RUHR-UNIVERSITÄT BOCHUM

RUB

RUBIN

WISSENSCHAFTSMAGAZIN

Schwerpunkt

VERBRECHEN

FORENSIK:
Wie Maden einen Mord aufdecken

PARTNERINNENTÖTUNG:
Warum die Strafen so milde sind

TRAUMA:
Wenn die Bilder immer wiederkommen

#32
Jahrgang

Nr. 1 | 2022

Linguistik

DIE ZERSTÖRERISCHE KRAFT DER SPRACHE

Künstliche Intelligenz kann Schimpfwörter gut identifizieren. Aber kann sie auch versteckte Formen sprachlicher Gewalt erkennen?

„Verpiss dich, du Schlampe!“ „Dich Penner werde ich bekommen. Ich stech’ dich ab.“ „Die sollte man alle abknallen.“ Nur ein paar Beispiele für die Form, die Sprache in den Sozialen Medien annehmen kann. Menschen werden beleidigt, bedroht oder zu Straftaten angestachelt. Was Hassrede und andere Formen von schädigender Sprache aus linguistischer Perspektive auszeichnet und wie man sie automatisch erkennen kann, interessiert Prof. Dr. Tatjana Scheffler. Sie forscht an der Ruhr-Universität Bochum im Bereich Digitale Forensische Linguistik.

„Die Sprachverarbeitung allgemein hat in den vergangenen Jahren große Sprünge gemacht“, sagt Scheffler. Wer heute Übersetzungsprogramme wie den Google Translator oder Sprachassistenten wie Siri nutzt, erzielt deutlich bessere Ergebnisse als vor ein paar Jahren. Auch das Klassifizieren von Texten klappt mittlerweile ganz gut. Künstliche-Intelligenz-Algorithmen können lernen, Aussagen verschiedenen Kategorien zuzuordnen. So können sie etwa entscheiden, ob eine Textpassage eine direkte Beleidigung enthält oder nicht. Die Algorithmen lernen die Kategorien anhand von großen Trainingsdatensätzen, die Menschen zuvor klassifiziert haben. Später können sie das Wissen über die erlernten Kategorien dann auf neue Daten übertragen.

„Direkte Beleidigungen und Schimpfwörter sind so schon gut zu identifizieren“, weiß Tatjana Scheffler. Oft reicht ein Abgleich mit einer Wortliste, die häufig verwendete Beleidigungen enthält. Unter schädigender Sprache versteht die Forscherin aber viel mehr als offensichtliche Hassrede, die sich gegen einzelne Personen richtet. „Es gibt implizitere Formen, die gar nicht an einen bestimmten Adressaten oder eine Adressatin gerichtet sind“, sagt sie. „Man kann auch Schaden anrichten, indem man auf eine gewisse Weise über andere spricht oder eine bestimmte Stimmung herstellt.“

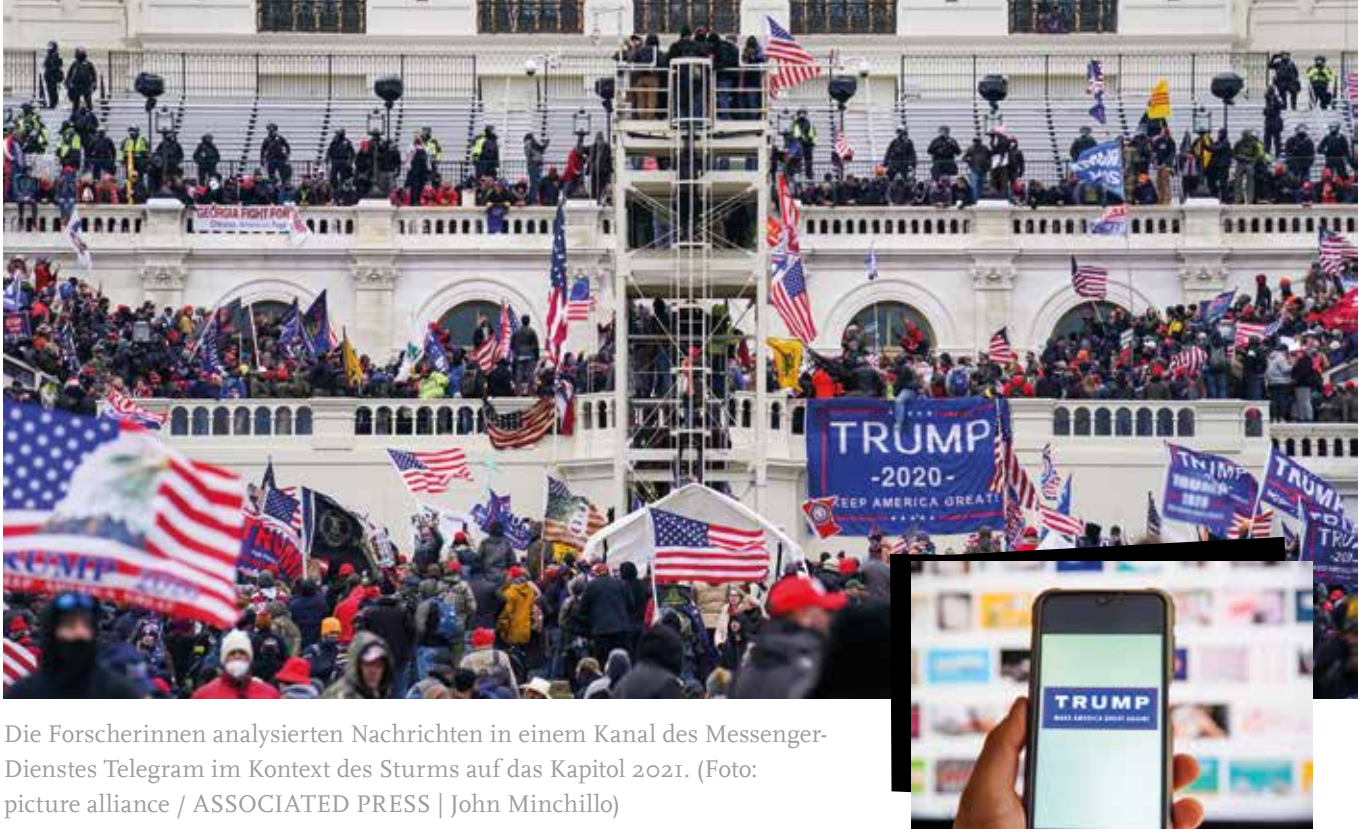
Schlimmstenfalls können solche Stimmungen in echte Handlungen umschlagen. Ein populäres Beispiel ist der Sturm auf das Kapitol durch Anhänger des damaligen US-Präsidenten Donald Trump am 6. Januar 2021. Die sozialen Medien werden mit dafür verantwortlich gemacht, dass die Lage so eskalieren konnte.

Genau mit diesem Beispiel, dem Sturm auf das Kapitol, hat Tatjana Scheffler sich zusammen mit zwei Kolleginnen aus Berlin befasst. Die Gruppe arbeitete mit 26.431 Nachrichten von 521 Nutzerinnen und Nutzern des Messenger-Dienstes Telegram. Die Nachrichten wurden zwischen dem 11. Dezember 2016 und dem 18. Januar 2021 in einem öffentlichen Kanal gepostet, in dem sich Menschen mit extrem rechter Gesinnung austauschen. Inhaltlich startete ihre Diskussion mit der theoretischen Idee, die Regierung zu stürzen, und entwickelte sich schrittweise zu den konkreten Plänen, das Kapitol zu stürmen.

Das Team um Tatjana Scheffler überprüfte, wie gut bereits existierende Algorithmen in diesem Datensatz schädigende Sprache identifizieren konnten. Um die Trefferquote der Algorithmen bewerten zu können, analysierten die Forscherinnen etwa ein Fünftel der Nachrichten von Hand und ▸

Tatjana Scheffler ist Expertin für Digitale Forensische Linguistik. (Foto: km)

Foto: rs



Die Forscherinnen analysierten Nachrichten in einem Kanal des Messenger-Dienstes Telegram im Kontext des Sturms auf das Kapitol 2021. (Foto: picture alliance / ASSOCIATED PRESS | John Minchillo)

Der ehemalige US-Präsident Donald Trump wird beschuldigt, durch seine aufwiegelnde Sprache den Sturm auf das Kapitol mitverantwortet zu haben. (Foto: rs)

verglichen ihre Ergebnisse mit denen der automatisierten Verfahren. Sie unterschieden dabei fünf verschiedene Formen von schädigender Sprache.

Die erste Kategorie umfasste aufwiegelnde Sprache, etwa Passagen wie „violence is 100000% justified now“ (Gewalt ist jetzt zu 10.0000 % gerechtfertigt). Die zweite Kategorie beinhaltete abwertende Begriffe wie „scum“ (Abschaum) oder „retarded“ (zurückgeblieben). In der dritten Kategorie fasste das Team Ausdrücke zusammen, die an sich nicht abwertend sind, aber in dem Kontext, in dem sie auftraten, abfällig gemeint waren – etwa „they are a sickness“ (sie sind Krankheiten). Eine vierte Kategorie war dem Othering gewidmet: Bemerkungen, die genutzt werden, um eine Gruppe Menschen von einer anderen abzugrenzen, wie in dem Beispiel: „Are women banned from this chat? If not, why the fuck not?“ (Sind Frauen aus diesem Chat ausgeschlossen? Falls nicht, warum verdammt noch mal nicht?). Die letzte Kategorie umfasste Insiderformulierungen, die eine Gruppe von Gleichgesinnten verwendet, um sich von anderen abzugrenzen und das Gruppengefühl zu stärken. Trump-Anhänger nutzen den Begriff „Patriot“ etwa auf eine bestimmte Art.

Die auf diese Weise kodierten Kommentare ließen die Forscherinnen auch von automatisierten Verfahren labeln, wie Tech-Firmen sie nutzen, um Hassrede oder beleidigende Sprache ausfindig zu machen. 4.505 Nachrichten gingen in den Vergleich ein. 3.395 davon stuften sowohl die Wissenschaftlerinnen als auch die automatisierten Verfahren als nicht schädigend ein, bei 275 waren sie sich einig, dass sie schädigende Sprache enthielten. 835 Nachrichten hingegen bewerteten Mensch und Maschine unterschiedlich: Etwa die Hälfte stuften die Algorithmen fälschlicherweise als Hassrede oder Beleidigung ein; den Rest erkannten sie – anders als die Wissenschaftlerinnen – nicht als schädigende Sprache.

Gerade bei aufwiegelnden Kommentaren, Insiderbegriffen und Othering lagen die automatisierten Verfahren oft daneben. „Wenn wir sehen, in welchen Fällen etablierte Methoden Fehler machen, hilft uns das, künftige Algorithmen besser zu machen“, resümiert Tatjana Scheffler. Mit ihrem Team entwickelt sie auch selbst automatisierte Verfahren, die schädigende Sprache noch besser erkennen sollen. Dazu braucht es zum einen bessere Trainingsdaten für die Künstliche Intelligenz. Zum anderen müssen auch die Algorithmen selbst optimiert werden. Hier kommt wieder die Linguistik ins Spiel: „Bestimmte grammatische Strukturen können zum Beispiel ein Hinweis darauf sein, dass ein Begriff abwertend gemeint ist“, erklärt Scheffler. „Wenn ich sage ‚Du Lauch‘ ist das etwas anderes als wenn ich nur ‚Lauch‘ sage.“

Nach solchen sprachlichen Merkmalen sucht Tatjana Scheffler, um die Algorithmen der nächsten Generation mit weiterem Hintergrundwissen zu füttern. Auch Kontextinformationen könnten den Maschinen helfen, schädigende Sprache zu finden. Welche Person hat den Kommentar abgegeben? Hat sie sich früher schon abfällig über andere geäußert? Wer wird adressiert – eine Politikerin oder ein Journalist? Diese Gruppen sind besonders häufig verbalen Angriffen ausgesetzt. Auch solche Informationen könnten die Trefferquote einer Künstlichen Intelligenz erhöhen.

Ohne maschinelle Unterstützung wird sich das Problem der schädigenden Sprache nicht in den Griff bekommen lassen, davon ist Tatjana Scheffler überzeugt. Zu groß ist das Volumen an Kommentaren, als dass Menschen sie ohne Unterstützung alle sichten und bewerten könnten. „Ohne die Expertise des Menschen wird es aber auch nicht gehen“, stellt die Forscherin klar. Denn es wird immer Fälle geben, in denen die Maschinen irren oder sich nicht sicher sind.

jwe

DESINFORMATION ERKENNEN

Falsche Informationen sind zu einer mächtigen Waffe im Internet geworden, zum Beispiel um den Ausgang von Wahlen zu beeinflussen. Aufgrund der schieren Masse an Texten im Netz ist es unmöglich, alle Quellen einem Faktencheck durch Menschen zu unterziehen. Ziel eines neuen Forschungsprojekts ist es, automatische Methoden zur Sprachanalyse zu entwickeln, die Menschen beim Aufspüren von Desinformation unterstützen. Zu diesem Zweck kooperiert Prof. Dr. Tatjana Scheffler mit der Arbeitsgruppe Kognitive Signalverarbeitung, geleitet von Prof. Dr. Dorothea Kolossa, dem Recherchenetzwerk CORRECTIV und einem Team der Technischen Universität Dortmund.

In dem Projekt soll ein Algorithmus entstehen, der beim Aufspüren von Desinformation hilft. Für Journalistinnen und Journalisten ist es unmöglich, diese Aufgabe allein zu verrichten – zu groß ist die Menge an Texten. Der Algorithmus soll daher nicht journalistisch ausgebildete Crowdworkerinnen und Crowdworker bei Faktenchecks unterstützen.„Natürlich geht es nicht ohne menschliche Kontrolle“, sagt Scheffler. „Aber wenn ein Algorithmus bestimmte Stellen in Texten markieren würde, könnten trainierte Crowdworkerinnen und Crowdworker speziell diese Passagen überprüfen.“ Große Textmengen könnten so effizienter gecheckt werden.

Der Algorithmus soll bestimmte Merkmale von Desinformationen identifizieren und Daten aus einem neu zu bewertenden Text mit schon bekannten Texten abgleichen. „So können wir zum Beispiel feststellen, dass ein Text in ähnlicher Form schon einmal gecheckt wurde, und bestimmte Informationen wiederverwenden“, erklärt die Bochumer Forscherin.

CORRECTIV bringt die notwendige journalistische Expertise und eine gewisse Datenbasis an bereits bewerteten Texten in das Forschungsprojekt ein. Außerdem wird das Netzwerk die Crowdworkerinnen und Crowdworker rekrutieren, ausbilden und betreuen, und entwickelt auch die Arbeitsplattform für die Faktenchecks weiter.

Wenn Falschinformationen schließlich identifiziert sind, stellen sich rechtliche Fragen: Was macht man mit den gefundenen Desinformationen? Bürgerinnen und Bürger sollen nicht falsch informiert werden, gleichzeitig muss die Meinungsfreiheit gewahrt werden. Mit diesen rechtlichen Aspekten beschäftigen sich die Partner aus Dortmund.

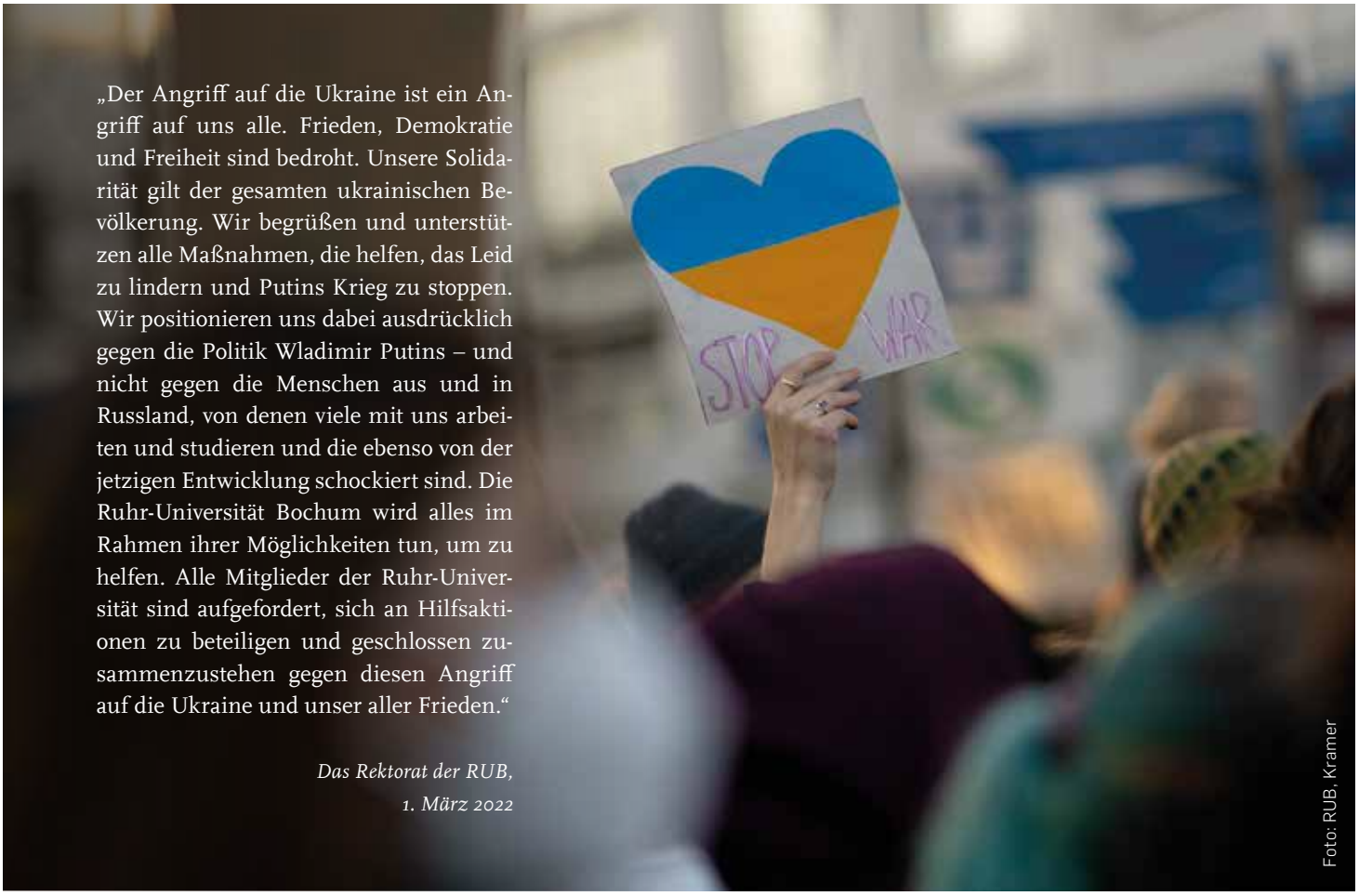
Dorothea Kolossa koordiniert das Projekt „noFake: KI-unterstütztes Assistenzsystem für die Crowdsourcing-basierte Erkennung von über digitale Plattformen verbreiteter Desinformation“. Es läuft von Dezember 2021 bis November 2024 gefördert vom Bundesministerium für Bildung und Forschung.

Text: jwe, Foto: rs

> „OHNE DIE EXPERTISE DES MENSCHEN WIRD ES NICHT GEHEN.“
>
> Tatjana Scheffler

Der russische Staatssender RT steht im Verdacht, gezielt Falschinformationen zu verbreiten. Europa hat Anfang 2022 daher ein Ausstrahlungsverbot verhängt.

REDAKTIONSSCHLUSS



„Der Angriff auf die Ukraine ist ein Angriff auf uns alle. Frieden, Demokratie und Freiheit sind bedroht. Unsere Solidarität gilt der gesamten ukrainischen Bevölkerung. Wir begrüßen und unterstützen alle Maßnahmen, die helfen, das Leid zu lindern und Putins Krieg zu stoppen. Wir positionieren uns dabei ausdrücklich gegen die Politik Wladimir Putins – und nicht gegen die Menschen aus und in Russland, von denen viele mit uns arbeiten und studieren und die ebenso von der jetzigen Entwicklung schockiert sind. Die Ruhr-Universität Bochum wird alles im Rahmen ihrer Möglichkeiten tun, um zu helfen. Alle Mitglieder der Ruhr-Universität sind aufgefordert, sich an Hilfsaktionen zu beteiligen und geschlossen zusammenzustehen gegen diesen Angriff auf die Ukraine und unser aller Frieden."

Das Rektorat der RUB,
1. März 2022

Foto: RUB, Kramer

IMPRESSUM

HERAUSGEBER: Rektorat der Ruhr-Universität Bochum in Verbindung mit dem Dezernat Hochschulkommunikation der Ruhr-Universität Bochum (Hubert Hundt, v.i.S.d.P.)

WISSENSCHAFTLICHER BEIRAT: Prof. Dr. Thomas Bauer (Fakultät für Wirtschaftswissenschaft), Prof. Dr. Gabriele Bellenberg (Philosophie und Erziehungswissenschaften), Prof. Dr. Astrid Deuber-Mankowsky (Philologie), Prof. Dr. Constantin Goschler (Geschichtswissenschaften), Prof. Dr. Markus Kaltenborn (Jura), Prof. Dr. Achim von Keudell (Physik und Astronomie), Prof. Dr. Dorothea Kolossa (Elektrotechnik/Informationstechnik), Prof. Dr. Günther Meschke (Prorektor für Forschung und Transfer), Prof. Dr. Martin Muhler (Chemie), Prof. Dr. Franz Narberhaus (Biologie), Prof. Dr. Sabine Seehagen (Psychologie), Prof. Dr. Roland Span (Maschinenbau), Prof. Dr. Martin Tegenthoff (Medizin), Prof. Dr. Martin Werding (Sozialwissenschaft), Prof. Dr. Marc Wichern (Bau- und Umweltingenieurwissenschaft), Prof. Dr. Peter Wick (Evangelische Theologie)

REDAKTIONSANSCHRIFT: Dezernat Hochschulkommunikation, Redaktion Rubin, Ruhr-Universität Bochum, 44780 Bochum, Tel.: 0234/32-25228, Fax: 0234/32-14136, rubin@rub.de, news.rub.de/rubin

REDAKTION: Dr. Julia Weiler (jwe, Redaktionsleitung); Meike Drießen (md); Lisa Bischoff (lb)

FOTOGRAFIE: Damian Gorczany (dg), Schiefersburger Weg 105, 50739 Köln, Tel.: 0176/29706008, damiangorczany@yahoo.de, www.damiangorczany.de; Roberto Schirdewahn (rs), Offerkämpe 5, 48163 Münster, Tel.: 0172/4206216, post@people-fotograf.de, www.wasaufdieaugen.de

COVER: Damian Gorczany

BILDNACHWEISE INHALTSVERZEICHNIS: Teaserfotos für die Seiten 18, 36 und 62: rs; Teaserfoto für die Seiten 40 und 50: dg

GRAFIK, ILLUSTRATION, LAYOUT UND SATZ: Agentur der RUB, www.rub.de/agentur

DRUCK: LD Medienhaus GmbH & Co. KG, Feldbachacker 16, 44149 Dortmund, Tel.: 0231/90592000, info@ld-medienhaus.de, www.ld-medienhaus.de

ANZEIGEN: Dr. Julia Weiler, Dezernat Hochschulkommunikation, Redaktion Rubin, Ruhr-Universität Bochum, 44780 Bochum, Tel.: 0234/32-25228, rubin@rub.de

AUFLAGE: 3.500

BEZUG: Rubin erscheint zweimal jährlich und ist erhältlich im Dezernat Hochschulkommunikation der Ruhr-Universität Bochum. Das Heft kann kostenlos abonniert werden unter news.rub.de/rubin/abo. Das Abonnement kann per E-Mail an rubin@rub.de gekündigt werden.

ISSN: 0942-6639