Tagungsbericht

Thomas Mandl*, Sylvia Jaki, Daphné Çetta, Ulrich Heid, Wolf J. Schünemann, Stefan Steiger und Johannes Schäfer

Interdisziplinäre Perspektiven auf Hate Speech und ihre Erkennung (IPHSE)

Bericht zur Online-Tagung am 8. Februar 2021

https://doi.org/10.1515/iwp-2021-2165

Über das Projekt

Hate Speech stellt in Online-Medien ein erhebliches Problem dar und wird zunehmend als gesellschaftliche Bedrohung wahrgenommen. Hasserfüllte Botschaften erschweren einen sachlichen öffentlichen Diskurs und gefährden so die demokratische Meinungsbildung. Das kürzlich ergangene Urteil im Fall der Politikerin Künast zeigt die Schwierigkeiten bei der Abgrenzung, der rechtlichen Bewertung und der Abwägung gegenüber der Meinungsfreiheit; es fördert aber auch die öffentliche Meinungsbildung zu der Thematik. Angesichts unablässiger Ströme nutzergenerierter Inhalte haben Regierungen und private Intermediäre ein gesteigertes Interesse an skalierbaren automatisierten Werkzeugen zur Content-Regulierung, die auch KI-basierte Methoden beinhalten. Längst entwickeln große Plattformen Verfahren der Künstlichen Intelligenz, um problematische Inhalte zu erkennen. Das im Oktober 2019 ergangene Urteil des EUGH sieht den Einsatz von automatisierten Verfahren als notwendig an.

Um die Forschung im Bereich der Künstlichen Intelligenz sinnvoll einordnen zu können, ist ein interdisziplinärer Dialog notwendig. An der Universität Hildesheim kooperieren Forschende aus Informationswissenschaft, Computerlinguistik, Sprachwissenschaft und Politikwissenschaft in dem vom Niedersächsischen Wissenschaftsministerium geförderten Projekt namens: „Das Phänomen Hate Speech und seine Erkennung durch KI: interdisziplinär – international – erklärbar? (HASeKI)“. In einer ersten Projekttagung „Interdisziplinäre Perspektiven auf Hate Speech und ihre Erkennung (IPHSE)“ am 8. Februar 2021 wurde dabei in 12 Fachvorträgen der aktuelle Stand der Forschung präsentiert. Die Tagung war ein großer Erfolg und zweitweise waren über 100 TeilnehmerInnen im virtuellen Konferenz-Raum.

Linguistik und Hate Speech

Im Eröffnungsvortrag bot Sylvia Jaki von der Universität Hildesheim einen Forschungsüberblick zu sprachwissenschaftlichen Untersuchungen von Hate Speech und verwandten Phänomenen, die eine große Bandbreite an Diskursen, sprachlichen Charakteristika und Zielgruppen behandeln. Im Fokus stehen dabei häufig lexikalische Aspekte. Jaki erläuterte diese Art der Charakteristika exemplarisch anhand von sprachlicher Kreativität und Entmenschlichungsmetaphern in zwei eigenen Fallstudien zu rechtsextremer Hate Speech auf Twitter und zum frauenfeindlichen Forum incels.me. Sie zeigte außerdem ein Desideratum der sprachwissenschaftlichen Forschung

*Kontaktperson: Prof. Dr. Thomas Mandl, Universität Hildesheim, Institut für Informationswissenschaft und Sprachtechnologie, Universitätsplatz 1, 31141 Hildesheim, E-Mail: mandl@uni-hildesheim.de
Dr. Sylvia Jaki, Universität Hildesheim, Institut für Übersetzungswissenschaft und Fachkommunikation, Universitätsplatz 1, 31141 Hildesheim, E-Mail: jakisy@uni-hildesheim.de
Daphné Çetta, Universität Hildesheim, Institut für Informationswissenschaft und Sprachtechnologie, Universitätsplatz 1, 31141 Hildesheim, E-Mail: daphne.cetta@uni-hildesheim.de
Prof. Dr. Ulrich Heid, Universität Hildesheim, Institut für Informationswissenschaft und Sprachtechnologie, Universitätsplatz 1, 31141 Hildesheim, E-Mail: ulrich.heid@uni-hildesheim.de
Prof. Dr. Wolf J. Schünemann, Universität Hildesheim, Institut für Sozialwissenschaften, Universitätsplatz 1, 31141 Hildesheim, E-Mail: wolf.schuenemann@uni-hildesheim.de
Stefan Steiger, Universität Heidelberg, Universitätsrechenzentrum, Im Neuenheimer Feld 330, 69120 Heidelberg, E-Mail: stefan.steiger@urz.uni-heidelberg.de
MS. Johannes Schäfer, Universität Hildesheim, Institut für Informationswissenschaft und Sprachtechnologie, Universitätsplatz 1, 31141 Hildesheim, E-Mail: johannes.schaefer@uni-hildesheim.de

auf, und zwar eine intensivere Befassung mit der Multimodalität von Hate Speech in sozialen Medien.

Konstanze Marx von der Universität Greifswald setzte die linguistische Sektion fort und behandelte Wendungen, die ursprünglich neutral oder positiv besetzt waren, aber sukzessive, insbesondere durch eine ironische Verwendung, eine negative Konnotation erhalten haben und nun vermehrt mit Hate Speech in Verbindung gebracht werden. Ein Beispiel ist Beatrix von Storchs Aussage „*Wir schaffen das*" auf Twitter als Reaktion auf eine Amokfahrt in Münster im Jahr 2018. Das rekontextualisierte Merkel-Zitat impliziert hier ein Versagen der deutschen Flüchtlingspolitik und stellt einen indirekten Aufruf dar, sich weiter diskreditierend zu äußern. Marx bezeichnete dies als „semantisches Trittbrett". Eine Interpretation erfolgt durch eine Kombination aus Bottom-up- und Top-down-Prozessen.

Klaus Geyer von der Syddansk Universitet unterwarf zunächst die Begriffe *Hate Speech* und *Xenophobie* einer kritischen Prüfung und plädierte für eine gründliche Terminologiearbeit im Bereich Hate Speech, um zu einer genauen Bezeichnung dessen zu gelangen, was man jeweils untersucht. Er stellte das von der Velux-Stiftung geförderte Projekt XPEROHS (*Towards Balance and Boundaries in Public Discourse: Expressing and Receiving Online Hate Speec*h, 2018-2021) vor, das von der süddänischen Universität in Odense geleitet wird und in dem Hate Speech auf der Basis von Facebook- und Twitterdaten untersucht wird.

Erkennung von Hate Speech durch KI

In den folgenden Vorträgen rückte die automatische Erkennung ins Zentrum. Johannes Schäfer und Ulrich Heid von der Universität Hildesheim gingen auf Diagnosewerkzeuge aus der Computerlinguistik ein. Sie klärten in einem ausführlichen Überblick, auf welchen sprachlichen Ebenen und mit welchen Ressourcen die unterschiedlichen Verfahren ansetzen.

Roman Klinger von der Universität Stuttgart erweiterte den Blick auf die Erkennung von Emotionen in Texten. Er stellte mehrere Korpora für diese Aufgabe vor, erläuterte deren Annotation und kam dann auf die Technologien zu sprechen. Vor allem neuronale Netze erweisen sich als erfolgversprechend, um Merkmale aus Emotionskomponenten und Beurteilungen zu lernen. Klinger verwies darauf, dass die Erkennungsqualität bei den unterschiedlichen Emotionen oft abweicht. Besonders gut erkannt wird das Gefühl „Ekel".

Philipp Dufter von der Ludwig-Maximilians-Universität München warf dann einen Blick in die momentan modernste Architektur für Textklassifikation, das BERT-Modell. BERT wird mit sehr großen Sprachkorpora vortrainiert und lernt dabei, ausgehend von einem Satz den folgenden Satz im Text vorherzusagen. Dufter zeigte, welche Details dieses ursprünglich von Google entwickelten Systems auch zu unerwünschtem Verhalten führen können. Am Beispiel der Position von Wörtern im Satz konnte Dufter vorführen, dass teilweise schon die Änderung der Reihenfolge von Wörtern das maschinelle Lernen täuschen kann und wie man sie in dieser Hinsicht robuster gestalten kann.

Evaluierung von KI-Systemen zur Erkennung von Hate Speech

Thomas Mandl von der Universität Hildesheim leitete über zu den Problemen der Evaluierung von KI-Verfahren. Er ging insbesondere auf die vielfältigen Gefahren des Bias in Trainingsmengen ein. Die Zusammenstellung von zahlreichen Beispielen für das maschinelle Lernen bleibt eine soziale Aktivität, die den Erfolg maßgeblich beeinflusst.

Melanie Siegel von der Hochschule Darmstadt knüpfte daran an und präsentierte Einsichten aus dem ersten Benchmark für das Deutsche, der GermEval Initiative. Siegel zeigte, welche Anstrengungen im Einzelnen unternommen werden, um sicherzustellen, dass die Systeme wirklich Hassrede erkennen und nicht z. B. lernen die Autorinnen und Autoren einfach anhand von Stilmerkmalen zu unterscheiden.

Jana Neitsch von der Universität Stuttgart und Oliver Niebuhr von der Syddansk Universitet berichteten von Experimenten mit Probandinnen und Probanden, in denen die Wahrnehmung von und die physiologischen Reaktionen auf Hassrede untersucht wurden. In dieser innovativen Arbeit wurden heterogene Formen des Ausdrucks von Hass wie Ironie oder rhetorische Fragen vorgelesen. Abhängig vom Ziel konnten Neitsch und Niebuhr sehr unterschiedliche Wirkungen bei den Teilnehmenden messen.

Mehrfach wurden auch die Tagungsteilnehmenden eingebunden und nach ihrer Ansicht gefragt. Erste Einschätzungen einer Kurzumfrage ergaben, dass eine deutliche Mehrheit die bisher von den Plattformen ergriffenen Maßnahmen zur Bekämpfung von Hassrede als unzureichend empfindet (73 %). Ein Fünftel der Teilnehmenden hielten sie für angemessen und nur sieben Prozent empfanden die Eingriffe als zu weitreichend. Nach Einschätzung der Mehrheit der Teilnehmenden (62 %) wird in zehn Jahren Hassrede viel stärker reguliert werden. 38 Prozent gehen auch in zehn Jahren von einem ähnlichen Regulationsniveau aus. Niemand rechnete mit einem Rückbau von Regulationsmaßnahmen. Schließlich empfanden 61 Prozent auch die Maßnahmen der Regierungen als eher unzureichend.

Politikwissenschaftliche Perspektive auf Hate Speech

Wolf Schünemann und Stefan Steiger von der Universität Hildesheim präsentierten in ihrem Vortrag einen disziplinenübergreifenden Überblick über die Forschung zur Regulierung von Internetinhalten. Sie gingen dabei von der übergeordneten Fragestellung aus, ob bzw. inwiefern sich ein Paradigmenwechsel im grundlegenden Regulierungsansatz liberaler Demokratien zeigt. Sie argumentierten, dass die Zurückhaltung mit Blick auf staatliche Maßnahmen zur Inhaltsregulierung in den letzten Jahren substanziell nachgelassen hat. Insbesondere in Europa wurden und werden verstärkt Maßnahmen zur Bekämpfung von Hassrede erlassen.

Murat Karaboga vom Fraunhofer-Institut für System- und Innovationsforschung (ISI) richtete seinen Blick basierend auf einem grundlegenden theoretischen Schema von Regulationsverhältnissen auf eine neue Relation zwischen Gesetzgebern und Intermediären. Er erkannte deren Rolle im Hinblick auf das Phänomen Hate Speech und potentielle Schädigungen in der Zirkulationsdimension und der Technologiedimension. Mit Blick auf die Zirkulationsdimension stellte Karaboga das Risiko einer Fragmentierung von Kommunikationsräumen durch nationalstaatliche Rechtsetzung heraus und diskutierte kritisch auch den aktuell im Rechtsetzungsprozess befindlichen Digital Services Act (DAS) der EU, als Ansatz zur Plattformregulierung. Mit Blick auf die Technologiedimension stellte Karaboga das KI-Weißbuch vor, in dem die Transparenz von Algorithmen aus seiner Sicht aber noch ein randständiges Thema darstellt.

Jürgen Sirsch von der Universität Bamberg und Doris Unger von der Johannes Gutenberg-Universität Mainz gingen der Frage nach, ob und unter welchen Bedingungen die Regulierung von Hassrede in liberalen Demokratien rechtfertigbar ist. Dazu stützten sie ihre Argumentation auf liberale Ansätze der Gerechtigkeitstheorie. Die Anwendung erfolgte ausgehend von einer engeren Definition von Hassrede in Bezug auf Hass gegen Gruppen und Minderheiten und dem potentiellen Schaden und Rechtsverletzungen gegen diese Gesellschaftsgruppen. Während bei Aufrufen zu Gewalt klar ein höheres Rechtsgut gefährdet ist und somit restriktive Maßnahmen gerechtfertigt werden können, ist die Einschränkung der Meinungsfreiheit im Fall von anderen Gefährdungen, etwa mit Blick auf die soziale Stellung, als höherer Schaden zu gewichten, so dass die Rechtfertigung von Einschränkungen schwerer fällt. Um höheres Anforderungen an die Rechtfertigung zu erfüllen, müssen Regulierung und ihre Legitimierung sehr kontextsensibel sein.

What's next?

Im Herbst 2021 wird eine zweite Fachtagung stattfinden, die dann international ausgerichtet sein wird. Zudem bietet das vom Ministerium für Wissenschaft und Kultur im Rahmen der Ausschreibung „Zukunftsdiskurse" geförderte Projekt HASeKI auch zwei Tagungen, bei denen der Forschungsstand außerhalb des Fachpublikums mit der Zivilgesellschaft diskutiert werden soll. Weitere Ergebnisse und Termine werden online bekanntgegeben: https://www.uni-hildesheim.de/fb3/institute/iwist/forschung/forschungsprojekte/aktuelle-forschungsprojekte/haseki

Deskriptoren: Tagung, Interdisziplinär, Informationswissenschaft, Linguistik, Computerlinguistik, Politikwissenschaft, Social Media, Hate Speech

Prof. Dr. Thomas Mandl
Universität Hildesheim
Institut für Informationswissenschaft und Sprachtechnologie
Universitätsplatz 1, 31141 Hildesheim
mandl@uni-hildesheim.de

Dr. Sylvia Jaki
Institut für Übersetzungswissenschaft und Fachkommunikation
jakisy@uni-hildesheim.de

Daphné Çetta
Institut für Informationswissenschaft und Sprachtechnologie
daphne.cetta@uni-hildesheim.de

Prof. Dr. Ulrich Heid
Institut für Informationswissenschaft und Sprachtechnologie
ulrich.heid@uni-hildesheim.de

Prof. Dr. Wolf J. Schünemann
Institut für Sozialwissenschaften
wolf.schuenemann@uni-hildesheim.de

Stefan Steiger
Universität Heidelberg
Im Neuenheimer Feld 330
69120 Heidelberg
stefan.steiger@urz.uni-heidelberg.de

MS. Johannes Schäfer
Institut für Informationswissenschaft und Sprachtechnologie
johannes.schaefer@uni-hildesheim.de