Tagungsbericht Mediendokumentation 2020

Marlies Ockenfeld*

Ohne Dokumentation geht's nicht

5. Symposium zur wissenschaftlichen und Mediendokumentation am 4. November 2020

<https://doi.org/10.1515/iwp-2020-2138>

Das Volontariat zur wissenschaftlichen Dokumentation ist eine zweijährige postgraduale Ausbildung, die seit 2015 kooperativ von der h_da mit Partnern durchgeführt wird, darunter der Großteil der ARD, das ZDF, RTL, die FAZ, Deutsche Welle, Deutschlandradio, das DRA, Leibniz-Forschungsinstitute wie das FIZ in Karlsruhe oder das DIPF in Frankfurt am Main und weitere Medien und Verbände. Diese entsenden die Volontäre für den akademischen Teil. Aktuell liegt der Schwerpunkt in der Ausbildung von Elitenachwuchs für die Mediendokumentation und im Forschungsdatenmanagement. Das kontinuierlich weiterentwickelte Programm, ein sogenanntes „lebendes Curriculum“, wird vom Programmdirektor Prof. Geribert E. Jakob geleitet und auch planerisch wie technisch zusammen mit etwa 15 Dozenten im Wesentlichen aus den Studiengängen der Information Science an der h_da, durchgeführt. Im Rahmen dieses Volontariats fand am 4. November 2020 zum 5. Mal in ununterbrochener Folge das Symposium zur wissenschaftlichen und Mediendokumentation der Hochschule Darmstadt statt. Die Tagesveranstaltung mit sieben Vorträgen und jeweils anschließender Diskussion wurde 2020 als Web-Konferenz durchgeführt. Knapp 60 Rechner waren zugeschaltet, Fragen an den Moderator Prof. Geribert Jakob, der immer wieder den Bezug zum Medienbereich und zur Dokumentation herstellte, konnten entweder per Mikrofonzuschaltung mündlich oder per Chatfunktion schriftlich gestellt werden.

Empfehlungssysteme

Im einleitenden Vortrag gab Midhal Blazevic, Masterabsolvent aus Darmstadt und angehender Doktorand, einen Überblick über Funktionsweise und Einsatzgebiete

Abbildung 1: Prof. Geribert Jakob moderierte das Symposium zur wissenschaftlichen und Mediendokumentation als Online-Konferenz (Screendump: Marlies Ockenfeld).

von modernen Empfehlungssystemen, die er mit anschaulichen Grafiken erläuterte. Er unterschied nicht-personalisierte Empfehlungssysteme, die etwa die zehn meistverkauften Smartphones anzeigen, halb-personalisierte, die den Standort des Geräts als Filter nutzen, und personalisierte, die aufgrund der gesammelten Nutzungsdaten maßgeschneiderte Empfehlungen geben. Dabei kommen unterschiedliche Methoden zum Einsatz, Collaborative Filtering bei der die Ähnlichkeit zwischen zwei Personen berechnet wird, Content based Filtering, bei dem gleichartige Produkte und deren Bewertungen Basis der Empfehlung sind, und Kombinationen beider Verfahren, das Hybrid Filtering. Anspruchsvoller und teurer sind wissensbasierte Systeme, die eine Domain-Ontologie nutzen, um entsprechend der angegebenen Vorlieben passende Empfehlungen zu machen. Bei Community-basierten Systemen kommt es häufiger zu nicht passenden Empfehlungen, weil sie lediglich Daten über das gesellschaftliche Umfeld sammeln und etwa Produkte empfehlen, die unseren Netz-Freunden gefallen. Demographische Systeme schließlich berechnen die Wahrscheinlichkeiten aufgrund demographischer Informationen über die Person. Bei der Analyse des Nutzerverhaltens muss der Datenschutz beachtet werden.

Dokumente, wie etwa Rundfunksendungen oder Zeitschriftenbeiträge, haben die Besonderheit, dass sie mittels linguistischer Werkzeuge analysiert und zusammenge-

*Kontaktperson: Marlies Ockenfeld, Deutsche Gesellschaft für Information und Wissen e.V., Windmühlstraße 3, 60329 Frankfurt am Main, E-Mail: ockenfeld@dgi-info.de, <https://orcid.org/0000-0002-6354-9610>

fasst werden können, um sie anschließend beim Content based Filtering einzusetzen, um etwa auf inhaltlich ähnliche Dokumente hinzuweisen.

Online-Only-Videos

Um die Archivierung von Online-Only-Videos ging es im Vortrag von B. Lichtenthaeler vom SWR. Aufgrund des neuen Telemedien-Staatsvertrags wird „Online only“ nicht mehr die Ausnahme, sondern zunehmend die Regel sein. Online Content gehört zum Programmvermögen und seine langfristige Sicherung gewinnt eine immer größere Bedeutung, nicht nur aus programmlicher, sondern auch aus rundfunkhistorischer Sicht und aus Gründen der Rechtssicherung. Herausforderungen sind u. a. das andere Codierungsformat und je nach Ausspielweg unterschiedliche Seitenverhältnisse. Es gibt kein automatisiertes Online-Monitoring, und anders als beim linearen Programm, weder Programmplanung noch zentrale Zuständigkeit. Es sind daher manuelle Eingaben in die Fernsehdatenbank FESAD erforderlich. Durch eine Dienstanweisung muss sichergestellt werden, dass das Material mit Metadaten versehen in einer archivierungsfähigen Qualität zugeliefert wird. Die Entscheidung über die Archivierung fällt im Videoarchiv. Derzeit wird ein Arbeitsprozess modelliert und implementiert, der auf Seiten der Archivierung das CMS Sophora zur automatischen Übernahme der Metadaten einsetzt und über einen VMS-Uploader einen Media Loopster genannten Massenspeicher für die Videos speist sowie die Metadaten in FESAD einträgt. Die Bereitstellung erfolgt dann über einen Multimedia Share, der sowohl das Fernsehproduktionssystem beliefern kann als auch andere Zielsysteme.

Erprobt wird das System bisher mit drei Formaten, darunter swr heimat. Nach und nach sollen weitere hinzukommen. Ferner sollen Datenfelder für weitere Mitwirkender angelegt und die ARD Normdatenbank eingebunden werden.

Fake News

Nach einem allgemeinen Übersichtsvortrag von Reinhard Karger (DFKI) über beispielhafte KI-Anwendungen und ihre Beschränkungen bei Bilderkennung, der maschinellen Übersetzung, in Chatbots, in der produzierenden Industrie und im Gesundheitssektor zur Früherkennung von Demenz aufgrund linguistischer Analysen, befasste sich Mina Schütz, Masterabsolventin der Hochschule Darmstadt und jetzt Promotionsstudentin, in ihrem Beitrag mit der

Erkennung von Nachrichten, die absichtlich und nachweislich falsch sind. Es bestehe ein hoher Bedarf an Werkzeugen, die es Redakteuren und Nutzern ermöglichen, derartige Nachrichten zu identifizieren. Schwierig sei die Behandlung von Satire und Scherzen. Ihre Arbeitsdefinition für Fake News ist deshalb umfassend: „Artikel, die eine verzerrte Sicht der realen Welt propagieren, unabhängig von der Absicht, die dahintersteckt“. Diese Absichten können psychologisch und gesellschaftlich sehr unterschiedliche sein. Bei der Verbreitung von Falschnachrichten spielen z. B. nicht nur böswillige Absichten eine Rolle, sondern auch mangelndes Bewusstsein für die Verantwortung, die man beim Weiterleiten oder Teilen einer Nachricht hat, zumal wenn oft nur noch die Überschriften gelesen werden, bevor man eine Aktion ausführt.

Zum Erkennen von Falschnachrichten gibt es verschiedene Ansätze, Schütz hat eine linguistische Auswertung durchgeführt. Ein anderer, aber kaum realisierbarer Ansatz wäre der Abgleich mit einer Datenbank, die verifizierte Fakten enthält. Als Forschungsfrage für ihre Promotion will sie herausbekommen, in welchem Umfang ein vortrainiertes Sprachdatenmodell für die inhaltsbasierte Erkennung gefälschter deutschsprachiger Nachrichten nützlich ist und ob es vielversprechende Ergebnisse bei der Vorhersage der Klassifizierung von einerseits Fließtexten und andererseits nur den Titeln von Nachrichtenartikeln erzielt. Das Training für das Sprachmodell soll anhand der deutschen Wikipedia erfolgen. Die eingesetzten Transformer-Modelle arbeiten bisher wie eine Black Box, es ist nicht klar, wie die Entscheidung zustande kommt. Ein Schwerpunkt der künftigen Arbeit soll deshalb auf Explainable Artificial Intelligence ausgerichtet werden. Ziel ist es, auch Medienanstalten beim Aufdecken von Narrativen in Fake News zu unterstützen. In der Diskussion wurde deutlich, dass es nicht immer eindeutig ist, was falsch und was richtig ist, weil es auch auf die jeweilige Perspektive ankommt, etwa bei der Aussage „Die Krim ist russisch“. Wie stets liegt auch hier die letztendliche Bewertung beim Menschen, der die Information verwerten muss.

Digitaler Zwilling

Dr. Achim Steinacker (empolis) warf einen Blick in den aktuellen Stand der Entwicklung im Bereich der technischen Dokumentation für die Industrie 4.0 und zeigte, wie die bedarfsgerechte Auswahl und Aufbereitung der in einer bestimmten Situation benötigten Daten aus den riesigen Mengen an verfügbaren Sensordaten und Dokumentationen im industriellen Umfeld realisiert werden kann. Vo-

raussetzung sind intellektuell erarbeitete und gepflegte Ontologien. Dieses in einem Wissensgraphen dargestellte vollständige semantische Modell einer ganz bestimmten Maschine wird als deren digitaler Zwilling bezeichnet, ist aber momentan noch eine Wunschvorstellung. Anstatt eine umfangreiche „tote“ PDF-Datei herunterzuladen, um darin die passende Information zu finden und mit den aktuellen Maschinendaten abzugleichen, soll der Maschinenbediener mit seinem Endgerät einzelne Teile anvisieren können und dann sofort interaktiv Hinweise bekommen, was zu tun ist. Damit soll auch vorhersehbar sein, welches Maschinenteil in absehbarer Zeit ausgetauscht werden muss, Modebegriff „predictive maintenance“. Um einen Austausch zwischen den unterschiedlichen Komponenten und Beteiligten zu gewährleisten, müssen standardisierte Metadaten und herstellerunabhängige Austauschformate verwendet werden. Die Rolle des technischen Redakteurs wandelt sich zum Informationsmanager, der sich grundsätzlich dazu Gedanken macht, wie sich ein digitaler Zwilling aufbauen lässt. An die Stelle von statischen Dokumenten treten situationsangepasste Informationsprodukte, wobei aufgabenspezifische Handbücher aus dem Informationsmodell automatisch generiert werden können. Doch eine Ontologie anzulegen, ist ein enormer zeitlicher und intellektueller Aufwand und der Prozess ist nie abgeschlossen.

Clustering zur Produktionsdatenanalyse

Die Weiterentwicklung des 2019 vorgestellten Redaktionsportals erläuterte Stefan Doğanay (RTL) im nächsten Beitrag. Ab 2021 werden alle tausend journalistisch arbeitenden Personen in der neuen Firma RTL Deutschland News beschäftigt sein mit der eigenständigen Abteilung Content Management. Das Redaktionsportal ist Suchmaschine und Nachrichtenarchiv. Clustering und Produktionsdatenanalyse findet in allen Stationen des Kreislaufs von Themenfindung über Planung und Realisation, Veröffentlichen und Erfolgsmessung statt. Sämtliche Bestandteile einer MAZ, eines konfektionierten Beitrags, werden in einem Cluster zusammengeführt. Dadurch lässt sich nachvollziehen, welche Schnipsel für einen Beitrag verwendet worden sind, und es lassen sich gezielt Lizenzbestimmungen anwenden, etwa Beiträge sperren, wenn die Lizenz aufgebraucht ist, oder die Lizenzgebühren abrechnen. Außerdem entstehen auf diese Art und Weise Themencontainer, die verwendet werden können, um den Redaktionen Empfehlungen für z. B. Sondersendungen zu geben. Die

Cluster erlauben auch Analysen der Geschäftstätigkeit im Sinn von Business Intelligence, etwa welche Themenschwerpunkte gesetzt worden sind. Allerdings funktioniert das nur für den Medientyp Video. Zur Rückverfolgung von Texten müssen die Videodaten mit Daten aus dem Produktionsdatensystem angereichert werden. Das Clustering wird mit linguistischen Analysemethoden unterstützt. Anschließend erfolgt eine Entitätenextraktion (Namen, Organisationen, Orte, Schlüsselbegriffe) aus den Texten zur Anreicherung der Metadaten.

In einem übersichtlichen Cockpit lassen sich dann Story-basierte Analysen anzeigen: wieviel Rohmaterial wurde produziert, wie viele Beiträge sind daraus entstanden, wieviel Agenturmaterial wurde verwendet, wieviel Archivmaterial, wieviel Eigenproduktionen, wie hoch war die Produktionseffizienz. Die Datenbasis lässt sich auch nutzen, um ein komplettes Cluster für eine Sendungsempfehlung zusammenzustellen. „Wir sind die Wächter, dass die Datenqualität hoch bleibt“, so Doğanay abschließend.

Datenforensik

Abschließend nahm Jörn Ratering, Absolvent des ersten Volontariats und jetzt seit knapp vier Jahren in der Verifikation beim ZDF tätig, das Publikum unter dem Motto „Das Böse übernimmt das Internet“ mit in die Internet-Hölle und gab einen Einblick in seine oftmals belastende Arbeit als Datenforensiker. Zum Tummelplatz für Menschen mit krassen, rassistischen und radikalen Ansichten hat sich inzwischen das 2003 gegründete 4chan entwickelt. Es handelt sich um ein einfaches bildbasiertes Schwarzes Brett, unterteilt in eine Vielzahl thematischer Foren, in denen jeder Kommentare veröffentlichen und Bilder austauschen kann, ohne sich registrieren zu müssen. 99 Prozent der Nutzer, die sich alle Anonymous nennen, sind männlich. Es herrscht in vielen Foren ein eigener vulgärer Ton, nicht nur, wie erwartbar, im Forum Politically Incorrect. Oft werden Bilder oder Kommentare binnen kurzem wieder gelöscht, sodass zur Quellensicherung alles, was eventuell verwendet werden soll, sofort archiviert werden muss. Das aus Sicht von Ratering einzige empfehlenswerte Board ist die Statistik unter 4stats.io, auf dem man nachvollziehen kann, was gerade auf 4chan viral geht. Bei Eilmeldungen ist 4chan oft schneller als Twitter, die „Arbeitssichere“ Variante ist unter 4channel.org erreichbar, die deutsche Version unter kohlchan.com. Auch pr0gramm.com ist ein deutschsprachiges Bulletin Board für Verschwörungsmythen und Radikalismus, bei dem man sich allerdings registrieren muss. Am Beispiel der Polizeimeldung über die Festnahme von mut-

maßlichen Mitgliedern sowie Unterstützern einer rechtsterroristischen Vereinigung zeigte Ratering wie er mittels professioneller, fast kriminalistischer Suchen in Facebook und ProGramm zunächst Posts des Gründers sowie den Gründungsakt der Gruppe und ein Foto sämtlicher Mitglieder aufspüren konnte.

Ein sehr hilfreiches Werkzeug für Recherchen ist Intelligence X, eine außergewöhnliche Suchmaschine die mit spezifischen Suchbegriffen wie E-Mail-Adressen, Domains, URLs, IPs, CIDRs, Bitcoin-Adressen, IPFS-Hashes usw. auch im Darknet, auf Plattformen zur gemeinsamen Nutzung von Dokumenten, in Whois-Daten, öffentlichen Datenlecks und anderen Orten sucht und die gefundenen Seiten auch in einem eigenen Archiv speichert. Ratering nutzt etwa 60 verschiedenen Werkzeuge zur Recherche und betonte, dass man genau wissen müsse, was zitierfähig und verwertbar sei und was nicht.

Telegram, das in der Vergangenheit bevorzugt von Informatikern und auf ihren Persönlichkeitsschutz bedachten Personen genutzt wurde, hat sich im Lauf der Pandemie zu einem Messenger-Dienst entwickelt, auf dem Verschwörungsmymen über Covid-19 gedeihen. Die Breitenwirkung der Telegram-Kanäle einiger prominenter Leugner der Pandemie sei erschreckend, die follower

würden unentwegt mit Propaganda befeuert. An zwei weiteren Beispielen gab Ratering einen Einblick in das Verifizieren von kompromittierenden oder mutmaßlich gestellten Fotos von Prominenten, die im Internet aufgetaucht sind. Eines der Werkzeuge, die er dazu nutzt, ist Jeffrey's Image Metadata Viewer, und wenn er Glück hat, wurden die Metadaten vor dem Hochladen der Bilder nicht gelöscht, sodass er Aufnahmedatum, Aufnahmegerät etc. feststellen kann. Gesichter lassen sich ebenfalls mit speziellen Suchmaschinen bestimmten Personen zuordnen, was jedoch im Fall von PimEyes kürzlich einiges Erschrecken ausgelöst hat, als bekannt wurde, dass dort bereits 100 Millionen Gesichter gespeichert sind. Ob PimEyes auch die Fotos speichert, die man im Zuge einer Suche hochlädt, ist unklar.

Ausblick

Insgesamt hat Geribert Jakob wieder einen anregenden und abwechslungsreichen Vortragstag organisiert und man kann bereits auf das Symposium am 5. November 2021 gespannt sein, vielleicht dann wieder auf dem Medien-campus in Dieburg mit gemeinsamen Kaffeepausen. Einige der Vortragenden aus diesem Jahr werden dann erneut dabei sein und über die aktuellen Weiterentwicklungen in ihren Arbeitsbereichen berichten.

Deskriptoren: Tagung, Hochschule Darmstadt, Volontariat wissenschaftliche Dokumentation, Medien, Recherche, Verifikation, Informationssystem, Redaktionssystem, Daten

President Donald J. Trump works in the Presidential Suite at Walter Reed National Military Medical Center in Bethesda, Md. Saturday, Oct. 3, 2020, after testing positive for COVID-19. (Official White House Photo by Joyce N. Boghosian)

President Donald J. Trump works in his conference room at Walter Reed National Military Medical Center in Bethesda, Md. Saturday, Oct. 3, 2020, after testing positive for COVID-19. (Official White House Photo by Joyce N. Boghosian)

04.11.2020 43

Abbildung 2: Mit Jeffrey's Image Metadata Viewer lässt sich ermitteln, dass die beiden Fotos von Präsident Trump im Abstand von nur zehn Minuten aufgenommen worden sind. (Darstellung Jörn Ratering)

Marlies Ockenfeld

Deutsche Gesellschaft für Information und Wissen e.V.

Windmühlstraße 3

60329 Frankfurt am Main

ockenfeld@dgi-info.de