Wühlen in Worten

Dass Rechner riesige Datenmengen verarbeiten und daraus Muster erkennen und ableiten können, ist bekannt. Längst funktioniert das auch mit Texten – doch noch haben Statistik und Geisteswissenschaften nicht recht zusammengefunden. Das ändert sich gerade: Wissenschaftler aus den Bereichen Journalistik, Statistik und Informatik haben ein virtuelles Institut gegründet, um im großen Stil Medienberichterstattung zu analysieren.

In Kürze

Die Suche

DoCMA ist ein interdisziplinäres Forschungsprojekt, in dem Algorithmen riesige Mengen an Zeitungsartikeln und Texten in sozialen Medien verarbeiten und dabei Muster erkennen.

Die Bedeutung

Mit den Ergebnissen kann die Wissenschaft aufkommende Trends in sozialen Netzwerken erkennen, die Entwicklung eines Themas nachverfolgen oder vergleichen, wie in verschiedenen Medien oder Ländern über ein Thema berichtet wird.

Prof. Dr. Henrik Müller studierte Volkswirtschaftslehre an der Christian-Albrechts-Universität zu Kiel und promovierte an der Universität der Bundeswehr Hamburg. Er absolvierte die Deutsche Journalistenschule München und arbeitete viele Jahre als Redakteur bei Zeitungen und Zeitschriften. Zuletzt war er stellvertretender Chefredakteur des Manager Magazins. Seit Oktober 2013 ist er Professor für Wirtschaftspolitischen Journalismus an der TU Dortmund. Müller ist Autor diverser Bücher zu wirtschaftspolitischen Themen.

Prof. Dr. Jörg Rahnenführer studierte Mathematik mit Nebenfach Psychologie an der Universität Düsseldorf, wo er auch promovierte. Nach Forschungsaufenthalten in Wien und Berkeley arbeitete er am Max-Planck-Institut für Informatik in Saarbrücken an statistischen Methoden in der Bioinformatik. Seit 2007 ist er Professor an der Fakultät Statistik. Dort entwickelt er statistische Methoden zur Analyse genomischer Daten für die Diagnose und Therapie von Krankheiten, Methoden zur biologischen Interpretation von genomischen Daten und statistische Methoden in der Toxikologie.

Prof. Dr. Carsten Jentsch studierte Mathematik mit Nebenfach Betriebswirtschaftslehre an der TU Braunschweig, wo er 2010 auch promovierte. Nach einem Forschungsaufenthalt an der University of California in San Diego wurde er 2011 Postdoc an der Fakultät für Rechtswissenschaft und Volkswirtschaftslehre der Universität Mannheim. Nach Professurvertretungen an den Universitäten Bayreuth und Mannheim lehrt und forscht er seit dem Sommersemester 2018 an der TU Dortmund, wo er die Professur für Wirtschafts- und Sozialstatistik an der Fakultät Statistik innehat. Seine Forschungsinteressen liegen im Bereich der mathematischen Statistik und der Ökonometrie.

Prof. Dr. Erich Schubert studierte Mathematik und Informatik an der Ludwig-Maximilians-Universität (LMU) München, wo er 2013 zum Thema Ausreißererkennung in hochdimensionalen Daten promovierte. Von 2016 bis 2018 war er Postdoc und Professurvertreter an der Universität Heidelberg. 2018 folgte er dem Ruf an die TU Dortmund auf die Professur für Data Mining. Seine Forschungsschwerpunkte sind das unüberwachte Lernen, insbesondere die Clusteranalyse, die Anomalieerkennung und das Text Mining, sowie Datenbanktechniken zur Beschleunigung solcher Analyseverfahren. Schubert beschäftigt sich seit 2014 zunehmend mit der Analyse von Textdatenströmen.

Im „Dortmund Center for data-based Media Analysis“ arbeiten vier Professoren zusammen: Ein Wirtschafts- und Kommunikationswissenschaftler, zwei Statistiker und ein Informatiker – sowie deren Mitarbeiterinnen und Mitarbeiter.

Wenn er an sein erstes Semester am Institut für Journalistik zurückdenkt, dann erinnert sich Prof. Henrik Müller auch an TTIP. Es war um die Jahreswende 2013/14, Müller war frisch berufen auf die Professur für Wirtschaftspolitischen Journalismus, und er fragte seine Studierenden in einem Seminar, welche Themen sie in der Berichterstattung für vernachlässigt hielten. „Die Verhandlungen zu TTIP“, antwortete einer, die geplante Transatlantische Handels- und Investitionspartnerschaft zwischen Europa und den USA komme in den Medien kaum vor. „Ich fiel aus allen Wolken“, erinnert sich Henrik Müller. Denn Müller kam gerade aus der Praxis: Vor seinem Ruf nach Dortmund war er stellvertretender Chefredakteur des Manager Magazins. „TTIP war für uns tatsächlich kein Thema. Wir glaubten, es gehe vor allem um technische Fragen – also langweiliges Zeug.“ Kurze Zeit später wusste dann jeder Zeitungsleser und jede Fernsehzuschauerin, wofür die Abkürzung steht. Die Medien waren voll von TTIP – vor allem von den Protesten dagegen.

Es war ein Aha-Moment für Henrik Müller. Denn die Debatte um das Freihandelsabkommen wurde, von der allgemeinen Öffentlichkeit weitgehend unbemerkt, zuerst in den sozialen Medien geführt. Dort war auch der Student aus Müllers Seminar auf das Thema aufmerksam geworden. „Die Protestbewegung gegen TTIP hat sich dort aufgebaut, bis sie auch die klassischen Medien erreichte. Dieses Phänomen hat sich seitdem bei verschiedenen Themen wiederholt, wobei die Zeiträume bis zum Erreichen der klassischen Medien immer kürzer werden“, sagt Müller.

Die Medien als Seismograph zu nutzen, um gesellschaftliche Entwicklungen früher zu erkennen – davon hatte Müller schon lange geträumt. In seiner Antrittsvorlesung zeichnete er das Bild fruchtbarer Kooperationen zwischen Ökonomie und Journalismusforschung. Mit DoCMA ist das nun Wirklichkeit geworden: Seit 2015 existiert das Dortmund Center for data-based Media Analysis – ein virtuelles Institut der TU Dortmund, unter dessen Dach inzwischen vier Hochschullehrer zusammenarbeiten. Dort ist Henrik Müller als Wirtschafts- und Kommunikationswissenschaftler in der Minderheit: Neben ihm besteht das Team aus zwei Statistikern und einem Informatiker sowie deren Mitarbeiterinnen und Mitarbeitern.

Denn um öffentliche Kommunikation zu sichten, womöglich sogar in Echtzeit, und daraus Trends und Entwicklungen abzuleiten, braucht es die Analyse riesiger Textmengen – und dazu wiederum die Techniken, die aus dem Datentrümmerfeld les- und verwertbare Ergebnisse machen. Sprich: Es braucht Data Mining. „Ich bin Ökonom – von Data Mining hatte ich zu diesem Zeitpunkt keine Ahnung“, gibt Müller zu. „Was ich kannte, war die Frequenzanalyse, das reine Wörterzählen: Wie oft taucht etwa das Wort ‚Rezession‘ in der Berichterstattung auf? Data Mining allerdings kann unendlich viel mehr. Und es ist viel komplizierter, als ich damals gedacht habe.“ Ein Jahr, dachte Müller bei seinem Start als Hochschullehrer, würde es wohl brauchen, um erste Ergebnisse zu erzielen. Das war vor fünf Jahren. Ergebnisse gibt es tatsächlich – doch der Entwicklungsprozess dauert an. Dafür waren die Resultate substanzieller als ursprünglich erhofft.

Unmerkliche Veränderungen sichtbar machen

Die Experten für das Thema sitzen nur wenige hundert Meter von Müllers Büro an der Emil-Figge-Straße 50 entfernt, nämlich im Mathe-Tower. Jörg Rahnenführer ist Professor für „Statistische Methoden in der Genetik und Chemometrie“ an der Fakultät Statistik. Mit großen Datenmengen kennt er sich aus – allerdings waren dies bei ihm bislang eher Zahlen denn Buchstaben. Sein Kollege Prof. Carsten Jentsch forscht in der mathematischen Statistik und ihrer Anwendung in den Wirtschafts- und Sozialwissenschaften. Er bringt Erfahrung in der Analyse politischer Textdaten mit und stieß 2018 zum DoCMA-Team. Data-Mining-Fachmann Prof. Erich Schubert aus der Informatik kam 2019 dazu.

Die Debatte um das Freihandelsabkommen TTIP wurde 2014 zuerst in den sozialen Medien geführt. Die Protestbewegung hat sich dort aufgebaut, bis sie auch die klassischen Medien erreichte.

Die Analyse von Textdaten mit Data Mining gewinnt rasant an Bedeutung. Noch vor wenigen Jahren wäre kaum denkbar gewesen, was Konzernen wie Google heute dank steigender Rechenkapazitäten und neuer Algorithmen möglich ist: riesige Mengen an Texten schnell und intelligent zu verarbeiten. Die Sozial-, Wirtschafts- und Geisteswissenschaften allerdings greifen bislang nur zögerlich auf Big-Data-basierte Methoden zurück. Es gab also kaum Vorbilder für das, was Müller vorhatte. Doch er war sich seiner Vision sicher: Gerade Massenmedien reagieren äußerst sensibel darauf, wenn sich Werte oder Kräfteverhältnisse in der Gesellschaft verändern. Und dem Journalismus selbst ist dies ebenso wenig bewusst wie dem Publikum.

Um solche fast unmerklichen Verschiebungen sichtbar zu machen, braucht es die Analyse der Medienberichterstattung über lange Zeitspannen. Dafür diente anfangs eine DVD mit dem SPIEGEL-Archiv, inzwischen unterstützt die Universitätsbibliothek die Forscher mit ständig aktualisierten Zeitungskorpora. Daten sind also vorhanden. Die Tools, um die Vielzahl an verfügbaren Texten untersuchen zu können, mussten allerdings noch entwickelt werden.

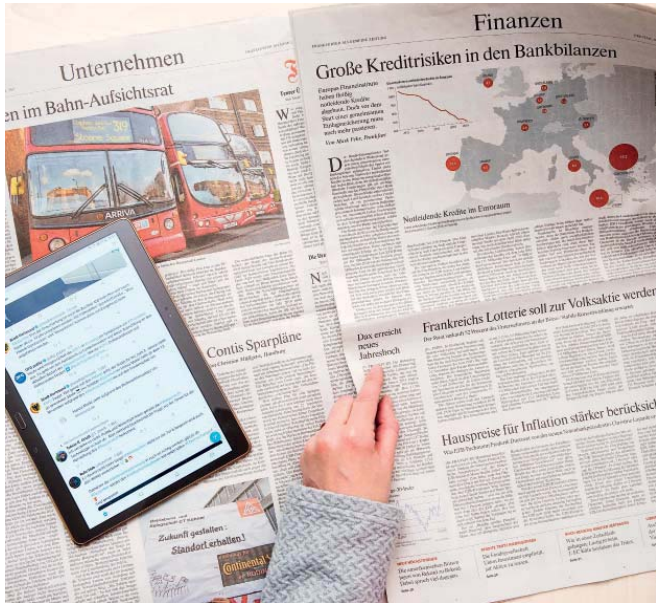
Glücklicherweise stieß Müller bei den Kollegen aus der Statistik nicht nur auf offene Ohren, sondern auf ein ähnlich gelagertes Forschungsinteresse. „Wir hatten Data Mining bislang in den Lebenswissenschaften betrieben und selbst gerade damit begonnen, uns mit Textanalysen zu beschäftigen. Wir waren auf der Suche nach einem interessanten Anwendungsgebiet – also nach großen Datensätzen, und nach einem Partner, der mit uns gemeinsam neue Methoden entwickeln will“, sagt Jörg Rahnenführer. „Für uns ist es die bislang intensivste Kooperation mit einem anderen Fach außerhalb der Lebenswissenschaften. Was ich dabei besonders interessant finde, ist der Arbeitsprozess. Wir werfen den Kollegen, die das Tool nutzen möchten, kein komplexes statistisches Modell hin, sondern es geht zwischen den Disziplinen hin und her, wird immer wieder angepasst.“

Statt Data Mining nun also Text Mining: Die Statistiker lassen Computer arbeiten, damit diese in der Berichterstattung Muster erkennen und Texte zu Kategorien oder Themen zusammenfassen. Das klingt simpel – ist es aber nicht. Denn die Algorithmen bilden die Zuordnungen anhand von Wahrscheinlichkeiten. Die Forschenden erhalten am Ende des Rechenprozesses beispielsweise die Information, dass ein Text zu einem Anteil von 87 Prozent von Fußball handelt. Anschließend müssen sie die vom Algorithmus automatisch gebildeten Cluster auf ihre Plausibilität überprüfen. Wenn die Computer gerechnet haben, fängt für die Forscherinnen und Forscher die mühsame Arbeit erst an.

Zehntausende Zeitungsartikel in der Auswertung

Doch es ist Mühe, die sich lohnt. „Wem wird die Schuld an der Finanzkrise nachgesagt?“, lautete eine der Forschungsfragen, die das DoCMA untersucht hat. Mehr als 50.000 Artikel aus mehreren europäischen Ländern kamen in der Auswertung – eine Größenordnung, an die im Zeitalter vor Big Data gar nicht zu denken war. „In einem anderen Projekt haben wir untersucht, wie das Thema TTIP auf Twitter behandelt wird – im Vergleich zu dem Diskurs, der dann in den klassischen Medien verhandelt wurde“, erzählt Müller.

„Der Mehrwert dieser Methode“, sagt Rahnenführer, „entsteht im Grunde über den Vergleich.“ Denn häufig sind die Ergebnisse des Text Mining auf den ersten Blick nicht besonders überraschend. Sie liefern zunächst oft eine Bestätigung bislang nur gefühlter Wahrheiten. So auch bei einer Analyse der Berichterstattung in den Jahren 2015 und 2016 über die sogenannte Flüchtlingskrise. Drei überregionale deutsche Tageszeitungen wurden verglichen mit der Zeitung „Junge Freiheit“, dem medialen Sprachrohr der Neuen Rechten. Der Algorithmus misst dabei die durchschnittliche Deutung und Themensetzung, keine extremen Schwankungen. „Aber als wir uns die Berichterstattung in den Medien einzeln angeschaut haben, erkannten wir durchaus Muster“, sagt Müller. Zum Beispiel hatte die „Junge Freiheit“ in der Debatte andere Aspekte betont als die Mainstream-Zeitungen. Sie berichtete nationaler und weniger europäisch – und es wurden in der Debatte seltener als bei den anderen Zeitungen wirtschaftliche Aspekte angeführt. „Das war tatsächlich überraschend – man hätte ja vermuten können, dass im rechten Diskurs stark mit den Kosten argumentiert wird, die die Aufnahme von Geflüchteten verursacht. Man kann hier zu Zahlenreihen geronnen erkennen, wie populistische Medien die komplizierteren Aspekte einer Debatte systematisch ausblenden und sich auf simple Narrative beschränken.“



Verschiedene Themen bauen sich in den sozialen Medien auf, bevor sie die klassischen erreichen – wobei die Zeiträume dazwischen immer kürzer werden.

Neben dem Vergleich liegt ein Mehrwert des Text Mining in der Chance, längere Zeitverläufe oder „Themenkarrieren“ darzustellen. So ist es möglich, die Entwicklung eines Themas in den Medien über lange Zeiträume hinweg lückenlos zu dokumentieren. Themenkarrieren abzubilden, zu beschreiben und zu vergleichen – das ist die Aufgabe des Kommunikationswissenschaftlers und seines Teams. Statistiker Rahnenführer erhofft sich von der Zusammenarbeit vielmehr, Standards für die Analyse von großen Datenmengen in den Kommunikationswissenschaften zu schaffen. Damit wäre es möglich, aus großen Textsammlungen nachvollziehbar und vor allem zuverlässig reproduzierbar Informationen zu gewinnen, die nicht von der Tagesform einer Wissenschaftlerin oder eines Wissenschaftlers abhängen.

Die Vision von der europaweiten Echtzeit-Analyse

DoCMA, das virtuelle Institut, hat inzwischen mehrere Promotionsarbeiten hervorgebracht, diverse Fachartikel veröffentlicht und Abschlussarbeiten von Studierenden ermöglicht. Aktuell arbeitet ein DoCMA-Team um Jentsch und Müller zusammen mit den Wirtschaftsforschungsinstituten RWI (Essen) und IW (Köln) im Auftrag des Bundeswirtschaftsministeriums an der Entwicklung neuartiger, medienbasierter Konjunkturindikatoren. Dadurch könnten Prognosen treffsicherer und schneller werden, so die Hoffnung der beiden Forscher.

Neue Forschungsfragen fallen dem Team noch viele ein, etwa die vergleichende Echtzeit-Analyse der Medienberichterstattung in verschiedenen europäischen Ländern. Mit Text Mining könnte man zu jedem Zeitpunkt herausfinden, welche Themen die einzelnen Nationen gerade bewegen. In der Klimadebatte etwa hat Polen ganz andere Interessen als Deutschland. Wo ist der gemeinsame Nenner? Gibt es überhaupt einen? „Unserer europäischen Demokratie fehlt die gemeinsame Öffentlichkeit“, so Müller, „stattdessen haben wir unterschiedliche Öffentlichkeiten, die einander kaum wahrnehmen. Und selbst bei gemeinsamen Herausforderungen wie der Eurokrise oder den Flüchtlingen haben wir unterschiedliche Diskurse. Unser Instrumentarium ist wie gemacht dafür, Diskursräume zu vergleichen – wir könnten dadurch zu einem besseren Verständnis zwischen den Ländern kommen und den Raum der Argumente vereinheitlichen“, schwärmt er. So könnten am Ende nicht nur Journalistik, Statistik und alle anderen Fächer profitieren, die mit Textdaten arbeiten – sondern ganz Europa.

Katrin Pinetzki