Beschreibung des Vorhabens

Wissenschaftliche Literaturversorgungs- und Informationssysteme (LIS)

Name

Automatic Quality Assessment: NLP-Verfahren zur semantischen Kartierung von lebenswissenschaftlichen Texten

Akronym

AQUAS

Laufzeit	3 Jahre
Geplanter Projektstart	1. März 2022
Förderumfang	1 WMA + 1 Hilfskraft + Sach-/Reisemittel
Förderprogramm	e-Research-Technologien

Antragsteller

Prof. Dr. Konrad U. Förstner

ZB MED – Information Centre for Life Sciences
Gleueler Straße 60
50931 Köln
foerstner@zbmed.de
Tel.: 0221 478-78909

Dr. Eva Seidlmayer, M.LIS

ZB MED – Information Centre for Life Sciences
Gleueler Straße 60
50931 Köln
seidlmayer@zbmed.de

1 Ausgangslage

Zielsetzung

Das wachsende Aufkommen von bewusst gestreuten Fehlinformationen stellt unsere demokratische Gesellschaft vor eine große Herausforderung. Sie werden zunehmend von politischen Interessensgruppen ausgesät, um den öffentlichen Diskurs zu bestimmen. Von den Rezipienten werden diese Falschinformationen mitunter nicht als solche erkannt. Da sich Desinformationen auch in wissenschaftlichen Informationsangeboten finden, betrifft diese Entwicklung auch Wissenschaftler:innen. In den medizinischen Anwendungen der Lebenswissenschaften (LeWi) kann dies gesundheitsgefährdende Auswirkungen haben.

In dem hier vorgestellten Vorhaben AQUAS wird der erste deutschsprachige Datensatz zu Desinformation in den Lebenswissenschaften erstellt. Auf dieser Basis soll mit modernen Machine-Learning-(ML)-Verfahren ein ML-Modell erstellt, das in der Lage sein wird, die semantische Nähe von unbekannten Texten zu den Klassen wissenschaftliche Texte, populärwissenschaftliche Texte und desinformierende Texte graduell einzuordnen. Gleichzeitig werden ergänzende Informationen zur guten wissenschaftlichen Praxis der Publikationen bereitgestellt. Mit der Anreicherung und Veröffentlichung der genannten Informationen (Basisset bzw. erweitertes Set an Merkmalen) strebt AQUAS die Unterstützung der Leser:innen an, eine informierte Einschätzung von Literatur zu treffen. Dabei geht es AQUAS nicht um eine abschließende Leseempfehlung der Inhalte oder Zensur.

Auf Grundlage des entwickelten Anreicherungs-Methoden wird im Rahmen von AQUAS ein Dienst implementiert, der über eine Programmierschnittstelle (API – Application Programming Interface) angesprochen werden kann. Als erste zentrale Anwendung werden wir diesen Dienst durch das ZB MED Discovery-System LIVIVO[1] nutzen, um die beschriebene Einstufung der Literatur den Nutzenden von ZB MED zur Verfügung zu stellen. Damit werden zunächst die Wissenschaftler:innen der LeWi und Praktiker:innen in Gesundheitsberufen sowie Studierende von der verbesserten Wissensinfrastruktur bei LIVIVO durch AQUAS profitieren. Der Datensatz, das Modell, der Workflow zum Training sowie die Software zum Betrieb des Dienstes werden nach Möglichkeit offen bereitgestellt und so auch für andere Themenfelder nutzbar gemacht.

1.1 Ausgangslage und eigene Vorarbeiten

Wachsendes Informationsaufkommen und Desinformationen als demokratische Herausforderung

Schon seit einigen Jahren warnen die Europäische Union, Internetaktivist:innen, Journalist:innen und Wissenschaftler:innen vor den ernsthaften Herausforderungen, die personalisierte Informationsangebote und gezielt gestreute Desinformationen für unsere demokratische Gesellschaft darstellen (Pariser 2011; Europäische Union u. a. 2013; Vaidhyanathan 2018; Polizzi 2019; Gensing 2020). Gleichzeitig stellen Forscher:innen die mangelnde Kompetenz der Menschen fest, die angezeigten Inhalte richtig einordnen zu können (Meßmer, Sängerlaub und Schulz 2021). Falschmeldungen nehmen nicht nur Einfluss auf Wahlen und Politik, sondern prägen auch den öffentlichen Diskurs (Epstein und Robertson 2015, Speed und Mannion 2017, **Neben Privatleuten und Politiker:innen sind Wissenschaftler:innen zunehmend von Falschmeldungen betroffen. Mitunter geraten sie auf Grundlage einseitiger durch Algorithmen lancierter Informationen sogar in eine „(re)search bubble“** (Pariser 2011 Ćurković und Košec 2018).

Dass auch im wissenschaftlichen Bereich Empfehlungsalgorithmen und Desinformationen das Informationsverhalten verändern und beeinflussen, darauf werfen die zunehmende Ablehnung

[1] https://www.livivo.de/

von Impfungen, das wachsende Interesse an Homöopathie oder das starke Wirken von Verschwörungserzählungen während der Corona-Pandemie ein Schlaglicht (European Commission, Directorate-General for Communications Networks, Content and Technology 2021). In diesem Kontext sah es etwa die Ärztekammer Schleswig-Holstein als geboten an, eine „Allianz gegen Corona-Leugner“ innerhalb der Ärzteschaft zu formieren.[2] Auch beim Thema Impfen etabliert sich eine skeptische Haltung selbst unter Mediziner:innen, die so weit geht, dass die Weltgesundheitsorganisation (WHO) die Impfverweigerung zu einer der zehn größten Gesundheitsgefahren weltweit erklärt hat.[3]

Desinformationen sind ubiquitär und können auch durch eine Qualitätskontrolle durch Informations-Anbieter nicht vollständig herausgefiltert und unterbunden werden. Dies betrifft sowohl generell wissenschaftliche Literaturportale als auch Plattformen der medizinischen Informationsversorgung im Besonderen (Holone 2016). **Eine Unterdrückung abweichender Positionen ist auch nicht anzustreben, da auch Grenzfälle, neue Strömungen oder besonders kritische Positionen betroffen sein könnten und dies so einer Zensur gleichkäme.**

Auch in dem von ZB MED betriebenen lebenswissenschaftlichen Discovery-System LIVIVO, das über 70 Millionen Literatur-Einträge enthält, finden sich Titel, die nicht als wissenschaftlich gelten können. Dies ist der Funktionsweise eines Discovery-Systems geschuldet, das Inhalte aus anderen Datenbanken zusammenführt. Insgesamt fehlt es in Online-Plattformen jedoch an unterstützenden Hinweisen für die Einschätzung von Inhalten (Meßmer, Sängerlaub und Schulz 2021). Weder LIVIVO noch PubMed, das Portal der US-National Library of Medicine (NLM), das weniger Quellen indexiert, aber international viel genutzt wird, bieten derzeit Unterstützung der Nutzer:innen in Fragen der Qualitätseinschätzung an. Wie groß jeweils der Anteil an Dokumenten mit desinformierenden Charakter ist, ist dabei unklar. Dass solche Dokumente jedoch unkommentiert in wissenschaftlichen Suchmaschinen auftauchen, ist insbesondere bedenklich, da sich etwa LIVIVO an Forscher:innen der LeWi, aber auch etwa an praktizierende Mediziner:innen richtet (B. Müller u. a. 2017). Nicht immer können nicht-wissenschaftliche Inhalte direkt erkannt werden; dies birgt das Risiko einer nicht optimalen Behandlung der Patient:innen. Unter Zeitdruck, der die Arbeitsrealität in Gesundheitsberufen ausmacht, stellt die inhaltliche Bewertung der Inhalte eine zusätzliche Belastung dar.

Für eine **reflektierte Umgangsweise mit Informationen und der Orientierung in der Informationswelt des 21. Jahrhundert ist vor allem eine Befähigung zur Bewertung von Informationen durch die Nutzenden zentral** (Polizzi 2019). Menschen, die nicht mit dem Internet aufgewachsen sind, haben Aufklärungsbedarf bezüglich der Informationskompetenz (Gensing 2020); hierzu gehören häufig auch Wissenschaftler:innen. Bibliotheken und Informationszentren fällt bei dieser Aufgabe eine besondere Verantwortung für die Stärkung der Informationskompetenz und der Erkennung von Desinformationen zu (Chartered Institute of Library Information Professionals (CILIP) 2018). Wir sehen es daher als geboten an, die Inhalte unserer Datenbank bei ZB MED mit modernen Textanalyse-Werkzeugen sowie maschinellen Lernverfahren mit Blick auf problematische Inhalte zu analysieren. So können den Nutzenden Zusatzinformationen zu Texten bereitgestellt werden, die eine mögliche semantische Nähe zu von Expertinn:en als Desinformation eingestuften Texten aufzeigen. Die erarbeiteten Verfahren sowie die Datensätze werden von ZB MED zur Nachnutzung bereit gestellt.

Identifikation von Desinformationen

Ein Ansatz dafür, Falschmeldungen zu begegnen, sind Dienste (Faktenchecker-Webseiten), die den Wahrheitsgehalt von Aussagen etwa in sozialen Netzwerken oder den Nachrichten untersuchen.

[2] https://www.aeksh.de/allianz-gegen-corona-leugnerIn, Abrufdatum 25.92.2022
[3] https://www.who.int/news-room/spotlight/ten-threats-to-global-health-in-2019, Abrufdatum 24.02.2022.

Für den Bereich Medizin gibt es die deutschsprachige Plattform MedWatch[4]. Das Überprüfen von einzelnen Aussagen ist zeitaufwändig und kostenintensiv – besonders, wenn es sich nicht nur um kurze Statements in Sozialen Netzwerken, sondern um längere Texte handelt. Die von diesen Initiativen z.B. durch Fach-Journalistinn:en kuratierten Daten können im Anschluss als Trainingsdaten für ML-Verfahren verwendet werden. Von den Ergebnissen dieser automatischen Einschätzungverfahren können wiederum die Faktencheck-Webseiten profitieren, indem etwa einzelne Aussagen durch ML-Techniken automatisiert mit Faktendatenbanken abgeglichen werden (Dale 2017, Babakar u. a. 2017, Oshikawa, Qian und Wang 2020).

Ein anderer Ansatz versucht, das Maß der inhaltlichen Kohärenz in Argumentationen in Texten zum Indikator von Wahrheit oder Falschheit von Dokumenten zu erheben (Singh, P. Deepak und Anoop 2020). Dabei wird die These verfolgt, dass Texte, die als Falschmeldungen gelten müssen, in sich weniger kohärent, d.h. fokussiert auf das eigentliche Thema sind. Mit computergenerierten Texten mit desinformierenden Inhalten beschäftigen sich Zhong u. a. 2020, die ebenfalls auf eine tendenziell andere Struktur in der Argumentation hinweisen. Von einer typischen Anordnung von desinformierenden Texten und sie illustrierenden Bildern wie auch deren Passung zum Thema machen dagegen Tan, Plummer und Saenko 2020 Gebrauch, um computergenerierte Falschmeldungen zu identifizieren.

Trotz der hohen gesellschaftlichen Relevanz für lebenswissenschaftliche Themen beschäftigen sich automatisierte Methoden des Identifizierens von Falschinformationen vorwiegend mit politischen Aussagen u.a. in sozialen Netzwerken (z.B. Vo und Lee 2020, Vogel und Jiang 2019). **Wissenschaftliche Texte und im besonderen Texte, die Gesundheitsinformationen transportieren, sind in den Forschungsansätzen der Textanalyse und des ML unterrepräsentiert (Kotonya und Toni 2020).**

Mit Blick auf Falschmeldungen im medizinischen Kontext verfolgen Ghosal, Padmanabhan Deepak und Jurek-Loughrey 2020 einen Ansatz, die zentralen Falschbehauptungen in Artikeln zu markieren. Kotonya und Toni 2020 stellen für den Bereich Public Health einen neuen Datensatz von 11 800 Aussagen aus Reuters News, Asssociated Press, Health News Review sowie Faktenchecker-Webseiten (u. a. Politifact, FactCheck, FullFact) vor, den sie mit den auf biomedizinische Themen vortrainierten Natural-Language-Processing-Modellen (NLP) auswerten. Sie errechnen dabei die Kohärenz zwischen Aussagen, um dadurch auf Korrektheit oder Falschheit einer Aussage zu schließen. **Für den deutschsprachigen Bereich gibt es nur wenige Ansätze, die sich inhaltlich jedoch nicht mit den LeWi befassen (Vogel und Jiang 2019, Chan, Schweter und Möller 2020).** Einen breiten Überblick über aktuelle Ansätze im Bereich Factchecking mit NLP-Methoden liefern Oshikawa, Qian und Wang 2020. Insgesamt fokussieren die genannten Ansätze auf eine konkrete Bewertung von Aussagen als eindeutig *wahr* oder *falsch*. Dieses Vorgehen ist derzeit für kurze Statements ausreichend, aber noch nicht für längere Abhandlungen, wie die mit denen wir es in der Regel in den LeWi zu tun haben, die komplexere Argumentationsstrukturen aufweisen und zB. nur unrichtig generalisieren oder unzulässig kontextualisieren.

AQUAS

Mit NLP in Kombination mit ML-Verfahren lassen sich typische Eigenschaften von Texten untersuchen und unbekannte Texte anhand der ermittelten Merkmale zuvor klassifizierten Textgattungen (target) zuordnen. Das hier vorgeschlagene Projekt *Automatic Quality Assessment: NLP-Verfahren zur semantischen Kartierung von lebenswissenschaftlichen Texten (AQUAS)* zielt darauf, einerseits Informationen zur Einordnung von Publikationen entsprechend einer guten wissenschaftlichen Praxis zu ermitteln und bereitzustellen und andererseits die Ähnlichkeit eines unbekannten deutschen oder englischen Textes zu den Klassen wissenschaftliche Texte, populärwissenschaftliche

[4] https://medwatch.de/, Abrufdatum 16.02.2022.

Texte oder desinformierende Texte anzugeben. Dabei wird je nach Projektverlauf das Basisset um die Zuordnung zu den Klassen erweitert. Im Basisset werden Angaben zu erfolgten Review-Verfahren, wissenschaftlichen Belegen und die Zitierung durch andere wissenschaftliche Akteure dokumentiert, aus denen sich für die Nutzenden die Compliance der Publikation mit einer „guten wissenschaftlichen Praxis“ ableiten lässt, wie sie die DFG vorsieht (DFG 2019). Im besten Fall gelingt es auf Grundlage der Modellierung der Trainings-Texte für jede Text-Gattung einen Wert errechnen, der die Wahrscheinlichkeit anzeigt, dass der unbekannte Text anderen Dokumenten dieser Klasse *ähnlich* ist.

Seit dem Erscheinen von Word2Vec (2013), haben sich die Embedding-Verfahren stark weiter entwickelt. State-of-the-Art-Verfahren nutzen vortrainierte Modelle, die dann anhand eines spezifischen Textkorpus für eine entsprechende Ausgabe des Modells nur noch justiert (fine tuning) werden müssen. Hier hat sich derzeit *BERT – Bidirectional Encoder Representation from Transformers* wegen seiner überragenden Performance und der Möglichkeit des Fine-Tunings auf relativ kleinen Datensätzen durchgesetzt (Devlin u. a. 2019). Im Rahmen von AQUAS werden wir die sehr gute Performance von BERT nutzen.

	deutsch	englisch
spezialisierte Texte	GMS Datenbank	PMC Datenbank
populärwissen-schaftliche Texte	Wikipedia (LeWi-Artikel) Apotheken-Umschau.de NetDoktor.de u.ä.	Wikipedia (LeWi-Artikel) MedlinePlus.gov MedHelp.org u.ä.
desinformierende Texte	de.sott.net pi-news.net anonymousnews.ru u.ä.	PUBHEALTH Datensatz HWB Datensatz

Abbildung 1: Datengrundlage und Klassen

Eine übliche Schwierigkeit solcher Ansätze, ist das Problem des Overfittings, das unbedingt vermieden werden muss, um eine konstruktive und seriöse Information der Nutzenden zu garantieren. In Zukunft sind im ML weitere verbesserte Methoden zu erwarten, die es auch im Falle eines Projektabschlusses, der die Zuordnung zu den Klassen aus Qualitätsgründen noch nicht kommunizierbar macht, rechtfertigen, die Datensätze und Service für spätere verbesserte Methoden vorzubereiten und zu etablieren. Im deutschen Sprachraum gibt es für den Bereich LeWi nach unserer Kenntnis keinen entsprechenden Datensatz. AQUAS füllt mit der Erstellung des Datensatzes eine Lücke.

Die ermittelten Informationen, die durch die Methoden von AQUAS erstellt werden, sollen dabei nicht das Ziel einer Angabe über die Wahrheit eines Text sein oder sogar vor dem Lesen *warnen*, wie das beispielsweise Tan, Plummer und Saenko 2020 explizit vorhaben (vgl. 1.1). Durch die detaillierte Anzeige von Informationen zum Text wird bei AQUAS stattdessen eine unterschwellige Zensur zu Gunsten der Befähigung zur Selbsteinschätzung der Nutzenden vermieden. Darin unterscheidet sich AQUAS von vielen anderen Vorhaben aus dem Bereich fact checking/stance detection. Wichtig ist dabei auch ein transparente Kommnikation über die Funktionsweise des ML-Modells (Explainability; vgl. AP4)

Vorarbeiten

Klassifizierung: spezialisierte, populärwissenschaftliche und desinformierende Texte

Die Klassifizierung und Bezeichnung der Textgattungen, deren semantische Merkmale mit Hilfe von AQUAS auseinander gehalten werden sollen, ist besonders sensibel, da sie eine wichtige Grundlage für die Auswahl der Daten und das Training der vorgeschlagenen Modelle darstellt.

Spezialisierte Texte und *populärwissenschaftliche Texte* lassen sich leicht abgrenzen. Spezialisierte Texte kommunizieren evidenzbasiertes Wissen, das sich besonders durch eine Fachsprache auszeichnet. In den LeWi ist dieses Wissen in der Regel durch Studien oder Experimente abgesichert und baut auf früheren Ergebnissen auf. Populärwissenschaftliche Texte machen dieses spezialisierte Wissen einer größeren Gruppe zugänglich, indem sie wissenschaftliche Themen vereinfachen oder in einem Überblick darstellen. Die dritte Gruppe der *Desinformationen* umfasst Texte zu lebenswissenschaftlichen Themen aus dem Bereich Verschwörungserzählungen, Pseudomedizin (z.b. Homöopathie), Esoterik oder auch Pseudowissenschaft. Charakteristisch ist hier der Gestus

eines Geheimwissens, das unterdrückt werde und noch nicht die Beachtung erfahre, die ihm zustehe. Für die Identifizierung typischer desinformierender Texte greifen wir auf die Einschätzung von Bundeszentrale für politische Aufklärung, ARD-Faktenfinder und MedWatch zurück. Auch stehen wir in Kontakt zu der Arbeitsgruppe des Institute of Electrical and Electronics Engineers (IEEE), die derzeit einen Standard für die Glaubhaftigkeit von Nachrichtenseiten erstellt.[5] Im Dezember 2022 soll der Standard publiziert werden, der, wenn er sich auch nicht auf wissenschaftliche Informationsquellen bezieht, einen Orientierung für den hier beschriebenen Kontext geben kann. Ein Standard für Wissenschaftlichkeit selbst liegt derzeit nicht vor und wird unseres Wissens nach auch nicht vorbereitet.

Datengrundlage

Der Erfolg des Vorhabens steht und fällt mit der Datengrundlage. Eine Gefahr ist, dass das ML-Modell formale Charakteristika einer Textgattung lernt und sie für inhaltliche Eigenschaften hält (Oshikawa, Qian und Wang 2020). Daher ist ein gründliches Preprocessing und Bereinigung der Daten im Vorfeld unabdingbar.

Das Training des beschriebenen Modells wird für die Kategorie der **Spezialisierten Texte** auf den Datenbanken ***PubMed Central* (PMC)** für den englischen und für den deutschen Sprachraum auf der Datenbank ***German Medical Science Portal* (GMS)** erfolgen.

GMS ist ein Publikationsportal, das in Zusammenarbeit mit der *Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e.V.* (AWMF), dem *Bundesinstitut für Arzneimittel und Medizinprodukte* (BfArM) und von ZB MED herausgegeben wird. Seit Gründung des GMS-Portals 2003 wurden mehr als 6 000 Zeitschriftenartikel veröffentlicht (Stand Februar 2022). Drei der Zeitschriften publizieren in deutscher Sprache, bei einigen anderen Zeitschriften liegen die Abstracts in deutsch und englisch vor. Die Inhalte von GMS werden unter einer CC-BY-4.0-Lizenz veröffentlicht. Die Jahrgänge bis 2015 stehen unter einer CC-BY-NC-ND-3.0-Lizenz.

PMC ist ein Repositorium für lebenswissenschaftliche, überwiegend englischsprachige Artikel und wird seit 2000 von der NLM betrieben. Die in PMC gesammelten Titel stehen überwiegend unter einer offenen Lizenz.

Die Nutzung von Webseiten und kostenpflichtigen Downloads etwa durch das Scraping der Inhalte ist seit dem 1.3.2018 durch §60d des **Urheberrechts-Wissensgesellschafts-Gesetzes** (UrhWissG) auch das Text- und Data-Mining für wissenschaftliche, nicht kommerzielle Zwecke geregelt. Hierdurch wird gestattet, aus urheberrechtlich geschützten Inhalten einen Datenkorpus zu erstellen, um ihn wissenschaftlich auszuwerten. Demnach ist eine umfassende Nutzung auch der Volltexte von PMC für AQUAS möglich. Für den Bereich der **populärwissenschaftlichen Texte** können **Wikipedia**, aber auch Verbraucher-Foren im Internet wie **Apotheken-Umschau**[6], **NetDoktor**[7] (beide deutsch), **MedlinePlus**[8] oder **Medhelp**[9] (beide englisch) genutzt werden. Wikipedia gilt als das größte allgemeine Nachschlagewerk im Internet, dessen englischsprachige Version mit 6,3 Millionen Einträgen die umfassendste ist. Die Artikel und ihre Versionsgeschichte können über eine Schnittstelle (API) heruntergeladen werden. Es stehen Artikel in Englisch und Deutsch zur Verfügung, wobei nicht jeder Artikel ein Pendant in der anderen Sprache hat. Die Nutzung der Inhalte aus Wikipedia ist durch die Lizenz CC-BY-SA-3.0 geregelt. Für die Eingrenzung auf lebenswissenschaftliche Artikel können wir auf die wikipedia-interne Inhaltliche Systematik

[5] https://development.standards.ieee.org/myproject-web/public/view.html#pardetail/6318, Abrufdatum 16.02.2022.
[6] https://www.apotheken-umschau.de/, Abrufdatum 16.02.2022.
[7] https://www.netdoktor.de, Abrufdatum 16.02.2022.
[8] https://medlineplus.gov, Abrufdatum 16.02.2022.
[9] https://www.medhelp.org, Abrufdatum 16.02.2022.

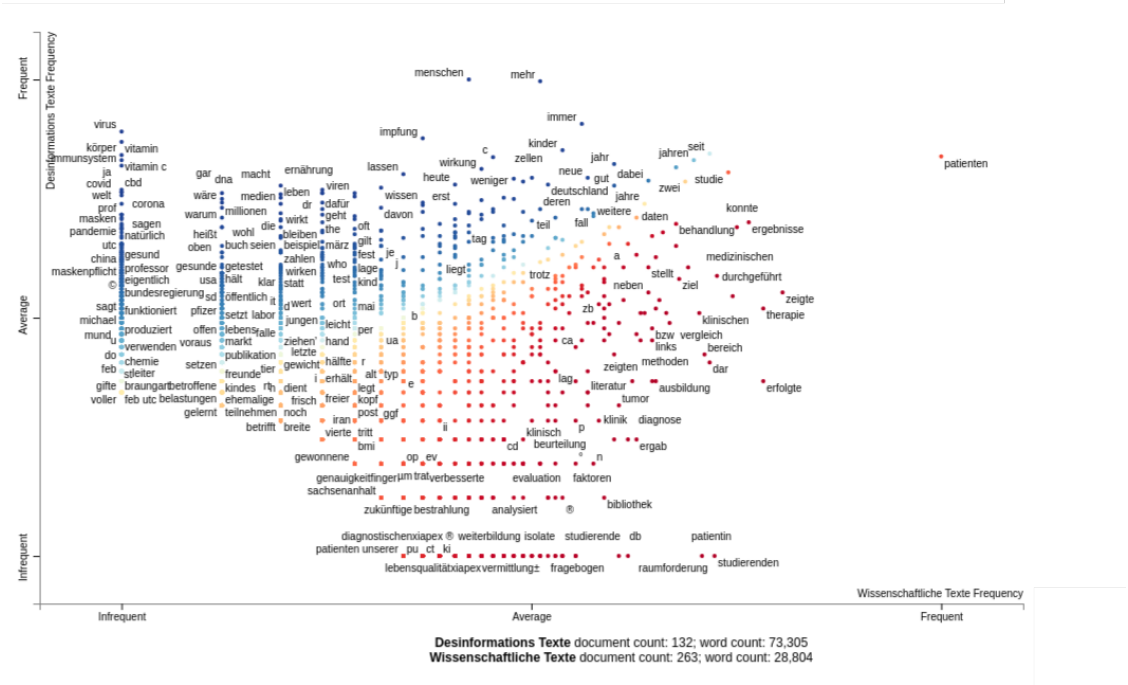


Abbildung 2: TF-IDF-Verteilung der Begriffe im deutschen Probedatensatz (desinformierende Texte: blau, spezialisierte Texte: rot)

(Sacherschließung) zurückgreifen.[10] Mit den *Kategorien* werden in Wikipedia Artikel verschiedenen inhaltlichen Bereichen (z.B. Biowissenschaften, Medizin, Ernährungswissenschaften) zugeordnet.

In der Apotheken-Umschau werden seit 1956 leicht verständliche Gesundheits-Artikel veröffentlicht, mit dem Ziel, die Lesenden wissenschaftlich fundiert und unabhängig zu informieren. Alle Artikel werden von Mediziner:innen vor Erscheinen überprüft. NetDoktor wird vom Burda-Verlag herausgegeben. Die Artikel der Webseite werden von Wissenschaftsjournalistinn:en mit Fachhintergrund verfasst. MedlinePlus ist das VerbraucherPortal der NLM mit Überblicksartikeln zu umfassenden Themen der Lebenswissenschaften. MedHelp ist eine der meist genutzten Webseiten für Gesundheitinformationen in den USA. Alle Artikel der Verbraucherinformationswebseiten können im Einklang mit dem UrhWissG aus dem Internet abgerufen werden.

Als **englischsprachigen Datensatz** für **Desinformation** können wir den kürzlich am Imperial College London zusammengetragenen Datensatz ***PUBHEALTH*** verwenden (Kotonya und Toni 2020). Der 11 800 Items starke Datensatz wurde journalistisch geprüft und die einzelnen Texte vier Kategorien *wahr, falsch, ungeprüft, gemischt* zugewiesen. Einen ähnlichen, etwas kleineren Datensatz (1 000 Items) für den Gesundheitsbereich ist das ***Health & Well Being (HWB)* Fake News Dataset** (Singh, P. Deepak und Anoop 2020). Die seriösen Nachrichten stammen aus Quellen wie CNN, New York Times oder New Indian Express, die nicht seriösen Artikel von einschlägigen Fehlinformations-Websites wie u.a. BeforeItsNews, Nephef oder MadWorldNews. Alle Aussagen wurden manuell auf ihren Wahrheitsgehalt überprüft. Beide Datensätze steht unter freien Lizenzen und können durch AQUAS nachgenutzt werden.

Eine größere Herausforderung stellt die Datengrundlage für das Training der Kategorie **Desinformation im deutschsprachigen Raum** dar. Hier gibt es nach unserem Wissen einen einzigen Datensatz – *GermanFakeNC* –, der inhaltlich aber vor allem Themen der Inneren Sicherheit und Migration bedient (Vogel und Jiang 2019). **Daher ist es nötig, einen neuen Datensatz für Des-**

[10] https://en.wikipedia.org/wiki/Wikipedia:Categorization, Abrufdatum 16.02.2022.

information im lebenswissenschaftlichen Feld aufzubauen. Für die Auswahl dessen, was als desinformierend gelten kann, können wir uns an Einschätzungen von *ARD-Faktenfinder*, *Bundeszentrale für politische Bildung* und *MedWatch* orientieren, bzw. dem oben genannten Standard zur Glaubhaftigkeit von Nachrichtenseiten des IEEE. Dieses Vorgehen hatte auch die Herausgeber:innen hinter HWB und GermanFakeNC gewählt (Singh, P. Deepak und Anoop 2020, Vo und Lee 2020). GermanFakeNC enthält z.T. ähnliche Quellen wie die von uns projektierten (s.u.).

Auch hier erlaubt es uns die Regelung des UrhWissG, die für das Training der Kategorie Desinformation benötigten Datengrundlage von vorhandenen Webauftritten abzurufen, für die Dauer der Analyse zu speichern und den benötigten Korpus zu erstellen. Artikel aus den Online-Plattformen *Anonymous News*[11], *Signs of the times)*[12], Politically Incorrect[13] oder *Compact Magazin online*[14] zu Gesundheits-Themen werden in den Korpus einbezogen. Auch das im gleichlautenden Verlag angebotene Magazin *WissenschafftPlus*[15] [sic], das sich einen wissenschaftlichen Anstrich gibt, sowie für Gesundheitsthemen relevante Artikel der im Umfeld der *Neuen Rechten* erscheinenden Zeitschrift *Sezession*[16] werden dem temporären Korpus zugefügt. Im Kopp-[17] sowie im Antaios-Verlag[18] sind viele Bücher erschienen, die als desinformierend oder pseudowissenschaftlich gelten müssen (Zywietz und Sachs-Hombach 2018). Einige Publikationen stehen nicht unter einer Open-Access-Lizenz und können aber wie beschrieben auf Grundlage des UrhWissG für die Dauer des Forschungsprojekts durch AQUAS genutzt werden. Vogel und Jiang 2019 hatten z.T. die gleichen Quellen gewählt.

Immer gilt, dass nicht notwendigerweise alle Texte einer Webseite oder Zeitschrift sich mit Themen aus dem Feld LeWi beschäftigen oder desinformierend sind. Dies erfordert eine nachträgliche manuelle Auswahl, um den inhaltlichen Fokus zu stärken. Bei der inhaltlichen Bewertung werden wir von den Wissenschaftsjournalist:innen von MedWatch unterstützt (vgl. Letter of Intent MedWatch)

Den erstellten **Datensatz zu Desinformationen im deutschsprachigen Raum werden wir im Anschluss im Einklang mit den F.A.I.R.-Prinzipien anderen wissenschaftlichen oder Infrastruktur-Vorhaben bereitstellen. Leider sieht das UrhWissG die Löschung der gescrapten Texte nach Abschluss des Forschungsprojekts vor.** Gleichwohl können wir einen Datensatz, in dem URLs auf entsprechende Texte verweisen, und Download-Skripte zur Nachnutzung zur Verfügung stellen, was ein automatische Reproduktion des Corpus ermöglicht. Durch die Nutzung von Zeitstempeln können dann mittels der Wayback Machine[19] des Internet Archives auch gelöschte oder verschobene Inhalte wieder zusammengetragen getragen werden. Dieses Arrangement ist kein wünschenswertes Vorgehen, jedoch das einzige rechtskonforme.

Erstes Anwendungsfeld Lebenswissenschaften (LeWi) und spätere Nachnutzung

Als Ansatzpunkt konzentriert sich AQUAS auf eine Bearbeitung von lebenswissenschaftlichen Texten. Der hohen Relevanz des Problems von Falschnachrichten im Gesundheitsbereich (Holone 2016, Speed und Mannion 2017, Singh, P. Deepak und Anoop 2020, European Commission, Directorate-General for Communications Networks, Content and Technology 2021), steht die eklatante Vernachlässigung des Themas im ML-Bereich gegenüber (Kotonya und Toni 2020).

[11] https://www.anonymousnews.ru, Abrufdatum 16.02.2022.
[12] https://de.sott.net, Abrufdatum 16.02.2022.
[13] https://www.pi-news.net, Abrufdatum 16.02.2022.
[14] www.compact-online.de, Abrufdatum 16.02.2022, Compact Magazin wird seit März 2020 vom Verfassungsschutz als rechtsextremer Verdachtsfall geführt.
[15] http://wissenschafftplus.de, Abrufdatum 16.02.2022.
[16] https://sezession.de, Abrufdatum 16.02.2022.
[17] https://www.kopp-verlag.de, Abrufdatum 16.02.2022.
[18] https://antaios.de, Abrufdatum 16.02.2022.
[19] https://archive.org/web/, Abrufdatum 16.02.2022.

Gleichzeitig gibt es zumindest für den seriösen Bereich gut kuratierte Textkorpora, die für das Text-Mining nutzbar sind.

Das erarbeite Multi-Label-Klassifikations-Modell und der darauf aufbauende Service können in der Folge auf andere Bereiche, wie z. B. auf gesellschaftspolitische Themenfelder, übertragen werden. Hierzu sind wir bereits mit anderen Institutionen im Gespräch (vgl. Letters of Intent der TIB, ZBW und GESIS). Möglich ist hier etwa auch die Einbindung des Tools in die Umgebung von Wikidata, das eine relativ einfache Umsetzung eines Frontends – unabhängig von LIVIVO – im Stile von existierenden Open-Source-Tools wie Scholia[20] (Nielsen et al. 2017) erlaubt (vgl. Letter of Intent Daniel Mietchen).

Institutionelle Ansiedlung bei ZB MED

Als zentrales Informationszentrum für die LeWi in Deutschland ist für ZB MED die Steigerung der Informationskompetenz ein genuines Anliegen.[21] Das beantragte Projekt wird in der Arbeitsgruppe Data Science and Services von Konrad Förstner realisiert, der zusätzlich Professor für *Data and Information Literacy* am *Institut für Informationswissenschaft* an der Technischen Hochschule Köln ist. In diesem Sinne versteht sich AQUAS als weiterer Beitrag von ZB MED zu einer Stärkung der demokratischen Öffentlichkeit in Zeiten von Desinformationskampagnen. Dieses Anliegen fügt sich in das Engagement von ZB MED für die F.A.I.R.-Prinzipien ein, Open Access und Transdisziplinarität (Seidlmayer und Poley 2017, Seidlmayer und Arning, Ursula 2019, Erdmann u. a. 2019, Seidlmayer, R. Müller und Förstner 2020).

Durch die Anbindung an das Forschungsinstitut ZB MED kann die Projektarbeit von der vorhandenen Infrastruktur bei ZB MED profitieren. Die erfahrene IT-Abteilung wird die benötigten Server betreuen, durch die eine dauerhafte Bereitstellung des entwickelten Service sichergestellt wird.

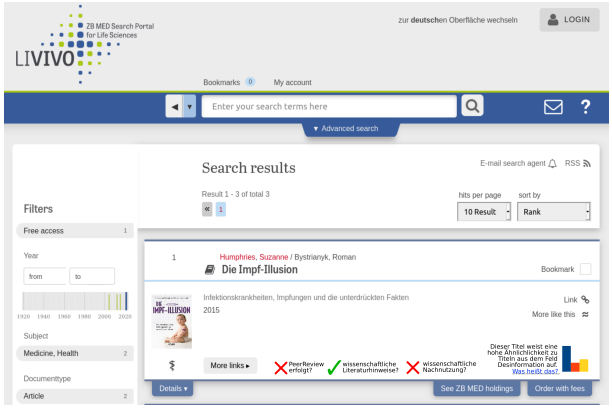


Abbildung 3: Mögliche Titelanzeige mit Ergebnisanzeige als Grafik in LIVIVO, hier Suzanne Humphries (2015): Die Impf-Illusion

Mit LIVIVO hat ZB MED, in Nachfolge der Suchmaschinen *MedPilot* und *Greenpilot*, seit über zwanzig Jahren Erfahrung im Anbieten von Discovery-Systemen (B. Müller u. a. 2017). LIVIVO stellt Inhalte der LeWi aus mehr als 50 Datenbanken bereit und hat im Jahr 2020 mehr als 2 Millionen Suchanfragen verarbeitet. Hiermit liegt eine sehr gute Arbeitsumgebung für die Implementation des beschriebenen Modells als Service von ZB MED vor. Schon in früheren Projekten wurde LIVIVO und die Datengrundlage der Knowledge Environment für Forschungszwecke genutzt, so im DFG-geförderten Projekt STELLA.[22] Selbstverständlich wird die im Projekt entwickelte Software unter einer permissiven OSI-konformen Open-Source-Lizenz auf GitHub bereitgestellt und auf Zenodo archiviert werden. Gleiches gilt für die Datenkorpora (siehe Abschnitt *Datengrundlage*), soweit möglich.

In verschiedenen Projekten haben wir Erfahrungen mit Metadaten und Information Retrieval im großen Stil sammeln können und auch State-of-the-Art ML-Verfahren eingesetzt (Galke, Melnychuk u. a. 2019, Seidlmayer, Galke u. a. 2019, Seidlmayer, Voß u. a. 2020, Melnychuk u. a. 2021, Galke,

[20] https://scholia.toolforge.org, Abrufdatum 16.02.2022.
[21] https://www.zbmed.de/ueber-uns/presse/neuigkeiten-aus-zb-med/artikel/zb-med-unterstuetzt-data-literacy-charta/, Abrufdatum 25.92.2022
[22] Projektnummer 407518790, https://www.zbmed.de/forschen/laufende-projekte/stella, Abrufdatum 16.02.2022.

Seidlmayer u. a. 2021. Durch die forschungsstarke Arbeitsumgebung bei ZB MED ist der kontiuierliche, direkte, inhaltliche Austausch zu NLP- und ML-Themen mit Kolleginn:en selbstverständlich (u.a. mit der Forschungsgruppe Wissensmanagement von Prof. Dr. Juliane Fluck (siehe Letter of Intent)).

Analyse zu Differenzierung der drei Texttypen

Um die Durchführbarkeit des Vorhabens zu belegen, wurde ein explorativer Datensatz erstellt, der 993 deutschsprachige Texte enthält. Die Kategorie populärwissenschaftlicher Texte enthält 300 Artikel aus Wikipedia zu Themen der Biowissenschaft, Medizin und Ernährungswissenschaft, sowie 111 Artikel aus Apotheken-Umschau und MedWatch. Weiterhin gingen 450 Abstracts als spezialisierte Texte von der medizinischen Fachplattform GMS in den explorativen Datensatz ein, daneben 132 Texte von einschlägigen Webseiten mit potentiell desinformierenden Inhalten wie: *Signs of the times*, Politically Incorrect, Anonymous News und Sezession. Die Artikel wurden dort nach einer Suche u.a. nach *Impfen, Nahrungsergänzungsmittel, queer, Krebs, vegan, Gesundheit, Medizin, Corona* aufgefunden. Die Webseite Signs of the times stellt einen eigenen Bereich *Gesundheit und Wohlbefinden* bereit, den wir für die Kategorisierung nutzen konnten.

Der hier vorgestellte Probedatensatz ist hinsichtlich der Anzahl der Dokumente für die einzelnen Kategorien derzeit noch unausgeglichen (unbalanced). Für eine erste Klassifizierung ist dies vernachlässigbar; im eigentlichen Training des Modells würde dies aber zu einer Verzerrung der Klassifizierung zu Gunsten der majority class führen. (Fernández u. a. 2018). Für das spätere Training der AQUAS-Modelle (deutsch/englisch) müssen die Anzahlen der Titel zahlenmäßig noch weiter angepasst bzw. der Mengenunterschied durch *Sampling* ausgeglichen werden (Fernández u. a. 2018).

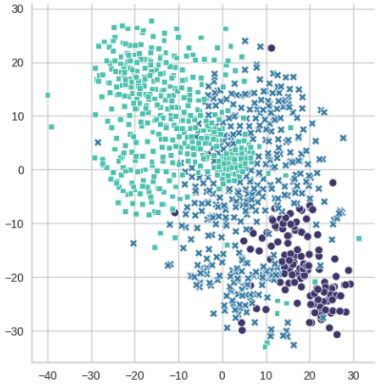


Abbildung 4: Unsupervised Clustering des deutschen Test-Datensatzes mit Doc2Vec: desinformierende (Kreise), populärwissenschaftliche (Kreuze) und spezialisierte Texte (Quadrate)

Um zu ermitteln, ob die Texte ausreichend unterschiedliche Merkmale aufweisen, was die Voraussetzung für das Training des Modells ist, haben wir zwei Unüberwachtes-Lernen-Analysen (*unsupervised classification*) mit TF-IDF (*term frequency – inverse document frequency*) und Doc2Vec durchgeführt.

Mit dem etablierten Maß der inversen Häufigkeit von Termen in Dokumenten, TF-IDF, können die relevanten Begriffe bestimmt und unterschiedliche Gruppen detektiert werden. Dazu wurden die Daten mit dem Visualisierungswerkzeug Scattertext[23] in einen speziellen Korpus verwandelt und mit spaCy[24] geparst. Stoppwörter wurden gerausgenomen, die Begriffe mit Stemming prozessiert. Die Abbildung 2 zeigt die Verteilung häufiger und weniger häufiger vorkommender Begriffe in 132 Texten der Desinformation und 263 wissenschaftlichen Abstracts.

Zum Abgleich haben wir mit *doc2vec* ein zweites Verfahren zur unüberwachten Klassifizierung der Daten getestet. Mit *gensim*[25] wurden die Texte vorprozessiert und Tokens aus einzelnen Wörtern erstellt. Im Anschluss erstellten wir mit *gensim* für jedes Dokument einen Doc2Vec-Vektor mit einer Länge von 40 Dimensionen und trainierten ein Modell in 60 Epochen. Die dadurch erzeugte Matrix für den Dokumenten-Korpus zeigt in der t-SNE-Projektion (siehe Abbildung 4) eine klare

[23] https://spacy.io/universe/project/scattertext, Abrufdatum 16.02.2022.
[24] https://spacy.io, Abrufdatum 16.02.2022.
[25] https://radimrehurek.com/gensim, Abrufdatum 16.02.2022.

Abgrenzung der Texttypen. Wie bei der ersten Analyse mit der TF-IDF-Methode wurden die Labels der einzelnen Text-Typen erst im Nachhinein zu den einzelnen Datenpunkten hinzugefügt.

Beide Verfahren zeigen eine klare Abgrenzung der Textarten; die drei Klassen sind ausreichend unterschiedlich, so dass ein Training von Modellen auf dieser Grundlage zielführend sein kann.

Projektbezogenes Publikationsverzeichnis Ihrer Arbeiten

Veröffentlichte Arbeiten aus Publikationsorganen mit wissenschaftlicher Qualitätssicherung

Erdmann, Christopher u. a. (2019). *Top 10 FAIR Data & Software Things*. https://zenodo.org/record/2555498.

Galke, Lukas, Tetyana Melnychuk, Eva Seidlmayer, Steffen Trog, Konrad U. Förstner, Carsten Schultz und Klaus Tochtermann (2019). „Inductive Learning of Concept Representations from Library-Scale Bibliographic Corpora“. In: *INFORMATIK*. doi: 10.18420/INF2019_26, S. 219–232.

Galke, Lukas, Eva Seidlmayer, Gavin Ludemann, Lisa Langnickel, Tetyana Melnychuk, Konrad U. Forstner, Klaus Tochtermann und Carsten Schultz (15. Dez. 2021). „COVID-19++: A Citation-Aware Covid-19 Dataset for the Analysis of Research Dynamics“. In: *2021 IEEE International Conference on Big Data (Big Data)*. 2021 IEEE International Conference on Big Data (Big Data). Orlando, FL, USA: IEEE, S. 4350–4355.

Melnychuk, Tetyana, Lukas Galke, Eva Seidlmayer, Konrad U. Förstner, Klaus Tochtermann und Carsten Schultz (2021). „Früherkennung wissenschaftlicher Konvergenz im Hochschulmanagement“. In: *Hochschulmanagement* 1, S. 24–28.

Seidlmayer, Eva und Arning, Ursula (2019). *Open-Access-Geschäftsmodell für PUBLISSO – ZB MED Publikationsplattform für die Lebenswissenschaften*. https://b-i-t-online.de/heft/2019-01-fachbeitrag-seidlmayer.pdf.

Seidlmayer, Eva, Lukas Galke, Tetyana Melnychuk, Carsten Schultz, Klaus Tochtermann und Konrad U Forstner (Sep. 2019). „Take it Personally - A Python library for data enrichment in informetrical applications“. In: *Posters and Demos at SEMANTiCS 2019*. CEUR Workshop Proceedings. urn:nbn:de:0074-2451-7. Alam, Mehwish; Usbeck, Ricardo; Pellegrini, Tassilo; Sack, Harald; Sure-Vetter, York, S. 5.

Seidlmayer, Eva, Rabea Müller und Konrad U. Förstner (2020). „Data Literacy for Libraries – A Local Perspective on Library Carpentry“. In: *BIBLIOTHEK – Forschung und Praxis*. doi: http://dx.doi.org/10.18452/22009.

Seidlmayer, Eva und Christoph Poley (2017). „One Health – Transdisziplinarität bei ZB MED“. In: *GMS Medizin – Bibliothek – Information; 17(3)*. doi: 10.3205/MBI000400.

Seidlmayer, Eva, Jakob Voß, Tetyana Melnychuk, Lukas Galke, Klaus Tochtermann, Carsten Schultz und Konrad U. Förstner (Nov. 2020). „ORCID for Wikidata – Data enrichment for scientometric applications“. In: *CEUR-WS*. CEUR Workshop Proceedings. urn:nbn:de:0074-2773-1. Lucie-Aimée Kaffee, Oana Tifrea-Marciuska, Elena Simperl, Denny Vrandečić.

2 Ziele und Arbeitsprogramm

2.1 Gesamtdauer des Projekts

Das Projekt wird auf drei Jahre für die Phase der anwendungsbezogenen Forschung und Entwicklung von e-Research-Technologien beantragt. In der ersten Phase wird die Infrastruktur aufgebaut und Methoden exploriert und in der zweiten Phase in einen Service implementiert, der gemeinsam mit der Fachcommunity evaluiert und angepasst wird (vgl. 2.2).

2.2 Ziele

Das vorgeschlagene Projekt AQUAS erarbeitet einen Service, der Wissenschaftler:innen sowie Arbeitnehmende in Medizinberufen bei einer informierten Recherche und Auswahl von relevanten Fachinhalten unterstützt.

Ziel 1: Erstellung des ersten deutschsprachigen Datensets zum Thema LeWi: Es werden zwei Datensätze (deutsch und englisch) erstellt, die sich aus Texten der Bereiche *spezialisierte Texte*, *populärwissenschaftliche Texte* und *desinformierende Texte* zusammensetzen. Die Texte, die aus urheberrechtlichen Gründen nicht veröffentlicht werden können, werden über URL-Zusammenstellungen automatisiert reproduzierbar gemacht. Für den deutschen Sprachraum ist dies der erste Datensatz.

Ziel 2: Informationsermittlung zu Inhaltseinschätzung: Das Basisset wird je nach Qualitäts-Abwägung des zu erarbeitenden ML-Modells um die ermittelten Klassifizierungergebnisse ergänzt.

Ziel 3: AQUAS-Service: AQUAS stellt einen Service mit einer API bereit, an den ein zu analysierender Text-String übergeben werden kann. Der Dienst liefert als Rückgabewert die beschrieben Informationen im JSON-Format aus. Die erzeugten Werte können von verschiedenen externen Applikationen genutzt werden – z.B. LIVIVO (vgl. Ziel 4).

Ziel 4: AQUAS-Applikation in LIVIVO: Im Rahmen von AQUAS wird der AQUAS-Service für die Suchmaschine LIVIVO genutzt. Zur leichteren Erfassung der Werte durch die Nutzenden werden sie als Grafik angezeigt (vgl. Abbildung 3). *Ziel 5: Sensibilisierung für das Thema Desinformation:* Wissenschaftler:innen in ihrer Informationskompetenz mit Blick auf Desinformationen zu stärken ist ein grundsätzliches Anliegen des Vorhabens. Dies geschieht im Projekt nicht nur durch die Erstellung des beschriebenen Werkzeuges, sondern auch durch den direkten Austausch mit den Wissenschaftler:innen in verschiedenen Formaten (Vorträge, Fachartikel, Podcasts).

Für das vorliegende Projektvorhaben sind im Kern drei **Zielgruppen** relevant, deren Arbeitsweise durch den Service bereichert werden soll: Zum einen **Forschende** der LeWi und medizinische Praktiker:n, die den Service über das Web verwenden; zum anderen **wissenschaftliche Forschungsinfrastruktureinrichtungen und Plattformbetreiber** von akademischen Suchsystemen, die die API in ihren eigenen Service integrieren; sowie **Infrastruktureinrichtungen und Forschende der Informationswissenschaften**, die die Methoden zur Erstellung eines Desinformations-Datensatzes in deutscher Sprache und unsere Optimierung des BERT-Modells nachnutzen wollen.

Exemplarisch für die zweite Gruppe soll im Projekt AQUAS ZBMED mit seinem Suchsystem LIVIVO stehen.

Arbeitsprogramm und Umsetzung

Im Folgenden werden sechs Arbeitspakete (AP) zur Umsetzung des Projektes beschrieben.

AP 0: Projektmanagement, Infrastruktur

Umfang AP 0 (1 Monat):

Um die Dokumentation und Organisation der Projektschritte zu gewährleisten sowie um das Projekt schon früh offen zugänglich zu machen, wird ein GitHub-Repositorium mit Webseite eingerichtet (ähnlich wie beim Projekt Q-Aktiv[26]). Auf dieser Webseite sollen unter anderem Vorgehensweise und Lösungen zu technischen Problemen dokumentiert werden.

Ergebnis AP 0 Das interne Projektmanagement und die Dokumentation sind sichergestellt.

AP 1: Erstellung der Datensätze und Preprocessing

Umfang AP 1 (12 Monate): Der Datensatz ist wichtige Grundlage für das Gelingen des Projektes. Die inhaltliche Fokussierung der ausgewählten Texte, die manuell durchgeführt wird, ist notwendig, um später im Training valide Ergebnisse zu erreichen (Kotonya und Toni 2020, vgl. 1.1). Bei der inhaltlichen Einschätzung der Titel unterstützt uns MedWatch. Um zu verhindern, dass das Modell die formalen Eigenheiten der Texte in seinen Lernprozess einbezieht, muss eine intensive Bereinigung der Texte durchgeführt werden (Oshikawa, Qian und Wang 2020).

Die Kompilierung der deutschen und englischen Datensätze für die *spezialisierten Texte* und *populärwissenschaftlichen Texte* von PMC und Wikipedia ist verhältnismäßig einfach, da sie frei heruntergeladen werden können. Die Abstracts von GMS liegen bereits intern als CSV-Daten

[26] https://q-aktiv.github.io, Abrufdatum 16.02.2022.

vor. Die englischen Texte für Desinformationen aus dem Feld der LeWi entnehmen wir dem PUBHEALTH- und dem HWB-Datensatz (Vgl.1.1). Die populärwissenschaftliche Textkategorie wird außerdem um Texte aus wissenschaftsjournalistischen Webseiten mit Verbraucherinformationen ergänzt. Die Inhalte von MedlinePlus und MedHelp sind hier für die englische Sprache maßgeblich. Für den deutschen Korpus nutzen wir Artikel aus Apotheken-Umschau, NetDoktor und MedWatch. Die Inhalte werden von den Webseiten heruntergeladen, für die inhaltliche Fokussierung der Texte aus PMC können MeSH-Terme sowie für jene aus Wikipedia die interne Wikipedia-Klassifizierung verwendet werden (vgl. 1.1).

Da es keinen Datensatz zu deutschsprachigen Desinformationstexten gibt, werden wir ihn über ein Webscrapingverfahren selbst erstellen (vgl. 1.1).

Trotz der erwartbaren Schwierigkeiten in der Zusammenstellung der Desinformationstexte streben wir einen vier- bis fünfstelligen Datensatz an. Für das Training des vorprozessierten Modells BERT ist dies ausreichend. Viele Datensätze, die im Bereich Fact Checking verwendet werden, bewegen sich in dieser Größenordnung oder sind kleiner (Oshikawa, Qian und Wang 2020, HWB, GermanFakeNC). Ein mögliches Missverhältnis zwischen den Klassen, wird durch Sampling Methoden ausgeglichen.

Ergebnis AP 1 Es liegen zwei zusammengestellte Datensätze (englisch/deutsch) mit Texten aus den drei genannten Bereichen vor.

AP 2: Exploration des ML-Modells (Training, Feinjustierung)

Umfang AP 2 (6 Monate)

Anders als in anderen ML-Verfahren, die Tausende von Items verarbeiten, ist für AQUAS ein kleinerer, das heißt vier- bis fünfstelliger, Datensatz zu erwarten. Hierfür eignet sich BERT sehr gut, das als vortrainiertes Modell für die entsprechende Fragestellung und das spezifische Datenset fine-getuned werden kann. Während wir für das Training des englischen Modells die BERT-Derivate HealthBERT oder BioBERT verweden werden, wird für das Training des deutschen Modells GermanBERT (Chan, Schweter und Möller 2020) genutzt, das speziell für die deutsche Sprache vorbereitet wurde. Weil wir mit unserem Modell nicht eine Klasse als Ergebnis erhalten möchten, sondern drei numerische Werte, verzichten wir auf den Softmax-Layer, der die Harmonisierung durchführen würde. Damit stellen wir sicher, drei Werte für drei Klassen ausgegeben zu erhalten.

Ergebnis AP 2: Zwei trainierte und für die LeWi optimierte Multi-Label-Klassifikation-Modelle (deutsch/englisch) liegen vor.

AP 3: Qualitätskontrolle des ML-Modells

Umfang AP 3 (4 Monat)

Der Erfolg des für unsere Fragestellung optimierten ML-Modells können wir über den Vergleich mit der Performance anderer Modelle z. B. in GermEval[27] abschätzen. Die Daten, die zur Evaluation genutzt werden, werden ein größeres Spektrum von Quellen kompilieren, um ein Overfitting an formale Charakteristika der Trainingsdaten ausschließen zu können. Zur weiteren Einschätzung der Güte des Ansatzes wird Menschen die gleiche Aufgabe der Zuordnung von Texten zu den genannten Klassen gestellt (Human in the Loop). Unter Berücksichtigung des Interrater-Agreements (also das Menschen sich oft nicht einig sind), lässt sich abwägen, wie brauchbar das erstellte Modell zur Einordnung von Texten in die genannten Klassen ist. Für den Fall, dass am Ende der Evaluation keine valide Zuweisung zu den Kategorien durch den ML-Klassifikator erreicht werden kann, werden die erweiterten Merkmale der Publikationen zunächst nicht an die Nutzenden

[27] https://projects.fzai.h-da.de/iggsa/germeval, Abrufdatum 16.02.2022.

kommuniziert. Da in den nächsten Jahren weitere Erkenntnisse im ML erwartbar sind, können die gemachten Vorarbeiten gesichert werden und die benötigte Service aufgebaut werden. Ein Folge-Projekt kann dann auf den Vorarbeiten aufbauen und neue Möglichkeiten des ML einbeziehen. In der Zwischenzeit lassen sich den Nutzenden bereits erste Informationen aus anderen Verfahren zur Verfügung stellen sowie ein nachhaltiger Service etablieren.

Ergebnis AP 3: Das Set an Informationen ist definiert.

AP 4:
Abschließende Aufbereitung der Publikations-Informationen

Umfang AP 4 (4 Monate)

Das Set als valide erkannter Kennzeichen für Publikationen wird die Nutzenden aufbereitet. Dieses soll die Nutzung wissenschaftlicher Literatur und die Nutzung durch wissenschaftliche Literatur dokumentieren. Hierzu wie auch für die Information, ob eine Publikation ein Peer-Review-Verfahren durchlaufen hat, können wir auf vorhandenen Workflows in LIVIVO aufbauen. Die Zitationsinformationen stehen Dank der Initiative for Open Citation (I4OC)[28] weitestgehend zur Verfügung und können von der Metadatenplattfrom Crossref bezogen werden.[29] Bei den kommunizierten Merkmalen orientieren wir uns an den Maßgaben der DFG zu einer guten wissenschaftlichen Praxis (DFG 2019). Im Falle einer vielversprechenden Zuweisung von Texten zu den genannten Kategorien durch das ML-Modell soll auch die Darstellung des Lösungsweges des Modells (Explainability) im Informationsset abgebildet bzw. darauf verwiesen werden. Das heißt, dass bei jeder Eingeschätzten Publikation auch eine Erklärung angezeigt wird, warum genau dieses Beispiel eine bestimmte Zuordnung bekommen hat. Gerade im Kontext von Desinformation ist es wichtig, dass die ML-Modelle für die Nutzenden nachvollziehbar sind. Mit dieser Transparenz-Offensive steht AQUAS im Einklang mit der Vorgabe 2021/0106 (COD) der EU Kommission (European Comission 2021). [30].

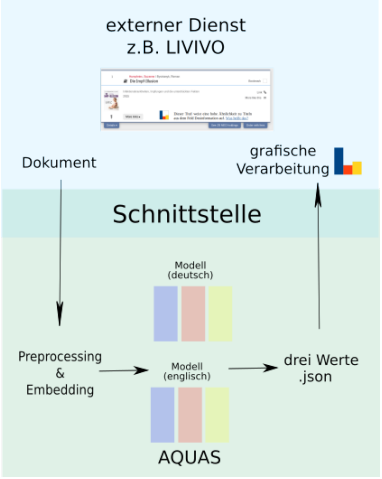


Abbildung 5: Funktionsweise

Ergebnis AP 4: Aufbereitung der Informationen der Informationen zu Publikationen abgeschlossen.

AP 5: Aufbau einer Schnittstelle und Integration in LIVIVO

Umfang AP 5 (4 Monate) Aufbau eines einfachen Webservices (mit dem Webframework fastAPI [31]), der auf Basis eines eingegebenen Text-Strings die Analysen durchführt und die beschriebenen Werte zu dem Dokument im JSON-Format zurückgibt. Auch von außen soll dieser Service von externen Suchmaschinen oder über ein GUI angesprochen werden können. Als ersten Use-Case integrieren wir den einfachen Webservice in das ZB-MED-Discovery-System LIVIVO.

Ergebnis AP 5: AQUAS kann durch LIVIVO genutzt werden

AP 6: Evaluierung und Usability des Webservices

Umfang AP 6 (6 Monate) Durch die Ansiedlung des Projekts bei ZB MED können die Ergebnisse direkt für die Nutzenden durch Implementierung und Evaluierung in Wert gesetzt werden. Für

[28] https://i4oc.org/
[29] https://www.crossref.org/
[30] https://blog.fiddler.ai/2021/07/eu-mandates-explainability-and-monitoring-in-proposed-gdpr-of-ai, Abrufdatum 25.02.2022
[31] https://fastapi.tiangolo.com, Abrufdatum 16.02.2022.

die Evaluierung kann das Feedback der Nutzenden direkt eingeholt werden und ggf. Umsetzung finden.

Ergebnis AP 6: Die Zielgruppen hatten Gelegenheit, den AQUAS-Service kennenzulernen. Hinweise wurden aufgenommen, ggf. integriert. Der auf AQUAS basierende Webservice und die API sind funktionstüchtig.

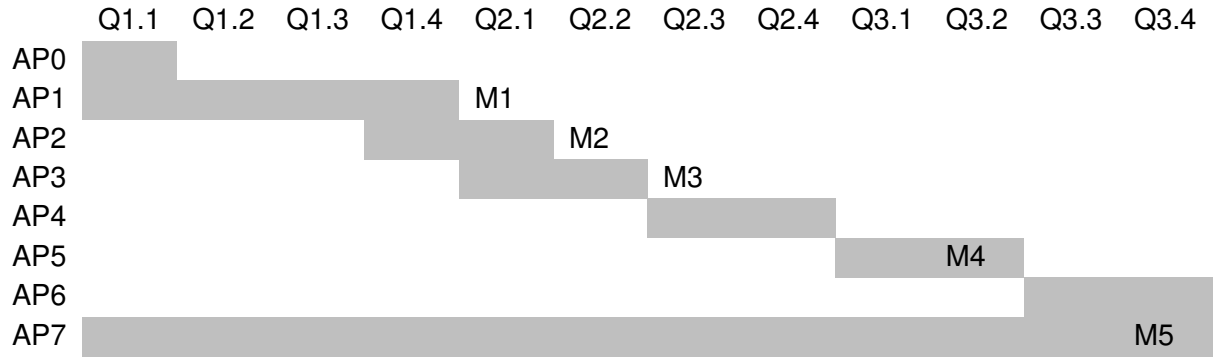
AP 7: Netzwerken, Vorstellung und Diskussion der Metrik in der Fach-Community

Umfang AP 7 (36 Monate) Ab dem Zeitpunkt des Projektstarts wird die Entwicklung des Projektes durch Fachvorträge und die Teilnahme an wissenschaftlichen Diskussionen begleitet. Der Antrag selbst wird bei Zenodo veröffentlicht. Abhängig von der Entwicklung der Corona-Pandemie wird die Dissemination digital oder im persönlichen Zusammentreffen stattfinden. Geplant ist eine aktive Teilnahme am *Bibliothekartag* sowie bei den internationalen Konferenzen *Semantic Web in Libraries* (SWIB), *International Conference on Cognitive Science, Natural Language Processing and Pattern Recognition (ICCSNLPPR)*, die 2023 in Berlin stattfindet, sowie die International Conference on Fake News, Media Manipulation and Disinformation, die 2023 und 2024 in London bzw. Rom durchgeführt werden. Für 2025 sind die Konferenzorte noch nicht festgelegt worden. Genauso wird der neue Service den Wissenschaftler:innen und Praktiker:innen aus dem Medizinbereich vorgestellt. Dies erfolgt zusammen mit den Kolleg:innen des LIVIVO-Teams bei ZB MED, die ohnehin auf Fachtagungen vertreten sind. Zusätzlich wird zum Ende des Projektes ein Online-Workshop für Praktiker:innen und Wissenschaftler:innen veranstaltet. Durch die Zusammenarbeit mit der medizinjournalistischen Plattform MedWatch erwarten wir auch die Beteiligung von Wissenschaftsjournalist:innen.

Ergebnis AP 7: Die Zielgruppen und insbesondere die Forschenden wurden durch verschiedenen Formate für das Thema Desinformation in wissenschaftlichen Datenbanken sensibilisiert.

Ausblick

Das Projekt soll im Q4 2022 beginnen und dem folgenden Zeitplan folgen:



Meilensteine (M): M1 (Monat 12) Finale Zusammenstellung der Trainings-Daten; **M2** (Monat 15) Exploration des ML-Verfahrens abgeschlossen; **M3** (Monat 18) Informations-Set definiert **M4** (Monat 30) Prototyp AQUAS-Webservice implementiert; **M5** (Monat 36) Abschluss des Projektes; Vorlegen der Evaluationsergebnisse, Erfassung des Wissenstransferkatalogs und Integration von Verbesserung des Webservices abgeschlossen.

3 Literaturverzeichnis

Babakar, Mevan, Nada Bakos, Hal Daumé, Alexios Mantzarlis, Djamé Seddah, Andreas Vlachos und Claire Wardle (2017). *Fake News Challenge*. http://www.fakenewschallenge.org/.

Chan, Branden, Stefan Schweter und Timo Möller (2020). „German's Next Language Model". In: *arXiv:2010.10906 [cs]*. arXiv: 2010.10906.

Chartered Institute of Library Information Professionals (CILIP) (2018). *Definitions of Information Literacy 2018*. https://infolit.org.uk/ILdefinitionCILIP2018.pdf.

Ćurković, Marko und Andro Košec (2018). „Bubble effect: including internet search engines in systematic reviews introduces selection bias and impedes scientific reproducibility". In: *BMC Medical Research Methodology* 18.1. doi: 10.1186/s12874-018-0599-2.

Dale, Robert (2017). „NLP in a post-truth world". In: *Natural Language Engineering* 23.2. doi: 10.1017/S1351324917000018, S. 319–324.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee und Kristina Toutanova (2019). „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: arXiv: 1810.04805.

DFG (2019). *Leitlinien zur Sicherung guter wissenschaftlicher Praxis*.

Epstein, Robert und Ronald E. Robertson (2015). „The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections". In: *Proceedings of the National Academy of Sciences* 112.33. doi: 10.1073/pnas.1419828112, E4512–E4521.

Europäische Union, Vīķe-Freiberga, Vaira, Däubler-Gmelin, Herta, Hammersley, Ben und Pessoa Maduro, Luís Miguel Poiares (2013). *A free and pluralistic media to sustain European democracy. The Report of the High Level Group on Media Freedom and Pluralism*. https://ec.europa.eu/information_society/media_taskforce/doc/pluralism/hlg/hlg_final_report.pdf.

European Comission (2021). *Proposal for a Regulation of the European Parliament and of the council laying down harmonised rules on artificial intelligence artifical intelligence act and amending certain union legislative acts*. 2021/0106 (COD).

European Commission, Directorate-General for Communications Networks, Content and Technology (2021). *European Commission Guidance on Strengthening the Code of Practice on Disinformation*. https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A52021DC0262.

Fernández, Alberto, Salvador Garcia, Mikel Galar, Ronaldo C. Prati, Bartosz Krawczyk und Francisco Herrera (2018). *Learning from imbalanced data sets*. New York, N.Y.

Gensing, Patrick (2020). *Fakten gegen Fake News oder Der Kampf um die Demokratie*. Schriftenreihe Band 10500. Bonn: Bundeszentrale für politische Bildung.

Ghosal, Soumya Suvra, Padmanabhan Deepak und Anna Jurek-Loughrey (2020). „ReSCo-CC: Unsupervised Identification of Key Disinformation Sentences". In: *arXiv:2010.10836 [cs]*. arXiv: 2010.10836.

Holone, Harald (2016). „The filter bubble and its effect on online personal health information". In: *Croatian Medical Journal* 57.3. doi: 10.3325/cmj.2016.57.298, S. 298–301.

Kotonya, Neema und Francesca Toni (Okt. 2020). „Explainable Automated Fact-Checking for Public Health Claims". In: *arXiv:2010.09926 [cs]*. arXiv: 2010.09926.

Meßmer, Anna-Katharina, Alexander Sängerlaub und Leonie Schulz (2021). *„Quelle: Internet"? Digitale Nachrichten- und Informationskompetenzen der deutschen Bevölkerung im Test*. Berlin.

Müller, Bernd, Christoph Poley, Jana Pössel, Alexandra Hagelstein und Thomas Gübitz (2017). „LIVIVO – the Vertical Search Engine for Life Sciences". In: *Datenbank-Spektrum* 17.1, S. 29–34.

Oshikawa, Ray, Jing Qian und William Yang Wang (2020). „A Survey on Natural Language Processing for Fake News Detection". In: *arXiv:1811.00770 [cs]*. arXiv: 1811.00770.

Pariser, Eli (2011). *The filter bubble: what the Internet is hiding from you*. New York.

Polizzi, Gianfranco (2019). „Information Literacy in the Digital Age: Why Critical Digital Literacy Matters for Democracy". In: *Informed Societies*. Hrsg. von Stéphane Goldstein. 1. Aufl. doi: 10.29085/9781783303922.003, S. 1–24.

Singh, Iknoor, P. Deepak und K. Anoop (2020). „On the Coherence of Fake News Articles". In: *ECML PKDD 2020 Workshops*. Hrsg. von Irena Koprinska u. a. Bd. 1323. Cham, S. 591–607.

Speed, Ewen und Russell Mannion (2017). „The Rise of Post-truth Populism in Pluralist Liberal Democracies: Challenges for Health Policy". In: *International Journal of Health Policy and Management* 6.5, S. 249–251.

Tan, Reuben, Bryan A. Plummer und Kate Saenko (2020). „Detecting Cross-Modal Inconsistency to Defend Against Neural Fake News". In: *arXiv:2009.07698 [cs]*. arXiv: 2009.07698.

Vaidhyanathan, Siva (2018). *Antisocial media: how facebook disconnects US and undermines democracy*. New York.

Vo, Nguyen und Kyumin Lee (2020). „Where Are the Facts? Searching for Fact-checked Information to Alleviate the Spread of Fake News". In: *arXiv:2010.03159 [cs]*. arXiv: 2010.03159.

Vogel, Inna und Peter Jiang (2019). „Fake News Detection with the New German Dataset 'GermanFakeNC'". In: *A. Doucet et al. (Hrsg.): Digital libraries for open knowledge: 23rd International Conference on Theory and Practice of Digital Libraries, TPDL 2019, Oslo, Norway, September 9–12, 2019: proceedings*. Lecture notes in computer science 11799, S. 288–295.

Zhong, Wanjun, Duyu Tang, Zenan Xu, Ruize Wang, Nan Duan, Ming Zhou, Jiahai Wang und Jian Yin (2020). „Neural Deepfake Detection with Factual Structure of Text". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, S. 2461–2470.

Zywietz, Bernd und Klaus Sachs-Hombach (2018). „Einführung: Propaganda, Populismus und populistische Propaganda". In: *Fake News, Hashtags & Social Bots*. Hrsg. von Klaus Sachs-Hombach und Bernd Zywietz. Wiesbaden, S. 1–11.

4 Begleitinformationen zum Projektkontext

4.1 Allgemeine ethische Aspekte

Häufig verfolgen Fakten-Prüfungs-Projekte, den Ansatz, die Korrektheit oder Falschheit einer Aussage festzustellen, die dann einer Leseempfehlung gleichkommt. Was für kurze Texte und abgegrenzte Aussagen umsetzbar scheint, erhält bei längeren und komplexeren Texten den Anstrich einer Zensur. Wir denken, dass die Nutzenden von AQUAS sehr unterschiedliche Erkenntnisinteressen verfolgen. Auf Grund der Vielseitigkeit von Fragestellungen halten wir eine absolute Bewertung der Qualität der Texte für nicht sinnvoll. Anstatt den Menschen eine Entscheidung über die Relevanz eines Titels abzunehmen, will die AQUAS-Metrik einen Hinweis zu diesem Dokument liefern, indem es semantische Ähnlichkeiten aufzeigt. Auf dieser Grundlage soll es den Nutzenden leichter gemacht werden, eine eigenständige Lese-Entscheidung und Bewertung des Textes durchzuführen. AQUAS zielt auf eine Befähigung der Menschen in ihrer Informationskompetenz, nicht auf eine Bevormundung.

4.2 Maßnahmen zur Erfüllung der Förderbedingungen und Umgang mit den Projektergebnissen

Um die e-Research-Infrastruktur von AQUAS zu etablieren, ist es notwendig die Projektergebnisse nachhaltig zu konzipieren und sie der Community frei zugänglich zur Verfügung zu stellen. AQUAS soll demnach technisch im Sinne einer Open-Source-Software nachnutzbar und weiterentwickelbar sein. Daher wird neben den Datensätzen und dem Modell auch der Quellcode für den AQUAS-Service offengelegt und unter eine Open-Source-Lizenz gestellt. Hierzu eignet sich das von außen zugängliche GitLab-Instanz von ZB MED oder auch GitHub (vgl. AP 0). Die Nutzung durch Dritte wird durch eine entsprechende Dokumentation unterstützt. Der Betrieb des AQUAS-Service wird für mindestens drei Jahre lang nach Projektende durch ZB MED aufrechterhalten. Hierzu wird die Infrastruktur von den Entwicklungsservern auf Server bei den Infrastruktureinrichtungen migriert. Für diese Zeit garantiert ZB MED den weiteren Betrieb.

4.3 Erklärungen zur Erfüllung der Förderbedingungen

Alle durch das Vorhaben zustande gekommenen Ergebnisse werden der Fachöffentlichkeit bekannt gemacht und kostenlos zur Nachnutzung zur Verfügung gestellt. Die produzierten Quellcodes sowie die Dokumentation werden unter Open-Source-Lizenzen veröffentlicht. Alle Veröffentlichungen im Rahmen des Projektes werden im (grünen oder goldenen) Open Access zur Verfügung gestellt.

5 Personen/Kooperationen/Finanzierung

5.1 Angaben zur Dienststellung

Name	Dienststellung
Dr. Eva Seidlmayer, M.LIS	Wissenschaftliche Mitarbeiterin, Data Science and Services, ZB MED (befristet bis 11/2022)
Prof. Dr. Konrad U. Förstner	Leiter des Bereichs Data Science and Services, ZB MED (befristet bis 05/2023)

5.2 Zusammensetzung der Projektarbeitsgruppe

Name	Projektaufgabe	Finanzierung
Dr. Eva Seidlmayer. M.LIS	Wissenschaftliche Mitarbeiterin	landesfinanziert
Prof. Dr. Konrad U. Förstner	Projektleitung ZB MED, wissenschaftliche Leitung	landesfinanziert

5.3 Institutionen oder Wissenschaftlerinnen und Wissenschaftlern in Deutschland, mit denen für dieses Vorhaben eine konkrete Vereinbarung besteht

Mit diesen Partnern und Institutionen soll im Projekt kooperiert werden (s. Absichtserklärungen):

- Sören Auer, Technische Informations Bibliothek (TIB), Hannover
- Klaus Tochtermann, Zentralbibliothek Wirtschaft (ZBW), Kiel
- Brigitte Mathiak, GESIS Köln
- Daniel Mietchen, University of Virginia
- Nicola Kuhrt, MedWatch – Magazin für evidenzbasierten Medizinjournalismus

5.4 Institutionen oder Wissenschaftlerinnen und Wissenschaftler im Ausland, mit denen für dieses Vorhaben eine konkrete Vereinbarung besteht

entfällt

5.4.1 Institutionen, Wissenschaftlerinnen und Wissenschaftler, mit denen in den letzten drei Jahren gemeinsame Projekte durchgeführt wurden

- Sören Auer, TIB
- Anke Becker, Univ. Marburg
- Peer Bork, EMBL
- Thomas Clavel, RWTH Aacher
- Alexander Goesmann, Univ. Gießen
- Evguenieva Hackenberg, Univ. Gießen
- Gabriele Klug, Univ. Gießen
- Stefanie Kuerten, Univ. Würzburg
- Brigitte Mathiak, GESIS Köln
- Manja Marz, Univ. Jena
- Alice McHardy, HZI
- Eva Medina, HZI
- Kai Papenfort, Univ. Jena
- Jörg Overmann, DSMZ
- Isabell Peters, ZBW
- Daniel Mietchen, University of Virginia
- Ulisses Nunes da Rocha, UFZ
- Philipp Schaer, TH Köln
- Ruth Schmitz-Streit, Univ. Kiel
- Alexander Sczyrba, Univ. Bielefeld
- Nicolai Siegel, LMU
- Jörg Soppa, Univ. Frankfurt
- Wolfgang R. Streit, Univ. Hamburg
- Gisela Storz, NIH, USA
- Jens Stoye, Univ. Bielefeld
- Carsten Schultz, Univ. Kiel

- Klaus Tochtermann, ZBW
- Jörg Vogel, HIRI
- Jakob Voss, GBV
- Alexander Westermann, HIRI
- Wilma Ziebur, Univ. Würzburg

5.5 Projektrelevante Zusammenarbeit mit erwerbswirtschaftlichen Unternehmen

entfällt

5.6 Projektrelevante Beteiligungen an erwerbswirtschaftlichen Unternehmen

entfällt

5.7 Weitere Antragstellungen

entfällt

5.8 Eigenleistung

Konrad U. Förstner ist institutionell bzw. durch Landesmittel finanziert. Laufende Mittel für Sachaufgaben, Hardware (weitgehend) sowie Software für die Projektmitarbeitenden erbringt ZB MED in Eigenleistung. Zudem wird das zentrale Knowledge Environment, als Datenbank, durch ZB MED zur Verfügung gestellt. Auch LIVIVO als erster Anwendungsfall für die Implementation des Webservices wird von ZB MED betrieben und kann frei von AQUAS genutzt werden.

6 Beantragte Module/Mittel

6.1 Basismodul

6.1.1 Personalmittel

Gewünschter Beginn der Finanzierung ist der 01.12.2022.

Für die ZB MED werden beantragt:

- **1 wissenschaftliche Mitarbeiter:in (Postdoc) (Fachgebiet Informationswissenschaft) für 36 Monate**, Anteil an der regelmäßigen Arbeitszeit: 100 %

- **1 wissenschaftliche Hilfskraft (WHK mit Bachelor) mit 20 h/Monat für 24 Monate**
  Kosten: mtl. 276 Euro, gesamt 6.624 Euro
  Aufgaben: Unterstützung u.a. bei der Zusammenstellung des Datensatzes, der Implementation und bei der Evaluierung des Webservice.

6.1.2 Sachmittel

Angaben zu den für das Projekt zur Verfügung stehenden größeren Geräten (ggf. auch Großrechenanlagen, wenn Rechenleistung benötigt wird) Für die Umsetzung dieses Services ist die Nutzung eines Servers mit GPU notwendig, da mit der GPU weitaus schnellere Rechenleistung umgesetzt werden kann. Der Server wird entsprechend der eingeholten Kostenvoranschläge etwa 26 359,- € kosten.

Sachmittel	Laufzeit	monatliche Kosten	Gesamt
Miete Server-Infrastruktur inklusive Software	12 Monate -	pauschal	**ZB MED 26 359,-**

Publikationskosten Für die Publikation von Fach-Artikeln in einschlägigen Fachzeitschriften werden insgesamt 2250,- € veranschlagt. Die Publikation erfolgt als Open-Access-Publikation. Für Korrekturen professionelle Korrekturen der zu veröffentlichenden wissenschaftlichen Artikel werden pauschal 200,- € pro Artikel fällig.

Sachmittel	Laufzeit	monatliche Kosten	Gesamt
Publikationkosten über drei Jahre 2023-2025	pauschal		**2250,-**
wissenschaftliches Korrektorat	2x Pauschale		**400,-**

Mittel für Reisen Projektergebnisse sollen im internationalen und nationalen Rahmen vorgestellt werden. Damit wird die Sichtbarkeit des Projektes verbessert sowie der Dialog mit Expertinnen und Experten sowie Kooperationspartnerinnen und Kooperationspartnern gesucht.

Reisen	Ort	Reisepauschale	Gesamt
3 Konferenzen, (Semantic Web In Library (SWIB), Bibliothkartag, International Conference on Cognitive Science, Natural Language Processing and Pattern Recognition (ICCSNLPPR))	Deutschland	600,-	1800,-
2 Konferenzen (17. & 18. International Conference on Fake News, Media Manipulation and Disinformation), international	Vereinigtes Königreich/Italien	1200,-	2400,-
		Summe	**4200,-**