

10 Consumer Health Search und die Authentizität von Informationen – Eine multidimensionale Betrachtung

Marco Viviani, Elisabeth Mess

Abstrakt: In der heutigen Gesellschaft greifen immer mehr Menschen auf online Informationen zurück, um ihren Informationsbedarf zu decken. Dieses Phänomen hat in verschiedenen Bereichen zugenommen, so auch im Bereich der Gesundheitsinformationen. Einerseits könnte sich das positiv auf die Gesundheitskompetenz der Menschen auswirken, was der Gesundheit des Einzelnen und der Gemeinschaft zugutekäme. Andererseits könnten Einzelpersonen und die Gesellschaft mit schwerwiegenden Folgen konfrontiert werden, wenn eine beträchtliche Menge an falschen Gesundheitsinformationen online kursiert. In diesem Zusammenhang verfolgt dieser Artikel zwei Ziele: (i) einen globalen Überblick über die oben genannten Probleme zu geben und (ii) aus technischer Sicht einige der aktuellen Forschungsrichtungen und Lösungen im Bereich Information Retrieval vorzustellen. Diese Technologien können Online-Nutzer:innen dabei helfen Gesundheitsinformationen zu finden die sowohl vertrauenswürdig sind, als auch ihrem Informationsbedürfnis gerecht werden.

Schlüsselwörter: Gesundheitssuche für Verbraucher, Consumer Health Search Gesundheitsfehlinformation, Authentizität von Informationen, Information Retrieval, Suchaufgabe, Mehrdimensionale Relevanz, Gesundheitskompetenz

10.1 Einleitung

In den letzten Jahren hat sich deutlich gezeigt, dass Menschen immer häufiger auf die Informationen von Online-Suchanfragen zurückgreifen [1]. Man kann Nachrichten über bestimmte Ereignisse suchen, die Meinung anderer über bestimmte Produkte oder Dienstleistungen einholen (Rezensionen) und Informationen über Krankheiten, Symptome, Behandlungen usw. suchen. Im letztgenannten Bereich beziehen wir uns speziell auf den Bereich der *Consumer Health Search* (CHS), d.h. die Suche nach Gesundheitsratschlägen im Internet, welche durch Laien durchgeführt wird [2]. Bereits 2013 wies das Pew Research Center¹, darauf hin, dass ein großer Teil der US-Bevölkerung Gesundheitsinformationen online sucht und konsultiert. Das geht sogar so weit, dass Ärzte aus dem Entscheidungsprozess ausgeschlossen werden [3]. Gesundheitsbezogene

¹ Das Pew Research Center ist ein „unparteiischer Fact Tank“, der die Öffentlichkeit über Themen, Einstellungen und Trends informiert, welche die Welt prägen.

Suchanfragen bei Google sind mittlerweile so populär, dass sich der Begriff „Dr. Google“ etabliert hat [4].

Bei einer Eurostat Erhebung [5] zeigte sich, dass insgesamt 52 Prozent der EU-Bürger nach gesundheitsbezogenen Informationen suchen (Bsp.: Verletzungen, Krankheiten, Ernährung etc.). Zu den Spitzenreitern zählen beispielsweise Finnland mit 81 Prozent, die Niederlande mit 78 Prozent und Dänemark mit 71 Prozent. In Deutschland sind es hingegen nur 37 Prozent. Für weitere Informationen siehe Abb. 10.1.

Individuals who sought health information online, 2022

(% of individuals)

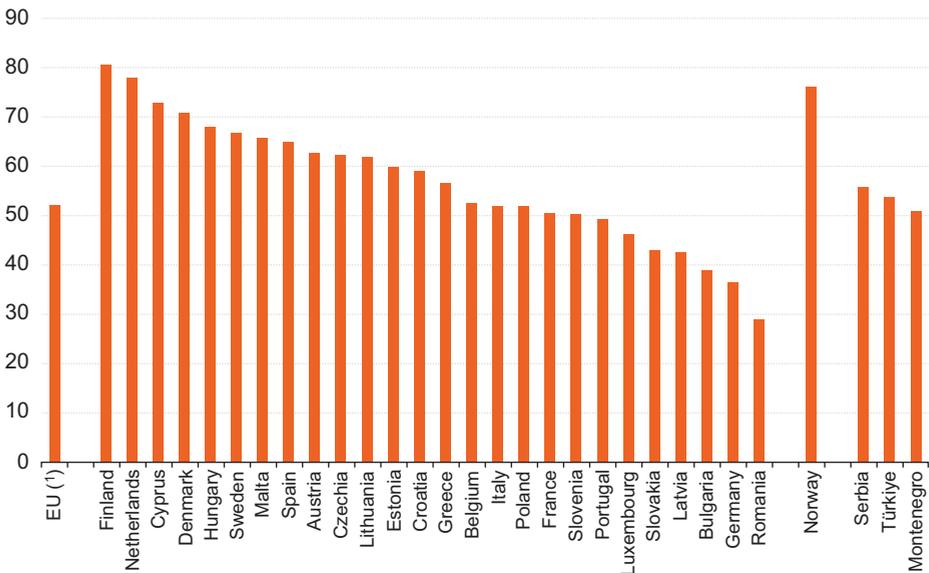


Abb. 10.1: Auszug aus der Eurostat-Erhebung wie viele Personen im Jahr 2022 online nach Gesundheitsinformationen gesucht haben [nach 5].

In der arabischen Welt nutzten rund 86 Prozent der Befragten Gesundheitsinformationen aus dem Internet. Die Recherchen reichten dabei von der Ernährung bis hin zu den Nebenwirkungen des Rauchens [6]. Ähnliche Studien zeigten eine Nutzung von 87 Prozent in China [7] und 92 Prozent in Indien [8]. In Australien sind etwa 17 Millionen Australier:innen aktiv online. Davon haben fast 80 Prozent nach Gesundheitsinformationen im Internet gesucht [9].

Diese Entwicklung bringt jedoch nicht nur positive Aspekte mit sich. Ein großer Teil der online verbreiteten Informationen erweist sich als falsch oder irreführend [10]. Gerade in einem so sensiblen Bereich wie dem Gesundheitswesen kann das schwerwiegende Folgen haben. Gurus, die „Wunderheilungen“ von Krebs oder anderen Krankhei-

ten versprechen, oder die Menge an *Fake News*² die in den letzten Jahren im Zusammenhang mit COVID-19 aufgetaucht sind. Zwei von vielen Beispielen die mit schwerwiegenden Folgen für den Einzelnen und die Gesellschaft einhergehen können [11, 12]. Hinzu kommt, dass Laien aufgrund ihrer mangelnden Gesundheitskompetenz – im Englischen als *health literacy* [13, 14] bezeichnet – in den meisten Fällen nicht in der Lage sind, richtige von falschen Gesundheitsinformationen zu unterscheiden. Der Begriff *health literacy* wurde 1988 in das Glossar der Weltgesundheitsorganisation (WHO) aufgenommen und bezeichnet „[...] *the ability of a citizen to obtain, process, and understand basic health information in order to make informed choices [...]*“ [15]. Zu Deutsch: die Fähigkeit eines Bürgers, grundlegende Gesundheitsinformationen zu recherchieren, zu verarbeiten und zu verstehen, um fundierte Entscheidungen treffen zu können.

Eine mögliche Lösung könnte beispielsweise sein, dass klinische Expert:innen die verfügbaren Informationen im Internet überprüfen. Aufgrund der großen Menge (Volume) an verfügbaren Informationen und der Geschwindigkeit (Velocity) in welcher diese generiert werden, ist es für klinische Expert:innen unmöglich jede einzelne Information zu überprüfen und zu bewerten. Deshalb ist es notwendig, automatisierte Lösungen oder Werkzeuge zu entwickeln, die es auch Laien ermöglichen, richtige von falschen Informationen zu unterscheiden. Nur so können möglicherweise gesundheitsgefährdende Verhaltensweisen vermieden werden [10]. In diesem Beitrag sollen besonders auf gesundheitsbezogene Fehlinformationen fokussiert werden, die aufgrund fehlender wissenschaftlicher Evidenz als Tatsachenbehauptungen im Internet verbreitet werden und derzeit als falsch gelten [16]. In diesem Zusammenhang konzentrieren wir uns auf das Forschungsgebiet des *Information Retrieval (IR)*, zu Deutsch: Informationsbeschaffung, und insbesondere auf den Bereich der *Consumer Health Search (CHS)*. Durch die Betrachtung beider Bereiche ist es möglich, Lösungen für die Beschaffung von Gesundheitsinformationen zu entwickeln, die sowohl die aktuelle Relevanz gesundheitsbezogener Inhalte als auch die Authentizität der Informationen selbst berücksichtigen.

10.2 Hintergrund

Es gibt mehrere Forschungsbereiche, welche sich mit der Entwicklung von automatisierten Lösungen für den Zugang und Abruf von Informationen beschäftigen. Derartige Systeme werden in den Bereichen Datenbanken [17], Information Retrieval [18] und Informationsfilterung [19] entwickelt. Da sich dieser Artikel auf Information Retrieval-Lösungen konzentriert, die den Zugang, die Suche und auch die Glaubwürdigkeit von Gesundheitsinformationen umfassen, geben wir in Abschnitt 2.1 eine kurze Einführung in diese Disziplin. Ebenfalls geben wir eine Übersicht über den Ablauf einer gesundheitsbe-

² Zu Fake News zählen falsche oder irreführende Informationen, welche mit böswilliger Absicht in sozialen Netzwerken veröffentlicht und verbreitet werden.

zogenen Suchanfrage. In Abschnitt 2.2 beleuchten wir, was unter der Glaubwürdigkeit von Informationen zu verstehen ist, wobei auch hier der Schwerpunkt auf dem Gesundheitsbereich liegt. In Kapitel 10.3 und 10.4 werden Nicht-IR-Lösungen und IR-Lösungen vorgestellt.

10.2.1 Information Retrieval

Information Retrieval (IR) ist ein Konzept, welches auf die 1950er Jahre zurückgeht. Die Disziplin etablierte sich jedoch erst in den 1970er Jahren. Im Laufe der Jahre wurden mehrere Definitionen von IR festgelegt. Nach [20] umfasst IR das Untersuchen, Organisieren, Speichern und die Suche nach Daten und Informationen; oder, wie es in [18] heißt: „[...] *IR is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need within large collections (usually stored on computers) of data and information used to search and develop tasks [...].*“ Das bedeutet, dass IR sich vor allem auf unstrukturierte Informationen konzentriert, die ein bestimmtes Informationsbedürfnis erfüllen sollen. Das Ziel von IR ist daher die Entwicklung von *Information-Retrieval-Systemen* (IRS), oder einfacher – Suchmaschinen. Diese sollen den Nutzern dabei helfen, *zuverlässige* und *wertvolle* Informationen zu finden. Das bedeutet wiederum, dass die Informationen für die Bedürfnisse der Nutzer *relevant* sind [21]. Der Aufbau bzw. die Struktur solcher Systeme basiert auf der Lösung eines Entscheidungsproblems: *Wie können Informationen identifiziert werden, die den Präferenzen des Nutzers entsprechen?*

Um diese Frage zu beantworten, ist es notwendig, sowohl den Inhalt der Informationen (z. B. Texte, Bilder, Videos, Audios), als auch die Bedürfnisse des Nutzers zu interpretieren, die in der Regel durch eine *Suchanfrage* ausgedrückt werden. Die Struktur einer einfachen Suchanfrage (*query*) ist in Abb. 10.2 dargestellt. Hier werden zwei Sammlungen verglichen. Es gibt eine Sammlung von Dokumenten, welche in einer Indexierungsphase identifiziert wurde (*Indexing*) und eine weitere Sammlung, die nach einer Suchanfrage verarbeitet wird (*Query Processing*). Diese beiden Sammlungen werden miteinander verglichen (*Matching*) und basierend auf ihrer Relevanz und der Suchanfrage eingestuft.

Der Begriff *Relevanz* wurde ursprünglich mit dem Wort *Aktualität* gleichgesetzt. Aktualität meint die Beurteilung, ob ein Dokument mit dem Themenbereich des Informationsbedarfs des Nutzers zusammenhängt. Diese Aufgabe wird in der Matching-Phase berechnet (siehe Abb. 10.2) und hierfür können verschiedene mathematische IR-Modelle genutzt werden [18]. Obwohl der Begriff Aktualität noch immer ein Relevanzkriterium darstellt, kann er heute als einer von vielen Facetten der Relevanz betrachtet werden. Man spricht hierbei von einem mehrdimensionalen Konzept der Informationsbewertung [23]. Mehrere Relevanzdimensionen (oder -kriterien) können als Merkmale betrachtet werden, die sich auf verschiedene Informationselemente beziehen. Diese Merkmale spielen bei der Bewertung der Nützlichkeit der Information für den Nutzer eine wichtige Rolle.

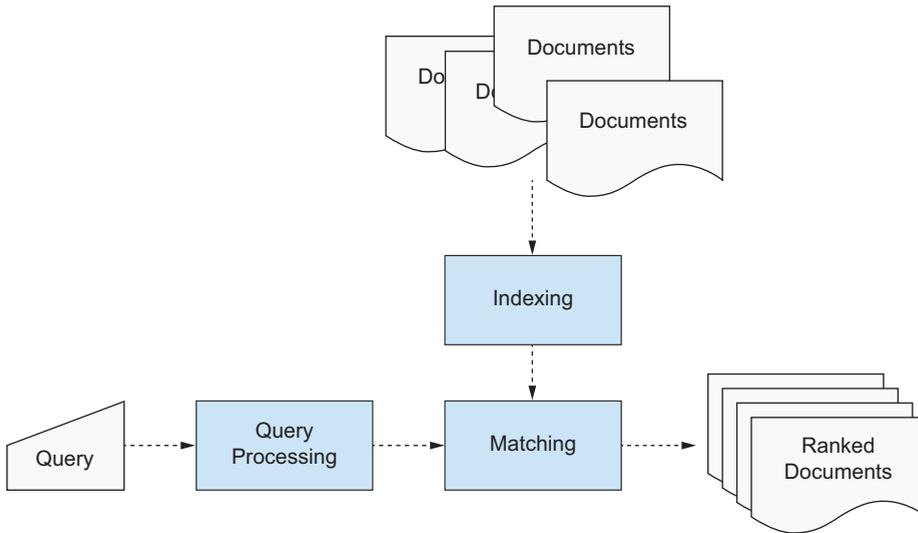


Abb. 10.2: Übersicht einer Architektur einer einfachen Suchanfrage basierend auf [modifiziert nach 22].

Mit Informationselement ist die kleinste abrufbare Informationseinheit (d. h. im Allgemeinen ein Dokument) gemeint. Diese kann im Falle einer Websuche eine Webseite, bei einer Suche in sozialen Medien (Social Media Information Search) ein Post oder bei einer akademischen Suche ein wissenschaftlicher Artikel sein.

Die Relevanzdimension Aktualität, im Englischen auch *topicality* genannt, bezeichnet eine Suchanfrage abhängig von der Eigenschaft eines Artikels. Es drückt aus, wie sehr der Inhalt einer Information thematisch der Suchanfrage ähnelt. Beispiele für weitere Relevanzdimensionen sind *recency* (Bsp.: Wie aktuell ist eine Nachricht?), *popularity* (Bsp.: Wie oft wurde ein Beitrag geteilt?) und *understandability* (Bsp.: Wie gut ist ein Zeitschriftenartikel für einen Laien lesbar?). *Recency* und *popularity* sind hierbei nicht von der Suchanfrage abhängig, während *understandability* vom Nutzer abhängig ist.

Die globale Relevanz einer Information für einen Nutzer kann also auf der Grundlage einer Teilmenge verschiedener Relevanzdimensionen bewertet werden, die je nach Suchanfrage ausgewählt werden können.

Es gibt verschiedene Arten von Suchaufgaben: von einfachen Navigations- oder Faktenfindungsaufgaben bis hin zu komplexeren oder domänenabhängigen Suchaufgaben, wie z. B. die Suche nach Produktrezensionen, die Suche von Nachrichten, die Literatursuche und die Suche nach Gesundheitsinformationen für Verbraucher (Consumer Health Search, CHS). In Bezug auf die letztgenannte Suchaufgabe (CHS), die wir in diesem Artikel untersuchen, sind weitere Dimensionen der Relevanz laut [24, 25] von Bedeutung. Hierzu zählen beispielsweise *supportiveness* (z. B.: Hilft mir die In-

formation Entscheidungen zu treffen?), die *credibility* (Bsp.: Wie glaubwürdig ist die Information?), wie auch die *usefulness* (z.B.: Wie nützlich ist die Information für mich?).

10.2.2 Die Echtheit von Informationen

In den letzten Jahren wurde der Wunsch immer größer, die Verbreitung von falschen Informationen in ihren unterschiedlichen Formen im Internet (Fake News, Gerüchte, Scherze etc.) einzudämmen. Ein Begriff, der hierbei in der englischen Auseinandersetzung mit dem Thema häufig auftaucht, ist *information disorder* [26]. Damit ist die Überflutung des Internets mit Fehlinformationen gemeint.

Bisher wurden bereits unterschiedliche Lösungsansätze entwickelt, welche die Echtheit von Informationen bestimmen sollen. In einigen aktuellen Arbeiten wird unter anderem auf die Dimensionen *reliability* (Zuverlässigkeit), *truthfulness* (Wahrhaftigkeit), *trustworthiness* (Vertrauenswürdigkeit), *credibility* (Glaubwürdigkeit) und *veracity* (Datenqualität) verwiesen. Diese Begriffe können jedoch unterschiedlich interpretiert werden – je nachdem ob sie sich auf die Informationsquelle, die Information selbst, das Kommunikationsmedium oder andere theoretische Aspekte beziehen.

Darüber hinaus gibt es beispielsweise auch unterschiedliche Interpretationen von falschen Informationen. So wird im Englischen beispielsweise unterschieden zwischen *Desinformation*, *Misinformation* und *Malinformation*. Der Begriff *Desinformation* beschreibt falsche Informationen, die mit einer böswilligen Absicht verbreitet werden. Mit der *Misinformation* werden falsche Informationen kategorisiert, die ohne böswillige Absicht verbreitet wurden. Als *Malinformation* werden die Informationen bezeichnet, die zwar richtig bzw. korrekt sind, jedoch weitergegeben werden, um Schaden anzurichten. Mit letzterem sind die Informationen gemeint, die privat bleiben sollten, aber dennoch an die Öffentlichkeit gelangen (im Englischen *Doxing* genannt) [26].

In diesem Artikel konzentrieren wir uns auf den Begriff der *Misinformation* – also falsche Informationen, die ohne böswillige Absicht verbreitet werden. In den folgenden Kapiteln 3 und 4 werden einige Projekte vorgestellt, welche für das Problem der *Misinformation* Lösungen entwickelt haben oder vorschlagen.

10.3 Nicht-IR-Lösungen

Ein großer Teil der Nicht-IR-Lösungen geht auf Arbeiten zurück, die sich mit dem Problem der Erkennung von *Misinformation* im Allgemeinen befasst und ihre Lösungen dann auf den Gesundheitsbereich übertragen. Dabei handelt es sich hauptsächlich um Arbeiten, welche die Erkennung von *Misinformation* als eine binäre Klassifizierungsaufgabe (*binary classification*) behandeln [10]. Sprich, die Informationen werden nach richtig oder falsch kategorisiert.

Andere Arbeiten konzentrieren sich mehr auf die Besonderheiten des Gesundheitsbereichs und entwickelten Ad-hoc-Lösungen. Zu den Besonderheiten zählen beispielsweise das gleichzeitige Vorhandensein von Informationen mit sehr guter oder sehr schlechter Qualität, die Verwendung einer spezifischen und von der Fachdisziplin abhängigen Sprache wie auch die Möglichkeit sich auf Expertenwissen zu stützen. Diese Lösungen basieren auf dem Input von gesundheitsbezogenen Inhalten in Form von Webseiten [27, 28] und Posts aus den sozialen Medien [29–31].

Da Gesundheitsinformationen durch Expert:innen überprüft werden können, wird derzeit untersucht, wie wissenschaftliche Beweise im Zusammenhang mit der Erkennung von Fehlinformationen im Gesundheitswesen automatisch genutzt werden können. In diesem Zusammenhang haben einige Studien bereits Expert:innen und Laien in die Bewertung der Echtheit von Gesundheitsinformationen einbezogen.

DISCERN [32] ist zum Beispiel ein kurzer Fragebogen, der im Rahmen des DISCERN-Projekts entwickelt wurde. Dieser soll den Nutzern eine mögliche, zuverlässige Methode zur Bewertung der Qualität schriftlicher Informationen über Behandlungsmöglichkeiten für ein bestimmtes Gesundheitsproblem anbieten. Die Bewertung der Informationen erfolgt mithilfe einer Bewertungsskala von 1–5. Ein Auszug des Fragebogens ist in Abb. 10.3 zu sehen.

Die Stiftung *Health on the Net Foundation* (HON) [34] hat einen Verhaltenskodex und ein Qualitätssiegel für medizinische Webseiten herausgegeben. Der Kodex arbeitet mit verschiedenen Attributen, wie Offenlegung der Autorenschaft, Quellennachweis, Aktualisierung der Informationen, Offenlegung der Redaktions- und Veröffentlichungsrichtlinien sowie Vertraulichkeit und Datenschutz.

Der als HC-COVID [35] bekannte Ansatz konzentriert sich auf die Erkennung von COVID-bezogenen Fehlinformationen im Gesundheitsbereich, wobei ein auf Crowdsourcing basierendes Wissensnetz als Grundlage genutzt wurde. Dieses Wissensnetz wird durch Expert:innen und Nicht-Expert:innen erstellt und überprüft.

Alle drei Ansätze (DISCERN, HON, HC-COVID) haben den Nachteil, dass sie auf ein hohes Maß an menschlicher Intervention angewiesen sind. Dazu gehört beispielsweise die manuelle Zuweisung von Qualitätsindikatoren zu jedem neuen Inhalt (Annotation), die Rekrutierung von Expert:innen und Nicht-Expert:innen und die Gewährleistung der Qualität der Kommentare.

Andere Studien definierten evidenzbasierte, gesundheitsbezogene Konzepte durch die Verwendung von Ontologien (geordnete Darstellung einer Menge von Begriffen) oder *knowledge graphs* (Deutsch: Systematik, nach der Informationen gesucht und miteinander verknüpft werden, auch Wissensnetz genannt) und führten diese in Wissensdatenbanken zusammen. Ein einfaches Beispiel für ein Wissensnetz ist in Abb. 10.4 zu sehen.

In dem Projekt MedFact [30] entwickelten die Autoren einen Algorithmus, zur Überprüfung von Beiträgen aus den sozialen Medien auf der Grundlage der evidenz-

Abschnitt I

Ist die Publikation zuverlässig?

1 Sind die Ziele der Publikation klar? Anleitung zur Bewertung					
	Nein		Teilweise		Ja
	1	2	3	4	5

Hinweis: Suchen Sie am Anfang der Publikation nach klaren Angaben:

- worum es in der Publikation geht,
- welche Themen behandelt werden (und welche nicht behandelt werden),
- wer die Publikation hilfreich finden könnte.

Wenn Sie Frage 1 mit "Nein" beantwortet haben, gehen Sie gleich zu Frage 3.

2 Erreicht die Publikation ihre selbstgesteckten Ziele? Anleitung zur Bewertung					
	Nein		Teilweise		Ja
	1	2	3	4	5

Hinweis: Bedenken Sie, ob die Publikation die Informationen enthält, die den in Frage 1 dargestellten Zielen entsprechen.

3 Ist die Publikation für Sie bedeutsam? Anleitung zur Bewertung					
	Nein		Teilweise		Ja
	1	2	3	4	5

Hinweis: Bedenken Sie, ob:

- die Publikation die Fragen anspricht, die die Leser stellen könnten,
- Empfehlungen und Vorschläge, die die Behandlungsalternativen betreffen, realistisch oder angemessen sind.

Abb. 10.3: Auszug des DISCERN Fragebogens [vgl. 3].

basierten Medizin [37]. Das bedeutet die Integration von individueller klinischer Expertise mit der besten verfügbaren externen klinischen Evidenz aus systematischer Forschung und vertrauenswürdigen medizinischen Informationsquellen wie der Datenbank *Turning Research Into Practice* (Trip).³

In einem anderen, recht aktuellen Modell namens DE-TERRENT [38], konzentrieren sich die Autoren auf die Identifizierung erklärbarer Fehlinformationen im Gesundheitswesen. Hierfür wird ein medizinischer Wissensgraph namens *Know-*

³ Trip ist eine klinische Suchmaschine, die es den Nutzern ermöglichen soll, schnell und einfach hochwertige Forschungsergebnisse zu finden und diese für ihre Praxis zu nutzen. URL: <https://www.tripdatabase.com/Home>.

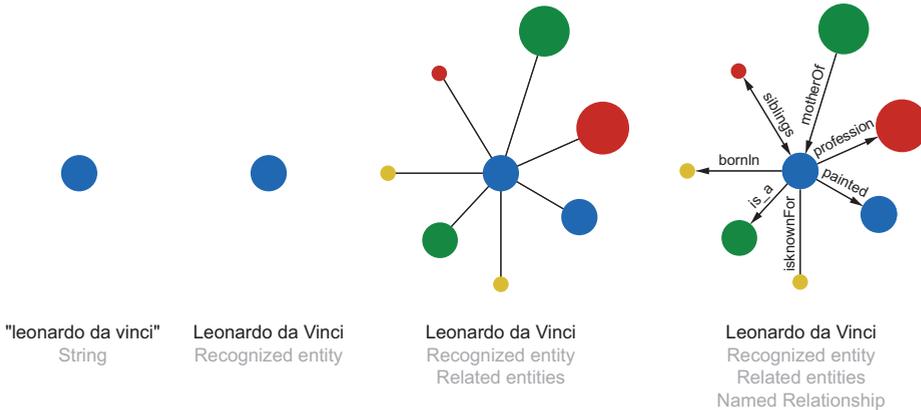


Abb 10.4: Ein einfaches Beispiel wie ein Wissensnetz aufgebaut ist [vgl. 36].

life [39] genutzt. Dieser greift auf medizinische Inhalte aus PubMed⁴, und anderen Gesundheitsportalen wie Mayo Clinic,⁵ RxList,⁶ dem Wikipedia-Medizinportal,⁷ und MedlinePlus⁸ zurück.

Die potenziellen Nachteile der letztgenannten Ansätze (MedFact, DETERRENT) vor allem in der sehr komplexen Formalisierung der Datenbank (Beschreibung und Darstellung der Informationen). Ein automatischer Aufbau einer solchen Datenbank, inklusive ständiger Aktualisierung der Inhalte, wäre wünschenswert, stellt jedoch aktuell eine zu große Herausforderung dar.

10.4 IR-Lösungen

Die bisher dargestellten Lösungen in Kapitel 3 zeigen auf, wie komplex das Problem der Informationsbeschaffung im Gesundheitswesen ist. Bisher hat jedoch keine der genannten Arbeiten eine befriedigende Lösung gefunden, welche die Relevanz von Informationen bewerten und darauf aufbauend z. B. eine Priorisierung vornehmen kann. Hinzu kommt, dass sich erst seit kurzem mehr und mehr Arbeiten auch mit der

⁴ PubMed ist eine Datenbank der National Library of Medicine und umfasst biomedizinische Literatur, biowissenschaftliche Fachzeitschriften und Online-Bücher. URL: <https://pubmed.ncbi.nlm.nih.gov/>.

⁵ Mayo Clinic ist ein englischsprachiges Onlineportal durch welches Arzttermine gebucht und auch Informationen zu Krankheiten gesucht werden kann. URL: <https://www.mayoclinic.org/>.

⁶ RxList ist eine Suchmaschine mit welchen Informationen zu Medikamenten und Nahrungsergänzungsmitteln erhalten werden können. URL: <https://www.rxlist.com/>.

⁷ Portal:Medizin ist eine Wissensdatenbank der freien Enzyklopädie Wikipedia. URL: <https://de.wikipedia.org/wiki/Portal:Medizin>.

⁸ MedlinePlus ist eine online Gesundheitsdatenbank und Teil der National Library of Medicine. URL: <https://medlineplus.gov/>.

Authentizität von Informationen beschäftigen. Eine Reihe von Beispielen wird im Folgenden aufgezeigt.

(1) Vera [40] – eine Softwarelösung, die hilfreiche und womöglich schädliche Informationen identifiziert, indem sie eine mehrstufige Rankingstruktur durchläuft. Hier werden Dokumente, die aufgrund ihres Contents mit dem BM25Model [41] als hoch relevant eingestuft werden, mit den Modellen mono-T5 und duo-T5 [42] nochmals klassifiziert. Dabei werden innerhalb der Dokumente die Passagen mit der höchsten Relevanzwahrscheinlichkeit herausgesucht. Anschließend wird ein *Label-Prediction-Model* zur Kategorisierung trainiert. Hierfür werden die Track Daten der Evaluierungsinitiative *TREC 2019 Medical Misinformation* [43] verwendet, um die Glaubwürdigkeit der Dokumente zu überprüfen und sie anhand dieses Kriteriums neu zu klassifizieren.

(2) In [44] berücksichtigen die Autoren neben der mit BM25 berechneten Relevanz eine weitere Dimension – die Informationsqualität. In dieser Arbeit wird die Qualitätsschätzung durch Training eines Multi-Label-Klassifikators durchgeführt, der einen probabilistischen Score für zehn Qualitätskriterien liefert (z. B. „Wird der Nutzen der Intervention in der Darstellung angemessen quantifiziert?“, „Wird über die Verfügbarkeit der Behandlung/des Tests/des Produkts/des Verfahrens aufgeklärt?“, usw.) [45].

(3) Speziell das RoBERTabase Model [46] wird auf den in [47] vorgestellten *Health News Review*-Datensatz trainiert, das nach bestimmten Qualitätskriterien gekennzeichnet ist. Sobald eine eindeutige Rangfolge auf der Grundlage von Bewertungen der thematischen Relevanz und der Informationsqualität vorhanden ist, werden diese mithilfe von *Reciprocal Rank Fusion* [48] zusammengeführt.

(4) Die in [49] beschriebene Arbeit zu Suchanfragen in sozialen Medien verwendet das *Query-Likelihood-Model* [50] zur Berechnung der thematischen Relevanz, einen *Multi Criteria Decision Making* (MCDM)-Ansatz zur Berechnung der Authentizität von Informationen [51] und eine einfache lineare Kombination, um den endgültigen Relevanzwert zu erhalten.

Leider haben auch diese Arbeiten (1–4) wesentliche Schwächen: (i) Datensätze müssen initial gelabelt werden, um den Authentizitätswert der Informationen berechnen zu können [42, 44]. Diese Datensätze sind jedoch nicht immer verfügbar oder beinhalten bestimmte Vorurteile (*biases*) bzw. Verzerrungen die abhängig von der jeweiligen Domäne und auch der subjektiven Einschätzung der menschlichen Gutachter sind. (ii) Die menschliche Intervention hat somit einen starken Einfluss auf das Berechnungsmodell und damit folglich auch auf die Authentizität der Informationen [49]. Ein möglicher Lösungsansatz für diese Probleme könnte der Vergleich von Gesundheitsinformationen aus verschiedenen Gesundheitsdokumenten und medizinischen Zeitschriftenartikeln sein. In einem IR-Modell könnte dann, basierend darauf, die Authentizität der Information berechnet werden [52]. Diese Informationen müssen jedoch aus zuverlässigen Quellen mit wissenschaftlichen Beweisen für eine bestimmte Suchanfrage stammen. Durch diesen Vergleich können zwei anfrageabhängige Relevanzwerte (Aktualität und Echtheit der Information) durch eine geeignete Gruppierung zusammengeführt und ein allgemeiner Re-

levanz-Score ermittelt werden. Hierfür sind jedoch wieder menschliche Intervention, komplexere Wissensdatenbanken oder gelabelte Datensätze notwendig.

10.5 Experimentelle IR-Evaluierungsinitiativen

Zur Bewertung von IR-Modellen oder -Konzepten gibt es allgemeine Bewertungskriterien die bei der Evaluation berücksichtigt werden können. Hierzu zählen beispielsweise,

- *Retrieval effectiveness* (Standard-IR-Evaluation), bezogen auf die Relevanz der Suchergebnisse,
- *System quality*, bezogen auf die Indizierungsgeschwindigkeit (z. B.: Wie viel Zeit für das Sammeln, Parsen und Speichern von Daten erforderlich ist),
- *Search speed* (die Suchgeschwindigkeit),
- *Coverage* (Umfang und Vielfalt der Dokumentensammlung),
- *Expressiveness* (Ausführlichkeit der Sprache über welche die Suchanfrage gemacht wurde),
- *User utility* (Nutzen für den Benutzer), erfasst durch die *Zufriedenheit* der Benutzer auf der Grundlage von *Relevanz*, *Geschwindigkeit*, und *Benutzeroberfläche*, *Benutzerrücklaufquote* und *A/B-Testings* (geringfügige Änderungen an einem System sind für einen bestimmten Teil der Nutzer sichtbar).

Da eine benutzerbasierte Evaluierung aufwendig und teuer ist, wird am häufigsten die Effektivität anhand der Relevanz der abgerufenen Dokumente bewertet. Relevanz ist jedoch ein Konzept, das eine *fließende* und nicht binäre Funktion ist. Sie ist *subjektiv* und hängt damit von der Perspektive des Benutzers ab; sie ist *kontextabhängig* und hängt von den Anforderungen des Benutzers ab; sie ist *kognitiv* und wird vom Benutzer wahrgenommen und erlebt; und sie ist *dynamisch* und verändert sich im Laufe der Zeit.

Um diese Komplexität zu minimieren, wurde das *Cranfield Evaluierungsparadigma* für IR-Systeme entwickelt [53]. Das Konzept basiert auf der Definition einer IR-Testsammlung für eine bestimmte Suchaufgabe, die sich zusammensetzt aus einer *Dokumentensammlung*, einer Menge von *Informationsbedürfnissen* (Themen) und einer Menge von *Relevanzbeurteilungen* für jedes Themen-Dokumenten-Paar, die durch menschliche Gutachter abgegeben wird. Insgesamt wird von drei Hauptannahmen ausgegangen: (i) an die Relevanz kann sich angenähert werden; (ii) eine Beurteilung durch eine einzelne menschliche Intervention kann repräsentativ für eine ganze Nutzergruppe stehen; (iii) die Listen der relevanten Dokumente sind für jedes Thema vollständig (alle relevanten Dokumente sind bekannt). Basierend auf diesen Annahmen erstellt jede durch eine IRS durchgeführte Suchstrategie eine Rankingliste für jedes Themengebiet. Diese Listen sind nach der Wahrscheinlichkeit sortiert, wie die Dokumente abgerufen werden. Hierfür gibt es mehrere Bewertungsmaßstäbe [54].

Im Kontext der *Consumer Health Search* wurden erst in jüngster Zeit einige Bewertungsinitiativen, die auf Grundlage des Cranfield-Paradigmas basieren, gegründet. Hierzu zählen die Text *REtrieval Conference* (TREC), die *Conference and Labs of the Evaluation Forum* (CLEF) und ein paar jüngere wissenschaftliche Forschungsgruppen. Erst diese Initiativen ermöglichen es den Forscher:innen die Authentizität von Informationen bei der Relevanzbewertung zu berücksichtigen. Herkömmliche Bewertungsinitiativen wie das *Forum for Information Retrieval Evaluation* (FIRE) und die *NII Testbeds and Community for Information Access Research* (NTCIR) berücksichtigen nur die Authentizität von Informationen (insbesondere in den Labors UrduFake [55] und Lab-PoliInfo [56]). Im Folgenden wird auf die Initiativen TREC und CLEF eingegangen.

10.5.1 TREC

Die Evaluierungsinitiative TREC hat erstmals den *Health Misinformation Track* im Jahr 2019 integriert. Die Teilnehmer an diesem Track mussten u.a. Systeme entwickeln, die „relevante und glaubwürdige Informationen im Gesundheitsbereich bereitstellen, welche den Nutzern dann dabei helfen, richtige Entscheidungen zu treffen“. Je nach Track-Ausgabe wurden über die Authentizität der Informationen hinaus weitere Kriterien berücksichtigt, die in den folgenden Abschnitten kurz erläutert werden.

Datenerhebung: Der *Health Misinformation Track* 2019 verwendete den ClueWeb12-B13 Datensatz als Grundlage [57]. Dieser Datensatz besteht aus englischen Webseiten, welche im Jahr 2012 zu verschiedenen Gesundheitsthemen gesammelt wurden und sowohl richtige als auch falsche Informationen mit unterschiedlicher Glaubwürdigkeit und Qualität enthalten. Der Track 2020 verwendete einen von Common Crawl bereitgestellten Datensatz, der sich insbesondere auf verschiedene Nachrichten fokussiert, die in den ersten vier Monaten des Jahres gesammelt wurden. Aus diesem Datensatz wurden 74 COVID-19-bezogene Themen ausgewählt und entsprechend gefiltert. In der aktuellen Ausgabe 2021 wurde die „noclean“-Version des von Google genutzten C4-Datensatzes verwendet [58].

Menschliche Intervention: In den verschiedenen Ausgaben des *Health Misinformation Track* wurden die Dokumente von menschlichen Gutachtern in Bezug zu verschiedenen Kriterien bewertet: *Relevanz, Wirksamkeit und Glaubwürdigkeit* (in 2019), *Nützlichkeit, Korrektheit und Glaubwürdigkeit* (in 2020), und *Nützlichkeit, Glaubwürdigkeit und Unterstützungswirkung* (in 2021). Wirksamkeit bezieht sich auf das Vorhandensein „richtiger“ Informationen in dem Dokument in Bezug zur Behandlung des Themas. Dieses Kriterium ist dem Kriterium der Korrektheit ähnlich, das 2020 benutzt wurde.

Sowohl die Wirksamkeit als auch die Korrektheit wurden auf einer Drei-Punkte-Skala bewertet – einschließlich eines „nicht bewertet“ Labels. Mit der Unterstützungswirkung ist gemeint, dass das Dokument eine bestimmte Behandlung im Kontext der

Suchanfrage empfehlen oder davon abraten kann. Dieses Kriterium wurde auf einer dreistufigen Skala bewertet, die auch einen neutralen Wert enthielt. Dokumente, die als nicht relevant (oder als nicht nützlich) eingestuft wurden in Bezug auf zusätzliche Kriterien nicht weiter bewertet.

Die Glaubwürdigkeit wurde auf einer dreistufigen Skala dargestellt (einschließlich der Kennzeichnung „nicht bewertet“) und in den beiden letzten Ausgaben auf einer binären Skala bewertet. Die menschlichen Gutachter/Bewerter wurden gebeten, ein Glaubwürdigkeitslabel zu vergeben, das unter anderem auf folgenden Aspekten basiert: Umfang des Fachwissens, Autorität und Vertrauenswürdigkeit des Dokuments, die Angabe eines Autors oder eines Instituts, welcher/welches das Webdokument veröffentlicht hat, das Vorhandensein von Zitaten zu vertrauenswürdigen/glaubwürdigen Quellen, der Schreibstil (gut oder schlecht geschrieben) und der Zweck für welchen das Dokument verfasst wurde (zur Information oder zu Werbezwecken). In jeder Ausgabe wurden etwa 20.000 gekennzeichnete Dokumente mit über 50 Themen zur Verfügung gestellt.

Basisdaten und Evaluationsmethoden: Sowohl für den Health Misinformation Track 2019 als auch für den Health Misinformation Track 2020 wurden die Basisdaten des BM25-Retrievalmodells unter Verwendung des Anserini-Toolkits⁹ mit Standardparametern implementiert. Sowohl die Basisdaten als auch die durchgeführten Durchläufe/Runden wurden in Bezug auf folgende Kriterien bewertet [59]:

- (1) *Normalized Local Rank Error* (NLRE): die Rangpositionen der Dokumente werden paarweise verglichen und auf „Fehler“ geprüft, die als Fehlplatzierung von Dokumenten gelten, d. h. relevante oder glaubwürdige Dokumente, die als nicht-relevante oder nicht-glaubwürdige Dokumente erfasst werden;
- (2) *Normalized Weighted Cumulative Score* (nWCS): aus den verschiedenen Kriterien wird ein einziges Label generiert, und der *Standard Normalized Discounted Cumulative Gain* (NDCG)-Maß wird berechnet [60];
- (3) *Convex Aggregating Measure* (CAM): jedes Kriterium wird separat betrachtet und eine Rangfolge berechnet.
- (4) Schließlich wird die *Mean Average Precision* (MAP) oder der NDCG Wert errechnet.

10.5.2 CLEF

Diese Bewertungsinitiative umfasst seit 2018 Aufgaben im Zusammenhang mit der automatischen Identifizierung und Überprüfung von Behauptungen in sozialen Medien, mit dem *CheckThat! Lab* [61]. Allerdings wird erst seit 2020 das Konzept der Authentizität von Informationen im Kontext von CHS im eHealth Lab berücksichtigt. Hierbei

⁹ Das Aserini-Toolkit ist eine Open Source Projekt für reproduzierbare Informationen in IR- Forschung. URL: <https://github.com/castorini/anserini>.

wird bewertet inwiefern Systeme relevante, lesbare und echte Informationen abrufen können. Zusätzlich gibt es eine Unteraufgabe, die sich speziell auf die Vorhersage der Authentizität von Informationen konzentriert.

Datenerhebung: Für die E-Health-Ausgaben 2020 und 2021 wurden wochenlang Webseiten durch wiederholte Übermittlung von CLEF eHealth 2018-Abfragen an Microsoft Bing APIs gesammelt. Die Liste der erhaltenen Webdokumente wurde um weitere glaubwürdige und unglaubwürdige Websites erweitert. Die Erweiterung basierte auf Websites, die zuvor von Gesundheitseinrichtungen und -agenturen zusammengestellt wurden. Darüber hinaus wurden in der Ausgabe 2021 auch Social Media Inhalte von Reddit und Twitter berücksichtigt. Diese Inhalte wurden zu 150 gesundheitsbezogenen Themen gesammelt. Aus diesen Themen wurden manuell Abfragen erstellt, mit denen Beiträge und Tweets von Reddit bzw. Twitter gefiltert wurden.

Ein Reddit-Dokument besteht aus einem sogenannten „Beitrag“. Das bedeutet einem Beitrag mit einem Titel und einer Beschreibung, in dem im Allgemeinen eine Frage gestellt wird, und einem „Kommentar“, d. h. einer „Antwort“ auf den Beitrag. Im Vergleich dazu besteht ein Beitrag bei Twitter lediglich aus einem einzelnen Tweet und den zugehörigen Metadaten.

Menschliche Intervention: Im CLEF eHealth wurden die Dokumente nach drei Kriterien gekennzeichnet, nämlich nach (thematischer) Relevanz, Lesbarkeit (oder Verständlichkeit) und Glaubwürdigkeit. Relevanz und Lesbarkeit wurden auf einer dreistufigen Skala bewertet, d. h. nicht relevant/lesbar, teilweise relevant/lesbar, relevant/lesbar. Im Hinblick auf die Glaubwürdigkeit wurde es für sinnvoll erachtet, ein viertes Label einzuführen, nämlich „nicht beurteilbar“. Bei der Bewertung der Glaubwürdigkeit von Websites und sozialen Inhalten mussten die menschlichen Gutachter/Bewerter insbesondere fünf Aspekte berücksichtigen.

1. Die Verfügbarkeit von Indikatoren für die Vertrauenswürdigkeit der Quelle (z. B. Fachwissen, Webreputation usw.),
2. Die syntaktischen und semantischen Merkmale des Inhalts (z. B. der Schreibstil),
3. Die Emotionen, die der Text hervorrufen soll,
4. Das Vorhandensein von überprüfbaren Fakten und Behauptungen (z. B. durch das Vorhandensein von Zitaten oder externen Links) und
5. Die Analyse der sozialen Beziehungen des Verfassers eines Beitrags (im Falle von sozialen Inhalten) beurteilen.

Basisdaten und Evaluationsmethoden: Im Rahmen des CLEF eHealth 2020 haben die Organisatoren Basisdaten und Methoden entwickelt, die basierend auf dem BM25-Retrievalmodell und einer durch Reinforcement Learning optimierten Anfrageerweiterung (*query expansion*) funktionieren. Das *query expansion model* wurde mit den TREC-CAR-, Jeopardy- und Microsoft Academic-Datensätzen aus [62] trainiert und die erweiterten Abfragen wurden als Input für das BM25-Modell verwendet.

Im CLEF eHealth 2021 wurden sechs Basissysteme basierend auf den Retrievalmodellen BM25, des *Dirichlet Language Model* (DirichletLM) und *Term Frequency-Inverse Document Frequency* (TF-IDF) mit Standardparametern bereitgestellt. Weitere Einzelheiten sind in [25] zu finden. Die folgenden Bewertungsmaßstäbe wurden zur Beurteilung der Basisdaten und der eingereichten Läufe verwendet um die Genauigkeit der (binären) Klassifizierung der Dokumente nur in Bezug auf die Glaubwürdigkeit zu bewerten: *MAP*; *NDCG*; *BPref* [63]; *uRBP* und *cRBP*. Diese Systeme bewerteten die Dokumente unter Berücksichtigung von fünf Kriterien: *accuracy* (Genauigkeit), *F1-Score* und die AUC (Area Under the ROC Curve) *uRBP* (understandability Rank Biased Precision) und *cRBP* (credibility Rank Biased Precision). *uRBP* wurde in [64] eingeführt, während *cRBP* zum ersten Mal (basierend auf *uRBP*) in der 2020er Ausgabe des CLEF eHealth [2] auftaucht.

10.6 Schlussfolgerungen

In diesem Artikel wollten wir einen allgemeinen Überblick über die Problemstellung von Gesundheitsfehlinformationen geben. Hierbei fokussierten wir die zunehmende Internetrecherche von Gesundheitsinformationen durch Laien.

Im Besonderen wurden die Herausforderungen und möglichen Lösungen im Bereich Information Retrieval und der Recherche von Gesundheitsinformationen zusammen mit automatisierten Lösungen skizziert. Diese automatisierten Lösungen sollen einen Zugang zu echten Gesundheitsinformationen gewährleisten, wobei wir die Besonderheiten und kritischen Aspekte dieses Bereichs berücksichtigt haben. Wir haben zum Beispiel gesehen, dass die Benutzung von rein manuellen Werkzeugen oder die intensive Hinzuziehung von Expert:innen keine praktikablen Lösungen sind. Jedoch stellen insbesondere neuere Ansätze, basierend auf algorithmischen Lösungen, die das von den Expert:innen selbst bereitgestellte Fachwissen oder zumindest deren wissenschaftliche Arbeiten beinhalten, einen möglichen Lösungsansatz dar.

Bitte ohne Absatz direkt nach dem vorherigen Satz. Entwicklung neuer technologischer Lösungen erfordert jedoch gleichzeitig die Entwicklung und Verbesserung von Bewertungsstrategien, von welchen wir die wichtigsten beschrieben haben. Bei diesem Fokus handelt es sich jedoch um einen Forschungsbereich, der angesichts zahlreicher offener Probleme noch vorrangig untersucht werden muss. Hierzu gehören Fragestellungen zu den Themen Subjektivität, Datenschutz, Nichtdiskriminierung, Voreingenommenheit bei vorgeschlagenen Modellen und Forschungsergebnissen sowie die Berücksichtigung zahlreicher kontextbezogener Aspekte bei der Entwicklung wirksamer Lösungen.

Danksagung

Dieser Artikel ist das Ergebnis der Forschungsarbeit, die im Rahmen des Lehrstuhls für *Information and Knowledge Representation, Retrieval, and Reasoning* (IKR3) Laboratory, zusammen mit Gabriella Pasi [0000-0002-6080-8170] und Rishabh Upadhyay [0000-0001-9937-6494], im Rahmen des EU Horizon 2020 ITN/ETN DoSSIER-Projekts zu domänenspezifischen Systemen für Information Extraction and Retrieval (H2020-EU.1.3.1., ID: 860721) durchgeführt wurde. Die in diesem Artikel behandelten Fragestellungen und Lösungen wurden auf folgendem Workshop vorgestellt: Personalisierte Services für ältere Menschen. Interdisziplinäre und interkulturelle Diskussion von Gestaltungsmöglichkeiten. Villa Vigoni, Deutsch-Italienisches Zentrum für den Europäischen Dialog. Lovenovo di Menaggio (Como), Italien. 28. November 2022 – 02. Dezember 2022 [65].

Der Beitrag der Autoren wird nach der CReDit-Taxonomie aufgeschlüsselt: Marco Viviani [0000-0002-2274-9050] trug zur Konzeptualisierung, formalen Analyse, Untersuchung und zum Verfassen des ursprünglichen Artikelentwurfs bei. Elisabeth Mess [0000-0003-1205-9141] trug zur Konzeptionalisierung, zum Schreiben des ursprünglichen Entwurfs und zur Überprüfung wie auch Redaktion des vorliegenden Artikels bei. Ein weiterer Schwerpunkt war die Visualisierung.

Literatur

- [1] Metzger, Miriam J, und Andrew J Flanagin. 2015. „Psychological Approaches to Credibility Assessment Online“. *The Handbook of the Psychology of Communication Technology*, 445–66. <https://doi.org/10.1002/9781118426456.ch20>.
- [2] Suominen, Hanna, Lorraine Goeuriot, Liadh Kelly, Laura Alonso Alemany, Elias Bassani, Nicola Brew-Sam, Viviana Cotik, u. a. 2021. „Overview of the CLEF eHealth Evaluation Lab 2021“. *Lecture Notes in Computer Science*, 308–23. https://doi.org/10.1007/978-3-030-85251-1_21.
- [3] Fox, Susannah, und Maeve Duggan. 2013. „Health Online 2013“. 2013. <https://www.pewresearch.org/internet/2013/01/15/health-online-2013/>.
- [4] Millenson, Michael L., Jessica L. Baldwin, Lorri Zipperer, und Hardeep Singh. 2018. „Beyond Dr. Google: the evidence on consumer-facing digital tools for diagnosis“. *Diagnosis* 5 (3): 95–105. <https://doi.org/10.1515/dx-2018-0009>.
- [5] Eurostat. 2022. „Digital economy and society statistics – households and individuals“. 2022. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Digital_economy_and_society_statistics_-_households_and_individuals#Purpose_of_the_use_of_internet.
- [6] Bahkali, Salwa, Reem Almaiman, Mamoun El-Awad, Huda Almohanna, Khaled Al-Surimi, und Mowafa Househ. 2016. „Exploring the Impact of Information Seeking Behaviors of Online Health Consumers in the Arab World“. *Studies in Health Technology and Informatics* 226: 279–82. <https://pubmed.ncbi.nlm.nih.gov/27350525/>.
- [7] Wong, David Ka-Ki, und Man-Kuen Cheung. 2019. „Online Health Information Seeking and eHealth Literacy Among Patients Attending a Primary Care Clinic in Hong Kong: A Cross-Sectional Survey“. *Journal of Medical Internet Research* 21 (3): e10831. <https://doi.org/10.2196/10831>.

- [8] Akerkar, S.M., M. Kanitkar, und L.S. Bichile. 2005. „Use of the Internet as a resource of health information by patients: a clinic-based study in the Indian population“. *Journal of Postgraduate Medicine* 51 (2): 116–18.
- [9] Chen, Yimin, Niall J Conroy, und Victoria L Rubin. 2015. „Misleading Online Content: Recognizing Click-bait as ‚False News‘“. *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection – WMDD '15*. <https://doi.org/10.1145/2823465.2823467>.
- [10] Viviani, Marco, und Gabriella Pasi, 2017b „Credibility in social media: opinions, news, and health information-a survey“. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 7 (5): e1209. <https://doi.org/10.1002/widm.1209>.
- [11] Barua, Zapan, Sajib Barua, Salma Aktar, Najma Kabir, und Mingze Li. 2020. „Effects of misinformation on COVID-19 individual responses and recommendations for resilience of disastrous consequences of misinformation“. *Progress in Disaster Science* 8: 100119. <https://doi.org/10.1016/j.pdisas.2020.100119>.
- [12] Enders, Adam M, Joseph E Uscinski, Casey Klofstad, und Justin Stoler. 2020. „The different forms of COVID-19 misinformation and their consequences“. *Harvard Kennedy School Misinformation Review* 1 (8). <https://doi.org/10.37016/mr-2020-48>.
- [13] Chang, Yung-Sheng, Yan Zhang, und Jacek Gwizdka. 2021. „The effects of information source and eHealth literacy on consumer health information credibility evaluation behavior“. *Computers in Human Behavior* 115: 106629. <https://doi.org/10.1016/j.chb.2020.106629>.
- [14] Schulz, Peter J, und Kent Nakamoto. 2022. „The perils of misinformation: when health literacy goes awry“. *Nature Reviews Nephrology*, 1–2. <https://doi.org/10.1038/s41581-021-00534-z>.
- [15] Kickbusch, I S. 2001. „Health literacy: addressing the health and education divide“. *Health Promotion International* 16 (3): 289–97. <https://doi.org/10.1093/heapro/16.3.289>.
- [16] Chou, Wen-Ying Sylvia, April Oh, und William M P Klein. 2018. „Addressing Health-Related Misinformation on Social Media“. *JAMA* 320 (23): 2417. <https://doi.org/10.1001/jama.2018.16865>.
- [17] Coronel, Carlos, und Steven Morris. 2018. *Database systems : design, implementation, and management*. 13. Aufl. Cengage Learning.
- [18] Manning, Christopher D., Prabhakar Raghavan, und Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Higher Education from Cambridge University Press.
- [19] Aggarwal, Charu C. 2016. *Recommender Systems*. Cham: Springer International Publishing. <https://doi.org/10.1007/978-3-319-29659-3>.
- [20] Salton, Gerard, und Donna Harman. 2003. *Information retrieval*. <https://dl.acm.org/doi/10.5555/1074100.1074478>.
- [21] Onwuchekwa, Edeama O. 2012. „Organisation of Information and the Information Retrieval System“. In *Library and Information Science in Developing Countries*, 275–92. IGI Global. <https://doi.org/10.4018/978-1-61350-335-5.ch020>.
- [22] Jurafsky, Daniel, und James H. Martin. 2006. *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition. Section 23.1*. [https://pages.ucsd.edu/~bakovic/compphon/Jurafsky,%20Martin.-Speech%20and%20Language%20Processing_%20An%20Introduction%20to%20Natural%20Language%20Processing%20\(2007\).pdf](https://pages.ucsd.edu/~bakovic/compphon/Jurafsky,%20Martin.-Speech%20and%20Language%20Processing_%20An%20Introduction%20to%20Natural%20Language%20Processing%20(2007).pdf).
- [23] Xu, Yunjie (Calvin), und Zhiwei Chen. 2006. „Relevance judgment: What do information users consider beyond topicality?“ *Journal of the American Society for Information Science and Technology* 57 (7): 961–73. <https://doi.org/10.1002/asi.20361>.
- [24] Clarke, Charles, Maria Maistro, Saira Rizvi, Mark Smucker, und Guido Zuccon. 2020. „Overview of the TREC 2020 Health Misinformation Track“. <https://trec.nist.gov/pubs/trec29/papers/OVERVIEW.HM.pdf>.

- [25] Goeuriot, Lorraine, Hanna Suominen, Gabriella Pasi, Elias Bassani, Nicola Brew-Sam, Gabriela González-Sáez, Liadh Kelly, u. a. 2021. „Consumer Health Search at CLEF eHealth 2021 CLEF 2021 - Conference and Labs of the Evaluation Forum“. <https://ceur-ws.org/Vol-2936/paper-62.pdf>.
- [26] Wardle, Claire, Grace Greason, Joe Kerwin, und Nic Dias. 2018. „Information Disorder: The Essential Glossary“.
- [27] Kinhead, Laura, Ahmed Allam, und Michael Krauthammer. 2020. „AutoDiscern: rating the quality of online health information with hierarchical encoder attention-based neural networks“. *BMC Medical Informatics and Decision Making* 20 (1). <https://doi.org/10.1186/s12911-020-01131-z>.
- [28] Upadhyay, Rishabh, Gabriella Pasi, und Marco Viviani. 2021. „Health Misinformation Detection in Web Content: A Structural-, Content-based, and Context-aware Approach based on Web2Vec“. *Proceedings of the Conference on Information Technology for Social Good*. <https://doi.org/10.1145/3462203.3475898>.
- [29] Bal, Rakesh, Sayan Sinha, Swastika Dutta, Rishabh Joshi, Sayan Ghosh, und Ritam Dutt. 2020. „Analysing the Extent of Misinformation in Cancer Related Tweets“. *Proceedings of the International AAAI Conference on Web and Social Media* 14: 924–28. <https://doi.org/10.1609/icwsm.v14i1.7359>.
- [30] Samuel, Hamman, und Osmar Zaiane. 2018. „MedFact: Towards Improving Veracity of Medical Information in Social Media Using Applied Machine Learning“. *Advances in Artificial Intelligence*, 108–20. https://doi.org/10.1007/978-3-319-89656-4_9.
- [31] Zhao, Yuehua, Jingwei Da, und Jiaqi Yan. 2021. „Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning based approaches“. *Information Processing & Management* 58 (1): 102390. <https://doi.org/10.1016/j.ipm.2020.102390>.
- [32] Charnock, D, S Shepperd, G Needham, und R Gann. 1999. „DISCERN: an instrument for judging the quality of written consumer health information on treatment choices“. *Journal of Epidemiology and Community Health* 53 (2): 105–11. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1756830/>.
- [33] Lerch, Magnus, Günter Ollenschläger, und Sylvia Sängler. 2000. „Das DISCERN-Instrument“. <http://www.discern.de/instrument.htm>.
- [34] Boyer, C, M Selby, J.-R Scherrer, und R D Appel. 1998. „The Health On the Net Code of Conduct for medical and health Websites“. *Computers in Biology and Medicine* 28 (5): 603–10. [https://doi.org/10.1016/s0010-4825\(98\)00037-7](https://doi.org/10.1016/s0010-4825(98)00037-7).
- [35] Kou, Ziyi, Lanyu Shang, Yang Zhang, und Dong Wang. 2022. „HC-COVID“. *Proceedings of the ACM on Human-Computer Interaction* 6 (GROUP): 1–25. <https://doi.org/10.1145/3492855>.
- [36] Dery, Sebastian. 2016. „Challenges of Knowledge Graphs From Strings to Things – An Introduction“. 2016. <https://medium.com/@sderymail/challenges-of-knowledge-graph-part-1-d9ffe9e35214>.
- [37] Sackett, David L. 1997. „Evidence-based medicine“. *Seminars in Perinatology* 21 (1): 3–5. [https://doi.org/10.1016/s0146-0005\(97\)80013-4](https://doi.org/10.1016/s0146-0005(97)80013-4).
- [38] Cui, Limeng, Haeseung Seo, Maryam Tabar, Fenglong Ma, Suhang Wang, und Dongwon Lee. 2020. „DETERRENT: Knowledge Guided Graph Attention Network for Detecting Healthcare Misinformation“. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 492–502. <https://doi.org/10.1145/3394486.3403092>.
- [39] Ernst, Patrick, Cynthia Meng, Amy Siu, und Gerhard Weikum. 2014. „KnowLife: A knowledge graph for health and life sciences“. In *2014 IEEE 30th International Conference on Data Engineering*, 1254–57. IEEE. <https://doi.org/10.1109/ICDE.2014.6816754>.
- [40] Pradeep, Ronak, Xueguang Ma, Rodrigo Nogueira, und Jimmy Lin. 2021. „Vera: Prediction Techniques for Reducing Harmful Misinformation in Consumer Health Search“. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2066–70. <https://doi.org/10.1145/3404835.3463120>.
- [41] Robertson, Stephen. 2010. „The Probabilistic Relevance Framework: BM25 and Beyond“. *Foundations and Trends® in Information Retrieval* 3 (4): 333–89. <https://doi.org/10.1561/1500000019>.

- [42] Pradeep, Ronak, Rodrigo Nogueira, und Jimmy Lin. 2021. „The Expando-Mono-Duo Design Pattern for Text Ranking with Pretrained Sequence-to-Sequence Models“. *arxiv.org*. <https://doi.org/10.48550/arXiv.2101.05667>.
- [43] Abualsaud, Mustafa, Christina Lioma, Maria Maistro, Mark D Smucker, und Guido Zuccon. o. J. „Overview of the TREC 2019 Decision Track“. <https://lemurproject.org/clueweb12/>.
- [44] Schlicht, Ipek Baris, Angel Felipe Magnossão de Paula, und Paolo Rosso. 2021. „UPV at TREC Health Misinformation Track 2021 Ranking with SBERT and Quality Estimators“. *arXiv:2112.06080 [cs]*. <https://arxiv.org/abs/2112.06080>.
- [45] Schwitzer, Gary. 2007. „Medicine and society HealthNewsReview.org: criteria for excellence in health and medical journalism“. *American Medical Association Journal of Ethics*. Bd. 9. www.virtualmentor.org.
- [46] Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, und Veselin Stoyanov. 2019. „RoBERTa: A Robustly Optimized BERT Pretraining Approach“, Juli.
- [47] Zuo, Chaoyuan, Qi Zhang, und Ritwik Banerjee. 2021. „An Empirical Assessment of the Qualitative Aspects of Misinformation in Health News“. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, 76–81. Stroudsburg, PA, USA: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.nlp4if-1.11>.
- [48] Cormack, Gordon v, Charles L A Clarke, und Stefan Buettcher. 2009. „Reciprocal rank fusion outperforms condorcet and individual rank learning methods“. *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 758–59. <https://doi.org/10.1145/1571941.1572114>.
- [49] Putri, Divi Galih Prasetyo, Marco Viviani, und Gabriella Pasi. 2020. „Social Search and Task-Related Relevance Dimensions in Microblogging Sites“. *Lecture Notes in Computer Science*, 297–311. https://doi.org/10.1007/978-3-030-60975-7_22.
- [50] Croft, Bruce, Donald Metzler, und Trevor Strohman. 2010. „Search Engines Information Retrieval in Practice“.
- [51] Viviani, Marco, und Gabriella Pasi, 2017a „A Multi-criteria Decision Making Approach for the Assessment of Information Credibility in Social Media“. *Fuzzy Logic and Soft Computing Applications*, 197–207. https://doi.org/10.1007/978-3-319-52962-2_17.
- [52] Upadhyay, Rishabh, Gabriella Pasi, und Marco Viviani. 2022. „An Unsupervised Approach to Genuine Health Information Retrieval Based on Scientific Evidence“. *Web Information Systems Engineering – WISE 2022*, 119–35. https://doi.org/10.1007/978-3-031-20891-1_10.
- [53] Voorhees, Ellen M. 2002. „The Philosophy of Information Retrieval Evaluation“. *Lecture Notes in Computer Science*, 355–70. https://doi.org/10.1007/3-540-45691-0_34.
- [54] Bama, SSathya, und MSIrifan Ahmed. 2015. „A SURVEY ON PERFORMANCE EVALUATION MEASURES FOR INFORMATION RETRIEVAL SYSTEM“. *International Research Journal of Engineering and Technology*. www.irjet.net.
- [55] Amjad, Maaz, Grigori Sidorov, Alisa Zhila, Alexander Gelbukh, und Paolo Rosso. 2022. „UrduFake@FIRE2020: Shared Track on Fake News Identification in Urdu“. *arXiv:2207.12406 [cs]*, 37–40. <https://arxiv.org/abs/2207.12406>.
- [56] Kimura, Yasutomo, Hideyuki Shibuki, Hokuto Ootake, Yuzu Uchida, Keiichi Takamaru, Kotaro Sakamoto, Madoka Ishioroshi, u. a. 2019. „Final Report of the NTCIR-14 QA Lab-PoliInfo Task“. *NII Testbeds and Community for Information Access Research*, 122–35. https://doi.org/10.1007/978-3-030-36805-0_10.
- [57] Lioma, Christina, Mark Smucker, und Guido Zuccon. 2019. „TREC Decision Track (2019)“. 2019. <https://trec-health-misinfo.github.io/2019.html>.
- [58] Clarke, Charles, Maria Maistro, und Mark Smucker. 2021. „TREC Health Misinformation Track (2021)“. 2021. <https://trec-health-misinfo.github.io/2021.html>.

- [59] Lioma, Christina, Jakob Grue Simonsen, und Birger Larsen. 2017. „Evaluation Measures for Relevance and Credibility in Ranked Lists“. *arXiv:1708.07157 [cs]*. <https://arxiv.org/abs/1708.07157>.
- [60] Wang, Yining, Liwei Wang, Yuanzhi Li, Di He, Tie-Yan Liu, und Wei Chen. 2013. „A Theoretical Analysis of NDCG Type Ranking Measures“. *arXiv:1304.6480 [cs, stat]*. <https://arxiv.org/abs/1304.6480>.
- [61] Barron-Cedeno, Alberto, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, u. a. 2020. „Overview of CheckThat! 2020: Automatic Identification and Verification of Claims in Social Media“. *arXiv:2007.07997 [cs]*. <https://arxiv.org/abs/2007.07997>.
- [62] Nogueira, Rodrigo, und Kyunghyun Cho. 2017. „Task-Oriented Query Reformulation with Reinforcement Learning“. *arXiv:1704.04572 [cs]*. <https://arxiv.org/abs/1704.04572>.
- [63] Lui, Ling, und Tamer M. Özsu. 2020. *Encyclopedia of Database Systems*. Herausgegeben von Ling Lui und Tamer M. Özsu. Springer.
- [64] Zuccon, Guido. 2016. „Understandability Biased Evaluation for Information Retrieval“. *Lecture Notes in Computer Science*, 280–92. https://doi.org/10.1007/978-3-319-30671-1_21.
- [65] Viviani, Marco. 2022. „Consumer Health Search in the Social Web“. <https://doi.org/10.13140/RG.2.2.32670.00323>.