Frühwarnsystem für Falschmeldungen

Fake News in sozialen Medien effizienter und treffgenauer bekämpfen: **Manuel Gomez Rodriguez** vom **Max-Planck-Institut für Softwaresysteme** kombiniert Verfahren der künstlichen Intelligenz mit der Auswertung von Signalen, in denen sich menschliches Urteil widerspiegelt.

Faktencheck vor dem Internetzeitalter: Wenn Pinocchio log, war das der Märchenfigur an der Nase anzusehen. Gegen Falschmeldungen in den sozialen Medien würde das aber auch nicht helfen.

Foto: shutterstock

Pope Francis shocks world, endorses Donald Trump for president

Die Behauptung, Papst Franziskus befürworte die Wahl von Donald Trump zum Präsidenten, verbreitete sich millionenfach, war aber völlig frei erfunden. Das wäre einfach aufzudecken gewesen: Die Webseite *WTOE 5 News*, die sie in die Welt setzte, bezeichnet sich selbst als Seite für Fantasienachrichten.

TEXT **RALF GRÖTKER**

Falschmeldungen sind gefährlich, manchmal sogar für Leib und Leben. Am 4. Dezember 2016 etwa drang ein Mann mit einem Sturmgewehr in die Pizzeria Comet Ping Pong in Washington, D. C., ein. Sein Vorhaben: Er wollte die angeblich in dem Restaurant festgehaltenen und missbrauchten Kinder befreien. Wie Millionen andere Internetnutzer hatte er über die sozialen Medien *Reddit* und *4chan* davon erfahren, dass der Keller dieser Pizzeria der Stützpunkt eines Pädophilenrings sei. Im Zentrum des Rings, so die Legende, habe die damalige Präsidentschaftskandidatin Hillary Clinton gestanden. Zu denjenigen, die die Falschmeldung mit verbreitet hatten, zählten der zwischenzeitliche Nationale Sicherheitsberater von Donald Trump, Michael T. Flynn und dessen Sohn.

„Pizzagate" markiert einen der vorläufigen Höhepunkte von Fake News. Zahlreiche soziale Netzwerke haben mittlerweile begonnen, ihre Nutzer um Hinweise auf falsche Meldungen zu bitten. Einige sind auch Kooperationen mit journalistischen Organisationen, die Fakten überprüfen, eingegangen, in Deutschland zum Beispiel mit correctiv.org.

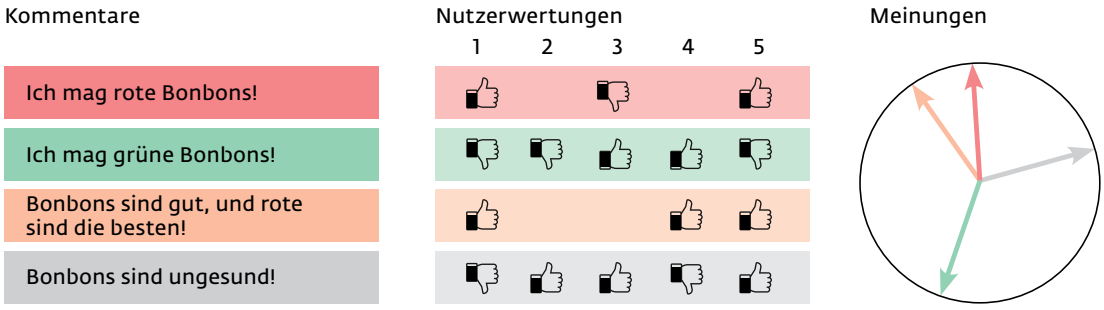
Manuel Gomez Rodriguez, Gruppenleiter am Max-Planck-Institut für Softwaresysteme in Kaiserslautern, arbeitet mit seinem Team an ausgeklügelten Verfahren, damit sich Falschnachrichten treffgenauer und effizienter identifizieren lassen. Die Methoden greifen dabei wie die Teile eines Puzzles ineinander, um die verschiedenen Aspekte und Informationen, die sich aus dem Nachrichtenstrom herauslesen lassen, im Zusammenhang zu analysieren. „Wir verfolgen einen hybriden Ansatz", erklärt Gomez Rodriguez. „Wir kombinieren Verfahren der künstlichen Intelligenz mit der Auswertung von Signalen, in denen sich menschliches Urteil widerspiegelt."

Als ein zentrales Ergebnis ihrer Arbeit haben die Forscher „Curb" präsentiert, ausgesprochen wie das englische Wort für „Drosselung". Der Algorithmus priorisiert, welche Inhalte die nur begrenzt verfügbaren menschlichen Faktenchecker am dringendsten überprüfen und gegebenenfalls als falsch kennzeichnen müssen. Ziel ist, dass möglichst wenige Menschen Falschmeldungen lesen, bevor diese als solche markiert sind.

EIN VERFAHREN MIT EINEM DYNAMISCHEN SCHWELLENWERT

Als wesentliche Information wertet das Verfahren auf ausgeklügelte Weise aus, wie Nutzer mit Inhalten umgehen. Zum einen, in welchem Maß Nutzer Inhalte weiterleiten und in welchem Tempo sich diese folglich verbreiten, zum anderen, wie viele Nutzer einen Beitrag als Fake markieren. Dies sind wichtige Kriterien dafür, wie schnell sich eine eventuelle Falschnachricht verbreitet. Gomez Rodriguez: „Während die meisten sozialen Medien momentan lediglich die Anzahl von Beanstandungen durch Nutzer auswerten, verwendet unser Verfahren einen dynamischen Schwellenwert, der sich über die Zeit hin verändert und der auf die Viralität einer Nachricht reagiert sowie auf die Wahrscheinlichkeit, mit der es sich um Fake News handelt."

Konkret nimmt der Algorithmus, den Gomez Rodriguez und sein Team entwickelt haben, zunächst die Relation zwischen Beanstandungen auf der einen Seite und Weiterleitungen (*shares*) ohne Beanstandung auf der anderen Seite in den Blick. Je öfter, im Verhältnis, eine Nachricht ohne Beanstandung geteilt wird, desto größer die Wahrscheinlichkeit, dass sie *nicht* falsch ist. Allerdings: Je schneller sich eine Nachricht verbreitet, desto größer der potenzielle Schaden in dem Fall, dass es sich doch um eine falsche Meldung handelt. Curb löst



Textanalyse über Nutzerwertungen: Wie differenziert oder polarisiert Aussagen im Netz sind, analysieren Max-Planck-Forscher aus Kaiserslautern an der Zustimmung (Daumen hoch) und Ablehnung (Daumen runter), die Sätze in einer Sequenz erfahren (Mitte). Daraus ermitteln sie Vektoren, die eine Aussage im Meinungsraum verorten (rechts). In der Darstellung ist zu erkennen, dass die Aussagen „Ich mag rote Bonbons" (roter Vektor) und „Ich mag grüne Bonbons" (grüner Vektor) entgegengesetzte Meinungen ausdrücken. Am Bonbon-Beispiel demonstrieren die Forscher ihr Vorgehen.

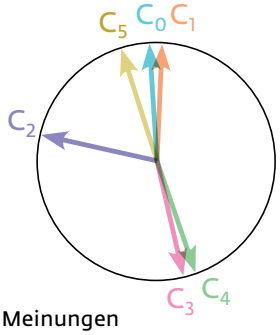
dieses Problem, indem die Informationen über die Verbreitungsgeschwindigkeit und über die Wahrscheinlichkeit, dass es sich um Fake News handelt, nebeneinander betrachtet und dabei immer wieder aktualisiert werden. Aufgabe des Algorithmus ist es, zwischen den beiden Kriterien optimal abzuwägen.

Ein Beispiel: Angenommen, eine Nachricht wird zehnmal pro Stunde geteilt und die Wahrscheinlichkeit, dass sie falsch ist, liegt der Nutzerbewertung zufolge bei fünfzig Prozent. Dann kann man rechnerisch davon ausgehen, dass pro Stunde fünf Nutzer einer Falschmeldung ausgesetzt werden. Diese Rechnung wird nun jedes Mal angepasst, wenn ein Nutzer die betreffende Nachricht weiterleitet und sie entweder als falsch markiert (*flaggt*) oder als mutmaßlich solide Nachricht nicht beanstandet. Auf diese dynamische Weise schafft der Algorithmus eine optimale Balance: zwischen dem Bemühen, möglichst wenige Menschen mit undeklarierten Falschmeldungen zu konfrontieren einerseits, und der Effizienz beim Einsatz der menschlichen Faktenchecker andererseits.

TEST MIT DATEN VON TWITTER UND WEIBO

Der finale Test für Curb war das Experiment mit realen Daten, die Wissenschaftler der koreanischen KAIST-Universität bereits via Webcrawling aus den Netzwerken Twitter und Weibo gesammelt und öffentlich zur Verfügung gestellt hatten. Der größere der beiden Datensätze aus dem chinesischen Social-Media-Netzwerk Weibo bestand aus mehr als 4600 einzelnen Nachrichtenmeldungen, die 2,8 Millionen Nutzer in Form von Posts oder Weiterleitungen 3,7 Millionen Mal gesendet hatten. „Wir kannten die Netzwerkstrukturen innerhalb des Datensatzes, also wie viele Follower die einzelnen Nutzer hatten, und wir wussten auch, welche der Nachrichtenmeldungen die Faktencheck-Organisation Snopes als Fake News deklariert hatte", erläutert Gomez Rodriguez."

Nicht bekannt war, wie und wann die Nutzer in dem Datensatz die Nachrichten geflaggt hatten. Hier mussten sich die Forscher mit einem Trick behelfen. Sie griffen dabei auf andere Untersuchungen zurück, wie oft Nutzer tatsächlich falsche Nachrichten als solche markiert hatten – so konnten sie begründete Annahmen darüber anstellen, wie gut Nutzer falsche Nachrichten erkennen und wie oft sie diese im Schnitt dann auch markieren. „Wir haben unseren Algorithmus einfach ein breites Spektrum an plausiblem Flagging-Verhalten ausprobieren lassen", erläutert Gomez Rodriguez.



C_0	[...] [Donald Trump] has [...] Enquirers[a] [which] he considers a treasure trove of information.
C_1	He should change his name to Donald J Dubious.
C_2	[...] Trump] can be an #$%$, and Islam can be cancer [...] they are not mutually exclusive [...]
C_3	Why not? Try anything. Terrorism has got to stop now!
C_4	It is a great idea.
C_5	Trump familiy motto: "It's not a lie if you believe it."

[a] National Enquirers is a well known entertainment magazine in US.

Orientierungshilfe im politischen Meinungsspektrum: Wie Nutzer die Kommentare in einer Onlinediskussion über Donald Trump bewerten, gibt Aufschluss über die politische Haltung hinter den Kommentaren, die sich in Vektoren im Meinungsraum erfassen lassen. So stammen die Aussagen C_0, C_1 und C_5 offensichtlich von Menschen mit völlig anderer politischer Einstellung als die Aussagen C_3 und C_4.

Grafiken: Manuel Gomez Rodriguez/MPI für Softwaresysteme

Vote Hillary from home! Save time & avoid the line!

Die Werbung, Unterstützer von Hillary Clinton könnten ihre Stimme als Textnachricht abgeben, sollte Wähler in die Irre führen. Die Anzeige mit dem Logo der Clinton-Kampagne wurde über Twitter verbreitet. Wer ihr folgte, verschenkte seine Stimme jedoch, denn per SMS zu wählen, war nicht möglich.

In dem Experiment mit den realen Daten aus Twitter oder Weibo testeten die Forscher aus Kaiserslautern, wie effektiv ihr Algorithmus, verglichen mit anderen Methoden, verdächtige Meldungen zum Faktencheck lotst. Gegen Curb trat unter anderem das Pseudoverfahren Oracle an, welches im Testszenario ganz schlicht Zugang zu der Information hatte, ob eine Nachricht tatsächlich falsch war oder nicht, und die Meldung dementsprechend zum Faktencheck schickte.

Andere Vergleichsmethoden benutzten einfache Faustregeln: einmal ein Algorithmus, der – wie die Methode des Kaiserslauterner Teams – aus dem bloßen Verhältnis zwischen der Anzahl der *flags* und der Anzahl von Weiterleitungen die Dringlichkeit für den Faktencheck ermittelt; dann ein Algorithmus, der eine Nachricht dem Faktencheck überstellt, sobald eine bestimmte Zahl von *flags* erreicht ist; schließlich ein Algorithmus, der allein das Ausmaß der Verbreitung einer Nachricht heranzieht, um eine Meldung für den Faktencheck zu priorisieren.

WEITERE ANWENDUNGEN FÜR DIE ALGORITHMEN VON CURB

Das Resultat des Vergleichstests: Curb verhinderte fast ebenso gut wie Oracle die Verbreitung von Falschinformationen, die nicht als solche indiziert waren. Die drei Faustregeln vermochten dies nicht.

Trotz des Testerfolgs kann Gomez Rodriguez die Aussicht von Curb, in der Praxis umgesetzt zu werden, noch nicht einschätzen: „Ob Curb hier als Lösung am Ende infrage kommt oder lediglich Komponenten unseres Verfahrens sich als interessant für kommerzielle Anbieter erweisen, wird man sehen müssen“, sagt der Forscher. „Einer der Entwickler von Curb hat aber vor Kurzem im Fake-News-Team bei Facebook angeheuert.“

Ähnliche Algorithmen wie Curb lassen sich, davon abgesehen, auch auf anderen Feldern einsetzen. „Sprachlern-Software zum Beispiel könnten Verfahren wie Curb optimieren, indem sie helfen, besser zu prognostizieren, welche Inhalte den Lernenden wiederholt präsentiert werden müssen, damit sie diese im Gedächtnis behalten“, sagt Gomez Rodriguez. Ein anderes Anwendungsfeld ist das virale Marketing. Für diese Anwendung haben die Forscher das Grundgerüst von Curb ursprünglich auch entwickelt: um herauszufinden, wie Nachrichten in sozialen Medien am effektivsten verbreitet werden.

Ein Problem lässt Curb jedoch ungelöst: Was passiert, wenn Nutzer das System gezielt sabotieren, indem sie solide Nachrichten als Fake markieren oder bewusst Falschmeldungen verbreiten? Bei solch extremem Verhalten dürfte Curb schwerlich noch richtig einschätzen, wie dringend eine Meldung zum Faktencheck muss. Um dieses Problem anzugehen, haben Gomez Rodriguez und seine Kollegen „Detective“ entwickelt.

Auch der Detective-Algorithmus dient dem Ziel, die Verbreitung von Falschinformationen zu reduzieren. Gomez Rodriguez’ Team hat das Verfahren auf der Web Conference in diesem Frühjahr in Lyon präsentiert. Während Curb alle Nutzer für gleich seriös hält, versucht Detective herauszufinden, wer Fake News besonders zuverlässig beanstandet und wer solide Meldungen vorsätzlich als Fake brandmarkt, um das System zu unterlaufen.

Zu diesem Zweck berücksichtigt der Algorithmus von Detective die Resultate des Faktenchecks, mit deren Hilfe er einschätzt, in welchem Maß Nutzer im Erkennen und Markieren von Fake News zuverlässig sind. „Wir beobachten eine Nutzerin oder einen Nutzer über einen gewissen Zeitraum hinweg“, erklärt Gomez Rodriguez. „Dabei übergeben wir Nachrichten, die sie oder er verfasst oder teilt, immer wieder an den Faktencheck.“

Auch Detective muss dabei einen Zielkonflikt lösen. Um die Zuverlässig-

Mit künstlicher Intelligenz gegen Fake News: Manuel Gomez Rodriguez und sein Team entwickeln unter anderem Methoden, um die Verbreitung von Falschmeldungen, die nicht als solche zu erkennen sind, effizient zu verhindern.

keit möglichst vieler Nutzer beurteilen zu können, sollten die Faktenprüfer einerseits Nachrichten, die von möglichst vielen unterschiedlichen Personen weitergeleitet wurden, validieren. Auch solche, bei denen es sich den Nutzermarkierungen zufolge wahrscheinlich nicht um Falschmeldungen handelt. So erfahren sie etwas darüber, welche Nutzer Informationen vertrauenswürdig beurteilen. Andererseits sollte die begrenzte Zeit der menschlichen Faktenchecker auch hier am besten wieder nur für Nachrichten verwendet werden, die vermutlich Fake sind. Dazu wäre es am effizientesten, einfach dem Urteil jener Nutzer zu vertrauen, die sich bereits als verlässlich erwiesen haben. Doch damit weitere Nutzer diesen Status erlangen, müssen die Verfahren des maschinellen Lernens, die bei Detective zum Einsatz kommen, das Verhalten möglichst vieler Personen kennenlernen. Eine Leistung von Detective besteht darin, mithilfe des maschinellen Lernens den optimalen Kompromiss zwischen den beiden Erfordernissen zu finden.

Wie Curb bestand auch Detective den Test mit empirischen Datensets mit Bravour. Im Experiment lieferte die Methode annähernd so gute Resultate wie ein Pseudo-Algorithmus, der das Flagging-Verhalten der Nutzer kannte. In der praktischen Anwendung dürfte Detective in Kombination mit Curb hilfreich sein für Administratoren, die mithilfe der Algorithmen den Einsatz menschlicher Faktenchecker möglichst effizient planen wollen.

Zudem könnten Administratoren auf Basis der Detective-Wertung Nutzern Informationen darüber zugänglich machen, wie verlässlich andere Personen innerhalb ihres sozialen Netzwerkes sind, wenn es um die Markierung von Nachrichten als falsch geht. „Praktisch setzt hier allerdings der Datenschutz Grenzen“, räumt Gomez Rodriguez ein. Schon dass „Freunde“ oder Follower sehen, welche Likes man setzt, sei für viele Nutzer nicht akzeptabel. „Eine Nachricht als Fake News zu markieren, kann ebenso problematisch sein, weil man dabei oft etwas von seiner eigenen politischen Orientierung preisgibt.“ Deshalb müssten Resultate von Detective entsprechend anonymisiert werden. „Zehn Prozent der vertrauenswürdigen Personen in Deinem Netzwerk haben diese Nachricht als ‚fake‘ geflaggt: So eine Information könnte man schon einspielen“, meint Gomez Rodriguez.

POLARISIEREN NACHRICHTEN IN SOZIALEN MEDIEN?

Manche Personen als besonders vertrauenswürdig darzustellen, könnte aber auch das Gegenteil des gewünschten Effekts bewirken: Nutzer, die zu Verschwörungstheorien neigen, könnten bewusst solchen Personen folgen, die absichtlich solide Nachrichten als Fake markieren und selbst Fake News in Umlauf bringen – weil sie glauben, dass es sich hier um eine lediglich vom Mainstream unterdrückte Wahrheit handelt. Allerdings erwies sich Detective gegen eine solche vorsätzliche Verbreitung falscher Informationen als ziemlich robust – gerade weil der Algorithmus die Vertrauenswürdigkeit der Nutzer berücksichtigt.

Neben dem Bemühen, Falschnachrichten effektiv aufzudecken, beschäftigt sich das Team von Gomez Rodriguez auch mit der Frage, wie sehr Nachrichten – ob Fake oder nicht – tatsächlich zu einer Polarisierung von Meinungen in den sozialen Medien beitragen. Für die Antwort darauf haben die Forscher ebenfalls einen Algorithmus entwickelt. Dieser wertet Urteile wie etwa „Daumen hoch!“ oder „Daumen runter!“ zu Textbeiträgen wie etwa Kommentaren in Onlinediskussionen aus.

Anstelle von Meinungen zu einzelnen Fragen betrachten die Forscher dabei aber ganze Meinungssequenzen. Was damit gemeint ist, veranschaulicht Gomez Rodriguez mit diesen Aussagen: „Ich mag rote Bonbons!“; „Ich mag grüne Bonbons!“; „Bonbons sind gut, und rote Bonbons sind die besten!“ und „Bonbons sind ungesund.“ Die jeweilige Meinung hinter einem einzelnen

Foto: Oliver Dietze

Kommentar ist mit einer Software, die den Text etwa auf bestimmte Wörter analysiert und mit anderen Aussagen vergleicht, nicht zuverlässig zu ermitteln. Anders ist das mit den Meinungen, die Nutzer ausdrücken, indem sie die Kommentare in einer solchen Aussagekette durch Zustimmung oder Ablehnung bewerten. Genau diese Urteile verschiedener Nutzer analysierten die Wissenschaftler und berechneten daraus auch die Meinung, die ein einzelner Kommentar widerspiegelt.

Bei der Analyse der Meinungen, die sich sowohl in einem einzelnen Kommentar als auch in den Bewertungen einer Aussagensequenz widerspiegeln, fokussieren Gomez Rodriguez und seine Kollegen auf zwei Merkmale.

Zum einen betrachten sie den Grad von Komplexität oder die Anzahl von Achsen, anhand derer sich der Meinungsraum darstellen lässt. Ein Beispiel: Wenn alle Teilnehmer an einer Diskussion entweder die gleiche Meinung oder genau die jeweils entgegengesetzten Meinungen bezüglich einer einzelnen Fragen vertreten, lassen sich die Antworten anhand von *einer* Achse sortieren – solche Diskussionen werden also buchstäblich eindimensional geführt.

Zum anderen ermittelten die Forscher, wie weit die einzelnen Meinungen voneinander entfernt sind. Zu diesem Zweck werden die Haltungen hinter den Kommentaren, aus denen sich die Sequenz zusammensetzt, als Vektoren in einem Meinungsraum dargestellt. Den jeweiligen Vektor ermittelt der Algorithmus, indem er auswertet, wie andere Nutzer einen Kommentar bewerten. Die Anordnung der Vektoren gibt Aufschluss über die Diversität der Meinungen. „Wir können Textbeiträge, die sich in ihrem semantischen Inhalt stark voneinander unterscheiden, die ganz unterschiedliche Worte verwenden und die vielleicht sogar Ironie enthalten, im Meinungsraum zueinander positionieren“, betont Gomez Rodriguez.

Die Analyse eines großen Datensatzes von Onlinediskussionen auf den Seiten von Yahoo News, Yahoo Finance, Yahoo Sports und der Yahoo Newsroom App hat ergeben: 75 Prozent der Onlinediskussionen bewegen sich auf zwei oder mehr Achsen im Meinungsraum, sie wurden also nicht polarisiert geführt. „Dies ist ein deutliches Zeichen dafür, dass die Diskussionen auf diesen Onlineseiten nicht dem Spiel vom Demagogen zum Opfer gefallen sind“, meint Manuel Gomez Rodriguez.

Der Algorithmus ermöglicht es also, Debatten in Onlineforen oder sozialen Medien zu bewerten, und wirkt mit seinen bisherigen Ergebnissen dem Eindruck entgegen, dass diese in der Anonymität des Internets stets undifferenziert geführt werden und überwiegend von Demagogen polarisiert werden. Wie Curb und Detective zeigt er mithin, dass ein hybrider Ansatz aus künstlicher Intelligenz und menschlichen Bewertungen hilft, solche Diskussionen zu versachlichen. ◂

Paid fake protesters were bussed in to the anti-Trump protests in Austin, Texas.

Die Fotos von zahlreichen Bussen führten fragwürdige Nachrichtenseiten als einzigen Beleg dafür an, dass bezahlte Demonstranten zu einem Protestmarsch gegen Donald Trump in Austin gebracht wurden. Die Busse waren jedoch für eine Veranstaltung in einem Kongresszentrum unterwegs, das mehrere Kilometer vom Startpunkt des Demonstrationszuges entfernt liegt. Für Zahlungen an die Teilnehmer gibt es keinerlei Beleg.

www.mpg.de/podcasts/digitale-gesellschaft

AUF DEN PUNKT GEBRACHT

- **Ein hybrider Ansatz aus künstlicher Intelligenz und menschlichen Bewertungen kann in verschiedener Hinsicht helfen, Debatten im Internet zu versachlichen.**
- **Der Algorithmus Curb priorisiert, wie dringend ein Inhalt einem Faktencheck unterzogen werden muss, damit sich eine eventuelle Falschmeldung nicht undeklariert verbreitet. Er analysiert dafür immer wieder neu, wie schnell sich eine Meldung verbreitet und wie viele Nutzer sie als Fake News markiert haben.**
- **Der Algorithmus Detective soll ebenfalls die Verbreitung von Falschmeldungen verhindern, berücksichtigt dabei aber, wie vertrauenswürdig die Nutzer sind, die eine Meldung als falsch markieren.**
- **Ein weiterer Algorithmus wertet aus, wie differenziert Diskussionen im Internet geführt werden. Demnach finden sie in 75 Prozent der Fälle nicht polarisiert statt – ein Indiz, dass sich die Nutzer mehrheitlich nicht von Demagogen leiten lassen.**