**Die Projektarbeit:**

*„Spezifikation und PoC der KI-gestützten „Beta-Klassen-Recognition“ in der Presserecherche.“*

**Abstract:**

**1. Zielsetzung**

**Was war die Aufgabenstellung was waren die Rahmenbedingungen**

Die Recherche nach Interviews, Chronologien oder Kommentaren in Presseartikeln gehört zum Alltag am IDA-Desk des Südwestrundfunks (SWR). Im jetzigen System zur automatischen Verschlagwortung von Pressetexten werden nur Inhaltsklassen und Entitäten maschinell indexiert. Präsentationsformen werden bislang nur in Einzelfällen händisch indexiert. Dadurch wird die Recherche nach eben diesen erschwert und ist mit einem erhöhten Zeitaufwand verbunden. Die konkrete Suche nach Präsentationsformen ermöglicht es, schnell inhaltliche Schneisen zu schlagen. PAN soll ab dem 01.01.2023 durch MDH:Presse abgelöst werden. Dabei werden die bestehenden Daten in das neue System importiert und die PAN-spezifischen Präsentationsformen auf Normdatenbank-konforme Formen gemappt. Mit dem Umzug in MDH:Presse wird auch das aktuell eingesetzte Textmining abgelöst. In der zukünftigen Crossmedialen Suche in Medas kommen diverse Mining-Services zum Einsatz.   
Der Service soll die Recherche nach Beta-Klassen ermöglichen und dadurch die Qualität der Rechercheergebnisse für Rechercheure\*innen steigern. Dazu soll ein Proof of concept (PoC) erstellt werden, indem verschiedene in Frage kommende Technologien zur Kategorisierung von Textdaten getestet werden. Geplant ist das Testen von Modellen auf Basis von Support Vektor Maschinen (SVM) und Logistischer Regression, sowie Deep-Learning Netzwerken auf Transformer-Architektur wie zum Beispiel BERT. Als Trainingsmaterial dienen zunächst Presseartikel aus PAN. Die Modellanforderungen sind bislang noch nicht spezifiziert, da beim Einsatz von KI bzw. Machine Learning in den seltensten Fällen hun­dert­pro­zen­tige Genauigkeit erreicht werden kann.

**2. (vereinbarte) Ergebnistypen**

Folgende Ergebnistypen wurden als relevant für die erfolgreiche Durchführung des Projekts identifiziert und vereinbart:

* Auswahl, Testung und Finetuning verschiedener KI-Modelle zur Erkennung von Präsentationsformen in Pressetexten
* Ausarbeitung eines PoCs zur Weitergabe an die Entwickler\*innen der Mining-Plattform
* Erstellung einer Spezifikation des KI-Modells
* Einbettung des Modells in einen Service, welcher in die Mining-Plattform von Medas implementiert werden kann

**3. (gelieferte) Ergebnistypen**

Folgende Ergebnistypen wurden erarbeitet und geliefert:

* Tests verschiedener KI-Modelle zur Erkennung von Präsentationsformen in Pressetexten
* Ausgearbeiteter Proof of Concept inklusive einer Spezifikation des KI-Modells mit weiteren Handlungsempfehlungen
* Dokumentierte Python Skripte zur Implementierung in die Mining-Plattform

**[Erläuterung und Bewertung der Ergebnisse aus der Spiegelstrichliste]** Der Proof of Concept mit Handlungsempfehlungen und dokumentierten Python-Skripten konnte den Stakeholdern erfolgreich übergeben werden. Die Prüfung steht noch aus, da die Implementierung und weitere Tests auf das Jahr 2023 verschoben worden sind. Die vorläufigen Ergebnisse sind vielsprechend und zeigen, dass die Bestimmung von Beta-Klassen auf Basis der Pressetexte eingeschränkt möglich ist und somit die Recherche erleichtern kann.

**4. Change Requests**

Die meisten Ergebnistypen konnten geliefert werden und das Projekt konnte in großen Teilen fertiggestellt werden. Verzögerungen im Prozessverlauf und vorstellbar-kritische Zwischenergebnisse führten zu einer Anpassung der Ergebnistypen. Die Implementierung in den Miningservice, sowie Tests von weiteren KI-Modellen wurden auf 2023 verschoben. Stattdessen wurden die programmierten Skripte dokumentiert und sind zur Weitergabe an die Entwickler\*innen der Mining-Plattform bereit.

**Strukturierte Beschreibung:**

* Bearbeitungszeitraum: April 2022 bis November 2022
* Auswahl der KI-Modelle, Bereinigung der Pressedaten: März - Mai 2022
* Entwicklungsphase: Mai – Juni 2022
* Testphase und anschließende Präsentation der Evaluationsergebnisse: Juli – September 22
* Übergabe in das Entwicklerteam zur Einbettung des PoCs in den Service
* November 2022: Abgabe der Projektarbeit
* Bearbeitungsdauer: 9 Monate
* Abgabe: November 2022

**Relevante Qualifikationen der berufspraktischen Phase des Volontariats:**

Von Januar 2021/April 2022 bis Dezember 2022/März 2023 hat Herr Timo Schumacher Qualifikationen in folgenden, nach thematischen Gruppen geordneten dokumentarischen Aufgabenfeldern bei/m NN erworben:

1. Projektmanagement und Vorgehensmodelle, Requirements Management, Geschäftsprozesse, Qualität, Führung und Organisation

* Nachhaltige Kenntnisse und Erfahrungen
  1. im Prototyping und Testing durch Beiträge zum Projekt "IDA-Bot: Automatisierter PAN-Zugriff"
  2. in agilen im Projekt „Spezifikation und PoC der KI-gestützten „Beta-Klassen-Recognition“ in der Presserecherche“ inklusive Entwicklungsmethoden inklusive Requirements Management und Requirements Analyse
* Umfassende Kenntnisse in agilen Vorgehensmodellen durch die Arbeit in einer agil-strukturierten Hauptabteilung (Teilnahme an Retrospektiven, Übernahme von unterschiedlichen Teamaufgaben) und durch die Organisation, Durchführung und Moderation von abteilungsinternen Veranstaltungen wie der Reihe „Erzählcafé: Wie war das damals bei der IDA“Grundlegende Kenntnisse in agilen Managementprozessen durch Beitrag im Projekt „IDA-Kartenspiel“, einem Tool zur schnellen agilen Entscheidungsfindung

1. Erwerbung, Import, Ingest, Retrodigitalisierung, Langzeitarchvierung

* Grundlegende Kenntnisse und Erfahrungen
  1. im Erwerb von Presseartikeln aus internen und externen Datenbanken (Agenturen, Genios, PMG) zur redaktionellen Nutzung
  2. in der Bearbeitung und Archivierung von Radiobeiträgen im Audiomassenspeicher (AMS)
  3. über das Retrodigitalisierungsprojekt „KARO“ im SWR
  4. über die Archivierung von Online-Only-Bewegtbildinhalten mithilfe von einem teilautomatisierten Workflow
* Nachhaltige Kenntnisse und Erfahrungen
  1. in der Lektorierung von deutschsprachigen, digitalen Zeitungen und Zeitschriften zur Auswahl von Artikeln für die Langzeitarchivierung in der Pressedatenbank PAN
  2. in der Archivierung von Fernsehsendungen und -beiträgen im Videomassenspeicher (VMS) sowie in der Essenzverlinkung mit FESAD und VPMS

1. Erschließung, Wissensrepräsentierung und Regelwerke/MD-Standards

* Umfassende und professionelle Erfahrungen und Kenntnisse
  1. in der inhaltlichen Erschließung von Bewegtbild-Magazinsendungen und Nachrichten wie SWR Aktuell, SWR Sport oder Made in Südwest mithilfe vom Regelwerk Mediendokumentation (REM) und der ARD-Sachklassifikation sowie beim sportlichen Großereignis Fußball-Weltmeisterschaft 2022
  2. der Inhaltserschließung von Hörfunkwortbeiträgen wie beispielweise SWR2 Aktuell und von diversen Magazinbeiträgen mithilfe vom Regelwerk Mediendokumentation (REM) und der ARD-Sachklassifikation.
* Grundlegende Kenntnisse und Erfahrungen
  1. in der Übernahme von Metadaten aus den Redaktionsplanungs- sowie Programm- und Sendeablaufsystemen für die inhaltliche Erschließung von Bewegtbild (OpenMedia) und Wort (Webmerlin)
  2. in der Indexierung von Zeitungsartikeln mithilfe eines semi-automatischen Verfahrens
* Erstellung und Anpassung von Inhalten und Einträgen im hausinternen Wiki

1. Dokumentation, Semantik, Automatisierung

* Nachhaltige Kenntnisse und Erfahrungen in Automatisierung von Pressedatenbank-Zugängen durch Beteiligung am Projekt „IDA-Bot“
* Umfassende Kenntnisse und Erfahrungen in der maschinellen Verarbeitung natürlicher Sprache durch die Leitung und Durchführung des Projekts „Beta-Klassen-Recognition“ in der Presserecherche“
* Anwendungserfahrung mit den Audio-und Textminingservices in FESAD und HFDB, sowie in der zukünftigen Crossmedialen-Suche
* Anwendungserfahrung mit der Verwendung von Autohotkeys in der Bewegtbild- und Worterschließung

1. Essenzspeicher & CMS, (Metadaten-)Datenbanken, semantische Technologien

* Professionelle Kenntnisse und Erfahrungen in der regelmäßigen Nutzung der ARD-internen Datenbanken (FESAD/Bewegtbild, HFDB/Wort, HFDB/Musik, PAN/Presse, 42/Agenturen, Zeitlupe/Termine) und externen Datenbanken zu Dokumentations-, Archivierungs- und Recherchezwecken (z.B. GENIOS oder Northdata)

1. Retrieval/Recherche, Bereitstellung, Beratung und Wissensvermittlung incl. Visualisierung

* Umfassende und professionelle Kenntnisse und Erfahrungen
  1. in multimedialer Recherche (Bewegtbild, Schall und Text) in allen ARD-eigenen Archiv- und Redaktionssystemen, externen Datenbanken sowie Online und Social Media durch die Arbeit am standortübergreifenden IDA-Recherchedesk
  2. in der digitalen und physischen Bereitstellung von unterschiedlichen Essenztypen als Ergebnis von multimedialen Recherchen
* Umfassende Erfahrungen und Kenntnisse in der Erstellung von Themendossiers und Pressespiegeln zu diversen Themen wie z.B. Corona oder Wissenschaft
* Nachhaltige Erfahrungen und Kenntnisse in Social-Seeding-Recherchen zum Beispiel für das Format „Down the Road“ (erscheint 2022) und der Protagonist\*innen-Suche (z.B. für SWR MixTalk)
* Grundlegende Kenntnisse und Erfahrungen in der Visualisierung mit Datawrapper

1. Sonderaspekte der Dokumentation, Schulung, sowie CC und Doku-Marketing

* Präsentation der Dienstleistungen und Services der Hauptabteilung IDA im Rahmen eines offenen Stammtischs und einer Redaktionssitzung von DASDing
* Selbstständige Beratung von der Nutzung der Archivsysteme für Kund\*innen

1. Redaktionsarbeit

* Grundlegende Erfahrungen und Kenntnisse
  1. im Community-Management (Moderation, Löschung und Antworten auf Community-Beiträge) diverser SWR-Kanäle auf Instagram, Facebook und den SWR-Webseiten mithilfe von Swat.io
  2. beim Feedback von digitalen- und analogen Formatideen für den SWR
  3. in der Distribution von Inhalten auf diversen Social-Media-Plattformen (Pinterest, TikTok, Instagram, Facebook usw.)
  4. in der Führung von persönlichen Interviews, sowie Telefoninterviews mit Regierungspoltiker\*innen
  5. in die Kuratierung der SWR-Homepage mit Sophora, sowie der SWR-Seite in der ARD-Mediathek mit ARD-Composer
* Vorführung von Reddit als Social-Mediaplattform für die Distribution von SWR-Inhalten
* Nachhaltige Erfahrungen und Kenntnisse
  1. von der Durchführung von A/B-Tests auf der SWR.de-Startseite
  2. in der Zusammenstellung und Versendung der Online-Themenplanung
  3. in Recherchen für die SWR-Hörfunk Redaktion im ARD-Hauptstadtstudio Berlin
  4. in der Produktion von Nachrichtenminuten für das ARD-Sammelangebot
  5. in der Durchführung von TV-Interviews
* Befähigung zum Erstellen von Inhalten für die Homepage-Redaktion SWR.de und mit dem CMS Sophora
* Weiterentwicklung des „Prognose-Tools“, einem KI-Tool, dass mit Wiedergabe-Daten aus der Vergangenheit und weiteren Metadaten die Wiedergabe-Nutzung für kommende Formate in der ARD-Mediathek prognostiziert
* Erstellung von Teasern für die Startseite der ARD-Mediathek mit ARD-Composer der ARD-Mediathek
* Umfassende Kenntnisse und Erfahrung
  1. in der Erstellung von bebilderten Kurznachrichten für die Nachrichtensendungen des SWR
  2. in der Erstellung von TV-Beiträgen sowohl im Schnitt, der textlichen Konzeption und Vertonung

1. Monitoring und Analyse

* Umfassende Kenntnisse und Erfahrungen in der Erstellung von Monitoring-Dashboards im Tool „Crowdtangle“ für die Themenfindung von Podcasts (Pancake Politik/DASDing, Chaos²/SWR3)
* Nachhaltige Kenntnisse und Erfahrungen in der Aufbereitung und Visualisierung von Nutzer\*innen-Statistiken für diverse Youtube-Formate mithilfe von Chart.io und Excel
* Professionelles Monitoring kurioser Online-News für den IDA-Netzradar (Programmvorschläge für die Redaktionen) via Tiny Tiny RSS, Crowdtangle und Twitter-Listen

1. Forschungsdatenmanagement und Auswertungen, auch: Quellen und Analysen im Datenjournalismus

* Grundlegende Kenntnisse und Erfahrungen in der Analyse von IDA-Desk-Statistiken im Rahmen des IDA-Bot-Projekts
* Nachhaltige Kenntnisse und Erfahrungen in der Analyse und Visualisierung von Metadaten der ARD-Mediatheksinhalte mithilfe von Chart.io und Airtable

1. Urheberrechte, Lizenzen, Persönlichkeitsschutzrechte

* Grundlegende Kenntnisse von Rechten, Lizenzen und Persönlichkeitsrechten im Rahmen der inhaltlichen Erschließung und Recherche
* Prüfung von Material auf urheberrechtlich geschützte Stellen für die Online-Publikation
* Durchführung von Kund\*innen-Beratungen zur Wiederverwendung von rechtlich beschränkten Hörfunk- und Bewegtbildmaterialen

Herr Timo Schumacher hat im Rahmen des Volontariats an folgenden Fachkonferenzen teilgenommen:

* Veranstaltung h\_da-Symposium zur wissenschaftlichen und Mediendokumentation (Hochschule Darmstadt, 10.11.21, Teilnehmer)
* Frühjahrs-Tagung "Große Freiheit oder Quarantäne - Agile Mediendokumentation in Zeiten von Corona" (Verein für Medieninformation und Mediendokumentation (vfm), 26.04.21 bis 28.04.21, Teilnehmer)
* Frühjahrs-Tagung "Fit for Files - Medieninformation zwischen dokumentarischer und künstlicher Intelligenz" (25.04.22 bis 27.04.22, Verein für Medieninformation und Mediendokumentation (vfm), Teilnehmer)
* Veranstaltung h\_da-Symposium zur wissenschaftlichen und Mediendokumentation (Hochschule Darmstadt, 9.11.22, Teilnehmer)

Herr Timo Schumacher hat während des Volontariats an folgenden fachlich relevanten Weiterbildungen teilgenommen:

* Workshop Agiles Basistraining: Agile Methoden und Werte (SWR, 29.04.21)
* Workshop Online-Recherche: Suchmaschinen, Deep Web-Recherchen, Wikidata & Fact-Checking (SWR, 17.02.-23.02.21)
* Seminar Online Texten & SEO-Tools: CMS (Sophora) Textaufbau (z.B. Keywordfinder), Kuratierung/A-B-Testing (z.B. Chartbeat) (SWR, 24.09.21-01.10.21)
* Workshop Datenrecherche: Aufbereitung (Excel) und Visualisierung (DataWrapper) (SWR, 21.09-23.09.21)
* Seminar Presse- und Urheberrecht mit Online-Fokus (SWR, 04.10.21, 06.10.21)
* Workshop / Blended Learning „Erfolgreich Feedback“ geben (Kick-Off 18.11.21, Trainingseinheiten / Learning Nuggets 11.21-01.2~~1~~2 sowie 06.22-07.22)

*Herr Timo Schumacher hat während des Volontariats in den angegebenen Zeiträumen zu folgenden abteilungsinternen Projekten beigetragen:*

1. *Projekt „IDA-Kartenspiel“, Design-Thinking-Projekt & Prototyp für Tool zur agilen Entscheidungsfindung / Führung (SWR/SR, 05.21 bis 06.21)*
2. *Projekt „IDA-Bot“, intuitive automatisierte Nutzung des ARD-Pressearchivs auf Presse-Websiten zur Freischaltung zugangsbeschränkter Artikel, (SWR/SR, 05.21 bis 08.21)*
3. *Projekt „Neue IDA-Dienstleistung: Social-Media-Monitoring-Tool als Themen-Datenbank für Podcast Chaos2 / (SWR3)” (SWR, 05.21 bis 06.21, jetzt: „Lebenslänglich – der Mama-Podcast“)*
4. *Projekt “Social-Media-Monitoring-Tool zur Themenfindung für Podcast PancakePolitics (DasDing)” (SWR, 07.21)*
5. *Projekt “Neue IDA-Dienstleistung: Social Seeding-Pilot für ARD/SWR Format ‘Down the Road” (SWR Programmplanung und -entwicklung, 04.22 bis 05.22)*

*Projekt Erzählcafé: Wie war das damals bei der IDA? Vorbereitung und Moderation für das Austauschformat zwischen älteren und jüngeren Kolleg\*innen über Ihre damalige Arbeit bei der IDA*

Herr Timo Schumacher hat im Rahmen des Volontariats an folgenden Fachkonferenzen teilgenommen:

* VFM-Frühjahrstagung als Teilnehmer, 26.04.-28.04.2021
* Cutting Edge Symposium als Teilnehmer, 10.11.2021
* VFM- Frühjahrstagung als Teilnehmer, 25.04-27.04.2022

Herr Timo Schumacher hat während des Volontariats an folgenden fachlich relevanten Weiterbildungen teilgenommen und Projekte durchgeführt:

* Seminar Office365, 12.02.2021
* Schulung: Wikidate/Wikipedia, Browsererweiterungen, Suchmaschinenoperatoren, Webarchivierung, Verifikation, 17.02.-24.02.2021
* Agiles Basistraining, 29.04.2021
* Seminar: Feedbackkultur, 05.03.2021, 07.03.2021 und 17.05.2021
* Schulung: Excel, Datawrapper, Datenjournalismus, SEO-Optimierung, CMS Sophora, Monitoring und Kuratierung, 21.09.-01.10.2021
* Presserecht und Urheberrechtsseminar, 04.10.2021 und 06.10.2021
* Projekt „IDA-Kartenspiel“, Design-Thinking-Projekt für Tool zur agilen Entscheidungsfindung / Führung, 05.21 bis 06.21
* Faktenchecking in Telegram (dpa-Faktencheck), 15.03.2022
* Projekt „IDA-Bot“, intuitive automatisierte Nutzung des ARD-Pressearchivs auf Presse-Websiten, 05.21 bis 08.21
* Projekt „Social-Media-Monitoring-Tool als Themen-Datenbank für Podcast Chaos2” (SWR3), 05.21 bis 06.21
* Projekt “Social-Media-Monitoring-Tool zur Themenfindung für Podcast PancakePolitics” (DasDing), 07.21
* Follow-Up-Projekt „IDA-Bot, Version 2“, neue Features, 04.22 bis 05.22
* Projekt “Social Seeding für neues ARD/SWR Format ‘Down The Road’”, 04.22 bis 05.22