|  |  |  |
| --- | --- | --- |
| Projekt-  dokumentation |  |  |

Abschlussarbeit zur Vollendung des Volontariats   
und zum erfolgreichen Erwerb des Zertifikats zum  
„wissenschaftlicher Dokumentar / Information Specialist“

**Spezifikation und PoC der KI-gestützten**

**„Beta-Klassen-Recognition“ in der Presserecherche**



**Volontärjahrgang 2022**

|  |  |
| --- | --- |
| vorgelegt von: | Timo Schumacher B.A., M.A. |
|  | Turmstr. 89  55120 Mainz |
|  |  |
|  |  |
| Volontariats-/Traineepartner: | SWR |
| Referent: | Prof. Geribert E. Jakob |  |

Zusammenfassung / Abstract

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.   
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

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

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.  
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.

Analytic Design and Results

…

Synopsis

…

Schlagwörter

…

Vorwort/Persönliche Danksagung

…

Inhaltsverzeichnis

Zusammenfassung / Abstract 1

Vorwort/Persönliche Danksagung 3

1 Einleitung 1

2 Projektbeschreibung und Einordnung 2

2.1 Zielsetzung und projektbezogene Ergebnistypen 2

2.2 Stakeholderanalyse 2

2.3 SWOT-Analyse 2

3 Grundlagen und Forschungsstand 3

3.1 Natürliche Sprachverarbeitung 3

3.2 Vorprozessierung 3

3.2.1 Segmentierung und Tokenisierung 3

3.2.2 Stammformreduktion und Lemmatisierung 3

3.2.3 Stoppwortentfernung 3

3.2.4 Part-of-speech Tagging 3

3.3 Text als Zahlen 3

3.4 Textklassifkation 3

3.4.1 Machine Learning zur Textklassifikation 3

3.4.2 Logistische Regression 3

3.4.3 Support Vektor Machine 3

3.4.4 BERT / Deep Learning 3

3.4.5 Evaluation von Modellen 3

3.4.6 Warum Python 3

4 Datengrundlage 4

4.1 Struktur 4

4.2 Statistiken 4

5 Methodik 5

5.1 Pipeline 5

5.2 Parameter 5

6 Ergebnisse 6

6.1 Proof of Concept 6

7 Schluss 7

7.1 Best practices 7

7.2 Lessons learned 7

7.3 Ausblick 7

8 Projektbeschreibung und Einordnung 8

Dies ist eine Überschrift ohne Zahl – Ebene 1 9

8.1 Dies ist eine Überschrift x.x (z.B. 1.1.) – Ebene 2 9

8.2 Zielsetzung und projektbezogene Ergebnistypen 9

8.3 Konkrete Lieferobjekte 9

8.3.1 Dies ist eine Überschrift n.n.n – Ebene 3 9

8.4 Analyse, Ist-Zustand und Rahmenbedingungen 10

8.5 Brauch ich nicht Vorgehensweise und Methoden zur Bedarfsanalyse 10

9 Analyse- und Gestaltungsteil 11

9.1 Stakeholderanalyse 11

9.2 SWOT-Analyse 11

Quellen 12

Abbildungen 13

Tabellen 14

Abkürzungen und Akronyme 15

Anhang 16

Eidesstattliche Erklärung 17

# Einleitung

Kinderkram haben sie gesagt, ruf doch erstmal die Oma an. Bis hier hin ist alles gut geangen. Fick deine Mutter rap seit 20 Jahren. Kinderkram haben sie gesagt. Schlirum schlarum.

# Projektbeschreibung und Einordnung

## Zielsetzung und projektbezogene Ergebnistypen

## Stakeholderanalyse

Was ist Text in Medas

Was ist MDH?

## SWOT-Analyse

Vor und Nachteile von KI

Der menschliche Faktor

# Grundlagen und Forschungsstand

## Natürliche Sprachverarbeitung

## Vorprozessierung

### Segmentierung und Tokenisierung

### Stammformreduktion und Lemmatisierung

### Stoppwortentfernung

### Part-of-speech Tagging

## Text als Zahlen

Sprachmodelle

Vektorisierung und Gewichtung

## Textklassifkation

### Machine Learning zur Textklassifikation

### Logistische Regression

### Support Vektor Machine

### BERT / Deep Learning

### Evaluation von Modellen

### Warum Python

# Datengrundlage

## Struktur

Korpuserstellung

## Statistiken

# Methodik

## Pipeline

## Parameter

# Ergebnisse

## Proof of Concept

# Schluss

## Best practices

## Lessons learned

## Ausblick

# Projektbeschreibung und Einordnung

Dies ist eine Überschrift ohne Zahl – Ebene 1

Dies ist der Standardtext.

Auch das Inhaltsverzeichnis mit aktualisiert werden. Mit Klick auf „Felder aktualisieren“.

## Dies ist eine Überschrift x.x (z.B. 1.1.) – Ebene 2

So wird eine Beispielgrafik beschriftet (Rechtsklick und „Beschriftung einfügen „ auswählen:



Abbildung - Beispielgrafik - Überschriften Formatvorlagen [Quelle]

So sieht eine Beispieltabelle aus:

Rechtsklick und „Beschriftung einfügen „ auswählen, dann „Tabelle“ auswählen und den Text oben eingeben.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Name | Anzahl | xx | yy | zz | Bemerkung | Status |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

Tabelle – Beispieltabelle [Quelle]

## Zielsetzung und projektbezogene Ergebnistypen

## Konkrete Lieferobjekte

…

### Dies ist eine Überschrift n.n.n – Ebene 3

…

## Analyse, Ist-Zustand und Rahmenbedingungen

…

## Brauch ich nicht Vorgehensweise und Methoden zur Bedarfsanalyse

…

# Analyse- und Gestaltungsteil

…

## Stakeholderanalyse

…

## SWOT-Analyse

…

Quellen

Monografien

[M\_....] …

Sammelwerke

[S\_....]

Aufsätze

[A\_....] …

Elektronische Dokumente (aus gesicherten Quellen)

[E\_....]

Sonstige Quellen

[X\_....] …

Abbildungen

[Abbildung 1 - Beispielgrafik - Überschriften Formatvorlagen 1](#_Toc45521115)

Tabellen

[Tabelle 1 - Beispieltabelle 1](#_Toc45521377)

*Mit Klick auf „Felder aktualisieren“ aktualisiert sich die Ansicht entsprechend der Angaben weiter oben .*

Abkürzungen und Akronyme

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ABD | Archiv Bibliothek Dokumentation |  | MINT | Medien Integrierende Netzwerk Technologie (CMS zur Organisation aktueller Produktionsmaterialien) |
| BKN | Bürokommunikationsnetz |  | NiF | Nachricht im Film |
| CF | Clean Feed |  | PN | Produktionsnetz |
| DAS | Digitales Archivsystem |  | Prod.nr. | Produktionsnummer |
| EDL | Edit Decision List |  | RMZE | Rechtemanagement und Zentraleinkauf |
| FCC | File Conversion Center |  | RUKAS | Rechte- und Kostenauskunftssystem |
| FSDB | Fernsehdatenbank |  | SLA | Service-Level-Agreement |
| GF | Geschäftsfeld |  | ST | Sendeton |
| HiRes | High Resolution – Material in hochauflösender Qualität (Sende-Qualität) |  | TC | Timecode |
| IT | Internationaler Ton |  | USP | Unique Selling Proposition |
| LoRes | Low Resolution – Material mit geringer Auflösung (Ansichtsqualität) |  | vfm | Verein für Medieninformation und Mediendokumentare |
|  |  |  | Vgl. | Vergleiche |
|  |  |  | z.B. | Zum Beispiel |

Anhang

**…**

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig erstellt und keine anderen als die angegebenen Hilfsmittel und Quellen verwendet habe.

Soweit ich auf fremde Materialien, Texte oder Gedankengänge zurückgegriffen habe, enthalten meine Ausführungen vollständige und eindeutige Verweise auf die Urheber und Quellen.

Alle weiteren Inhalte der vorgelegten Arbeit stammen von mir im urheberrechtlichen Sinn, soweit keine Verweise und Zitate erfolgen.

Mir ist bekannt, dass ein Täuschungsversuch vorliegt, wenn die vorstehende Erklärung sich als unrichtig erweist.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_   
**Ort, Datum Unterschrift**