Maschinelle Beschlagwortung mit Algorithmen

Ein Blick in die Werkstatt des KI-Projektes der Deutschen Nationalbibliothek

Elisabeth Mödden

Abstract

Die Deutsche Nationalbibliothek nutzt Künstliche Intelligenz (KI), um ihre Erschließungsprozesse zu optimieren. Ein Schwerpunkt liegt dabei auf der Integration innovativer Technologien für die Erschließung von Dokumenten mit der Gemeinsamen Normdatei. Das Forschungsprojekt „Automatisches Erschließungssystem – Inhaltserschließung von Publikationen mit KI" untersucht, wie KI-Lösungen die maschinelle Beschlagwortung von Publikationen verbessern können. Ziel des im Rahmen der Nationalen Strategie Künstliche Intelligenz geförderten Projekts ist es, die Qualität maschinell generierter Erschließungsdaten zu verbessern. Das KI-Projekt konzentriert sich auf deutschsprachige wissenschaftliche Online-Publikationen und strebt an, die entwickelten Werkzeuge als Open-Source-Software zur Verfügung zu stellen.

The German National Library is using Artificial Intelligence (AI) to optimise its cataloguing processes. One focus is the integration of innovative technologies for subject cataloguing with the German Integrated Authority File (GND). The research project "Automatic Cataloguing System – Subject Cataloguing of Publications with AI" is investigating how AI solutions can improve automatic subject cataloguing of publications. The aim of the project, which is funded as part of the National Strategy for Artificial Intelligence, is to improve the quality of machine-generated indexing data. The AI project focuses on German-language online scientific publications and aims to develop open source software tools.

Im digitalen Zeitalter entwickelt die Deutsche Nationalbibliothek (DNB) ihre Erschließungsprozesse für Online- und Print-Publikationen kontinuierlich weiter, wobei Künstliche Intelligenz (KI) eine entscheidende Rolle spielt. Ein besonderer Schwerpunkt liegt dabei auf der Integration innovativer Technologien für die Beschlagwortung mit der Gemeinsamen Normdatei GND[1], was der effizienten Generierung von Metadaten für die vollständige und präzise thematische Erfassung aller Medienwerke, sowohl in digitaler als auch in gedruckter Form, dienen soll. Für die thematische Recherche spielt die Beschlagwortung mit der GND eine entscheidende Rolle. Die DNB setzt bereits seit 10 Jahren maschinelle Verfahren für die Beschlagwortung mit der GND ein und arbeitet intensiv daran, diese Verfahren noch weiter zu verbessern. Damit soll trotz wachsender Medienbestände eine möglichst einheitliche und vollständige Erschließung mit der GND erreicht werden. Angesichts der rasanten Entwicklung von Techniken der Künstlichen Intelligenz (KI) stellt sich die Frage, auf welche Weise diese aktuellen Innovationen für diese bibliothekarische Aufgabe genutzt werden können. Dieser Frage geht das Forschungsprojekt „Automatisches Erschließungssystem – Inhaltliche Erschließung von Publikationen mit KI"[2] nach. Hier werden systematische Untersuchungen durchgeführt, um zu identifizieren, welche KI-Lösungen Verbesserungen bei der maschinellen Beschlagwortung von natürlichsprachlichen Texten ermöglichen. Das intern als KI-Projekt bezeichnete Vorhaben mit einer Laufzeit von Oktober 2021 bis Dezember 2025 wird von der Staatsministerin für Kultur und Medien im Rahmen der Nationalen Strategie für Künstliche Intelligenz[3] gefördert und ist technologieoffen und Open Source orientiert. Der Forschungsschwerpunkt liegt auf der maschinellen Beschlagwortung deutschsprachiger wissenschaftlicher Online-Monografien und -Artikel mit dem Vokabular der GND, die über eine Million semantische Konzepte umfasst, darunter Sachschlagwörter, Personen, Körperschaften, Konferenzen, Geografika und Werke. Dabei wird auch untersucht, ob neue Techniken die Vollständigkeit und Genauigkeit maschinell generierter Erschließungsdaten messbar verbessern können und welcher Ressourcenaufwand dafür erforderlich ist. Im Rahmen des Projekts werden systematische Evaluationen durchgeführt und die erarbeiteten Werkzeuge nach Möglichkeit als Open-Source-Software zur Nachnutzung zur Verfügung gestellt.

Erschließungsmaschine EMa

Zum besseren Verständnis des Kontextes des KI-Projekts soll zunächst ein kurzer Rückblick auf die Entwicklungsgeschichte der maschinellen GND-Beschlagwortung in der DNB gegeben werden.

Mit der Entwicklung automatischer Erschließungsverfahren wurde 2009 im Rahmen des internen Projekts PETRUS[4]

1 Gemeinsame Normdatei (GND). https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd_node.html

2 Projekt „Automatisches Erschließungssystem – Inhaltliche Erschließung von Publikationen mit Künstlicher Intelligenz". https://www.dnb.de/DE/Professionell/ProjekteKooperationen/Projekte/KI/ki_node.html

3 Nationale Strategie für Künstliche Intelligenz. https://www.ki-strategie-deutschland.de/home.html

4 Projekt Prozessunterstützende Software für die digitale Deutsche Nationalbibliothek (PETRUS). https://files.dnb.de/pdf/petrus/dialog_2010_1_petrus.pdf

begonnen. In Zusammenarbeit mit dem Softwareentwicklungspartner Averbis GmbH wurde die Averbis Extraction Platform, eine Anwendung zur linguistischen und inhaltlichen Analyse elektronischer Texte, für die automatische Klassifikation und Beschlagwortung von Online-Publikationen angepasst, weiterentwickelt und in die Geschäftsprozesse der DNB integriert.[5] Die Averbis Extraction Platform nutzt für die Zuordnung von DDC-Sachgruppen und DDC-Kurznotationen[6] einen Support Vector Machine Ansatz (SVM). Die SVM ist ein Verfahren der künstlichen Intelligenz, das auf einem statistischen Ansatz basiert und als Algorithmus des maschinellen Lernens verwendet wird, um Publikationen bestimmten Klassen (DDC-Sachgruppen oder DDC-Kurznotationen) zuzuordnen. Für die GND-Beschlagwortung von Online-Publikationen wird seit 2014 ein wörterbuchbasiertes sog. lexikalisches Verfahren eingesetzt, das den Text linguistisch analysiert und mit den Konzepten der GND im Wörterbuch mit einem Matchingverfahren abgleicht. Zu diesem Zweck wurde die Averbis Extraction Platform um eine Komponente zur Wörterbuchpflege erweitert. Damit können beispielsweise Terme der GND, die maschinell häufig falsch vergeben wurden, deaktiviert werden.

Mit dem Abschluss des Petrus-Projekts wurde 2014 in der Organisationsstruktur der DNB ein neues standortübergreifendes Referat mit dem Namen „Automatische Erschließungsverfahren; Netzpublikationen" eingerichtet, das für die Sammlung von Netzpublikationen bzw. Online-Publikationen sowie den Routinebetrieb und die Weiterentwicklung der automatischen Erschließung zuständig ist.

2018 kündigte die Firma Averbis an, die Weiterentwicklung der Averbis Extraction Platform einzustellen. Da die Weiterentwicklung von Software für eine sinnvolle Fortführung der Optimierung maschineller Erschließungsprozesse jedoch unerlässlich ist, entschied sich die DNB für die Umstellung auf ein neues System. Zu diesem Zweck wurde das interne Projekt EMa ins Leben gerufen, dessen Ziel die Entwicklung einer neuen Erschließungsmaschine[7] war. Damit bot sich die Chance, auf der Basis von 10 Jahren Erfahrung deutliche Verbesserungen zu erzielen. Auch die neue, nun modular aufgebaute Software wird wieder in der eigenen IT-Infrastruktur der DNB betrieben. Der große Vorteil des modularen Konzepts: Die EMa ist flexibel erweiterbar und kann mit geringem Aufwand an den technologischen Fortschritt angepasst werden. Services oder Verfahren können einfach ausgetauscht oder erweitert, neue Funktionen jederzeit hinzugefügt werden. Mit dem erfolgreichen Abschluss des EMa-Projekts im Jahr 2022 wurde das Averbis-System weitgehend abgelöst.

Für die Klassifikation und Beschlagwortung sind KI-basierte Verfahren des Annif Toolkits[8] als Service in EMa integriert. Dieses vielseitige Toolkit für Bibliotheksanwendungen wurde von der Finnischen Nationalbibliothek entwickelt[9]. Annif ist eine Open-Source-Software, die Verfahren der natürlichen Sprachverarbeitung und des maschinellen Lernens umfasst. Bestehende KI-Verfahren wurden speziell für die Erschließung von Publikationen ausgewählt und für eine einfache bibliothekarische Nutzung aufbereitet. Da die meisten KI-Verfahren in Annif sprachunabhängig sind, kann jedes Fachvokabular im SKOS-Format für die Erschließung verwendet werden, so auch das Vokabular der GND. Mit der wachsenden Zahl an Bibliotheken, die sich für Annif interessieren, wächst auch eine Community, die sich aktiv an der Weiterentwicklung und Optimierung von Annif beteiligt. Diese Vielfalt an Perspektiven und Fachwissen kann zur kontinuierlichen Verbesserung von Annif beitragen.

Derzeit sind in der EMa angepasste Konfigurationen[10] folgender KI-Verfahren aus dem Annif-Toolkit im Einsatz:

- SVC (Linear Support Vector Classification): Hierbei handelt es sich um ein lernendes Verfahren, das für die DDC-Sachgruppenvergabe eingesetzt wird.
- Omikuji-Bonsai (beruht auf Entscheidungsbäumen): Dieses lernende Verfahren wird für die Vergabe der DDC-Kurznotationen eingesetzt.
- Omikuji-Bonsai und MLLM (Maui-like Lexical Matching) als Ensemble: Hier werden ein lernendes Verfahren und ein lexikalisches Verfahren im Ensemble für die Beschlagwortung der Online-Publikationen und Hochschulschriften mit der GND kombiniert.
- Omikuji-Bonsai, Omikuji-Attention, fastText, stwfsa und MLLM als Ensemble: Drei lernende und zwei lexikalische Verfahren werden für die Beschlagwortung der Kinder- und Jugendliteratur mit einem Ausschnitt der GND kombiniert.

Der Prozess der maschinellen Erschließung läuft vollständig automatisiert ab, siehe dazu Abbildung 1, die sche-

5 Mödden, Elisabeth; Schöning-Walter, Christa; Uhlmann, Sandro: Maschinelle Inhaltserschließung in der Deutschen Nationalbibliothek, in: BuB Forum Bibliothek und Information, 01/2018, S.30-3 https://zs.thulb.uni-jena.de/servlets/MCRFileNodeServlet/jportal_derivate_00333732/BuB_2018_01_030_035.pdf

6 Die DDC in der Deutschen Nationalbibliothek. https://www.dnb.de/DE/Professionell/DDC-Deutsch/DDCinDNB/ddcindnb_node.html

7 Uhlmann, Sandro et al.: In der DNB lesen jede Nacht die Maschinen. In: DNB Blog 23.10.2023. https://blog.dnb.de/in-der-dnb-lesen-jede-nacht-die-maschinen/

8 Annif - Tool for automated subject indexing and classification. http://annif.org/ Abgerufen am 19. April 2024.

9 Suominen, Osma; Inkinen, Juho; Lehtinen, Mona: Annif and Finto AI: Developing and Implementing Automated Subject Indexing. JLIS.It, 13 (2022), 1, 265-282. https://doi.org/10.4403/jlis.it-12740

10 Annif Verfahren und Ensembles unter backends in https://github.com/NatLibFi/Annif/wiki/Getting-started

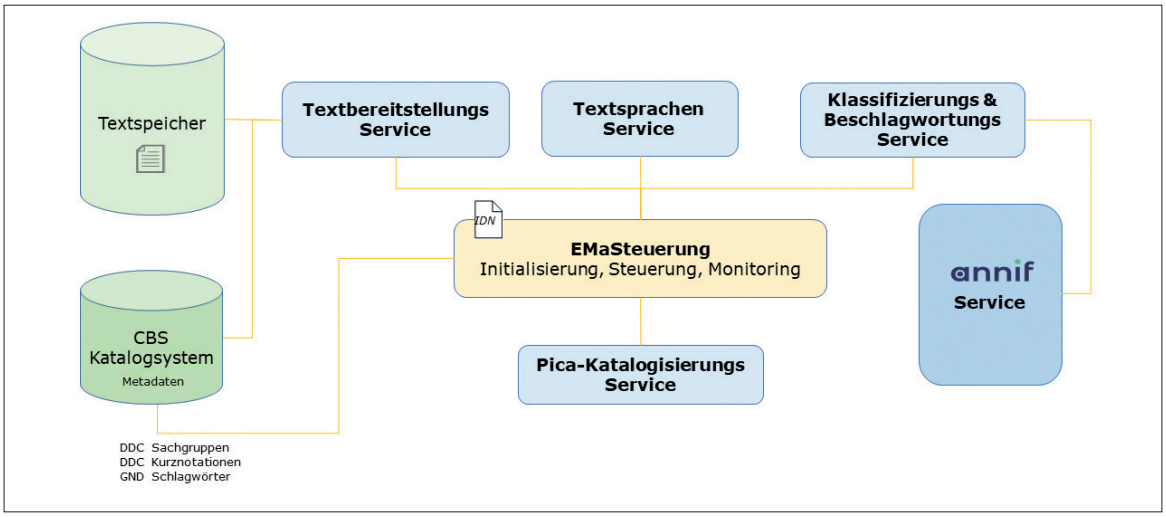

Abbildung 1: Schematische Darstellung des modularen Erschließungssystems EMa[12]

matisch den modularen Aufbau der EMa zeigt. Der tägliche Prozess wird durch eine IDN-Liste der am Vortag neu eingetroffenen Online-Publikationen[11] an die EMa-Steuerung gestartet. Die EMa-Steuerung steht im Zentrum. Sie initialisiert, steuert und überwacht den gesamten produktiven Betrieb. Über den Service zur Textbereitstellung werden dann die Online-Publikationen aus dem Textspeicher und die relevanten Metadaten aus dem Katalogsystem als Textgrundlage für die nachfolgenden Analysen abgerufen. Ein Service zur Textsprachenerkennung ermittelt zunächst die Sprache des Textes. Deutsch- oder englischsprachige Texte werden anschließend zusammen mit ihren Metadaten über den Service zur Klassifizierung und Beschlagwortung an den Service Annif übergeben. Dieser bietet vielfältige Möglichkeiten und Konfigurationen, die unterschiedlichen Arten der zu erschließenden Publikationen differenziert zu verarbeiten. Die so erzeugten maschinellen Erschließungsergebnisse – Sachgruppen, DDC-Kurznotationen oder GND-Schlagwörter – werden anschließend vom Katalogisierungs-Service in das Format Pica+ des Katalogsystems der DNB konvertiert. Zum Schluss werden sie von der EMa-Steuerung in den Metadatensatz des Medienwerks im Katalogsystem geschrieben. Im Datensatz wird gekennzeichnet, dass die Erschließungsdaten maschinell generiert wurden. Sie stehen sofort für die Recherche im DNB-Portal zur Verfügung.[12]

Der weitere Ausbau und die Ausweitung der maschinellen Erschließung auf weitere Publikationsformen ist ein fortlaufendes Ziel. Dazu werden die Funktionalitäten der EMa kontinuierlich weiterentwickelt, verbessert und ergänzt. Im Zusammenhang der derzeit intensiv geführten Diskussion über den Einsatz künstlicher Intelligenz ist von besonderem Interesse, welche neuen KI-Verfahren und technologischen Innovationen die Ergebnisse der maschinellen Beschlagwortung deutschsprachiger wissenschaftlicher Publikationen weiter verbessern können. Mit dieser Frage beschäftigt sich das Forschungsprojekt „Automatisches Erschließungssystem", das im Folgenden vorgestellt wird.

KI-Projekt: „Automatisches Erschließungssystem"

Es gibt wesentliche Einflussfaktoren, die die erfolgreiche Umsetzung von KI-Lösungen in einem Projekt maßgeblich beeinflussen. Entscheidend ist, dass bei der Initiierung eines KI-Projekts die Ziele klar und deutlich definiert werden. Ein wichtiger Ausgangspunkt ist dabei ein gutes Verständnis der zur Verfügung stehenden Daten und ein für maschinelle Verfahren zugängliches Datenmanagement. Weitere wichtige Faktoren sind verfügbare KI-Expert:innen und Entwickler:innen sowie die notwendige technische Infrastruktur, also entsprechende Rechenleistung wie Graphics Processor Unit (GPU) und Speicherkapazität. Außerdem sind in diesem Zusammenhang die geeigneten KI-Algorithmen von zentraler Bedeutung. Die Auswahl adäquater KI-Algorithmen und -Modelle ist entscheidend für den Projekterfolg. Je nach Zielsetzung, spezifischen Anforderungen und verfügbaren Daten können unterschiedliche Ansätze erforderlich sein.

Projektziel

Übergeordnetes Ziel des KI-Projekts der DNB ist die Generierung möglichst vollständiger und präziser GND-Schlagwörter zur inhaltlichen Beschreibung deutschsprachiger wissenschaftlicher Online-Monografien und -Artikel. Dabei werden innovative Verfahren der Künstlichen Intelligenz experimentell untersucht, um bestmögliche Ergebnisse zu erzielen. Diese neuen Verfahren sollen im Vergleich zu den derzeit eingesetzten Erschließungsmethoden messbar bessere Erschließungsergebnisse lie-

11 Dabei handelt es sich überwiegend um Online-Publikationen (Monographien und Artikel) und einige ausgewählte Print-Publikationen. Bei den Print-Publikationen werden die Titeldaten und die gescannten Inhaltsverzeichnisse oder nur die Titeldaten für die maschinelle Erschließung verwendet.

12 Uhlmann, Sandro et. Al.: Erschließungsmaschine gestartet. In: DNB Blog 04.05.2022. https://blog.dnb.de/erschliessungsmaschine-gestartet

Flexibler mit Foxis!

Regale, Büchertröge, Hocker und mehr: Platzieren Sie die farbenfrohen Foxis-Kindermöbel mit Schwenkrollen dort, wo Ihre Nutzer*innen sie brauchen!

Foxis-Möbel im Shop

Wir beraten Sie gerne: Telefon 07121 144-420 • bibliotheksausstattung@ekz.de • ekz.de

Besuchen Sie uns auf der BiblioCon 2024 in Hamburg im CCH, Halle H, Stand 050.

:ekz bibliotheks service

fern. Die Leistungsfähigkeit dieser Verfahren wird anhand von Metriken wie Precision (Genauigkeit), Recall (Trefferquote) und F-Score gemessen, um die Ergebnisse verschiedener Ansätze miteinander vergleichen zu können. Ziel ist es, einen möglichst hohen F-Score von über 0,4 zu erreichen. Mit dem lexikalischen Verfahren der Averbis Extraction Platform, das in der DNB zum Zeitpunkt des Projektantrags 2021 eingesetzt wurde, erreichte die DNB einen durchschnittlichen F-Score von 0,27 auf einer Skala von 0 bis 1. Der F-Score ist ein statistisches Maß und gibt das harmonische Mittel aus Precision und Recall an. Zur Berechnung des F-Score werden die Ergebnisse, die maschinell vorgeschlagenen Schlagwörter, mit den intellektuell vergebenen Schlagwörtern, die den sogenannten Goldstandard abbilden, abgeglichen. Die dabei gefundenen und erfolgreich erprobten Verfahren sollen für die Eignung im produktiven Einsatz evaluiert werden. Es ist geplant Werkzeuge, die im Rahmen des KI-Projektes entwickelt werden, auch anderen Institutionen mit ähnlichen Anforderungen zur Verfügung zu stellen und zu diesem Zweck als Open-Source-Software auf einer geeigneten Plattform, z.B. GitHub, bereitzustellen.

Ein weiteres wichtiges Ziel des KI-Projekts ist der gezielte Transfer von KI-Technologie und -Wissen in die bibliothekarische Praxis. Die DNB ist bestrebt, diese Kompetenzen in Richtung einer Expertise im Bereich der KI-Technologien weiter auszubauen und zu vertiefen. Dazu gehört nicht nur die Weiterbildung von Mitarbeiter:innen, sondern auch die Schaffung von Strukturen für ein KI-Kompetenzzentrum mit Expert:innen, die einen effektiven Wissensaustausch zur Forschung und damit den Umgang und Einsatz neuer Technologien ermöglichen. Auch mit europäischen Partnerbibliotheken wird über die Arbeitsgruppe „AI in Libraries Network Group – CENL“ eine Community of Practice für den gegenseitigen Austausch und die gemeinsame Entwicklung von technischen Lösungen aufgebaut.

Projektteam

Das Projektteam setzt sich aus vier Mitarbeiter:innen zusammen. Von Beginn an arbeitet ein Informatiker im Projekt mit. Darüber hinaus konnten ein Mathematiker und zwei Computerlinguistinnen als KI-Expert:innen für das Projekt gewonnen werden. Zeitweise gehörten auch ein Wirtschaftsinformatiker und ein weiterer Informatiker zum Team.

KI-Datenlabor

Im ersten Projektjahr wurde durch die Anschaffung geeigneter Hardware mit leistungsfähigen Grafikprozessoren eine effiziente Forschungsinfrastruktur geschaffen. Auf dieser Grundlage wurde ein wissenschaftliches Testlabor mit folgenden Spezifikationen aufgebaut: Die Hauptprogrammiersprachen des Teams sind Python und R, die im Bereich des maschinellen Lernens und der Datenanalyse weit verbreitet sind und sich als vielseitig und leistungsfähig erwiesen haben. Die Datenverarbeitungs- und Trainingspipelines werden mit dem Pipelining-Tool Data Version Control (DVC) modelliert und synchronisiert, das konfigurierbare Funktionen für die Zusammenstellung und Vorverarbeitung von Textkorpora bietet. DVC ermöglicht den Aufbau einer Pipeline, die alle Bearbeitungsschritte für die Verarbeitung der Daten von einem anfänglichen Meta-Daten-Dump im internen PICA-Katalog-Format, über linguistische Vorverarbeitung der Volltexte bis zum Training maschineller Verfahren und den Endergebnissen zur Evaluation umfasst. Die meisten dieser Bearbeitungsschritte, wie z.B. das Speichern und Aufbereiten der Zuordnungen zwischen Dokumenten und GND-Konzepten, müssen dabei nur einmal durchgeführt werden. Die erzeugten Dateien werden mittels DVC in einem temporären oder permanenten Speicher abgelegt und bei erneutem Bedarf automatisch abgerufen. Soll mit einem Verfahren ein neues Experiment durchgeführt werden, so müssen lediglich die für dieses Experiment charakteristischen Parameter der Pipeline geändert werden. So kann z.B. angegeben werden, ob das Experiment auf Basis des Titeldatenkorpus oder des Volltextkorpus durchgeführt werden soll. Nach der Durchführung eines Experiments kann der Arbeitsstand gespeichert werden. Damit können die Ergebnisse des Experiments gespeichert und bei Bedarf wieder aufgerufen werden. Das Team verwendet GitLab zur Versionsverwaltung, um die Experimente und deren Verlauf übersichtlich und strukturiert zu halten. Die Softwarepakete werden in reproduzierbaren virtuellen Umgebungen mit Conda oder Mamba verwaltet. Für die Optimierung der Hyperparameter wird z.T. das Framework Optuna genutzt. Dabei handelt es sich um ein Open-Source-Framework, mit dem die Suche nach Hyperparametern für die Modelle automatisiert werden kann, um die bestmöglichen Ergebnisse zu erzielen. Datenabfragen aus dem DNB-Katalog werden mit der Eigenentwicklung pica-rs[13] durchgeführt. Bei der Entwicklungsumgebung setzt das Team auf eine Auswahl bewährter Werkzeuge wie JupyterLab, VS-Code und RStudio, die alle ihre eigenen Stärken und Vorteile haben und es dem Team ermöglichen, effizient zusammenzuarbeiten.

Durch die beschriebene Toolchain wird eine konsistente und zuverlässige Entwicklungsumgebung im KI-Datenlabor zur Verfügung gestellt. Da aber besonders daten- und rechenintensive Experimente die Möglichkeiten dieser DNB-eigenen IT-Infrastruktur übersteigen, wird für solche

13 Wagner, Nico: Pica-rs. https://github.com/deutsche-nationalbibliothek/pica-rs. Abgerufen am 19. April 2024.

Experimente auch die Infrastruktur des Zentrums für Informationsdienste und Hochleistungsrechnen (ZIH) der TU Dresden[14] genutzt.

Daten

Ein genauerer Blick auf die Datenbasis zeigt, dass für die Experimente im Projekt ca. 200.000 Volltexte sowie ca. 1 Million Titeldaten mit Goldstandard, d.h. mit intellektueller GND-Beschlagwortung, zur Verfügung stehen. Die in der Regel als PDF vorliegenden Volltexte werden mit PDF-Extraktionstools in Rohtext umgewandelt. Anschließend erfolgen weitere Vorverarbeitungsschritte wie Textkürzungen oder die Entfernung von Sonderzeichen. Damit maschinelle Lernverfahren Textdaten verarbeiten können, müssen diese in der Regel in eine vektorisierte Form gebracht werden. Darunter versteht man im Allgemeinen die Umwandlung eines Textdokuments in eine Liste von Zahlenwerten.[15] Bei modernen Verfahren, wie z.B. AttentionXML, werden dazu sogenannte Word-Embeddings verwendet. Für viele Experimente in der DNB werden die Textdaten jedoch auch mit der klassischen TF-IDF-Vektorisierung (Term Frequency – Inverse Document Frequency) verarbeitet. Für die Experimente werden die Daten in Trainingsdaten, Validierungsdaten und Testdaten unterteilt. Die Trainingsdaten dienen als Beispiele, mit denen das lernende KI-Verfahren trainiert wird, um ein Modell zu entwickeln.[16] Die Validierungsdaten werden verwendet, um die Hyperparameter des Modells zu optimieren und somit eine Über- oder Unteranpassung des Modells zu erkennen.[17] Die Testdaten mit Goldstandard werden dann verwendet, um abschließend die Leistungsfähigkeit des Modells zu bewerten. Dies geschieht durch die Berechnung von Precision, Recall und F-Score. Für eine Bewertung der kompletten Vorschlagslisten eignen sich auch Precision-Recall-Kurven[18] und der darunter liegende Flächeninhalt (Area Under the Curve – AUC). Zusätzlich zu den Trainings-, Validierungs- und Testdaten steht außerdem für eine intellektuelle Bewertung der Ergebnisse ein Testset mit deutschsprachigen wissenschaftlichen Publikationen aus 18 ausgewählten DDC-Sachgruppen[19] ohne Goldstandard zur Verfügung.

Für die Beschlagwortung verwendet die DNB die Terminologie der GND. Die semantischen Konzepte der Normdatei repräsentieren Sachbegriffe, Personen, Körperschaften, Konferenzen, Geografika und Werke. Die GND bildet die Begriffe von Kultur und Wissenschaft und deren Synonyme in einem alle Themenbereiche umfassenden deutschsprachigen Vokabular ab. Die Begriffe sind durch Relationen unterschiedlichen Typs miteinander verbunden (z.B. Ober- und Unterbegriffe, verwandte Begriffe). Auch zu anderen kontrollierten Vokabularen bestehen Verknüpfungen. Durch die Beschlagwortung mit der GND werden die Texte thematisch eingeordnet und mit anderen Publikationen zum gleichen Thema vernetzt. Die GND ist somit ein wichtiger Beitrag zur Unterstützung der Suche in vernetzten Informationssystemen.

Die GND enthält 1,4 Millionen Terme bzw. Konzepte, die potenziell als Schlagwörter verwendet werden können. Allerdings sind davon derzeit nur ca. 340.000 verschiedene GND-Entitäten über die inhaltliche Erschließung mit mindestens einer Publikation im Katalog der DNB verknüpft. Bei den im KI-Projekt betrachteten deutschen wissenschaftlichen Publikationen wird sogar nur ein Teilbestand von ca. 204.000 GND-Entitäten für die inhaltliche Erschließung verwendet. Ein Großteil der GND-Entitäten wird also gegenwärtig in der DNB nicht genutzt und es existiert somit auch kein Goldstandard, an dem maschinelle Verfahren trainiert oder evaluiert werden können. Eine weitere Herausforderung besteht darin, dass die Verwendungshäufigkeit von GND-Entitäten sehr unterschiedlich ist: Einige wenige Konzepte bzw. Labels der GND sind sehr häufig, d.h. sie werden mehr als 10.000 Mal verwendet. Spitzenreiter ist das Schlagwort „Deutschland", das 260.000 Mal vergeben wurde. Unter den 340.000 vergebenen GND-Entitäten befinden sich aber auch fast 40 Prozent, die nur einmal vergeben wurden. Diese Unterschiede sind in einem Häufigkeitsverteilungsdiagramm (siehe Abbildung 2) dargestellt.

Die am häufigsten verwendeten GND-Entitäten werden als „Head-Labels" bezeichnet. Etwas seltener verwendete GND-Entitäten werden dem „Body" oder „Torso" zugeordnet. Selten oder nie verwendete GND-Entitäten werden als Tail-Labels bezeichnet. Aus der großen Menge an Tail-Labels, wie sie im Trainingsmaterial der DNB zu finden sind, leitet sich die Bezeichnung Long Tail Charakteristik ab. Diese Ausprägung ist insbesondere im Hinblick auf

14 Zentrum für Informationsdienste und Hochleistungsrechnen (ZIH) – TU Dresden. https://tu-dresden.de/zih

15 Hirschle, Jochen: Deep Natural Language Processing: Einstieg in Word Embedding, Sequence-to-Sequence-Modelle und Transformer mit Python, 1. Aufl., München: Carl Hanser Verlag GmbH & Co. KG. 2022, S. 41

16 Botsch, Benny: Maschinelles Lernen – Grundlagen und Anwendungen, 1. Aufl., Berlin, Heidelberg: Springer Berlin Heidelberg, 2023, S.16. https://doi.org/10.1007/978-3-662-67277-8

17 Botsch, Benny: Maschinelles Lernen – Grundlagen und Anwendungen, 1. Aufl., Berlin, Heidelberg: Springer Berlin Heidelberg, 2023, S.12. https://doi.org/10.1007/978-3-662-67277-8

18 Boyd, K.; Eng, K.H.; Page, C.D.: Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals. In: Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2013. Lecture Notes in Computer Science, vol 8190. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-40994-3_29

19 DNB-Glossar - DDC-Sachgruppen. https://www.dnb.de/DE/Service/Glossar/_functions/glossar.html?cms_lv2=56762&cms_lv3=326696

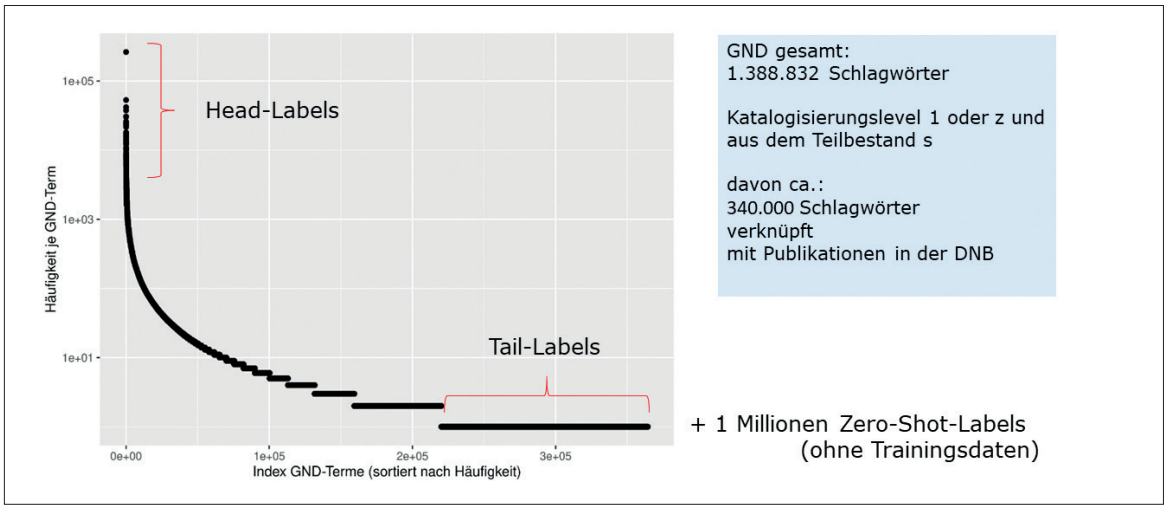

Abbildung 2: Long-Tail Charakteristik für die Trainingsdaten in der DNB[20]

lernende Verfahren herausfordernd, da möglichst viele Datensätze einer „Klasse" benötigt werden, um Muster zu lernen und entsprechende Vorhersagen berechnen zu können. Würde man pauschal sagen, dass nur Labels, die häufiger als 10 mal vorkommen, überhaupt sinnvoll durch ein statistisches Verfahren gelernt werden können, dann wären selbst in dem „großen" Trainingsset mit ausschließlich Titeldaten nur 2% aller GND-Terme für ein sinnvolles Training nutzbar.[21] Die Zuordnung von mehreren Labels aus einer großen Menge von Millionen potentiell möglicher Labels, wobei für die überwiegende Zahl der Labels nur ein oder meist gar kein Trainingsobjekt vorliegt, wird im maschinellen Lernen als Extreme Multi-Label-Text-Classification (XMLC) Problem eingestuft. Multi-Label-Classification bedeutet in diesem Fall, dass Textdokumente mit Labels aus einer a priori festgelegten Menge von Labels verknüpft werden sollen. Somit lässt sich die maschinelle Beschlagwortung von Texten mit Konzepten aus der GND als ein solches XMLC-Problem abstrahieren. Dabei ist die Menge der zutreffenden Labels pro Dokument unterschiedlich und nicht begrenzt. Charakteristisch für XMLC-Probleme ist zudem, dass zum einen eine große „Labelmenge", d.h. 10^5 - 10^6 Labels zur Verfügung stehen und zum anderen eine Long-Tail-Charakteristik vorliegt, wie sie oben auch für unsere Trainingsdaten beschrieben wurde. Die wissenschaftliche Fragestellung des KI-Projekts kann somit auf die Suche nach geeigneten algorithmischen Lösungen für das XMLC-Problem eingegrenzt werden. Das XMLC-Problem ist herausfordernd, da die Anzahl der möglichen Labels enorm groß ist und für jedes potentielle Label eine Vorhersage bezüglich der Relevanz für die zu erschließende Publikation generiert werden muss.

Hinzu kommt die Herausforderung, dass die Volltextverarbeitung der Online-Publikationen hochdimensionale „Features" als Input für die Modelle liefert. Dies erfordert sehr effiziente Algorithmen. XMLC wird in verschiedenen Bereichen eingesetzt, wie z.B. zum automatischen Taggen von Texten in sozialen Medien, zur automatischen Kategorisierung von Artikeln in Nachrichtenportalen, zur Kennzeichnung von Webseiten, zur Kommentierung von Nachrichten, zur Produktkategorisierung in E-Commerce-Plattformen und in Empfehlungssystemen für personalisierte Inhalte.[22] Zur Lösung dieser Aufgaben wurde und wird eine Vielzahl von Algorithmen bzw. Verfahren entwickelt, die nun in Experimenten auf ihre Eignung für die GND-Beschlagwortung deutschsprachiger wissenschaftlicher Online-Publikationen untersucht und verglichen werden sollen.

Experimente mit verschiedenen Verfahren

Um herauszufinden, welche Verfahren gute Ergebnisse für XMLC liefern, bieten verschiedene öffentliche Benchmark-Studien wertvolle Anregungen. In solchen Benchmark-Studien wird die Leistung verschiedener Verfahren unter gleichen Bedingungen verglichen und die Ergebnisse in Ranglisten aufgeführt, wie z.B. die Top 10 Verfahren im Wikipedia 500K Benchmark[23].

In weiterführenden Literaturstudien werden die besten Verfahren aus den Benchmark-Studien genauer untersucht, um festzustellen, ob diese auch für die spezifischen Anwendungsbereiche und Ziele des Projekts geeignet

20 Kähler, Maximilian: Erschließungsmaschine EMa und KI-Projekt der DNB. S.11. Präsentation bei Workshop 2022: Einsatz von KI und DH in Bibliotheken. https://wiki.dnb.de/display/FNMVE/Workshop+2022%3A+Einsatz+von+KI+und+DH+in+Bibliotheken+--+ein+Erfahrungsaustausch+auf+Werkstattebene?preview=/259630830/259637585/DNB_Projekte_Aktivitaeten_2022-11-03.pdf. Abgerufen am 19. April 2024.

21 Beim Training mit Volltexten sind es weniger als 0,7 % aller GND-Entitäten, die über 10 oder mehr Trainingsdatensätze verfügen.

22 You, Ronghui et al. "AttentionXML: Label Tree-based Attention-Aware Deep Model for High-Performance Extreme Multi-Label Text Classification." Neural Information Processing Systems (2018).

23 Quelle: The Extreme Classification Repository (http://manikvarma.org/downloads/XC/XMLRepository.html), Wikipedia-500K Benchmark. Abgerufen am 19. April 2024.

Universität Marburg

zAmbelli

EINFACH MACHEN. AUS METALL.

Zambelli Bibliotheken

Lernen und Wohlfühlen

Die Zambelli Bibliothekseinrichtungen begleiten wissenschaftliche und öffentliche Bibliotheken, die sich mit neuen Gegebenheiten auseinandersetzen und sich weiterentwickeln wollen. Wir helfen Ihnen Ihre Bibliothek so auszustatten, dass attraktive und funktionale Lernräume entstehen. Dabei können Sie sich auf in Sicherheit und Funktion bewährte Einrichtungslösungen verlassen.
Gemeinsam schaffen wir gestalterisch-kreative Raumkonzepte.

Nehmen Sie mit uns Kontakt auf!
regalsysteme@zambelli.com

112.BIBLIOCON 2024
offen.lokal.global.
04. – 07.06.2024 | Hamburg

Besuchen Sie uns:
Congress Center Hamburg | Halle H | Stand 011

Universitätsbibliothek Salzburg

Berlin-Brandenburg International School

www.zambelli.com/regalsysteme

sind. Wenn ein Verfahren oder Modell geeignet erscheint und ausgewählt wurde, wird es als DVC-Projekt in eine Test-Pipeline des KI-Labors integriert. Es werden verschiedene Experimente durchgeführt, um die Leistungs- und Anpassungsfähigkeit des ausgewählten Verfahrens zu überprüfen. Am Anfang steht die Definition der Hyperparameter, d.h. der Parameter, die das Verhalten des Modells beeinflussen, aber nicht direkt durch das Training gelernt werden. Anschließend wird das Modell mit den Trainingsdaten trainiert, wobei die Validierungsdaten zur Bewertung und Überprüfung der Modellleistung während des Trainings verwendet werden. Zur Feinabstimmung der Hyperparameter wird zunächst festgelegt, wie viele Optimierungsläufe durchgeführt werden sollen. Diese Optimierungsläufe können manuell definiert werden, z.B., wenn aufgrund der erforderlichen Rechenleistung für einen Trainingslauf nur wenige Durchläufe durchgeführt werden können. Die Hyperparameter können aber auch automatisiert optimiert werden, z.B. mit dem Hyperparameter-Tuning-Framework Optuna. Für jeden dieser Durchläufe schlägt ein Optimierungsalgorithmus eine bestimmte Kombination von Hyperparametern vor. Anschließend wird das Modell mit den vorgeschlagenen Parametern trainiert und die Ergebnisse werden anhand der Validierungsdaten ausgewertet und gespeichert. Die Hyperparameteroptimierung ist ein wichtiger Schritt im maschinellen Lernprozess, da sie dazu beiträgt, die Leistung des Modells zu maximieren. Abschließend wird mit Hilfe der Testdaten überprüft, wie gut das Modell auf neue, unbekannte Daten generalisiert. Wenn das trainierte Verfahren eine gute Qualität für die Testdaten vorhersagt, wird das Modell auf unbekannte Daten, d.h. Daten ohne intellektuelle GND-Erschließung, angewendet. Die daraus resultierenden Ergebnisse, also die maschinellen GND-Vorschläge, werden einer Stichprobenkontrolle unterzogen. Für 18 ausgewählte DDC-Sachgruppen führen die Fachreferent:innen der Inhaltserschließung der DNB, eine intellektuelle Bewertung durch. Dabei werden die maschinell vergebenen Schlagwörter den Bewertungskategorien "sehr nützlich", "nützlich", "wenig nützlich" bis "falsch" zugeordnet. Diese qualitative Beurteilung durch professionelle Inhaltserschließer:innen wird als sehr wertvoll angesehen. Gegenüber dem einfachen Abgleich mit dem Goldstandard zur Berechnung des F-Score, der als binäre Relevanzbewertung verstanden werden kann und bei dem jede Vorhersage eine Entweder-oder-Entscheidung ist, liefert die qualitative Beurteilung differenziertere Ergebnisse. Die Nützlichkeit eines Schlagworts wird auf einer Ordinalskala bewertet, was zum besseren Verständnis beiträgt, wie gut die maschinell generierten GND-Vorschläge den Inhalt der Publikation abbilden und inwieweit sie zur Verbesserung der Suche beitragen können.

Mit diesem Vorgehen wird systematisch untersucht, welche Methoden aus dem Bereich innovativer KI-Entwicklungen für die maschinelle Beschlagwortung mit der GND von deutschsprachigen wissenschaftlichen Publikationen genutzt werden können. Die gewonnenen Erkenntnisse sollen dazu dienen, die vorhandenen Tools in der Erschließungsmaschine zu ergänzen und auszubauen.

Zunächst wurden die Verfahren der Omikuji-Familie intensiv untersucht; sie sind auch in dem Open-Source-Tool Annif implementiert. Omikuji-Bonsai wird in der EMa als Teil des Ensembles für die Beschlagwortung eingesetzt. Die Verfahren Omikuji-Bonsai, Omikuji-Parabel und Omikuji-Attention-XML verwenden alle sogenannte Partitioned Label Trees, Baumstrukturen, in denen die Schlagwörter nach Ähnlichkeit in Cluster aufgeteilt werden.[24] Darüber hinaus wurden die Verfahren MLLM, DISMEC++ und ZestXML im oben beschriebenen eigenen KI-Labor und AttentionXML im externen ZIH der TU Dresden[25] untersucht.

Bei MLLM, Maui-like Lexical Matching handelt es sich um ein lexikalisches Verfahren.[26] DISMEC++[27] ist ein One-vs-All-Classifier, auch bekannt als One-vs-Rest-Classifier. Die One-vs-All-Klassifikation funktioniert, indem ein Mehrklassen-Klassifikationsproblem in mehrere binäre Klassifikationsprobleme umgewandelt wird. Jeder binäre Klassifikator wird darauf trainiert, zu unterscheiden, ob ein Label zutreffend ist oder nicht.[28]

Das XMLC-Verfahren ZestXML ist ein experimentelles Verfahren von Microsoft/Bing.[29] ZestXML ist für die Aufgabe von Generalized Zero-shot XML (GZXML) konzipiert, bei der aus allen verfügbaren Labels die besonders relevanten, sowohl mit als auch ohne Trainingsbeispiele, ausgewählt werden müssen. Es befasst sich also mit der XMLC-Herausforderung, unbekannte Labels ohne Trainingsbeispiele vorherzusagen, indem statistische Assoziationen zwischen Textmerkmalen und Labelmerkmalen gelernt werden. AttentionXML kombiniert Partitioned Label Trees

24 Konermann, Katja. KI-Projekt: Abschlussbericht für das Verfahren Omikuji, Deutsche Nationalbibliothek, August 2023, S.2, (nicht öffentlich zugänglich).

25 Zentrum für Informationsdienste und Hochleistungsrechnen (ZIH) – TU Dresden. https://tu-dresden.de/zih

26 Suominen, Osma; Koskenniemi, Ilkka. Annif Analyzer Shootout: Comparing text lemmatization methods for automated subject indexing. In Code4Lib Journal, Issue 54, 2022-08-29. https://journal.code4lib.org/articles/16719

27 DiSMEC++. A software package for large-scale linear multilabel classification. https://github.com/xmc-aalto/dismecpp. Abgerufen am 19. April 2024.

28 Wikipedia contributors. (2023-10-20). Multiclass classification. In Wikipedia, The Free Encyclopedia. Abgerufen am 24.3.2024. from https://en.wikipedia.org/w/index.php?title=Multiclass_classification&oldid=1181072194.

29 Gupta, N. (2021). Generalized Zero-Shot Extreme Multi-label Learning. In GitHub - nilesh2797/zestxml: This is the official codebase for KDD 2021 paper Generalized Zero-Shot Extreme Multi-Label Learning. Abgerufen am 24.3.2024.

und spezielle Typen Neuronaler Netze (sog. LSTM[30]) für die Vorhersage von Labels.[31]

Zusätzlich werden Experimente mit dem großen Sprachmodell Luminous[32] der Firma Aleph Alpha aus Heidelberg durchgeführt. Luminous ist vergleichbar mit Chat GPT, die Infrastruktur befindet sich jedoch in Deutschland und unterliegt damit den EU-Datenschutzrichtlinien. Der zukünftige Fokus von Luminous liegt nach Angaben der Firma auf der Erklärbarkeit, wodurch Transparenz und Nachvollziehbarkeit auch in Bezug auf ethische Fragen hergestellt werden soll. Die Experimente mit Luminous zur Beschlagwortung greifen auf die API des Herstellers zu, so dass alle eigentlichen Berechnungen nicht im KI-Labor, sondern direkt im Rechenzentrum der Firma Aleph Alpha stattfinden. Deshalb können aus urheberrechtlichen Gründen für diese Experimente nur die Titeldaten und freizugängliche Online-Publikationen verwendet werden. Im ersten Schritt werden mit Hilfe spezieller Prompts freie Schlagwörter generiert. Dabei werden dem Sprachmodell Beispiele mitgegeben, die verdeutlichen, in welcher Weise Schlagwörter vergeben werden sollen. Die genaue Zusammensetzung der Beispiele in den Prompts ist entscheidend für die Ergebnisse, die man erhält. Vereinfacht ausgedrückt: Wenn z.B. nur Beispiele verwendet werden, in denen die Schlagwörter genau im Text vorkommen, werden bei der Anwendung auf die Testdaten auch bei vielen Beispielen mehr Schlagwörter vorgeschlagen, die auch im Text vorkommen. Die Prompts lassen sich wie folgt skizzieren: „Dies ist der Titel: „Statistik ohne Angst vor Formeln: das Studienbuch für Wirtschafts- und Sozialwissenschaftler", er wurde mit folgenden Schlagwörtern: „Sozialwissenschaften; Statistik; Wirtschaftswissenschaften; Statistik", erschlossen". Dies ist der Titel: „Aktualität von Bibliotheksangst", er wurde mit folgenden Schlagwörtern:"…" usw." Am Ende der so formulierten Beispiele wird für die zu beschlagwortenden Titel die Frage gestellt: „Dies ist der Titel „…", wie lauten die Schlagwörter dazu?" usw. Das Sprachmodell vergibt dann freie Schlagwörter. Diese frei vergebenen Schlagwörter entsprechen nicht automatisch den GND-Konzepten. Daher müssen in einem zweiten Schritt die vom Luminous-Modell vergebenen Schlagwörter dem jeweils am besten passenden GND-Konzept zugeordnet werden. Dazu werden sowohl von den GND-Konzepten als auch von den freien Schlagwörtern über das Sprachmodell Vektorrepräsentationen (sog. Word-Embeddings) erzeugt. Diese Vektoren können mit gängigen Abstandsmaßen, z.B. mit Hilfe der Kosinusähnlichkeit, verglichen werden, so dass für jedes freie Schlagwort ein GND-Konzept gefunden werden kann, das diesen Vektorabstand minimiert.

Bei der Evaluierung der verschiedenen Verfahren werden die Experimente in zwei Kategorien bewertet: In der ersten Kategorie werden die Modelle nur mit Titeldaten trainiert. Dabei wird die Leistungsfähigkeit der Modelle mit sehr wenig Inputtext, aber vergleichsweise viel Trainingsmaterial getestet. In der zweiten Evaluationskategorie werden Modelle mit Volltexten (oder ggf. mit gekürzten Volltexten) trainiert. Abbildung 3 zeigt den Vergleich von Titeldatenmodellen mit den bisher untersuchten Verfahren und Abbildung 4 entsprechend den Vergleich mit Volltextmodellen. Innerhalb der beiden Evaluationskategorien "Titeldatenmodelle" und "Volltextmodelle" werden die Modelle auch übergreifend nach weiteren Evaluationsdimensionen verglichen, z.B. nach dem Abschneiden in 18 ausgewählten DDC-Sachgruppen der Deutschen Nationalbibliografie. Die Abbildung mit den Ergebnissen der Volltextmodelle zeigt, dass die Sachgruppen 340 Recht und 650 Management bei allen Verfahren außer bei MLLM am besten abschneiden, während die MINT-Sachgruppen[33] insgesamt schlechter abschneiden. Möglicherweise auch eine Folge fehlender intellektueller Erschließung in den sogenannten nichtbuchaffinen[34] Sachgruppen. Überraschend gut funktioniert MLLM für die Sachgruppe 100 Philosophie. Vergleicht man die Ergebnisse der Volltextmodelle mit denen der Titeldatenmodelle, so fallen einige Unterschiede auf. Dies gilt besonders für die Sachgruppe 530 Physik. DISMEC++ liegt bei fast allen Sachgruppen unabhängig vom Modell vorne. Das volle Potenzial der verschiedenen Verfahren ist jedoch bei der Hyperparameter Auswahl noch nicht ausgeschöpft.

Die Ergebnisse zeigen, dass jede Verfahrensfamilie unterschiedliche Eigenschaften aufweist. Im weiteren Projektverlauf werden die Stärken und Schwächen der Verfahren analysiert. Es ist unwahrscheinlich, dass es genau eine optimale Methode für das Beschlagwortungsproblem gibt. Ziel des KI-Projektes ist es, verschiedene Methoden zu vergleichen und Stärken und Schwächen zu identifizieren. Heterogene Ansätze können dann zu einem starken Ensemble kombiniert werden. Die gesammelten Erkenntnisse werden helfen, ein Ensemble von mehreren sich ergänzenden Methoden und Modellen zu bilden. Viele ein-

30 Long short-term memory

31 YOU, Ronghui, et al. Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. Advances in neural information processing systems, 2019, 32. Jg. https://doi.org/10.48550/arXiv.1811.01727

32 Hahn, Silke. Luminous schließt Europas KI-Lücke: Aleph Alpha auf Augenhöhe mit US-Anbietern. In Heise online, 21.02.2023. https://heise.de/-7521254

33 MINT-Fächer: Mathematik, Informatik, Naturwissenschaft und Technik.

34 Veränderungen in der Inhaltserschließung der Deutschen Nationalbibliothek ab 1. Juli 2019. (2021) https://dnb.de/SharedDocs/Downloads/DE/Professionell/Erschliessen/veraenderungenInhaltserschliessungDnbJuli2019.pdf?__blob=publicationFile

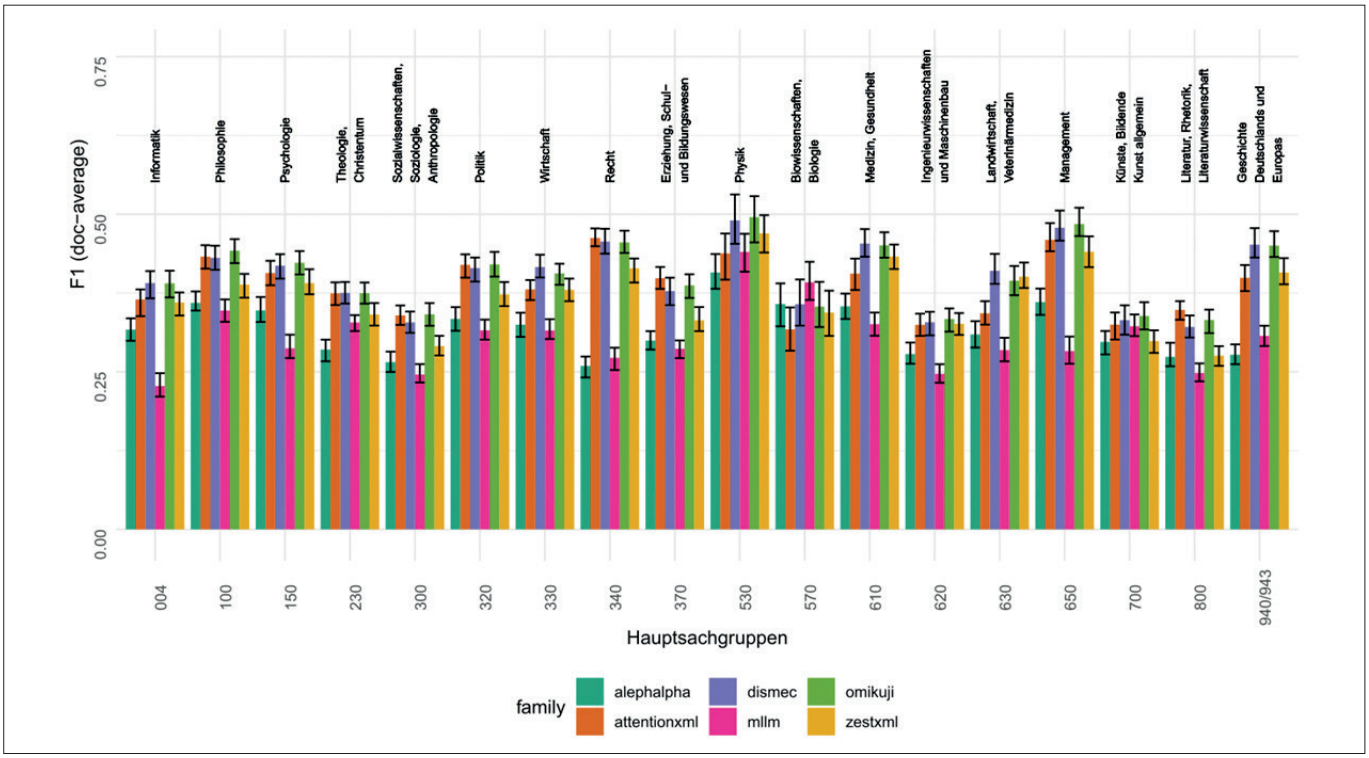

Abbildung 3: Ergebnisse der verschiedenen Verfahren mit Titeldatenmodellen in 18 ausgewählten DDC-Sachgruppen, F1 (doc-average) und 95%-Konfidenzintervalle [35]

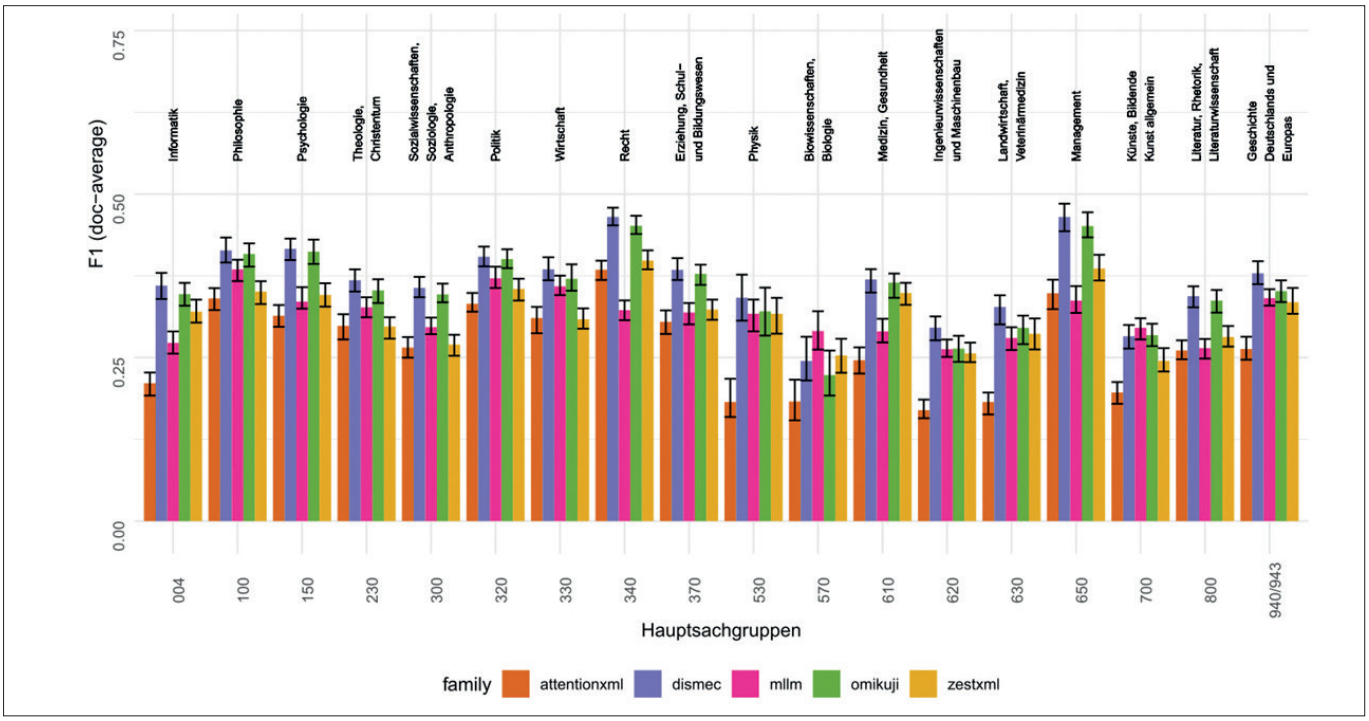

Abbildung 4: Ergebnisse der verschiedenen Verfahren mit Volltextmodellen in 18 ausgewählten DDC-Sachgruppen, F1 (doc-average) und 95%-Konfidenzintervalle[36]

zelne Erkenntnisse aus dem KI-Projekt sind bereits in die Erschließungsmaschine EMa eingeflossen, so dass dort im Produktivbetrieb mit einem Ensemble aus Titeldaten- und Volltextmodellen jeweils für Omikuji-Bonsai und MLLM bereits ein F1-Score ≥ 0,4 erreicht wird.

Im Laufe des Projekts werden noch weitere Verfahren er-

35 Kähler, Maximilian: Zwischenbericht: KI-Projekt – Vergleichende Evaluation zum Ist-Stand bislang getesteter KI-Verfahren. Deutsche Nationalbibliothek, März 2023, S.3, (nicht öffentlich zugänglich).

36 Kähler, Maximilian: Zwischenbericht: KI-Projekt – Vergleichende Evaluation zum Ist-Stand bislang getesteter KI-Verfahren. Deutsche Nationalbibliothek, März 2023, S.4, (nicht öffentlich zugänglich).

probt. Insbesondere ist geplant, die Ergebnisse großer offener Sprachmodelle wie Llama2[37] oder Mistral[38] miteinander und vor allem auch mit den Ergebnissen von Lumious zu vergleichen. Experimente mit Luminous deuten darauf hin, dass große Sprachmodelle Ergebnisse liefern, die sich nur wenig mit anderen klassischen Verfahren überschneiden. Es treten häufig andere Fehlertypen auf als z.B. bei lexikalischen Ansätzen oder 1VsAll- Classifier. Dies könnte eine wertvolle Ergänzung für ein zukünftiges Beschlagwortungs-Ensemble sein. Besonders bei GND-Entitäten, für die es wenig Trainingsmaterial gibt, könnte die Treffersicherheit erhöht werden. Es ist jedoch eine große Herausforderung, das Werkzeug des Sprachmodells so zu steuern, dass es für die Aufgabe der GND-Beschlagwortung geeignet ist. Das Thema Große Sprachmodelle hat seit Beginn des Projekts stark an Bedeutung gewonnen, viele neue Sprachmodelle wurden entwickelt. Welche davon geeignet sind und in das KI-Labor integriert werden können, muss nun geprüft werden. Die Entwicklungsdynamik auf dem Gebiet der Sprachmodelle ist jedoch so groß, dass abschließende Bewertungen zu diesem Thema selten langfristig gültig sind.

Schlussbemerkung

Da der XMLC-Ansatz auch für Suchmaschinenhersteller und E-Commerce-Anwendungen interessant ist, werden KI-Lösungen derzeit intensiv weiterentwickelt. Diese dynamische Entwicklung neuer Ansätze und Algorithmen verspricht auch für die maschinelle GND-Beschlagwortung deutliche Verbesserungen. Die Evaluation neuer Verfahren und deren Integration in bestehende Systeme ist keine zeitlich abgeschlossene Aufgabe, sondern eine kontinuierliche Verpflichtung. Das KI-Projekt dient nicht zuletzt auch dem Transfer geeigneter Technologien aus der Forschung und Entwicklung in die bibliothekarische Praxis. Ein intensiver Informations- und Erfahrungsaustausch mit Institutionen aus den Bereichen Forschung, Entwicklung und Anwendung ist dafür wichtig. In diesem Zusammenhang sind auch die jährlichen Veranstaltungen im Rahmen des „Netzwerks Maschinelle Verfahren in der Erschließung“[39]zu sehen, z.B. die Veranstaltung „KI in Bibliotheken – Neue Wege mit großen Sprachmodellen?“[40] im Dezember 2023 in Frankfurt. Es ist von entscheidender Bedeutung, stets auf dem neuesten Stand der Entwicklung zu sein und Anwendungen der künstlichen Intelligenz regelmäßig zu überprüfen und zu aktualisieren. Dies setzt eine ständige Bereitschaft zur Anpassung und Weiterentwicklung voraus. Allerdings erfordern die Entwicklung und der Einsatz von KI-Methoden qualifizierte Expert:innen, z.B. auf dem Gebiet der Mathematik und Informatik, sowie umfangreiche und spezialisierte Hardware-Ressourcen, wie z.B. Server mit leistungsfähigen GPU-Prozessoren. Eine zentrale Frage ist, ob die notwendige Hardware-Infrastruktur intern aufgebaut oder externe Hochleistungsrechner, z.B. als eine Art nationale Infrastruktur, genutzt werden können. Ein weiterer sehr wichtiger Aspekt ist die Weiterentwicklung eines qualitativ hochwertigen und aktuellen Goldstandards, der von den Fachreferent:innen der inhaltlichen Erschließung auch im Hinblick auf neue Themen und neues GND-Vokabular kontinuierlich aktualisiert werden muss. Die Zusammenarbeit zwischen den KI-Entwickler:innen und den Fachreferent:innen ist von entscheidender Bedeutung, um sicherzustellen, dass die KI-Verfahren die bestmöglichen Ergebnisse liefern können.

In einer zunehmend digitalisierten Welt stehen (National-)Bibliotheken vor der Herausforderung, ihre Rolle im Informationszeitalter neu zu definieren. Eine vielversprechende Möglichkeit ist die Bereitstellung von KI-Services. Diese können verschiedene Formen annehmen, von der Bereitstellung erprobter Methoden als Open-Source-Tools zur Nachnutzung über vortrainierte (Sprach-)Modelle bis hin zu Chatbots zur Unterstützung bei der Recherche. Durch die geschickte Verknüpfung von KI-Technologien mit etablierten bibliothekarischen Standards können Bibliotheken einen wichtigen Beitrag zur Modernisierung und Weiterentwicklung des Informationszugangs leisten. Dies könnte nicht nur die Effizienz der eigenen Bibliothek steigern, sondern auch die Wissensvermittlung und den Wissensaustausch auf globaler Ebene fördern. Mit diesem KI-Projekt geht die Deutsche Nationalbibliothek einen wichtigen Schritt in diese Richtung.

Elisabeth Moedden
studierte Bauingenieurwesen an der Technischen Universität Braunschweig und absolvierte an der Universität- und Landesbibliothek Darmstadt das Bibliotheksreferendariat. Seit 2007 arbeitet sie an der Deutschen Nationalbibliothek, zunächst als Fachreferentin für Informatik und Technik, seit 2014 leitet sie das standortübergreifende Referat Automatische Erschließungsverfahren, Netzpublikationen.
e.moedden@dnb.de

37 Large language model, Meta. https://llama.meta.com/llama2/. Abgerufen am 19. April 2024.

38 Large language model, Hugging Face. https://huggingface.co/docs/transformers/model_doc/mistral. Abgerufen am 19. April 2024.

39 Homepage zum „Netzwerk maschinelle Verfahren in der Erschließung“: https://wiki.dnb.de/display/FNMVE/Netzwerk+maschinelle+Verfahren+in+der+Erschliessung. Abgerufen am 19. April 2024.

40 Fachtagung „KI in Bibliotheken – Neue Wege mit großen Sprachmodellen?“: https://wiki.dnb.de/pages/viewpage.action?pageId=306153594. Abgerufen am 19. April 2024.