Die Projektarbeit

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

**Exposee**

**IST-Problem/-Aufgabenstellung**

**Projektaufgabe (Synopsis)**

Spezifikation, Entwicklung und Implementierung eines KI-basierten Services zur Bestimmung der Präsentationsformen von Texten für die Presserecherche.

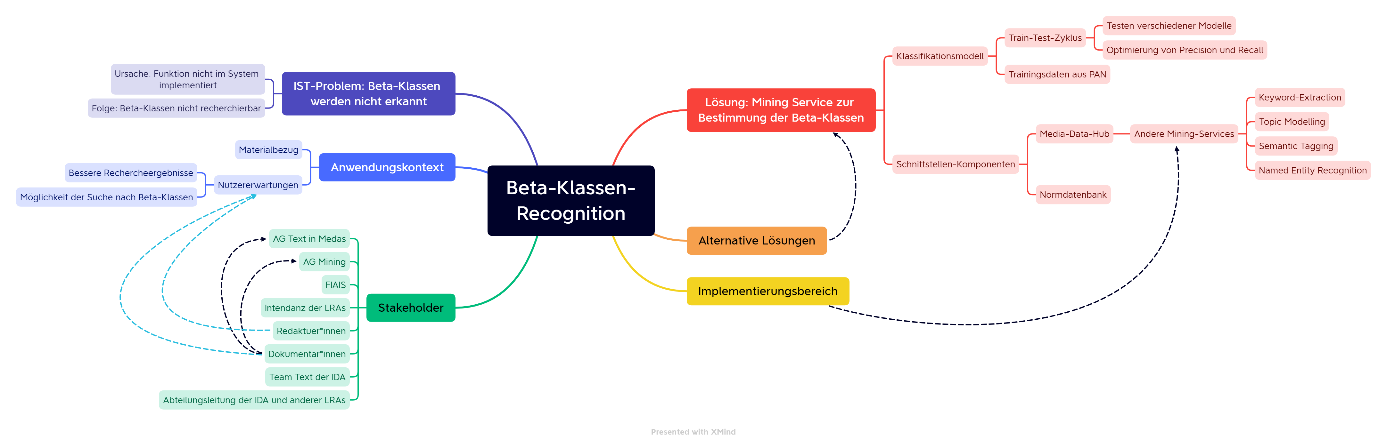
**IST-Problem**

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. Die Präsentationsform, eine formale bzw. Beta-Klasse erhält als Default-Einstellung immer die Klasse „Bericht“. Sie bildet mit über 3.217.851 Treffern (Stand 25.01.2022) rund 71,5 Prozent der Treffermenge, während alle anderen Präsentationsformen in der Pressedatenbank (29 Formen mit kumulierten 1.281.310 Treffern, Stand 25.01.2022) auf rund 28,5 Prozent kommen. Diese 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 im letzten Quartal 2022 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. Dieses stammt von der Firma Picturesafe. Da die Firma seit Jahren insolvent ist, erfolgen keine Updates mehr. In der zukünftigen Crossmedialen Suche in Medas kommen diverse Mining-Services zum Einsatz. Diese konzentrieren sich auf verschiedene Aufgaben wie *Keyword Extraction, Semantic Tagging, Topic Modelling* oder *Named Entity Recognition (NER)*. Für die einzelnen Mining-Services wird dabei auf verschiedene Technologien zurückgegriffen, die in Zusammenarbeit mit dem Frauenhofer-Institut für Intelligente Analyse- und Informationssysteme (FIAIS) entwickelt werden. Das Semantic Tagging wird durch das neuronale Netzwerk [*Starspace*](https://github.com/facebookresearch/StarSpace) von *facebookresearch* durchgeführt. Die Keyword Extraction nutzt die sogenannte *Term Frequency - Inverse Document Frequency* (*TF-IDF*) als Grundlage.Für die NER greift der Service auf die in [Standford entwickelte](https://nlp.stanford.edu/software/CRF-NER.html) und im [DKPro](https://dkpro.github.io/) implementierte Software zurück. Das Topic Modelling basiert auf der in [Mallet](https://mimno.github.io/Mallet/topics) verwendeten Implementierung des LDA-Algorithmus (Latent Dirichlet Allocation). Die Beta-Klassen-Recognition soll neben diesen bereits entwickelten Services die Textmining-Pipeline ergänzen.

**Lösungsansatz**

Entwicklung eines KI-basierten Services für die Mining-Plattform des Crossmedialen Mediendatensystems (Medas) zur Auszeichnung der Präsentationsformen (Beta-Klasse/Gattung) von Pressetexten. 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 auf NDB-konforme Präsentationsformen gemappt werden können. Im Produktivsystem können später weitere Daten zum Testen und Optimieren herangezogen werden. Die Modellanforderungen sind bislang noch nicht spezifiziert, da beim Einsatz von KI bzw. Machine Learning in den seltensten Fällen hun­dert­pro­zen­tige Genauigkeit erreicht werden kann. Das PoC wird am Ende der Testphase darüber entscheiden, welches Modell am geeignetsten ist.

**Mind Map**

******  
**Abb. 1: Mind Map** (eigene Darstellung)

**Auftraggeber, Verwender und Implementierungsbereich**

Auftraggeber:

* Nutzer\*innen (Redakteur\*innen, Dokumentar\*innen) der zukünftigen Textrecherche in Medas
* Implementiert als Service in die Mining-Plattform der Crossmedialen Suche in Medas

Implementierungsbereich:

* Nutzer\*innen (Redakteur\*innen, Dokumentar\*innen) der zukünftigen Textrecherche in Medas
* Implementiert als Service in die Mining-Plattform der Crossmedialen Suche in Medas

**Zielsetzung/vereinbarte Ergebnistypen**

**Zielsetzungen**

Das Projekt soll im Rahmen eines PoCs bestimmen, welche KI-Technologie für die Bestimmung der Präsentationsformen von Pressetexten geeignet ist. Nach der Testung von den vorausgewählten Technologien soll die am besten geeignete in die Mining-Plattform implementiert werden. Dort bietet der Service dann einen nachhaltigen Mehrwert für die Presserecherche, indem er die Suche nach Präsentationsformen ermöglicht.

**Vereinbarte Ergebnistypen**

* 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

**Erwarteter Mehrwert**

* Automatisierte Erschließung von weiteren Metadaten.
* Presseartikel werden durch bessere technisch-formale Metadaten schneller im Archiv recherchierbar und nutzbar.
* Geld- und Ressourcenersparnis, wenn automatisierbare Aufgaben, nicht mehr von hochqualifiziertem Personal erfüllt werden müssen
* Entwickeltes Modell kann möglicherweise für die automatisierte Erkennung weiterer Beta-Klassen genutzt werden.

**Erkenntnisse aus dem Requirements Management**

**Erkannte implizite Anforderungen (Auftraggeber-Grunderwartungen)**

* Zeit- und Kostenersparnis
* Automatisierung eines Arbeitsprozesses in der Pressedokumentation
* Bessere und vollständigere Metadaten von Presseartikeln
* Eigenverantwortliche Entwicklung des Modells

**Explizite Anforderungen (allgemein)**

* Kundenservice erhöhen

**Anforderungen je vereinbartem Ergebnistypen**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ET-Nr.** | **vereinbarter Ergebnistyp** | **Anf.-Nr.** | **Anf.-Typ** | **Anforderung** |
| 1 | Modell-Prototypentwicklung | 1 | implizit,  explizit | Zeit- und Kostenersparnis,  Kundenservice erhöhen,  Entwicklung eines Modell-Prototypen,  Eigenverantwortliche Entwicklung |
| 2 | Modell-Einbettung in Mining-Plattform | 2 | implizit, explizit | Zeit- und Kostenersparnis,  Kundenservice erhöhen,  Arbeitsprozessautomatisierung,  bessere Metadaten |
| 3 | Tests verschiedener KI-Modelle | 3 | explizit | Identifikation des am besten geeigneten Sprachmodells,  Eigenverantwortliche Entwicklung |
| 4 | Erstellung einer Spezifikation des KI-Modells | 4 | explizit | Erstellung einer Spezifikation des KI-Modells |
| 5 | Ausarbeitung eines PoCs | 5 | explizit | Ausarbeitung eines PoCs |

**Nachbarprojekte/Vorläuferprojekte**

Dieses Projekt entsteht in enger Zusammenarbeit mit den Projektverantwortlichen von *Text in Medas*, welche federführend für die Entwicklung der anderen Textmining-Services sind.  
Im diesjährigen Dieburg-Jahrgang beschäftigen sich gleich mehrere Arbeiten mit KI-basierten Verfahren. Zum Beispiel Jana Gierdens (SWR) Projekt zur KI-basierten Optimierung der Keyword-Extraction bei Wort-Beträgen oder die Entwicklung einer Testumgebung für das Testen von KI-Services von Kristoph Sachsenweger (SWR). 2018 nutzte Jan Max Milbradt (FIZ Karlsruhe) ähnliche Textmining-Methoden, aber auf andere Datengrundlage (Patentinformationen).

**Relevante (informationswissenschaftliche) Grundkonzepte**

* Agile Softwareentwicklungsmethoden und Techniken
* Qualitätsmanagement- und sicherung des Modells und der entstehenden Metadaten
* Requirementsmanagement zur Bestimmung der Anforderung an das Projekt
* Metadaten und Metadatenstandards
* KI-Verfahren insbesondere aus dem Bereich der natürlichen Sprachverarbeitung und Klassifikation
* Prozessmanagement
* Vektorraum- und semantische Modelle

**Relevante Schnittstellen**

**Organisatorisch**

* Projektgruppe Text in Medas, die die Realisierung einer gemeinsamen Presseanwendung der Rundfunkanstalten im MDH (Media Data Hub) zum Ziel hat
* EG Mining (WDR), die für die Qualitätssicherung der Mining-Services verantwortlich ist
* IT-Abteilung des SWR oder externer Kooperationspartner wie das FIAIS für die Implementierung in Medas

**Technisch**

* Schnittstelle zur NDB für das Mapping der PAN-Gattungen
* Schnittstelle zur Mining-Plattform für die Qualitätssicherung und Re-Trainings

**Stakeholder**

****

**Abb. 2: Influence/Interest - Matrix** (eigene Darstellung)

**Risk-Analyse**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Eintrittswahrscheinlichkeit** | häufig |  |  |  |  |
| wahrscheinlich |  | Kommunikationsprobleme | Probleme bei der Implementierung der Software allgemein an unterschiedlichen Stellen (KI-Modell, Schnittstelle zur Datenbank) |  |
| gelegentlich |  | Kein Interesse an der Entwicklung bei IDA |  | Fehlende Kapazitäten auf Entwicklerseite, um Miningservice zu implementieren |
| vorstellbar |  | Geringe Qualität der Trainingsdaten, die Datenbereinigung erfordern | Kritik an Ergebnissen des Modelltrainings (zu geringe Trefferquote) |  |
| unwahrscheinlich |  |  | Kein Interesse am fertigen Miningservice-Konzept unter den Medas-Verantwortlichen | Ausfalls der Datenbank / Schnittstellen  Einstellung der Entwicklung am Service |
| unvorstellbar |  |  |  |  |
|  | | unwesentlich | gering | kritisch | katastrophal |
| **Erwartetes Schadensausmaß** | | | |

**Abb. 3: Riskanalyse** (eigene Darstellung)

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Externe Faktoren** | | |
| **Interne Faktoren** |  | **O**pportunities (Chancen):   * Kostenersparnis durch Automatisierung * Automatische Erschließung von zusätzlichen Metadaten / Verbesserung der Metadatenqualität | **T**hreats (Risiken):   * Datenqualität kann vereinzelnd durch Automatisierung sinken * Technische Probleme von Software, Schnittstellen, Datenbanken |
| **S**trengths (Stärken):   * Know-how vorhanden * Mächtige IT-Struktur * Innovationswille vorhanden | Ausbauen:   * Fachwissen über Datengrundlage nutzen, um den Automatisierungsprozess optimal zu gestalten | Absichern:   * Know-how und IT-Struktur nutzen, um Qualitätsverlust gering zu halten und technische Probleme schnell zu lösen |
| **W**eaknesses (Schwächen):   * Historisch gewachsene Strukturen erschweren Projektmanagement und Entwicklung | Aufholen:   * Kommunikation mit Stakeholdern ausbauen | Meiden:   * Stilllegung des Projektes durch fehlende Kapazitäten * Ausfälle von Software, Schnittstellen oder Datenbanken |

**Abb. 4: SWOT-Matrix** (eigene Darstellung)

**Vorgegebenes oder vereinbartes Vorgehensmodell**

Die Entwicklung des Modells erfolgt nach agilen Softwareentwicklungsmethoden.

**Geplante Vorgehensweise (grob)**

Entwicklung eines Proof of concepts für die automatische Bestimmung der Präsentationsformen. Dazu muss im ersten Schritt ein Testdatensatz aus PAN aggregiert und mit Methoden der natürlichen Sprachverarbeitung vorverarbeitet werden. Anschließend werden die oben erwähnten Modelle in mehreren Iterationen trainiert, deren Leistung evaluiert und optimiert. Geplant ist, dass dieser Prozess Ende Juni 2022 abgeschlossen ist. Im Anschluss daran soll das ausgewählte Modell als Service in MDH:Presse integriert werden. Dazu werden die SWR-IT oder das FIAIS miteinbezogen, die die Entwicklung bei der Einbettung unterstützen.

**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

**Ideenspeicher**