Zusammenfassung / 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.