Umgang mit Hate Speech

Das Projekt HASeKI beinhaltet neben diesem Sammelband die Durchführung mehrerer Tagungen, und zwar sowohl von Tagungen fachlicher Natur als auch von Veranstaltungen für die Zivilgesellschaft. Die ersten beiden Konferenzen, also die erste Fachtagung und die erste Bürgertagung, wurden auch über die Diskussions- bzw. Fragerunden hinaus interaktiv gestaltet, indem eine Variante einer Delphi-Befragung durchgeführt wurde.

Delphi-Befragungen können als „ein Instrument zur verbesserten Erfassung von Gruppenmeinungen" gesehen werden (Häder, 2014, S. 19). Die Antworten beruhen dabei auf „intuitiv vorliegenden Informationen der Befragten" (Cuhls, 2019, S. 5) zu Fragen in Bereichen, in denen lediglich unsicheres Wissen vorliegt (Niederberger & Renn, 2018, S. 8). In diesem Fall handelt es sich um die Meinungen bzw. Einschätzungen der Tagungsteilnehmer*innen zu Hate Speech und ihrer Regulierung. Ein wichtiges Prinzip von Delphi ist, dass diese Meinungen in mehreren Wiederholungen eingeholt werden, die Antworten anonym erfolgen, sich die Fragen an Expert*innen richten und das Gesamtergebnis für jede Frage wieder an die Expert*innen zurückgespiegelt wird (vgl. Häder, 2014, S. 25). Typischerweise findet eine solche Befragung mittels standardisierter

Fragebögen statt, aber Varianten mit mündlichen Fragen, beispielsweise im Rahmen von Workshops, kommen ebenso zum Einsatz (vgl. Häder, 2014, S. 25). Das Delphi-Verfahren wird insbesondere bei „Fragestellungen zu Forschung, Technologie aber auch Organisation, Personal oder Bildung verwendet“ (Cuhls, 2019, S. 7).

Die Tagung fand als Videokonferenz auf der Plattform BigBlueButton statt und stellte somit weder eine genuin schriftliche noch eine Face-to-face-Kommunikationssituation dar. Stattdessen konnten die Befragten mittels einer in BigBlueButton integrierten Polling-Funktion anonym abstimmen. Ähnlich wie bei schriftlichen Fragebögen hat dies im Delphi-Verfahren den positiven Effekt, dass keine Antworten aus sozialer Erwünschtheit zu erwarten sind – anders als bei den Formen, die sonst außerhalb schriftlicher Fragebögen in mündlichen Kontexten bei Delphi angewendet werden (vgl. Cuhls, 2019, S. 7 f.). Bei den Expert*innen handelte es sich generell um Menschen mit unterschiedlichen Hintergründen und Graden an Expertise, wie für Delphi-Untersuchungen typisch (vgl. Cuhls, 2019, S. 20). Bei der Fachtagung im Februar 2021 waren überwiegend Wissenschaftler*innen zugegen, die sich mit der Beschreibung, Regulierung und automatischen Erkennung von Hate Speech befassen, also einen ausgewiesenen Status als fachliche Expert*innen aufweisen. Jedoch war die Tagung auch für Interessierte über diesen Kreis hinaus offen und das Angebot wurde beispielsweise von einigen Mitgliedern der Universität Hildesheim angenommen, die sich zwar für das Thema interessierten und daher in der Regel bereits Vorkenntnisse besaßen, aber nicht Expert*innen i. e. S. waren. Ebenfalls anders als beim klassischen Delphi-Verfahren (vgl. Niederberger & Renn, 2018, S. 11) wurden in unserem Fall bei der zweiten Befragung im Juni 2021 auch nicht mehr dieselben Konferenzteilnehmer*innen wie in der ersten Befragungsrunde befragt. Dies liegt daran, dass die Wiederholung auf einer zweiten Tagung durchgeführt wurde, bei der nur teilweise dieselben Teilnehmer*innen anwesend waren. Bei der Veranstaltung handelte es sich um eine Bürgertagung, zu der auch interessierte Menschen von außerhalb des universitären Bereichs eingeladen waren und bei der die Vortragenden bzw. Diskutant*innen auf dem Podium ebenfalls einen heterogenen Hintergrund aufwiesen.

Die Fragen beziehen sich allesamt auf den Umgang mit Hate Speech, und zwar durch verschiedene Akteure, wobei besonders die Rolle von Künstlicher Intelligenz im Fokus stand. Aufgrund des unterschiedlichen Publikums wurden beim zweiten Durchgang drei Fragen gestrichen, die ein Vorwissen im Bereich Natural Language Processing nötig machten. Im Folgenden werden nacheinander die Fragen sowie die Antworten in beiden Durchgängen vorgestellt.

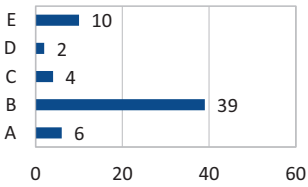


Abb. 1 Frage 1, 1. Durchgang (n = 61)

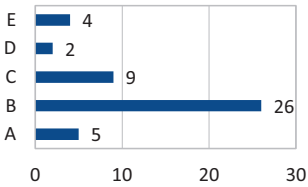


Abb. 2 Frage 1, 2. Durchgang (n = 46)

Die Fragen wurden in drei Blöcke aufgeteilt, die nach den jeweils thematischen Blöcken der Fachtagung geordnet waren und die sich auch in der Gliederung dieses Tagungsbandes wiederfinden (vgl. Abschn. 3). Den Teilnehmer*innen wurde freigestellt, ob sie auf die jeweiligen Fragen antworten. Der erste Block war der allgemeinste und Frage 1 lautete folgendermaßen: „Stimmen Sie der folgenden Aussage zu? Die Plattformen leisten genug, um Hassrede im Netz zu bekämpfen." Die Befragten konnten unter den fünf Antwortmöglichkeiten „A) Stimme gar nicht zu", „B) Stimme eher nicht zu", „C) Stimme eher zu", „D) Stimme vollauf zu" und „E) Weiß nicht" wählen. Antwort A) wurde im ersten Durchgang im Februar 2021 6 (10 %[1]) Mal gewählt und im zweiten Durchgang im Juni 2021 5 (11 %) Mal, Antwort B) jeweils 39 (64 %) und 26 Mal (57 %), Antwort C) jeweils 4 (7 %) und 9 Mal (20 %), Antwort D) jeweils 2 (3 %) und 2 (4 %) Mal und Antwort E) jeweils 10 (16 %) und 4 (9 %) Mal (vgl. Abb. 1 und 2). Wie die Antworten zeigen, waren sich beide Gruppen (Fachpublikum und Teilnehmer*innen der Bürgertagung) weitgehend einig, dass die Plattformen bei der Bekämpfung von Hate Speech nicht

[1] Das System in BigBlueButton gibt lediglich ganze Prozentzahlen an. Wir sind uns im Klaren darüber, dass bei der geringen Teilnehmer*innenzahl Prozentzahlen zum Teil irreführend sein können. Da diese jedoch die Übersichtlichkeit erhöhen und eine bessere Vergleichbarkeit zwischen den beiden Durchgängen gewährleisten, sind sie hier mit angegeben.

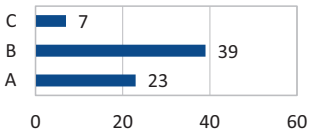


Abb. 3 Frage 2, 1. Durchgang (n = 69)

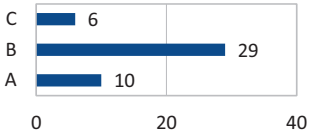


Abb. 4 Frage 2, 2. Durchgang (n = 45)

aktiv genug sind, wenngleich es bei der Bürgertagung anteilig mehr Stimmen gab, die mit der Regulierung von Seiten der Plattformen eher zufrieden sind.

Stärker auf den Bereich KI ging Frage 2 ein: „Sofern verfügbar und technisch ausgereift, sollten KI und automatisierte Detektionsverfahren standardmäßig auf user-generated content etwa in sozialen Netzwerken angewandt werden?“ Hier standen die drei Auswahlmöglichkeiten „A) Ja“, „B) Nur unter engen gesetzlich definierten Bedingungen und bei Verdachtsfällen“ und „C) Gar nicht“ zur Verfügung. Antwort A) erreichte im ersten Durchgang 23 (33 %) und im zweiten Durchgang 10 (22 %) Stimmen, B) jeweils 39 (57 %) und 29 (64 %) Stimmen und C) jeweils 7 (10 %) und 6 (13 %) Stimmen (vgl. Abb. 3 und 4). Auch bei dieser Frage lagen die beiden Gruppen nicht weit auseinander, wobei das Fachpublikum der Verwendung von KI im Bereich der Erkennung von Hate Speech etwas positiver gegenüberstand. Unter den Teilnehmer*innen befanden sich jedoch auch einige Personen, die selbst solche Verfahren entwickeln, was diesen Unterschied vermutlich erklärt.

Bei der dritten Frage war die Einschätzung der Expert*innen in Bezug auf die Entwicklung von Hate Speech gefragt: „Wie wird sich das Problem von Hate Speech in sozialen Netzwerken in zehn Jahren darstellen?“ Die drei Antwortmöglichkeiten waren „A) Sehr viel schlimmer“, „B) Ähnlich wie heute“ und „C) Weitgehend gelöst“. Beantwortet wurde die Frage in der ersten Abstimmung 12 Mal (18 %) und in der zweiten Abstimmung 9 Mal (21 %) mit A), jeweils 54 (79 %) bzw. 26 Mal (60 %) mit B) sowie 2 (3 %) und jeweils 8 Mal (19 %) mit C) (vgl. Abb. 5 und 6). Interessant ist hier, dass das Fachpublikum bei dieser Frage pessimistischer war als das Publikum der zweiten Tagung, da weniger Teilnehmer*innen der Meinung waren, das Problem sei in zehn Jahren weitgehend gelöst. Allerdings rechneten sie auch nicht stärker mit einer Verschlimmerung des Problems.

Abb. 5 Frage 3, 1. Durchgang (n = 68)

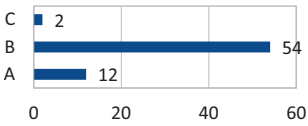
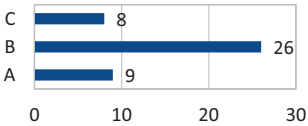


Abb. 6 Frage 3, 2. Durchgang (n = 43)



Der zweite Fragenblock beschäftigte sich ausschließlich mit der automatisierten Erkennung von Hassrede. Frage 4 zielte auf die Qualität von automatischer Hate-Speech-Erkennung ab: „Wie gut ist die Erkennung von Hate Speech durch KI Ihrer Einschätzung nach aktuell?“ Die Befragten konnten sich für „A) Fast perfekte Erkennung“, „B) Mittelmäßige Erkennung“ oder „C) Sehr schlechte Erkennung“ entscheiden. In der ersten Abstimmung antworteten eine Person (2 %) und in der zweiten Abstimmung 3 Personen (7 %) mit A), jeweils 36 (65 %) und 32 Personen (73 %) mit B) und jeweils 18 (33 %) bzw. 9 Personen (20 %) mit C) (vgl. Abb. 7 und 8). Ähnlich wie in der vorhergehenden Frage war das Fachpublikum etwas pessimistischer, wenn es um die Performanz automatischer Detektionsverfahren geht, wenngleich sich auch hier keine starken Diskrepanzen abzeichneten.

Abb. 7 Frage 4, 1. Durchgang (n = 55)

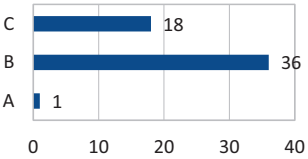


Abb. 8 Frage 4, 2. Durchgang (n = 44)

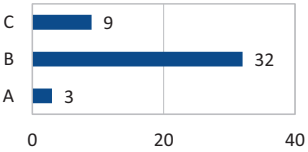
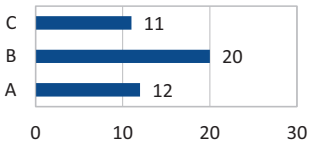


Abb. 9 Frage 5 (n = 43)



Frage 5 fokussierte die Akteur*innen der Hate-Speech-Erkennung und lautete „Wer bietet die besten Verfahren zur Erkennung von Hate Speech durch KI – Industrie/grosse Plattformen oder die Wissenschaft?“ Die drei Antworten waren „A) Wissenschaft deutlich besser“, „B) Etwa gleich gut“ und „C) Plattformen deutlich besser“. Diese Frage wurde lediglich auf der Fachtagung im Februar 2021 erhoben. 12 Personen (28 %) antworteten mit A), 20 (47 %) mit B) und 11 (26 %) mit C) (vgl. Abb. 9). Diese Antwort zeigt, dass die automatischen Verfahren der Industrie/Plattformen und der Wissenschaft als ähnlich leistungsstark eingeschätzt wurden.

Bei der sechsten Frage sollte eine Vorhersage zur Qualität von Hate-Speech-Erkennung in der Zukunft getroffen werden: „Wie wird die Erkennung von Hate Speech in zehn Jahren funktionieren?“ Ähnlich wie bei Frage 4 waren die drei Antwortmöglichkeiten „A) Fast perfekte Erkennung“, „B) Mittelmäßige Erkennung“ oder „C) Sehr schlechte Erkennung“. Diese Frage beantworteten im ersten Durchgang 12 (27 %) und im zweiten Durchgang 18 (43 %) Teilnehmer*innen mit A), 33 (73 %) bzw. 22 (52 %) mit B) und 0 (0 %) bzw. 2 (5 %) mit C) (vgl. Abb. 10 und 11). Das Ergebnis der vierten Frage zur aktuellen Performanz von Detektionssystemen spiegelt sich auch hier wider, denn an eine fast perfekte Erkennung von Hate Speech in zehn Jahren glaubten ebenfalls die Teilnehmer*innen der Bürgertagung stärker als die der Fachtagung.

Abb. 10 Frag 6, 1. Durchgang (n = 45)

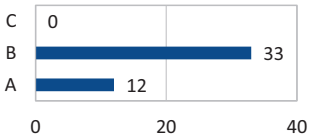
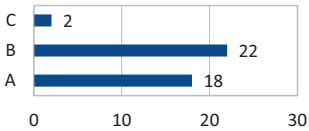


Abb. 11 Frage 6, 2. Durchgang (n = 42)



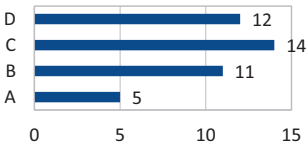


Abb. 12 Frage 7 (n = 42)

Wie Frage 5 wurde Frage 7 nur bei der ersten Erhebung gestellt. Sie lautete „Was wird Ihrer Meinung nach am meisten zu den Fortschritten bei der Erkennung von Hate Speech durch KI beitragen?“ Die Befragten konnten zwischen den vier Antworten „A) Netzwerkanalysen“, „B) Sprachanalyse“, „C) Maschinelles Lernen“ und „D) Größere Mengen an Trainingsdaten“ wählen. Antwort A) wurde 5 (12 %) Mal gewählt, B) 11 (26 %) Mal, C) 14 (33 %) Mal und D) 12 (29 %) Mal (vgl. Abb. 12). Folglich werden verschiedene Verfahren und Aspekte als vielversprechend für die Entwicklung besserer Systeme eingeschätzt, am wenigsten allerdings Netzwerkanalysen und am meisten Maschinelles Lernen.

Im letzten Block ging es vor allem um die Regulierung von Inhalten in den sozialen Medien. Bei Frage 8 sollten die Tagungsteilnehmer*innen die Regulierung der sozialen Medien beurteilen: „Wie bewerten Sie die Regulierung von Hate Speech in sozialen Netzwerken heute?“ Die Antwortmöglichkeiten waren „A) Zu starke Regulierung“, „B) Angemessen“ und „C) Deutlich zu wenig Regulierung“. In der ersten Erhebung antworteten 3 (7 %) und in der zweiten 4 Personen (9 %) mit A), 9 (20 %) bzw. 8 Personen (19 %) mit B) und 33 (73 %) bzw. 31 Personen (72 %) mit C) (vgl. Abb. 13 und 14). Bei dieser Frage fielen die Antworten bei den beiden Gruppen am ähnlichsten aus und beide bewerteten die aktuelle Regulierung von Hate Speech als zu gering.

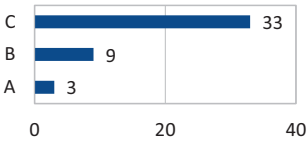


Abb. 13 Frage 8, 1. Durchgang (n = 45)

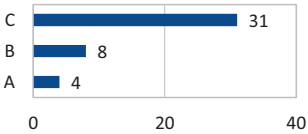


Abb. 14 Frage 8, 2. Durchgang (n = 43)

Abb. 15 Frage 9, 1. Durchgang (n = 47)

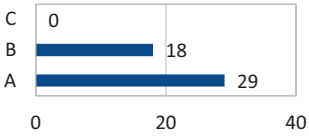
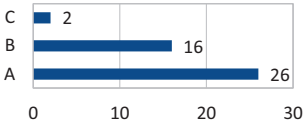


Abb. 16 Frage 9, 2. Durchgang (n = 44)



Auch Frage 9 zielte auf die Regulierungsperspektive ab, aber aus Zukunftsperspektive: „Wie wird sich die Regulierung von Hate Speech in sozialen Netzwerken in zehn Jahren darstellen?“ Die Befragten konnten sich zwischen „A) Viel stärkere Regulierung“, „B) Ähnlich wie heute“ oder „C) Sehr viel weniger Regulierung“ entscheiden. In der ersten Befragung reagierten 29 (62 %) und in der zweiten Befragung 26 (59 %) Personen mit A), 18 (38 %) bzw. 16 (36 %) Personen mit B und 0 (0 %) bzw. 2 (5 %) mit C) (vgl. Abb. 15 und 16). Wie in der vorherigen Frage waren die Antworten in beiden Runden ähnlich und die Befragten rechneten mehrheitlich mit einer stärkeren Regulierung.

Die zehnte und letzte Frage, die ebenfalls nur während der ersten Tagung erhoben wurde, lautete folgendermaßen: „Stimmen Sie der folgenden Aussage zu? Die Politik leistet genug, um Hassrede im Netz zu bekämpfen.“ Hier waren dieselben fünf Antwortmöglichkeiten wie bei Frage 1 gegeben, nämlich „A) Stimme gar nicht zu“, „B) Stimme eher nicht zu“, „C) Stimme eher zu“, „D) Stimme vollauf zu“ und „E) Weiß nicht“. 8 Personen (15 %) reagierten mit A), 24 (46 %) mit B), 9 (17 %) mit C), 3 (6 %) mit D) und 8 (15 %) mit E) (vgl. Abb. 17). Auch wenn hier die Meinungen relativ stark verteilt waren, wird dennoch deutlich, dass die größte Gruppe der Befragten der Meinung ist, dass die Politik eher nicht genug gegen Hate Speech unternimmt.

Zusammengefasst illustriert diese Kurzbefragung daher insgesamt von Seiten des Publikums einen Wunsch zu mehr Regulierung. Besonders die Befragten mit Vorkenntnissen in NLP standen dem Einsatz von KI zur Regulierung relativ offen gegenüber, waren aber, vermutlich aufgrund ihrer Erfahrung mit den Grenzen automatischer Erkennungssysteme, häufig etwas skeptischer, was die aktuelle und zukünftige Performanz solcher Systeme betrifft.

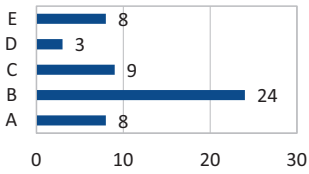


Abb. 17 Frage 10 (n = 52)

3 Die Beiträge

In Analogie zum Aufbau der Fachtagung, aus der dieser Band hervorgegangen ist, ist dieser Sammelband in drei größere Sektionen gegliedert, die die (rein beschreibende) linguistische Perspektive, die computerlinguistische und die politikwissenschaftliche widerspiegeln.

Die ersten beiden Beiträge befassen sich mit der Beschreibung beziehungsweise Rezeption von Hate Speech. Diese Sektion demonstriert nicht nur das Interesse der Linguistik an der Aufdeckung sprachlicher Muster, sondern kann auch als bedeutsame Grundlage gelten, auf der Gegenmaßnahmen für verschiedene kommunikative Äußerungen, die Hassrede enthalten, diskutiert werden können. In dem Beitrag von **Sylvia Jaki** geht es um einen Überblick über die linguistische Forschung zu Hate Speech, wobei verschiedene sprachliche Kontexte, Medien, Diskurse, Methoden und sprachliche Charakteristika berücksichtigt werden, und auch ein Blick auf die Multimodalität von Hate Speech geworfen wird. Anders als der erste Artikel beschäftigt sich der Beitrag von **Oliver Niebuhr und Jana Neitsch** mit der Rezeptionsperspektive von Hate Speech, die bislang noch als weniger gut erforscht gilt. Er präsentiert eine Übersicht zur Forschung der Autor*innen zu gesprochener versus geschriebener Hate Speech, die auch den sozialen Kontext der Rezeptionssituation mit in die Analyse einbezieht.

Die zweite Sektion ist der computerlinguistischen Perspektive gewidmet, wobei Verfahren zur automatischen Erkennung vorgestellt und diskutiert werden. **Christoph Demus, Dirk Labudde, Jonas Pitz, Nadine Probol, Mina Schütz und Melanie Siegel** stellen das Thema anhand von Prozessen und Vorgehensweisen sogenannter Shared Tasks vor, indem sie nicht nur einen Überblick über bereits durchgeführte Shared Tasks und deren Methoden bzw. Ergebnisse bieten, sondern auch in die mit dem Aufbau einer Datensammlung für einen Shared Task verbundenen Schritte und Schwierigkeiten einführen. **Johannes Schäfer** greift Problembereiche computerlinguistischer Forschung zu Hate Speech auf, unter

anderem die mangelnde Berücksichtigung des sprachlichen Kontextes. Der Beitrag stellt verschiedene Lösungsansätze für die automatische Erkennung von Hate Speech vor und unterscheidet dabei lexikonbasierte Systeme, erklärbare maschinelle Lernsysteme und neuronale Netzwerke. Die Auswahl der Trainingsdaten entscheidet maßgeblich über die Leistungsfähigkeit automatisierter Hate-Speech-Erkennung. **Thomas Mandl** untersucht daher die Bedeutung von Trainingsdaten auch mit Blick auf die Erklär- und Nachvollziehbarkeit der Ergebnisse automatisierter Verfahren. **Roman Klinger** stellt anhand verschiedener Textgattungen die Möglichkeiten der automatisierten Emotionsanalyse vor. Dabei plädiert er für eine Untersuchung mit Hilfe der Appraisaltheorien, um den Zusammenhang zwischen kognitiven Prozessen und Emotionen zu erklären.

Der dritte Teil dieses Bandes ist der Regulation von Hassrede im Internet gewidmet und fokussiert damit die politikwissenschaftliche Sicht. **Wolf Schünemann und Stefan Steiger** bieten einen Überblick über die bestehende politikwissenschaftliche Forschungslandschaft und gehen dabei der Frage nach, ob sich im Kontext der Regulation von Hate Speech ein Paradigmenwechsel auf Seiten der liberalen Demokratien abzeichnet, die lange Zeit vor der Regulation von Inhalten im Netz zurückschreckten, nun aber immer aktiver gegen die Verbreitung von Hate Speech vorgehen. Als logische Konsequenz dieser Überlegungen schließt sich der Beitrag von **Doris Unger und Jürgen Unger-Sirsch** an, in dem die Autor*innen die Ziele von Hate-Speech-Regulierung darlegen und auf dieser Basis konkrete Richtlinien herausarbeiten. Anhand dieser Richtlinien wird wiederum ein konkreter Fall, der Umgang mit gruppenspezifischen Schimpfwörtern, reflektiert. Deep Fakes, die sich im Spannungsfeld zwischen Desinformation und Hate Speech befinden, werden einschließlich ihres Gefahrenpotenzials im abschließenden Artikel von **Murat Karaboga** diskutiert. Der Beitrag behandelt mit dem Digital Services Act und dem KI-Regulierungsvorschlag der EU-Kommission insbesondere die Bemühungen zur Regulierung auf europäischer Ebene.

Danksagungen An dieser Stelle möchten wir, auch im Namen von Thomas Mandl, Ulrich Heid, Wolf Schünemann und Johannes Schäfer, dem Niedersächsischem Ministerium für Wissenschaft und Kultur herzlich danken, das das Projekt HASeKI und damit auch diesen Tagungsband im Rahmen der Förderlinie Zukunftsdiskurse gefördert hat. Überdies bedanken wir uns bei Daphné Cetta, die die Durchführung der beiden Tagungen durch ihr organisatorisches Talent erst möglich gemacht hat, und bei den studentischen Hilfskräften, die uns im Rahmen des Projekts tatkräftig unterstützt haben.

Literatur

Buchstein, H. (1996). Bittere Bytes: Cyberbürger und Demokratietheorie. *Deutsche Zeitschrift Für Philosophie, 44*(4), 583–608. https://doi.org/10.1524/dzph.1996.44.4.583.

Bundesgesetzblatt. (2017). Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (Netzwerkdurchsetzungsgesetz–NetzDG). https://www.bgbl.de/xaver/bgbl/start.xav?startbk=Bundesanzeiger_BGBl&jumpTo=bgbl117s3352.pdf. Zugegriffen: 28. Aug. 2021.

Cuhls, K. (2019). Die Delphi-Methode – eine Einführung. In M. Niederberger & O. Renn (Hrsg.), *Delphi-Verfahren in den Sozial- und Gesundheitswissenschaften: Konzept, Varianten und Anwendungsbeispiele* (S. 3–31). Springer VS.

Eickelmann, J., Grashöfer, K., & Westermann, B. (2017). #NETZDG #MAASLOS. *Zeitschrift Für Medienwissenschaft, 9*(17–2), 176–185. https://doi.org/10.14361/zfmw-2017-0218.

EU. (2016). Code of conduct on countering illegal hate speech online. https://ec.europa.eu/newsroom/just/document.cfm?doc_id=42985.

Häder, M. (2014). *Delphi-Befragungen: Ein Arbeitsbuch* (3. Aufl.). Springer VS.

Lee, R.K.-W., & Li, Z. (2021). Online Xenophobic behavior amid the COVID-19 pandemic. *Digital government: research and practice, 2*(1), 1–5. https://doi.org/10.1145/3428091

Niederberger, M., & Renn, O. (2018). *Das Gruppendelphi-Verfahren: Vom Konzept bis zur Anwendung*. Springer VS.

Uyheng, J., & Carley, K. M. (2020). Bots and online hate during the COVID-19 pandemic: Case studies in the United States and the Philippines. *Journal of Computational Social Science, 3*, 445–468. https://doi.org/10.1007/s42001-020-00087-4.