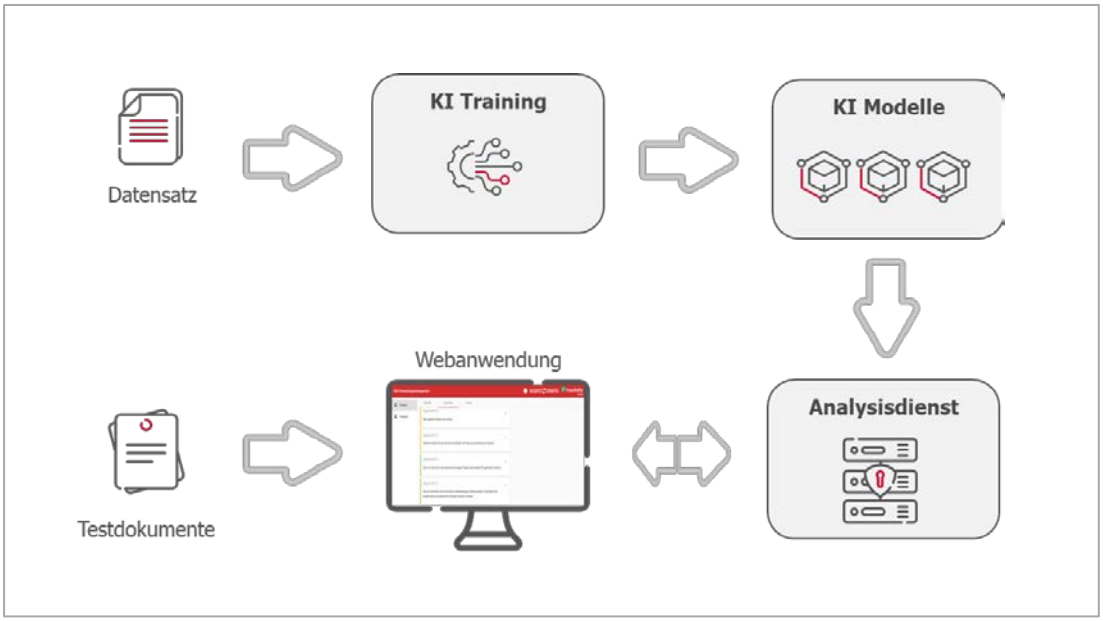
Deutsches Zentrum für
Schienenverkehrsforschung beim
Eisenbahn-Bundesamt

Berichte
des Deutschen Zentrums
für Schienenverkehrsforschung

Bericht 22 (2022)

Pilotstudie für ein KI-gestütztes Einwendungsmanagement



Berichte des Deutschen Zentrums
für Schienenverkehrsforschung, Nr. 22 (2022)
Projektnummer 2021-31-S-1202

Pilotstudie für ein KI-gestütztes Einwendungsmanagement

von

Dr. Jochen Harding, Dr. Wael Alkhatib, Benjamin Fadl
Sopra Steria SE, Hamburg

unter Mitwirkung von

Silvio Peikert, Thilo Ernst,
Fraunhofer-Gesellschaft e.V., Institut FOKUS, München

Im Auftrag des Deutschen Zentrums für Schienenverkehrsforschung beim Eisenbahn-Bundesamt

Impressum

HERAUSGEBER
Deutsches Zentrum für Schienenverkehrsforschung beim Eisenbahn-Bundesamt

August-Bebel-Straße 10
01219 Dresden

www.dzsf.bund.de

DURCHFÜHRUNG DER STUDIE
Sopra Steria SE
Hans-Henny-Jahnn-Weg 29
22085 Hamburg

mit dem Nachunternehmer
Fraunhofer-Gesellschaft e.V., Institut FOKUS
Hansastr. 27c
80686 München

ABSCHLUSS DER STUDIE
Dezember 2021

REDAKTION
DZSF
Dr. Thomas Buder, Forschungsbereich Sicherheit

BILDNACHWEIS
DZSF

PUBLIKATION ALS PDF
https://www.dzsf.bund.de/Forschungsergebnisse/Forschungsberichte

ISSN 2629-7973

doi: 10.48755/dzsf.220006.01

Dresden, März 2022

Inhaltsverzeichnis

Kurzbeschreibung

Im Rahmen des gesetzlich angeordneten Planfeststellungsprozesses haben Betroffene sowie Träger öffentlicher Belange die Möglichkeit, Einwendungen zu erheben und Stellungnahmen abzugeben. Dabei fallen große Mengen strukturell heterogener Daten in Textform an, deren fristgerechte Bearbeitung einen ressourcenintensiven Prozess bezüglich Personal, Zeit und Kosten darstellt. Im Rahmen der vorliegenden Studie wurde untersucht, inwieweit der Prozess des Einwendungsmanagements durch künstliche Intelligenz (KI)-basierte Verfahren für die maschinelle Textverarbeitung (MTV) in Teilbereichen automatisiert werden kann. Ziel ist es, die Mitarbeitenden zu entlasten und die Effizienz zu steigern. Für diesen Zweck wurden in der Studie grundsätzlich geeignete MTV-Verfahren recherchiert und Lösungsansätze entwickelt. Für die Lösungsansätze wurden unter Verwendung eines konkreten Beispieldatensatzes Experimente durchgeführt. Dabei erfolgte inhaltlich eine Fokussierung auf vier konkret vom Auftraggeber vorgegebene Aufgaben. Die Ergebnisse der Untersuchung lassen sich wie folgt zusammenfassen:

- bis zu 85 % der argumentativen Segmente können automatisch korrekt gekennzeichnet werden
- Vorschläge zur Klassifizierung von Einwendungen hinsichtlich eines vorgegebenen Themenkatalogs lassen sich mit hoher Genauigkeit von z. T. mehr als 80 % ableiten
- Musterschreiben und ähnliche Argumente lassen sich mit etablierten Automatismen effizient identifizieren
- Plausibilitätsprüfungen auf der Basis textueller Übereinstimmung liefern belastbare Indikatoren für potenzielle Unstimmigkeiten in Gegenüberstellungen

Darüber hinaus wurde ein Demonstrator entwickelt, mit dem im Sinne eines Proof-of-Concept (PoC) dargelegt wird, in welcher Form sich die Segmenterkennung und die Klassifizierung nach Themen in den Fachprozess integrieren lassen. So können in der entwickelten Web-Anwendung vorgegebene Testdokumente oder frei editierbare Texte unter Verwendung der im Rahmen der Studie vortrainierten KI-Modelle analysiert und die Ergebnisse der Analyse angezeigt und nachbearbeitet werden.

Insgesamt zeigen die Studienergebnisse, dass KI-basierte Lernverfahren geeignet sind, um wesentliche Aufgaben im Prozess des Einwendungsmanagements in Teilen zu automatisieren. Dabei ist eine Umsetzung auf Grundlage von Open-Source-Bausteinen und ohne Abhängigkeiten von externen Diensten oder Cloud-Anwendungen problemlos möglich, so dass auch Anforderungen hinsichtlich Datenschutz und Datensouveränität erfüllt werden. Die Teilautomatisierung liefert Vorschläge, die den Mitarbeitenden in der Anwendung zur Validierung präsentiert werden und somit zu einer deutlichen Entlastung beitragen. Die Entscheidungshoheit verbleibt dabei stets bei den Mitarbeitenden, wodurch Transparenz und Nachvollziehbarkeit gewahrt bleiben. Die erreichbaren Genauigkeiten sind ausreichend hoch, um MTV-Verfahren bereits heute als integralen Baustein in dem ganzheitlichen Prozess des Einwendungsmanagements umzusetzen. Somit kann die Studie als Entscheidungsgrundlage für Umsetzungsentscheidungen herangezogen werden. Folgende Handlungsempfehlungen zur Überführung in die Praxis lassen sich ableiten:

- Verwendung eines einheitlichen Themenkatalogs und Sammlung von Trainingsdaten bei der aktuellen (manuellen) Bearbeitung, um eine ausreichend große Datenbasis für das initiale Training der KI-Modelle und die spätere Umsetzung zu haben
- Umsetzung einer intuitiv bedienbaren Benutzerschnittstelle, die eine schnelle Erfassung der Vorschlagsgüte ermöglicht und komfortable Funktionen zur manuellen Validierung bereitstellt
- Nutzung der Informationen aus der manuellen Validierung der automatisch generierten Vorschläge als Eingangsgrößen für die fortlaufende Verbesserung der KI-Modelle
- Berücksichtigung von KI-Modellen bereits bei der initialen Konzeption von Prozessen und unterstützenden Systemen und entsprechende Einbettung in die IT- und Prozesslandschaft
- frühzeitige und proaktive Begleitung der Einführung im Sinne eines Change-Managements, um etwaige Vorbehalte gegenüber KI frühzeitig entkräften und Akzeptanz gewinnen zu können.

1 Einleitung

Das Eisenbahn-Bundesamt (EBA) ist die zuständige Planfeststellungsbehörde für Eisenbahnen des Bundes. Ein Bestandteil des Planfeststellungsprozesses ist die Erhebung von Einwendungen durch Betroffene sowie die Abgabe von Stellungnahmen durch Träger öffentlicher Belange, die Teile von eingehenden Schreiben in unterschiedlicher Form (Brief, Handschrift, E-Mail) sind.[1] Dabei fallen strukturell heterogene Datenmengen in Textform an, deren Bearbeitung einen ressourcenintensiven Prozess bezüglich Personal, Zeit und Kosten darstellt.

Der Arbeitsumfang durch Planfeststellungsprozesse wird perspektivisch noch umfangreicher, da das EBA seit Dezember 2020 zusätzlich auch für die Anhörungen in den Planfeststellungsverfahren zuständig ist, für die bisher die Landesbehörden zuständig waren. D. h., dass zusätzlich zu Einwendungen auch die Objekte „Erwiderungen“ zu betrachten sind, die als Antwort auf Einwendungen von den Vorhabenträgern erstellt und seitens des EBA zu validieren und zuzuordnen sind. Beide Textobjekte werden durch „Gegenüberstellungen“ (Synopsen) zueinander in Beziehung gesetzt, um sicher zu stellen, dass alle Einwendungen vollständig und einheitlich beantwortet werden.

Perspektivisch ist es daher von enormem Vorteil, den ganzheitlichen Textdaten-bezogenen Prozess (im Folgenden im Sinne der vorliegenden Studie als “Einwendungsmanagement” bezeichnet) von der Verwaltung über die Analyse und Bewertung bis zur Ergebniskommunikation und die zugrundeliegenden Arbeitsschritte erheblich konsistenter und effizienter zu gestalten, indem der Prozess durch eine zu entwickelnde Software unterstützt wird. Hierbei ist aufgrund der speziellen Gestalt der Datenstruktur zunächst nicht klar, inwieweit etablierte Methoden der maschinellen Textverarbeitung (im Folgenden als MTV-Verfahren bezeichnet) geeignet sind, manuelle Prozesse hinreichend zuverlässig zu automatisieren.

Im Rahmen der vorliegenden Studie soll eine wissenschaftlich fundierte Grundlage für eine praxistaugliche Automatisierung des Fachprozesses „Einwendungsmanagement“ unter Nutzung von KI erarbeitet werden. Dabei wird der Fokus auf die Kernelemente des Prozesses und spezifische Aufgaben gelegt, bei denen die Automatisierung nach initialer Einschätzung Mitarbeitende entlasten und die Effizienz erhöhen kann.

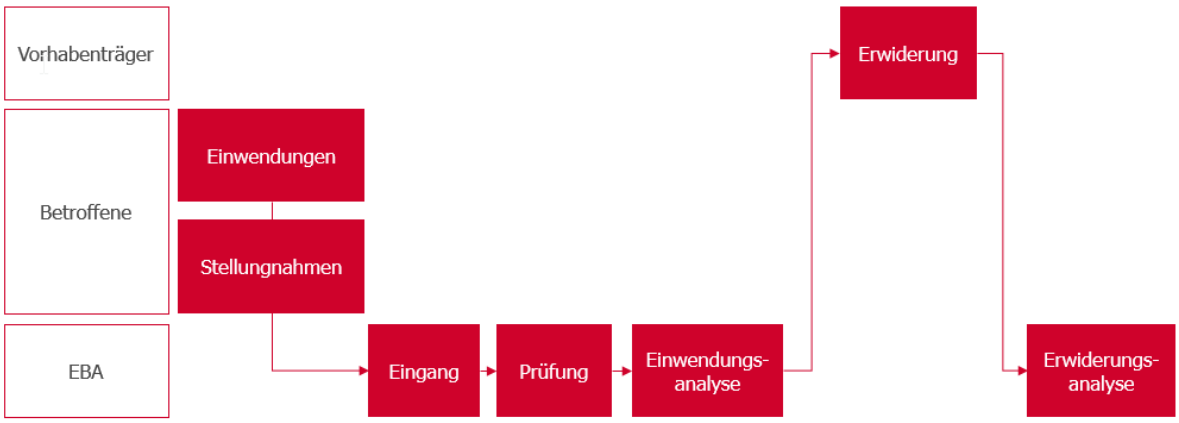


Abbildung 1: Fachprozess „Einwendungsmanagement“ – vereinfachte Darstellung

[1] Im Sinne der sprachlichen Vereinfachung werden Einwendungen und Stellungnahmen im Folgenden unter dem Begriff „Einwendungen“ subsummiert. Nur dort, wo explizit auf Unterschiede in der Verarbeitung eingegangen wird, erfolgt eine sprachliche Differenzierung.

2 Projektvorgehen

Das Vorgehen im Rahmen der vorliegenden Studie orientiert sich am „Cross Industry Standard Process for Data Mining“ (CRISP-DM) Industriestandard, der sich für datenzentrische Projekte bewährt hat[2]. Dabei erfolgte die Bearbeitung grundsätzlich in drei Schritten (Arbeitspaketen, kurz AP):

- AP 1: Datenaufbereitung und -annotation
- AP 2: Bewertung geeigneter MTV-Verfahren
- AP 3: Anwendung und Potentialabschätzung

Abbildung 2 verdeutlicht, wie sich die Vorgehensschritte (Arbeitspakete) in die Phasen des CRISP-DM Vorgehensmodells einordnen. Die Inhalte der einzelnen Arbeitspakete sind nachfolgend zusammengefasst.

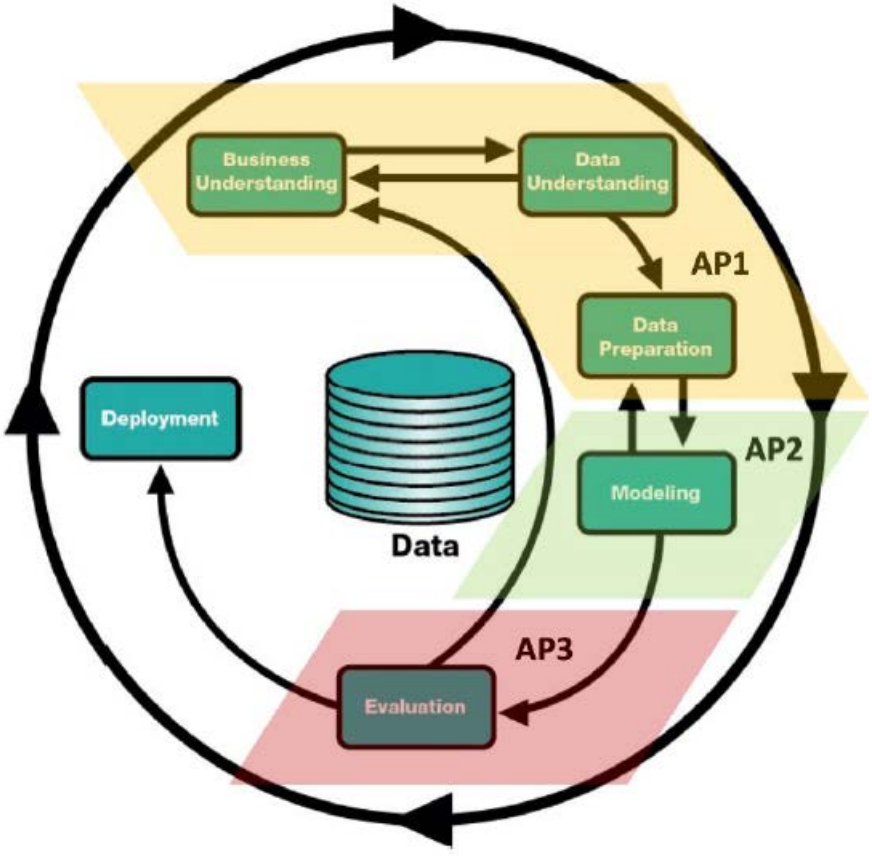


Abbildung 2: Vorgehensmodell in Anlehnung an CRISP-DM

[2] Siehe z. B.: A Systematic Literature Review on Applying CRISP-DM Process Model, Abrufdatum 07.01.2022, https://www.sciencedirect.com/science/article/pii/S1877050921002416

2.1 AP 1 – Datenaufbereitung und -annotation

Ziel des ersten Arbeitspakets ist es, für die nachfolgenden Arbeitspakete einen harmonisierten Datensatz zur Verfügung zu stellen. Für diesen Zweck wurden die folgenden drei Phasen des CRISP-DM Vorgehensmodells durchlaufen:

- **„Business Understanding“**: Initial galt es, ein Verständnis für den fachlichen Prozess zu schaffen und davon ausgehend, die in der Studie zu betrachtenden Aufgaben zu definieren und die relevanten Dokumenten- und Datentypen zu identifizieren. Zur Schärfung des Verständnisses wurde zu Projektbeginn ein Arbeitstreffen mit Vertretern der Fachabteilung des EBA durchgeführt. Im Ergebnis liegt eine Definition der im Rahmen der Studie zu betrachtenden Aufgaben vor (apitel 3).
- **„Data Understanding“**: In der zweiten Phase wurden die vorliegenden Daten (Beispieldatensatz, siehe Kapitel 4) systematisch untersucht und ein gemeinsames Datenverständnis erarbeitet. Dies erfolgte auf Basis der Ergebnisse des „Business Understandings“, dessen Ergebnisse in weiteren Arbeitstreffen validiert und punktuell nachgeschärft wurden. Ausgehend von dem Datenverständnis wurde ein harmonisiertes Datenformat (logisches Datenmodell) und dessen Abbildung in einem physischen Datenmodell entworfen.
- **„Data Preparation“**: In der dritten Phase erfolgte das methodische Überführen der Daten in das harmonisierte Format in Form eines ETL-Prozesses (Extract, Transform und Load). Für Details zur Datenaufbereitung siehe Kapitel 5.

2.2 AP 2 – Bewertung geeigneter MTV-Verfahren

Das zweite Arbeitspaket fokussiert sich in Bezug auf das CRISP-DM Vorgehensmodell auf das Modelling. Ziel des zweiten Arbeitspakets ist es, ausgehend von der Analyse der Datengrundlage geeignete MTV-Verfahren für die Automatisierung der zu betrachtenden Aufgaben zu identifizieren. Für den Zweck wurden grundsätzlich für die Bewältigung der Aufgaben in Betracht kommende Verfahren analysiert und systematisch hinsichtlich ihrer Eignung bewertet (siehe Kapitel 6).

2.3 AP 3 – Anwendung und Potentialabschätzung

Das dritte Arbeitspaket umfasst die Realisierung des für die verfügbaren Daten und das avisierte Anwendungsszenario bestgeeigneten MTV-Verfahrens in Form eines PoC sowie dessen Evaluierung (Phase des CRISP-DM Vorgehensmodells). Die Ergebnisse dieses Arbeitspakets sind in Kapitel 7 zusammengefasst. Ziel ist es, die Güte und Performanz des Verfahrens und der Umsetzung zu untersuchen sowie belastbare Erkenntnisse hinsichtlich des Potentials für eine spätere Umsetzung im Einwendungsmanagement zu erlangen.

Im anschließenden Ausblick (Kapitel 8) werden Vorschläge für das weitere Vorgehen zur Umsetzung eines mit maschinellen Lernverfahren unterstützten Einwendungsmanagements im EBA dargestellt und Optionen zur Weiterentwicklung unter Einbeziehung weiterer fachlicher Aufgaben aufgezeigt, um einem späteren Praxiseinsatz noch besser gerecht zu werden.

3 Betrachtete Aufgaben im Kontext des Fachprozesses

Im Sinne einer Fokussierung wurden in Abstimmung mit dem Auftraggeber vier spezifische Aufgaben definiert, für die Möglichkeiten der Automatisierung untersucht werden. Dabei handelt es sich um drei Aufgaben im Zusammenhang mit der maschinellen Verarbeitung von Einwendungen sowie eine Aufgabe im Zusammenhang mit der Erwiderungsanalyse, bei der es vornehmlich um einen Abgleich zwischen Einwendung und Erwiderung in den Gegenüberstellungen geht.

Abbildung 2 illustriert die betrachteten Aufgaben im Kontext des Fachprozesses. Die Definition der einzelnen Aufgaben und die damit einhergehenden Festlegungen für die weitere Bearbeitung im Rahmen der Studie sind nachfolgend zusammengefasst.

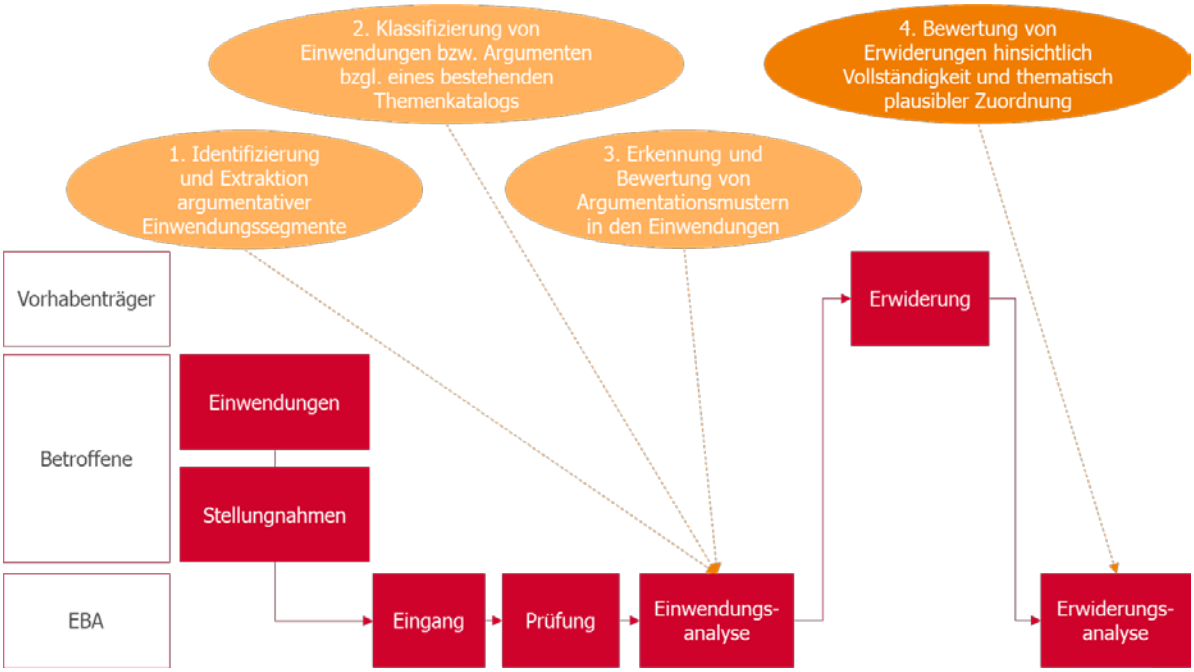


Abbildung 3: Betrachtete Aufgaben im Kontext des Fachprozesses

3.1 Identifizierung und Extraktion argumentativer Einwendungssegmente

Die Aufgabe besteht darin, die Texte von Einwendungen und Stellungnahmen zu zerschneiden und argumentative Segmente zu identifizieren. Bei der Bewertung von Segmenten wird im Rahmen der Studie wie folgt unterschieden:

- nicht bearbeitungsrelevant:
  Segmente, die ausschließlich Metadaten (z. B. Anschrift, referenziertes Planfeststellungsverfahren etc.), Begrüßungs- oder Verabschiedungsformeln etc. enthalten.

- bearbeitungsrelevant:
  - Hinweis: Segmente, die lediglich Informations-/ Abstimmungsforderungen oder die Nennung von Ansprechpartnern umfassen.
  - Argument: Segmente, die eine Antwort des Projektträgers zu dem vorgebrachten Sachverhalt erfordern. Dies sind z. B. ein Ausdruck oder ein Gefühl der Ablehnung bzw. Opposition (Einwendung) oder aber auch explizite Aufforderungen zur Begründung von Maßnahmen.

Abbildung 4 illustriert die Festlegungen zur Segmentierung.

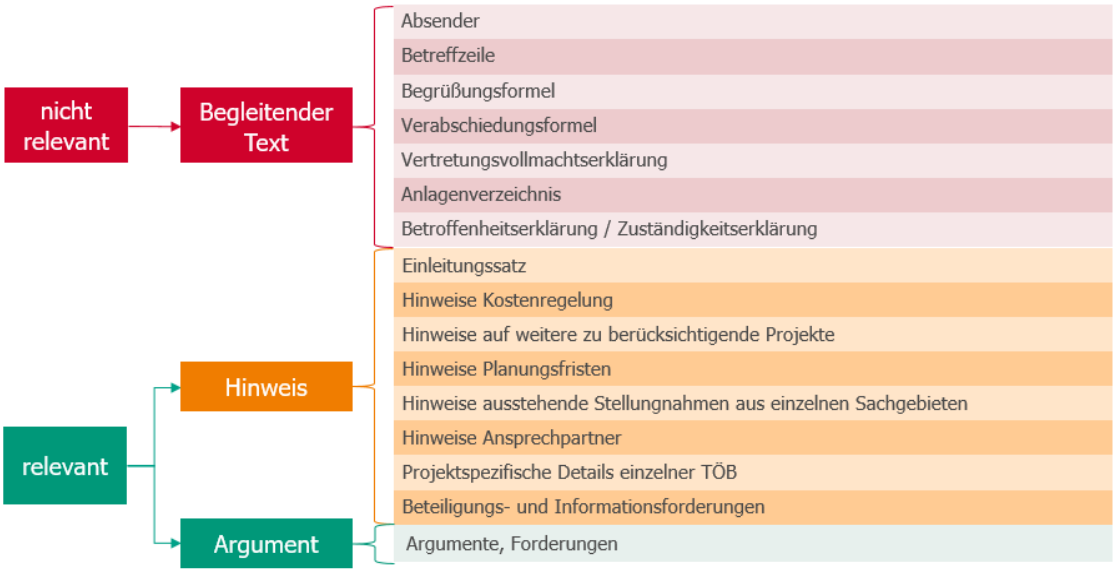


Abbildung 4: Illustration der Segmentbewertung

3.2 Klassifizierung von Einwendungen bzgl. eines Themenkatalogs

Die Aufgabe besteht darin, den bearbeitungsrelevanten Segmenten Themen eines vorgegebenen Katalogs zuzuordnen. Im Zusammenhang mit dem Fachprozess stellt der Themenkatalog insbesondere ein Strukturierungselement für die Gliederung des Planfeststellungsbeschlusses dar.

Während im Fachprozess i. d. R. ein editierbarer Themenkatalog zur Anwendung kommt, wird für die vorliegende Studie ein vorgegebener Themenkatalog verwendet (siehe Anhang 1, Tabelle 13). Dies dient dem Zweck der Reproduzierbarkeit von Ergebnissen.

Der für die Studie vorgegebene Themenkatalog besteht aus Oberthemen mit jeweils zugeordneten Unterthemen. Dies bietet den Vorteil, dass sowohl die Güte der Zuordnung zu übergeordneten Themen als auch zu spezifischen Sachverhalten analysiert werden kann. Für nicht eindeutig zuordenbare Argumente ist jeweils ein Ober- bzw. Unterthema „Sonstiges“ vorgesehen.

Zudem ist in dem vorgegebenen Themenkatalog die Anzahl der Ober- und Unterthemen begrenzt, um für den Zweck der Studie eine möglichst hinreichend große Menge an Beispieldaten pro Ober- bzw. Unterthema vorliegen zu haben.

In Summe umfasst der vorgegebene Themenkatalog 12 Oberthemen sowie 33 Unterthemen.

3.3 Erkennung und Bewertung von Argumentationsmustern in den Einwendungen

Die Erkennung von Argumentationsmustern dient dazu, Cluster ähnlicher oder gleichlautender Einwendungen zu identifizieren. Die Aufgabe besteht darin, inhaltlich ähnlich aufgebaute und strukturierte Einwendungen zu erkennen.

3.4 Bewertung von Erwiderungen hinsichtlich Vollständigkeit und thematisch plausibler Zuordnung

Der erste Teil der Aufgabe besteht darin, zu überprüfen, ob jedem Argument eine Erwiderung des Vorhabenträgers zugeordnet ist. Diese Überprüfung dient dazu, die Einhaltung des formalen Kriteriums der Vollständigkeit im Rahmen des Fachprozesses zu prüfen.

Gegenstand des zweiten Teils der Aufgabe ist der inhaltliche Abgleich zwischen Einwendung und der jeweils zugeordneten Erwiderung (Synopse). Der Abgleich erfolgt in Form einer Bewertung der thematischen Übereinstimmung. Für die Bewertung der Übereinstimmung kommt ein Ampelsystem zur Anwendung:

- grün = inhaltliche Übereinstimmung
- gelb = manuelle Nachprüfung erforderlich
- rot = keine inhaltliche Übereinstimmung erkennbar

4 Datenbasis

Grundlage für die Untersuchung ist ein vom Eisenbahn-Bundesamt in anonymisierter Form zur Verfügung gestellter Beispieldatensatz. Der Datensatz umfasst eine Mischung aus handschriftlichen Einwendungen und digitalisierten Gegenüberstellungen von Einwendungen und Erwiderungen (sogenannten Synopsen) im Word-, PDF- und Excel-Format. Tabelle 1 verdeutlicht die heterogene Struktur der Daten und Unterschiede in Bezug auf den Grad der Vorverarbeitung. Die grundlegenden Charakteristika der Daten lassen sich wie folgt zusammenfassen:

- handschriftliche Dokumente: anonymisierte Scans handschriftlich eingereichter Einwendungen privater Personen ohne weitere Vorverarbeitung
- Sammeleinwendungen: Sammelliste mit Auszügen von Einwendungen, denen jeweils die Erwiderung des Vorhabenträgers zugeordnet ist
- Ursprung gemischt: Sammelliste mit Auszügen von Einwendungen, denen jeweils die Erwiderung des Vorhabenträgers zugeordnet ist; teilweise ergänzt um eine Themenzuordnung sowie die Einschätzungen der Anhörungsbehörde über die Relevanz und Rechtmäßigkeit von Erwiderungen zu den Einwendungen
- Ursprung privat: In Segmente unterteilte Auszüge von Einzelschreiben privater Personen, denen jeweils die Erwiderung des Vorhabenträgers zugeordnet ist
- Ursprung Träger öffentlicher Belange (TöB): In Segmente unterteilte Auszüge von Einzelschreiben mit dem Ursprung TöB, denen jeweils die Erwiderung des Vorhabenträgers zugeordnet ist

TABELLE 1: SPEZIFIKA DER DATENBASIS

		handschriftlich		**digitalisiert**	
Quelle	privat	Sammeleinwendungen	Ursprung gemischt	Ursprung privat	Ursprung TöB
Format	PDF (Scan)	Word	Excel	Word	Word
Anzahl Dokumente	42 Dateien	6 Synopsen	2 Synopsen	19 Synopsen	77 Synopsen
Struktur	Einzelschreiben	Sammelliste	Sammelliste	Einzelschreiben	Einzelschreiben
Segmentierung	nein	ja	ja	ja	ja
Selektion nach Relevanz	nein	ja	ja	ja	ja
Themenzuordnung	nein	nein	teilweise	nein	nein

5 Datenaufbereitung

5.1 Festlegungen für das logische Datenmodell

Ausgehend von der Analyse des fachlichen Prozesses (Business Understanding) und den im Rahmen der vorliegenden Studie zu betrachtenden Aufgaben (vgl. Kapitel 3) sowie dem Abgleich mit den Spezifika der Datenbasis (vgl. Kapitel 4) wurden geeignete Entitäten und Attribute für ein harmonisiertes Datenformat (logisches Datenmodell) identifiziert.

Das logische Datenmodell umfasst drei Entitäten:

1. die Einwendung,
2. die Erwiderung und
3. Gegenüberstellungen.

Jede Entität umfasst mehrere Attribute, welche im Folgenden beschrieben werden.

1. Einwendungen umfassen:

<table>
<tr><th>Attribut</th><th>Beschreibung</th></tr>
<tr><td>Einwendung Nr.</td><td>eindeutige ID</td></tr>
<tr><td>Quelle</td><td>eine Referenz auf das Original</td></tr>
<tr><td>Ursprung</td><td>privat / öffentlich</td></tr>
<tr><td>Datum</td><td>Erstellungsdatum</td></tr>
<tr><td>Volltext</td><td>Referenz auf den extrahierten Text (maschinell lesbare Version des Originals)</td></tr>
<tr><td>Einwendungs-segment</td><td>Jede Einwendung ist in ein oder mehrere Textsegmente unterteilt. Jedes Segment hat mehrere Attribute.
<table>
<tr><td>Segmenttext</td><td>reiner Text eines Absatzes</td></tr>
<tr><td>Segment-Nr.</td><td>Sequenznummer, die die Reihenfolge der Segmente bestimmt</td></tr>
<tr><td>Bearbeitungsrelevanz</td><td>Kennzeichnung, ob das Segment im Einwendungsverfahren berücksichtigt werden sollte</td></tr>
<tr><td>Segmentbewertung</td><td>Zuordnung des Segments zu einer der drei Kategorien [Argument, Hinweis, Begleitender Text]</td></tr>
<tr><td>Themen</td><td>Ein Segment kann ein Thema oder mehrere Themen aus dem Themenkatalog enthalten. Reflektiert die Hauptthemen des Segments.</td></tr>
</table>
</td></tr>
</table>

2. Erwiderungen umfassen:

Attribut	**Beschreibung**
Erwiderung Nr.	eindeutige ID für jede Erwiderung
Quelle	eine Referenz auf das Original
Datum	Erstellungsdatum
Volltext	Referenz auf den extrahierten Text
Erwiderungs-segment	Jede Erwiderung ist in ein oder mehrere Textsegmente unterteilt, die eine Erwiderung auf das entsprechende Einwendungssegment sind. Segmenttext: reiner Text eines Erwiderungssegments Segment-Nr.: Sequenznummer, die die Reihenfolge der Segmente bestimmt

3. Gegenüberstellungen

Attribut	**Beschreibung**
Einwendung Nr.	Referenz für jede Einwendung
Erwiderung Nr.	Referenz für jede Erwiderung
Vollständigkeit	Angabe dazu, ob sich zu einem Segment eine Erwiderung findet
thematische Zuordnung	Analyseergebnisse zur Bewertung der inhaltlichen Zuordnung von Einwendung und Erwiderung Einwendungssegment-Nr.: Einwendungssegment Erwiderungssegment-Nr.: Erwiderungssegment thematische Zuordnung: Inhaltlicher Abgleich zwischen Einwendung und Erwiderung. Die Bewertung erfolgt anhand eines Ampelsystems: Grün = Inhaltliche Übereinstimmung Gelb = Manuelle Nachprüfung benötigt Rot = Keine inhaltliche Übereinstimmung

5.2 Festlegungen für das physische Datenmodell

Aus der gesamtheitlichen Betrachtung der o. g. Entitäten ergaben sich mehrere Anforderungen an das physische Datenmodell:

- Die Originaldokumente liegen in unterschiedlichen Formaten (PDF, MS Office-Formate etc.) vor, weshalb im logischen Datenmodell lediglich auf die Originaldokumente verwiesen wurde. Die aus den Originaldokumenten extrahierten Texte wurden im TXT-Format gespeichert und im Datenmodell hinterlegt.
- Redundanzen bei der Gegenüberstellung von Einwendungen und Erwiderungen müssen ausgeschlossen werden.
- Derzeit nicht abschätzbare Komplikationen späterer Arbeitsschritte erfordern eine hohe Flexibilität und Skalierbarkeit des logischen und, daraus abgeleitet, des physischen Datenmodells.
- Bei der Zusammenarbeit einzelner Teilnehmer des interdisziplinär aufgestellten Teams dieser Studie ist darauf zu achten, dass das gewählte Format universal zugänglich ist und keine kommerziell erhältliche Software benötigt.

Nach Abwägung der Vor- und Nachteile in Frage kommender Alternativen wurde für die vorliegende Studie das CSV-Format ausgewählt. Das CSV-Format erleichtert eine Überführung des Datensatzes z. B. in eine Datenbank oder eine sogenannte Ontologie, sollte dies zu einem späteren Zeitpunkt erforderlich sein.

5.3 Vorgehensweise bei der Überführung in das Datenmodell

Der Prozess der Überführung einzelner Dokumente in das o. g. Datenmodell wurde in Abhängigkeit vom Ursprung der Daten wie folgt umgesetzt:

- handschriftliche Dokumente: Diese Dokumente lagen zunächst eingescannt vor und wurden manuell in Text-Form überführt. Die daraus abgeleiteten Segmente wurden anschließend dem Datenmodell angepasst. Eine maschinelle Überführung unter Verwendung derzeit erhältlicher „Optical Character Recognition“ (OCR)-Software-Lösungen war aufgrund mangelhafter Genauigkeit bei der Erkennung handschriftlicher Texte nicht möglich.
- Sammeleinwendungen und Ursprung gemischt: Die Einwendungen und die entsprechenden Erwiderungen wurden automatisiert in das Datenmodell überführt.
- Ursprung privat und Ursprung Träger öffentlicher Belange (TöB): Die Dokumente wurden manuell geprüft, segmentiert und dann in das Datenmodell überführt.

5.4 Vorgehensweise bei der Segmentierung

Nach der Übertragung der Daten in das Datenmodell wurden die 5.286 einzelnen Segmente manuell annotiert. Hierzu wurde zunächst die Bearbeitungsrelevanz in den Kategorien „relevant“ und „nicht relevant“ festgelegt. Unter „nicht relevant“ fielen Segmente, die keine Antwort seitens des Eisenbahn-Bundesamtes bzw. der Vorhabenträgerin erfordern. Dazu zählen Textfragmente wie Absender, Betreffzeile, Begrüßungsformel, Verabschiedungsformel, Vertretungsvollmachtserklärungen, Anlageverzeichnisse, Betroffenheitserklärungen und Zuständigkeitserklärungen. Diese Fragmente wurden als „begleitender

Text“ spezifiziert. Im Datenmodell sind 596 Segmente als „nicht relevant“ bzw. “begleitender Text“ und 4.690 als „bearbeitungsrelevant“ bewertet.

Im nächsten ebenfalls manuellen Arbeitsschritt wurden die „bearbeitungsrelevanten“ Segmente in Hinweise und Argumente unterteilt.

Hinweise sind u. a. Bemerkungen zu folgenden Aspekten: Kostenreglungen, weitere zu berücksichtigende Projekt-/Planungsfristen, ausstehende Stellungnahmen aus einzelnen Sachgebieten, Ansprechpartner, projektspezifische Details einzelner TÖB, Beteiligungs- und Informationsforderungen. Des Weiteren wurde der Einleitungssatz als Hinweis klassifiziert, wenn dieser ausschließlich eine nachfolgende Einwendung ankündigt, wie „erhebe Einwendungen“, „Zustimmung“, „Ablehnung“.

Ein Argument zeichnet sich durch explizite Nennung von Argumenten oder Forderungen aus. Ein weiteres Indiz für ein Argument waren die Position im Text, die Nummerierung und bei einem kleinen Teil der Segmente die vorrangegangene Bewertung des EBA als Argument.

Insgesamt wurden 1.232 Hinweise und 3.459 Argumente identifiziert.

5.5 Vorgehensweise bei der Themenzuordnung

Im Rahmen der Themenzuordnung wurden zunächst die Segmente zugeordnet, die bereits mit Annotation übergeben wurden. Diese umfassten insgesamt 122 Themenbezeichnungen. Unter Mitwirkung der Fachexperten des EBA wurden die vorgegeben Einträge in den für die Studie verwendeten Themenkatalog einsortiert. 80 Themenbezeichnungen konnten somit eindeutig übernommen werden. Die nicht zuordenbaren Themenbezeichnungen konnten fachlich nicht in den für die vorliegende Studie verwendeten Themenkatalog einsortiert werden und wurden der Kategorie „Sonstiges“ zugeordnet. Die Themen einer vom AG bereitgestellten Datei (Excel Tabelle mit Namen „Data2“) wurden skriptbasiert in das Datenmodell übertragen.

Alle weiteren Segmente mussten manuell in den Themenkatalog einsortiert werden. Um eine hohe Datenqualität sicherzustellen, galt es dabei, die Gütekriterien der Validität, Reliabilität und Objektivität zu wahren. Vor dem Hintergrund wurden in einem nächsten Schritt auf Basis der Daten aus Data2 Schlüsselworte definiert, die als Codierer dienen und eine Hilfestellung bei der Abgrenzung von Themen (z. B. Naturschutz und Umweltschutz) erleichtern, d. h. die Schlüsselwörter dienten als Hilfestellung für die manuelle Annotation der Themen. Mittels dieser Annotationsrichtlinien konnte ein konsistentes Ergebnis der kollaborativen Annotation gewährleistet werden. Perspektivisch können anhand der Schlüsselwörter zu einem späteren Zeitpunkt Themen ggf. auch automatisiert zugeordnet werden.

Die Schlüsselworte wurden anhand einer „Natural Language Processing“ (NLP)-Extraktion aus den bereits annotierten Daten gewonnen: Mittels Analyse der „Term Frequency Inverse Document Frequency“ (TF-IDF)-Vektoren (siehe Abschnitt 6.1.2) dieser Segmente wurden für den Text relevante Worte anhand ihrer Häufigkeit bestimmt. Durch einen vorab trainierten Filter konnten Wörter aussortiert werden, die in allen Segmenten häufig genannt werden, wie z. B. die Vorhabenträgerin. Somit konnten Wörter identifiziert werden, die klar zu erkennen geben, zu welchem Ober- und Unterthemen das Segment gehört.

Die Analyse zeigte, dass bei dem gewählten Vorgehen bereits einzelne Oberthemen nur sehr geringe Fallzahlen aufweisen (2–60 Segmente). Bei derart geringen Fallzahlen ist davon auszugehen, dass die identifizierten Schlüsselwörter keine hohe Aussagekraft aufweisen, da zu wenig Einträge für die Ableitung eines sprachlichen Musters vorliegen.

Vor diesem Hintergrund wurden alle Schlüsselworte durch Begriffe mit semantischer Ähnlichkeit erweitert. Mit Hilfe der frei verfügbaren mehrsprachigen Wissensdatenbank „ConceptNet“ [28] wurden für jedes Thema Synonyme und sachverwandte Worte ergänzt. Bei Themen, die sehr wenige Schlüsselwörter hatten, wurde auch Wikipedia als Anhaltspunkt für die Erweiterung der Schlüsselwörter verwendet. Die resultierende Liste der Schlüsselwörter ist im Anhang 2 angeben.

5.6 Ergebnis der Datenaufbereitung

Im Ergebnis der oben beschriebenen Arbeitsschritte liegt ein einheitlich strukturierter digitaler Datensatz der vorliegenden Einwendungs-, Erwiderungs- und Metadaten in einem für die vorliegende Studie geeigneten Format vor. Details der Ergebnisse zur Segmentierung finden sich in Tabelle 2.

TABELLE 2: ERGEBNISSE DER SEGMENTIERUNG

	Insgesamt	**Handschriftlich**	**digitalisiert**			
Quelle	-	privat	Sammeleinwendungen	Ursprung gemischt	Ursprung privat	Ursprung TöB
Anzahl Einträge / Segmente	5.289	479	974	2.318	351	1.167
begleitender Text	598	252	0	1	52	293
bearbeitungsrelevant	4.691	227	974	2.317	299	874
Hinweis	1.232	32	236	460	89	415
Argument	3.459	195	738	1.857	210	459

6 Bewertung geeigneter MTV-Verfahren

Ausgehend von den in Kapitel 3 genannten Aufgaben wurden zunächst grundsätzlich geeignete MTV-Verfahren aus dem Stand der Technik recherchiert und eine Übersicht erstellt. Diese Übersicht umfasst sowohl erprobte Verfahren, die bereits erfolgreich in Produktivsystemen eingesetzt werden, als auch aussichtsreiche Verfahren, die gegenwärtig noch Forschungsgegenstand sind.

Auf Basis dieser Verfahren wurden potenzielle Lösungsansätze je Aufgabenstellung erarbeitet. Dabei wurden die genutzten Verfahren so ausgewählt und zusammengestellt, dass die gelieferte Lösungsskizze möglichst im Rahmen der vorliegenden Studie als PoC umsetzbar ist und mit den vorliegenden Daten das Potenzial dieser Verfahren aufgezeigt werden kann.

Zudem wurden, sofern dies notwendig schien, bereits konkrete Experimente auf den vorliegenden Daten durchgeführt, um die Realisierbarkeit und potenziell erreichbare Ergebnisse abzuschätzen.

Anschließend wurde eine Bewertung der vorgeschlagenen Lösungsskizzen vorgenommen. Dazu wurden Bewertungskriterien betrachtet, die den potenziellen Nutzen und Realisierbarkeitsaspekte abdecken. Zusammenfassend wurde anhand der wesentlichsten Kriterien eine Bewertungsmatrix erstellt in der die Lösungen gegenübergestellt werden.

6.1 Verfahrensübersicht

Das Gebiet der Computerlinguistik hat eine Vielzahl von Methoden hervorgebracht. Diese reichen von traditionellen Ansätzen, die maschinelles Sprachverständnis mittels Regelwerken und Wörterbüchern realisieren, über statistische Methoden bis hin zu Deep Neural Networks. Bei den vorliegenden Aufgabenstellungen handelt es sich technisch betrachtet im Kern meist um Textklassifikations- und Ähnlichkeitsprobleme. In Tabelle 3 sind die zur Lösung geeigneten Methoden sowie zur Realisierung geeignete Softwarebibliotheken zusammengefasst, die nachfolgend beschrieben werden.

6.1.1 Vorverarbeitung, Analyse und Filterung

Zur Bewältigung von Textverarbeitungsaufgaben ist in der Regel zunächst eine Vorverarbeitung nötig, um die zu verarbeitenden Dokumente in eine maschinell verarbeitbare Repräsentation zu überführen. In aller Regel werden Dokumente zunächst in einzeln zu verarbeitende Segmente geteilt. Hierfür können Reguläre Ausdrücke oder Tokenizer eingesetzt werden. Tokenizer setzen regelbasierte oder statistische Verfahren ein, um Texte in Absätze, Sätze und einzelne Token (Worte, Zahlen und Trennzeichen) zu zerlegen. Häufig ist es nützlich Flexionen aus dem zu verarbeitenden Text zu entfernen. Dieser Vorgang wird Lemmatisierung [1] oder „Stemming“ [2] genannt. Dazu werden entweder wörterbuchbasierte Methoden oder Regelsätze, die eine Reduktion auf Basis von Wortendungen vornehmen, eingesetzt.

Daneben gibt es maschinelle Verfahren, um Texte grammatisch und inhaltlich zu annotieren. Zu den grammatischen Annotationen gehören die Bestimmung von „Part-of-Speech-Tags“ [3] (Wortarten), morphologischer Merkmale (Numerus, Tempus, Genus, Modus) und die Bestimmung von grammatischen Zusammenhängen zwischen Worten innerhalb eines Satzes in Form von Dependenzbäumen.

Zur inhaltlichen Annotation können „Named-Entity-Recognition“ (NER) oder „Keywordextraction“-Verfahren eingesetzt werden. NER-Verfahren basieren entweder auf umfangreichen Stichwortlisten, die aus Wissensbasen wie DBpedia, Wikidata oder ConceptNet gewonnen werden oder verwenden „Machine Learning“-Verfahren, um Entitäten wie Personen, Organisation, Orte, Preise, Datumsangaben, anhand der

TABELLE 3: VERFAHRENSÜBERSICHT

	Vorverarbeitung, Analyse und Filterung	Repräsentation und Ähnlichkeit	Klassifikation und Topic-Modeling
Verfahren	Vorverarbeitung (RegEx, Tokenizer, Sentence-Splitter) Normalisierung (Stemmer [1], Lemmatizer [2]) Grammatische Annotation (Morphology, POS [3], Dependency-Trees) Inhaltliche Annotation (Entity-Recognition, Keyword-Extraction)	Dokumentvektorisierung (BoW, Termfrequency, TF-IDF) Wortvektoren (Word2Vec [4], GloVe [5], ELMo [6]) Sprachmodelle (Universal Sentence Encoder [7], Bert [8], XLNet [10], Electra) syntaktische Ähnlichkeitsmaße (BM25 [14], Levenshtein [13], Jaccard-Index [11], Cosinusähnlichkeit [12])	Regelsysteme Topic-Modeling (LDA [15], LSA [16], NMF [17], t-SNE) Shallow-Learning-Verfahren (SVM [18], KNN [19], Logit [20], Naive Bayes [21], Random Forests [22]) Deep-Learning-Verfahren (MLP [23], CNN [24], RNN [25], LSTM [26])
Software-Bibliotheken	NLTK SpaCy Syntok Re	Gensim SpaCy Transformers	Genism Scikit-learn Imbalanced-learn Keras TensorFlow PyTorch

Satzstellung und des semantischen Kontextes zu erkennen. „Keyword- oder Keyphraseextraction“-Verfahren versuchen die relevantesten Begriffe oder Phrasen in einem Text zu identifizieren. Diese arbeiten anhand relativer Worthäufigkeiten oder analysieren linguistische Merkmale. Bekannte Keywordextractionmethoden sind TF-IDF, TextRank, RAKE oder YAKE.

Zur Realisierung der benannten Vorverarbeitungsschritte kann auf eine Vielzahl von Python-Bibliotheken zurückgegriffen werden. Dazu gehören SpaCy, NLTK, PyStemmer, Syntok, PyTextRank, Python RAKE und YAKE.

6.1.2 Repräsentation und Ähnlichkeit

Statistische Modelle für die Lösung von Textklassifikations- und Ähnlichkeitsproblemen arbeiten numerisch, d. h. der zu verarbeitende Text muss zunächst in eine vektorielle Repräsentation konvertiert werden. Hierfür gibt es verschiedene Ansätze, die abhängig von der Problemstellung und dem gewählten Klassifikationsverfahren, unterschiedlich gute Ergebnisse erzielen. Die Vektorisierungsverfahren Bag of Words (BoW), Termfrequency (TF) und TF-IDF (Termfrequency weighted by Inverse Document Frequency) erzeugen hochdimensionale dünn besetzte Vektoren. Dabei wird für jedes Wort im Vokabular eine Vektorkomponente erzeugt. Im Falle von BoW werden dabei die Komponenten aller im Dokument vorkommenden Worte auf 1 und alle anderen auf 0 gesetzt. TF berücksichtigt zudem die Häufigkeit der Worte und TF-IDF nimmt zudem eine Normalisierung anhand der Worthäufigkeit im Korpus vor. Daneben gibt es

weitere Varianten von TF-IDF, die Normalisierungen - beispielsweise anhand der Dokumentlänge - vornehmen oder die Werte logarithmisch skalieren. Mittels dieser Varianten lassen sich die Ergebnisse für bestimmte Anwendungsfälle optimieren. Die genaue Wahl der Berechnungsformel erfolgt im Rahmen der Verfahrensoptimierung.

Andere Vektorisierungsverfahren basieren auf Wort- oder Dokumentenvektoren, auch als Embeddings bezeichnet. Bei diesen Embeddings handelt es sich um dicht besetzte Vektoren geringerer Dimensionalität. Dabei werden semantisch ähnliche Worte auf ähnliche Vektoren abgebildet. Bekannte Verfahren sind Word2Vec, fastText oder GloVe. Bei diesen Verfahren handelt es sich um statische Wortvektoren, d. h. Worte werden unabhängig von ihrem Kontext immer auf den gleichen Vektor abgebildet. Embeddings from Language Model (ElMo) generiert Wortvektoren kontextabhängig und kann entsprechend auch Ambiguitäten abbilden. Ein einfacher Ansatz, Dokumentembeddings aus Wortvektoren zu generieren, ist den Mittelwert über die Wortvektoren zu bilden. Andere Verfahren versuchen direkt Dokumentembeddings zu erstellen. Dazu gehören der Universal Sentence Encoder (USE), Bidirectional Encoder Representations from Transformers (BERT) und Electra.

Verfahren zur Ähnlichkeitsanalyse lassen sich in syntaktische und semantische Verfahren aufteilen. Zu den syntaktischen Verfahren gehören die Levenshteindistanz – ein zeichenbasiertes Verfahren, das die Anzahl der Bearbeitungsschritte misst, der Jaccard-Index – ein Vergleich auf Basis einer BoW-Repräsenation und Okapi BM25 – eine häufig in Suchmaschinenrankings verwendete TF-IDF-Variante. Semantische Verfahren nutzen dagegen eines der oben beschriebenen Dokumentembeddingverfahren und vergleichen die generierten Embeddings mittels Cosinusähnlichkeit, indem sie das Skalarprodukt aus den zu vergleichenden Dokumentenvektoren bilden.

6.1.3 Klassifikation und Topic-Modeling

Bei Klassifikationsverfahren kann grundsätzlich zwischen überwachten und nicht-überwachten Verfahren unterschieden werden. Überwachte Verfahren werden mittels eines Datensatzes bestehend aus einer Liste von Eingaben und entsprechend zugeordneten Klassen trainiert. Nicht überwachte Methoden identifizieren selbstständig Klassen, indem sie Gemeinsamkeiten und Unterschiede der Eingaben analysieren.

Im Bereich Textklassifikation werden zum unüberwachten Lernen häufig Softclusteringverfahren eingesetzt, die unter dem Oberbegriff „Topic Modeling“ zusammengefasst werden. „Softclustering“ heißt, dass diese Verfahren den Eingaben nicht genau ein Thema zuweisen, sondern eine gewichtete Mischung aus mehreren Themen. Bekannte Vertreter dieser Verfahren sind „Latent Dirichlet Allocation“ (LDA), „Latent Semantic Analysis“ (LSA) und „Non-negative Matrix Factorization“ (NMF).

Bei Verfahren zur Lösung von überwachten Klassifikationsproblemen lässt sich grundsätzlich zwischen regelbasierten Verfahren und Lernverfahren unterscheiden.

Regelbasierte Verfahren liefern sehr gute Ergebnisse, sofern es klare Merkmale (z. B. Textlänge, Textposition, Formatierungsmuster, vorkommen bestimmter Floskeln oder Signalworte) gibt, auf deren Basis zuverlässig auf die Klassenzuordnung geschlossen werden kann. Bei praktischen Textklassifikationsproblemen lassen sich derartig klare Merkmale und Schlussregeln jedoch häufig nicht herausarbeiten oder die während der Entwicklung erstellten Regelsätze verallgemeinern nur unzureichend auf die in der Praxis anzutreffenden Daten.

Lernverfahren verwenden anstelle eines Regelsatzes ein statistisches Modell, dessen Parameter anhand bekannter Trainingsbeispiele so optimiert werden, dass ein bestmögliches Klassifikationsergebnis erzielt wird. Innerhalb dieser Verfahren kann zwischen „Shallow- und Deep-Learning“-Verfahren unterschieden werden. Deren wesentlicher Unterschied besteht in der Anzahl der Modellparameter, die bei „Deep-Lear-

ning“ um Größenordnungen höher ist. Ein weiterer Unterschied besteht darin, dass „Deep-Learning“-Modelle die Reihenfolge der Wörter und den Kontext berücksichtigen, indem sie eine Folge von Wörtern und nicht nur einen Vektor von Merkmalen als Eingabe verwenden.

Die Anzahl der Parameter statistischer Modelle hat eine Vielzahl von Implikationen für den praktischen Einsatz dieser Techniken. Shallow-Learning-Methoden automatisieren in der Praxis den Schritt der Regelerstellung, während die Extraktion und Auswahl der zu betrachtenden Merkmale weiterhin manuell erfolgt. So kann analytisch gewonnenes Vorwissen über die Problemstellung in Form von Vorverarbeitungsschritten in das Verfahren eingebracht werden, wodurch auch mit einer überschaubaren Menge von Trainingsdaten praxistaugliche Ergebnisse erzielt werden können. Bekannte Klassifikationsmethoden aus dem Bereich Shallow-Learning sind K-Nearest-Neighbour (KNN), Naive Bayes, Logistic Regression (Logit) [20], Random Forest und Support Vector Machines (SVM). Im Python-Umfeld können diese mithilfe der Bibliotheken Scikit-learn und Imbalanced-learn realisiert werden.

Deep-Learning-Verfahren verfolgen dagegen einen Ende-zu-Ende-Lernansatz, d. h. die Erstellung des Klassifikators wird vollständig automatisiert. Hierfür ist einerseits eine sehr große Anzahl von Trainingsbeispielen erforderlich und andererseits erfordern diese Verfahren einen sehr hohen Einsatz von Hardwareressourcen. Sofern diese Anforderungen erfüllt werden können, erreichen sie allerdings die besten nach dem Stand der Technik erreichbaren Ergebnisse. In vielen praktischen Anwendungsszenarien gelingt es jedoch nicht, einen hinreichend umfangreichen und korrekt annotierten Datensatz zu erstellen oder die nötige IT-Infrastruktur bereitzustellen. Während „Shallow-Learning“-Verfahren in der Regel auch in häufig anzutreffenden virtualisierten IT-Infrastrukturen unproblematisch ausführbar sind, erfordern „Deep-Learning“-Verfahren eine leistungsstarke GPU, deren Bereitstellung in vielen Infrastrukturen Anpassungen erfordert.

Deep-Learning wird in der Regel mittels tiefer neuronaler Netze realisiert. Bei diesen lassen sich verschiedene Architekturen unterscheiden. Bekannte Architekturen sind „Multilayer Perceptron“ (MLP) [23], „Convolutional Neural Networks“ (CNN), „Recurrent Neural Networks“ (RNN), „Long Short-Term Memory“ (LSTM) und Transformer [27]. In Python können diese Lernverfahren mit den Bibliotheken PyTorch, Tensorflow, Keras, Transformers oder FARM realisiert werden. Die Klassifikationsverfahren werden in der Tabelle 4 gegenübergestellt.

6.2 Lösungsentwürfe je Aufgabe

6.2.1 Identifizierung und Extraktion argumentativer Einwendungssegmente

Die Identifizierung und Extraktion argumentativer Einwendungssegmente lässt sich in folgende Unteraufgaben zerlegen, die im Folgenden unabhängig voneinander betrachtet werden:

1. Zerlegung von Einwendungen in Segmente
2. Identifikation relevanter Segmente
3. Identifikation argumentativer Segmente

Ziel der ersten Teilaufgabe ist es, den Text innerhalb einer Einwendung anhand der Formatierung und ohne Berücksichtigung des Textinhaltes in Segmente für die weitere Analyse zu zerlegen. Um die Eignung von Verfahren für diese Aufgabe zu untersuchen, werden möglichst vollständige und weitestgehend unveränderte Einwendungen als Referenz benötigt. Hierfür stehen im vorliegenden Datensatz 41 digitale Transkripte handschriftlicher Einwendungen als Textdateien zur Verfügung, die als Referenz zur Beurteilung eines Segmentierungsalgorithmus dienen können.

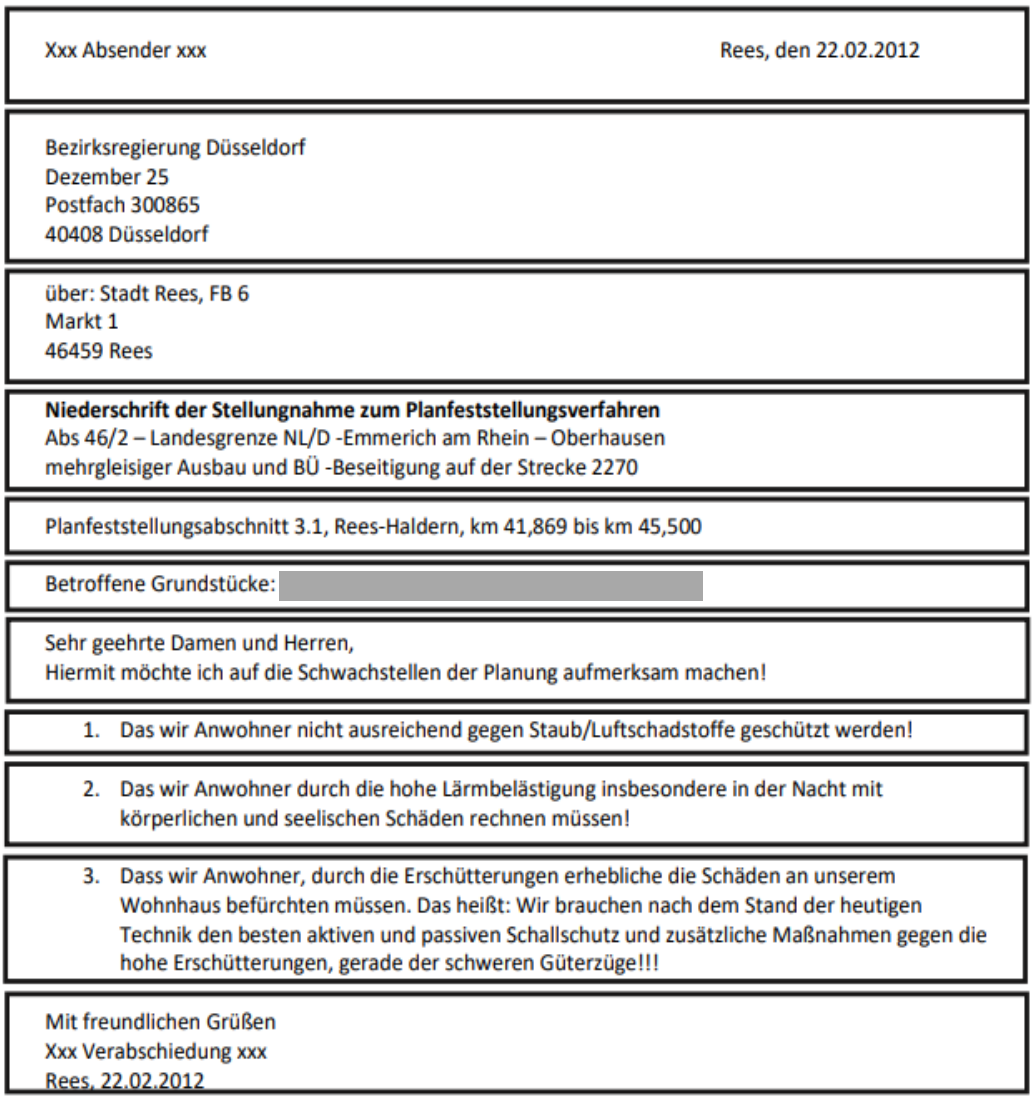
TABELLE 4: GEGENÜBERSTELLUNG DER KLASSIFIKATIONSVERFAHREN

	Regelbasierte Verfahren	Shallow-Learning	Deep-Learning
Verfahren	Mustererkennung - Reguläre Ausdrücke - Gazetteers Linguistische Merkmale - Wortmorphologie - PoS-Tags [1] - Syntaxbäume	Statistische Verfahren - Naive Bayes - KNN - Logit - Random Forests - SVM	Neuronale Netzwerke - MLP - CNN - RNN - LSTM - Transformer
Vorteile	Schnelle Berechnung Hohe Zuverlässigkeit Triviale Erklärbarkeit der Ergebnisse	Trainingsprozess automatisierbar Gute Verallgemeinerung der Modelle Gute Erklärbarkeit der Ergebnisse	Sehr hohe Ergebnisgüte Ende-zu-Ende Training Weitere Verbesserung mit künftiger Hardware zu erwarten
Nachteile	Verallgemeinert kaum bei komplexen Problemen Aufwändiges Regeldesign und hoher Anpassungsaufwand bei geänderten Anforderungen	Aufwändiges Featuredesign Erzielbare Ergebnisgüte ist methodisch begrenzt und profitiert kaum von gesteigerten Ressourceneinsatz	Benötigt sehr viele Trainingsbeispiele Ressourcenintensiver Trainingsprozess
Python-Bibliotheken	NLTK SpaCy	Scikit-learn Imbalance-learn	PyTorch Keras Tensorflow Transformers FARM

Grundsätzlich sind für die Segmentierung verschiedene Varianten möglich. Es ist eine Zerlegung in Unterabschnitte anhand von Teilüberschriften, eine absatzweise Zerlegung unter Berücksichtigung von Zeilentrennzeichen, Nummerierungen und Anstrichen oder eine satzweise Zerlegung mittels entsprechender Tokenisierer möglich. Die im Datensatz vorliegenden bereits segmentierten und annotierten Daten entsprechen am ehesten einer absatzweisen Zerlegung. Entsprechend wurde diese Variante angestrebt, sodass die erstellten Segmente möglichst ähnlich zu den vorliegenden Trainingsdaten sind und ein darauf basierend erstellter Klassifikator möglichst vergleichbare Ergebnisse erzielt.

Für die Mehrzahl der Referenzdokumente gelang es mittels eines Mustererkennungsalgorithmus mit fest definierten Mustern auf Basis regulärer Ausdrücke, die Zeilenumbrüche und strukturierende Elemente - wie Nummerierungen oder Anstriche – berücksichtigen, eine sinnvolle Segmenttrennung zu erreichen. Eine manuelle Sichtung der 41 getesteten Dokumente ergab, dass die Segmentierung in 36 Fällen sehr gute Ergebnisse liefert. In 4 Dokumenten wurden einzelne Absätze nicht zufriedenstellend getrennt und ein Dokument wies eine Vielzahl von Segmentierungsfehlern auf. Bei den entdeckten Fehlern wurden mehrere Argumente zu einem Segment zusammengefasst, sodass auffällig lange Segmente entstanden.

In den betroffenen Dokumenten werden Absätze unterschiedlich mit einzelnen und doppelten Zeilenumbrüchen getrennt. In diesen Fällen berücksichtigt der erstellte Regelsatz nur die Trennung mittels doppelter Zeilenumbrüche. Es ist anzunehmen, dass sich dies durch Verbesserung des Regelsatzes beheben lässt, sodass etwa 95 % der Dokumente durch ein derartiges Verfahren korrekt segmentiert werden können. Abbildung 5 zeigt ein Beispielergebnis der regelbasierten Segmentierung.



Xxx Absender xxx Rees, den 22.02.2012

Bezirksregierung Düsseldorf
Dezember 25
Postfach 300865
40408 Düsseldorf

über: Stadt Rees, FB 6
Markt 1
46459 Rees

Niederschrift der Stellungnahme zum Planfeststellungsverfahren
Abs 46/2 – Landesgrenze NL/D -Emmerich am Rhein – Oberhausen
mehrgleisiger Ausbau und BÜ -Beseitigung auf der Strecke 2270

Planfeststellungsabschnitt 3.1, Rees-Haldern, km 41,869 bis km 45,500

Betroffene Grundstücke:

Sehr geehrte Damen und Herren,
Hiermit möchte ich auf die Schwachstellen der Planung aufmerksam machen!

1. Das wir Anwohner nicht ausreichend gegen Staub/Luftschadstoffe geschützt werden!
2. Das wir Anwohner durch die hohe Lärmbelästigung insbesondere in der Nacht mit körperlichen und seelischen Schäden rechnen müssen!
3. Dass wir Anwohner, durch die Erschütterungen erhebliche die Schäden an unserem Wohnhaus befürchten müssen. Das heißt: Wir brauchen nach dem Stand der heutigen Technik den besten aktiven und passiven Schallschutz und zusätzliche Maßnahmen gegen die hohe Erschütterungen, gerade der schweren Güterzüge!!!

Mit freundlichen Grüßen
Xxx Verabschiedung xxx
Rees, 22.02.2012

Abbildung 5 Beispielergebnis der regelbasierten Segmentierung

Das Ziel der zweiten und dritten Unteraufgabe besteht darin, die so getrennten Segmente als „bearbeitungsrelevant“ oder „Begleittext“ zu kennzeichnen bzw. innerhalb der bearbeitungsrelevanten Segmente zwischen Argumenten und Hinweisen zu unterscheiden.

Der erstellte Datensatz enthält als Referenz zur Bewertung von Verfahren für diese Aufgabe 5.289 bereits getrennte Einwendungssegmente, deren Attribute *Bearbeitungsrelevanz* und *Segmentbewertung* eine Zuordnung zu den zuvor erwähnten Kategorien darstellen. Übertragen auf ein Datenverarbeitungsproblem handelt es sich bei den fachlichen Unteraufgaben zwei und drei also um ein binäres Klassifikationsproblem.

Der vorliegende Datenbestand umfasst 4.691 bearbeitungsrelevante Segmente und 598 Begleittexte bzw. 3.459 Argumente und 1.232 Hinweise, die zum Erstellen von Klassifikatoren genutzt werden können. Um eine Abschätzung der erreichbaren Ergebnisse zu ermöglichen, wurden zunächst verschiedene Shallow-Learning-Verfahren zur Textklassifikation erprobt. Diese eignen sich besonders, da mit diesen Methoden unter Verwendung der zur Verfügung stehenden Daten und mit angemessenen Aufwand voraussichtlich Ergebnisse erzielt werden können, die eine Potenzialabschätzung ermöglichen. Shallow-Learning-Verfahren bestehen aus einem Vektorisierungsverfahren, das den Text in eine numerische Repräsentation übersetzt und einem maschinellen Lernverfahren, das anhand dieser Repräsentationen der Trainingsdaten die Klassifikation erlernt. Hierfür wurden folgende Vektorisierer und Lernverfahren betrachtet:

- Vektorisierungsverfahren:
  - Bag of Words
  - Termfrequenz
  - Tf-Idf

- Lernverfahren:
  - K-Nearest-Neighbour
  - Logistic Regression
  - Naive Bayes
  - Random Forests

Um einen systematischen Vergleich dieser Verfahren zu ermöglichen, wurde der je Problemstellung zur Verfügung stehende Datensatz in Trainings- und Testdaten aufgeteilt, wobei 80 Prozent zum Training und 20 Prozent zur Bewertung der Methoden verwendet wurden. Die Häufigkeit der Klassen ist in beiden Datenteilen identisch. Zudem wurden im Datensatz enthaltene Duplikate entfernt, um zu vermeiden, dass es Überschneidungen zwischen Trainings- und Testdaten gibt.

Zur numerischen Bewertung von Klassifikationsverfahren gibt es verschiedene Metriken. Accuracy (ACC) gibt den Anteil korrekt klassifizierter Testdaten gegenüber Fehlklassifikationen an. Dieses Maß ist intuitiv interpretierbar und trifft eine Aussage über die praxisrelevante Korrektheitsquote. Insbesondere bei unbalancierten Datensätzen, d. h. in Fällen in denen sich die Häufigkeit der Klassen stark unterscheidet, ist diese Metrik allerdings ungeeignet, um Vergleiche zwischen verschiedenen Klassifikationsverfahren vorzunehmen. In diesen Fällen sind der F1-Score (F1) - das gewichtete harmonische Mittel aus Genauigkeit und Trefferquote - bzw. Mathews Correlation Coefficient (MCC) – ein Maß für die Korrelation zwischen dem Klassifikationsergebnis und der tatsächlichen Klasse – besser zur Güteeinschätzung geeignet. Allgemein werden MCC-Scores über 0,2 als schwach, über 0,3 als mäßig, über 0,4 als gut und über 0,7 als sehr gut eingeschätzt [29].

In Tabelle 5 und Tabelle 6 werden die Evaluationsergebnisse entsprechend der drei Metriken von allen Kombinationen aus den oben genannten Vektorisierungs- und Klassifikationsverfahren für die Unteraufgaben 2 und 3 wiedergegeben. Dabei erzielt ein BoW-Vektorisierer mit logistischer Regression für die Klassifikation relevanter Segmente die besten Ergebnisse (dunkelgrau hinterlegte Felder). Für die Klassifikation Argument/Hinweis erzielt KNN über TF-IDF-Vektoren geringfügig bessere Werte entsprechend der ACC- und F1-Metrik. Numerisch ist dieser Unterschied allerdings vernachlässigbar. Entsprechend können beide Problemstellungen mittels logistischer Regression erfolgreich bearbeitet werden. Dabei wird in diesem Experiment für die Unterscheidung von relevanten Segmenten und Begleittexten mit 98 Prozent Trefferquote ein sehr gutes Ergebnis erzielt, während bei der Unterscheidung von Argumenten und Hinweisen 78 Prozent der Testbeispiele korrekt zugeordnet wurden. Diese Trefferquoten können ggf. durch die Berücksichtigung weiterer Features, wie grammatischer oder morphologischer Merkmale, Segmentlänge, Segmentposition im Text oder Textkomplexitätsmetriken ggf. weiter verbessert werden.

TABELLE 5: ERGEBNISSE DER KLASSIFIKATION RELEVANTER SEGMENTE

Klassifikation relevant / nicht relevant		**BoW**	**TF**	**TFIDF**
K-Nearest-Neighbour	ACC	0,9412	0,9753	0,9608
	MCC	0,6384	0,8417	0,7675
	F1	0,9677	0,9865	0,9784
Logistic Regression	ACC	**0,9804**	0,9794	0,9598
	MCC	**0,8778**	0,8696	0,7302
	F1	**0,9893**	0,9887	0,9784
Random Forest	ACC	0,9773	0,9711	0,9701
	MCC	0,8554	0,8127	0,8054
	F1	0,9876	0,9843	0,9838
Naive Bayes	ACC	0,9639	0,9639	0,9619
	MCC	0,7610	0,7610	0,7456
	F1	0,9805	0,9805	0,9794

TABELLE 6: ERGEBNISSE DER KLASSIFIKATION VON ARGUMENTEN

Klassifikation Argument / Hinweis		**BOW**	**TF**	**TFIDF**
K-Nearest-Neighbour	ACC	0,6331	0,7622	**0,7860**
	MCC	0,1772	0,2715	0,3730
	F1	0,7300	0,8548	**0,8664**
Logistic Regression	ACC	0,7826	0,7735	0,7656
	MCC	**0,3863**	0,3063	0,2668
	F1	0,8603	0,8638	0,8615
Random Forest	ACC	0,7622	0,7622	0,7622
	MCC	0,2483	0,2487	0,2487
	F1	0,8594	0,8591	0,8591
Naive Bayes	ACC	0,7656	0,7690	0,7633
	MCC	0,3216	0,3285	0,3084
	F1	0,8514	0,8541	0,8508

6.2.2 Klassifizierung von Einwendungen bezüglich eines Themenkatalogs

Die Klassifizierung von Segmenten bezüglich eines Themenkataloges baut auf die Ergebnisse der vorangegangenen Aufgabe auf. Ziel dieser Aufgabe ist es, die als argumentativ gekennzeichneten Einwendungssegmente systematisch nach Themen zu katalogisieren.

Als Datengrundlage stehen hierfür 3.461 argumentative Segmente mit manuell annotierten Themenzuordnungen entsprechend dem zweistufigen Themenkatalog in Anhang 1 zur Verfügung. Von diesen sind 1.144 privaten Ursprungs (handschriftliche, private und Sammeleinwendungen), 460 aus Stellungnahmen von TÖBs und 1.857 gemischten Ursprungs.

Analog zu den Unteraufgaben 2 und 3 der Segmentidentifikation kann auch die Themenzuordnung unter Verwendung des vorgegebenen Themenkatalogs als überwachtes Klassifikationsproblem betrachtet werden. Entsprechend können ähnliche Methoden und Bewertungsmetriken angewendet werden.

Eine alternative Lösungsmöglichkeit ist die unüberwachte Klassifikation mittels Topic-Modeling-Verfahren. Hierbei ist der Themenkatalog nicht vordefiniert, sondern wird aus den Daten gewonnen. Ein Vorteil von Topic-Modeling ist, dass die maschinelle Zuordnung zu den so gewonnenen Themen häufig genauer erfolgen kann. Allerdings lassen sich die algorithmisch identifizierten Themen häufig nicht unmittelbar auf die im Fachprozess relevanten Themen abbilden. Entsprechend wird folgend auf die Verfahren des überwachten Lernens fokussiert.

Analog zur vorangegangenen Aufgabe wurde der Datensatz zur experimentellen Betrachtung der Verfahren ebenfalls um Duplikate bereinigt und in 80 Prozent Trainingsdaten und 20 Prozent Testdaten aufgeteilt. Folgende Tabelle gibt die Anzahl der Segmente je Thema in diesen Daten an. Für einige Unterthemen ist die Anzahl der Beispielsegmente zu gering, um Lernverfahren sinnvoll zu bewerten. Um zu vermeiden, dass die gemessenen Scores durch diese unterrepräsentierten Themen zu stark verfälscht werden, wurden Unterthemen mit weniger als 30 vorhandenen Segmenten aus der Verfahrensbewertung ausgeschlossen.

TABELLE 7: ANZAHL DER SEGMENTE JE THEMA IN DEN TRAININGS- UND TESTDATEN (UNTERREPRÄSENTATIVE THEMEN SIND DURCHGESTRICHEN)

	Train	**Test**	**Gesamt**
1.00 - Immissionen und Emissionen	906	226	1132
1.01 - Schall Lärm	574	143	717
1.02 - Erschütterungen	170	43	213
1.03 - sekundärer Luftschall	14	0	14
1.04 – Luftverunreinigungen Feinstaub	75	19	94
1.05 - Elektromagnetische Felder	76	19	95
1.06 - Immissionen und Emissionen - Sonstiges	51	13	64
2.00 - Naturschutz	184	47	231
2.01 - Natur und Landschaftspflege	108	27	135
2.02 - Gebietsschutz	7	0	7
2.03 - Artenschutz	60	15	75

2.04 - Naturschutz - Sonstiges	23	0	23
3.00 - Umweltschutz	213	53	266
3.01 - Bodenschutz	60	15	75
3.02 - Gewässerschutz	127	32	159
3.03 - Umweltschutz - Sonstiges	26	6	32
4.00 - Landwirtschaft	134	33	167
4.01 - Landwirtschaft	130	33	163
4.02 - Landwirtschaft - Sonstiges	4	0	4
5.00 - Forstwirtschaft	15	0	15
5.01 - Forstwirtschaft	15	0	15
6.00 - Denkmalschutz	69	17	86
6.01 - Bodendenkmal	45	11	56
6.02 - Kulturdenkmal	20	0	20
6.03 - Denkmalschutz - Sonstiges	13	0	13
7.00 - Abfallwirtschaft	33	6	39
7.01 - Abfallwirtschaft	31	6	37
7.02 - Abfallwirtschaft - Sonstiges: 2	2	0	2
8.00 - Brand- und Katastrophenschutz	182	45	227
8.01 - Brand- und Katastrophenschutz	159	40	199
8.02 - Brand- und Katastrophenschutz - Sonstiges	28		28
9.00 - Bauzeitliche Beeinträchtigungen	452	113	565
9.01 - Bauzeitliche Beeinträchtigungen	407	102	509
9.02 - Bauzeitliche Beeinträchtigungen - Sonstiges	45	11	56
10.00 - Anlagen Dritter	299	74	373
10.01 - Infrastrukturleitungen/Sonstige Anlagen Dritter	150	37	187
10.02 – (öffentliche) Ver- und Entsorgungsanlagen	34	9	43
10.03 – Straßen Wege Zufahrten	116	29	145
10.04 - Anlagen Dritter - Sonstiges	3	0	3
11.00 - Inanspruchnahme von Grundeigentum	139	35	174
11.01 - Eigentumsbeeinträchtigungen/Grunderwerb	126	31	157
11.02 - Verschattung	10	0	10
11.03 - Inanspruchnahme von Grundeigentum - Sonstiges	7	0	7
12.00 - Sonstiges	1044	261	1305

Eine regelbasierte Lösungsmöglichkeit besteht in einer einfachen Suche nach Stichworten in den Segmenten, die eine Themenzuordnung ermöglichen. Hierzu wird zu jedem Thema eine Liste von Stichworten benötigt. Die Zuordnung eines Segmentes zu einem Thema geschieht dann, indem diese Stichworte in den Texten gesucht und die entsprechenden Themen zugeordnet werden. Entsprechende Stichwortlisten wurden bereits als Bestandteil der Annotationsrichtlinien zur Erstellung des Datensatzes definiert und sind im Anhang gelistet.

Zum Einsatz der Lernverfahren aus dem vorangegangenen Abschnitt für die Themenidentifikation sind zwei Anpassungen erforderlich. Bei der Themenzuordnung handelt es sich einerseits nicht um eine binäre Entscheidung, da der Themenkatalog 12 Oberthemen bzw. 33 Unterthemen umfasst. Andererseits besteht die Anforderung, einem Segment ggf. mehrere Themen zuordnen zu können. Von den bereits erwähnten Methoden sind die K-Nearest-Neighbour- und die Random-Forest-Methode bereits methodisch bedingt für derartige Probleme einsetzbar. Zum Einsatz der anderen Methoden kann die Problemstellung in eine binäre Klassifikation umformuliert werden. Dazu wird ein Klassifikator je Thema trainiert, indem alle Trainingsdaten mit diesem Thema als Positivbeispiel und die sonstigen als Negativbeispiele genutzt werden. Für die finale Bewertung eines Segmentes werden dann alle Klassifikatoren ausgeführt. Sollten dabei Fälle auftreten, bei denen keiner der Klassifikatoren eine Positivzuordnung vornimmt, kann mittels der OneVsRest-Strategie auch in diesen Fällen eine Themenzuordnung, anhand der von den Klassifikatoren geschätzten Wahrscheinlichkeiten, vorgenommen werden.

Um ein geeignetes Lernverfahren auszuwählen, wurde zunächst mittels der Vektorisierungs- und Klassifikationsverfahren aus dem vorangegangenen Abschnitt ein Satz von Klassifikatoren für alle Ober- bzw. Unterthemen trainiert. Dabei wurden die Texte mittels eines Filters um Stoppworte bereinigt. Stoppworte sind Worte, die aus syntaktischen Gründen zur Sprache gehören, jedoch kaum eine semantische Bedeutung tragen, wie beispielsweise Artikel. Entsprechend kann durch Entfernen von Stoppworten das Rauschen in den Daten reduziert und so eine höhere Klassifikationsgüte erzielt werden.

Tabelle 8 und Tabelle 9 zeigen die Ergebnisse für die Klassifikation von Ober- bzw. Unterthemen mittels dieser Verfahren. Die angegebenen F1-Scores wurden mittels Makro-Averaging aggregiert. Dabei werden die Ergebnisse der Klassifikation je Thema zusammengefasst, indem zunächst das arithmetische Mittel der Genauigkeiten und Trefferquoten bestimmt wird und der F1-Score anhand dieser Mittelwerte berechnet wird.

Aus den Ergebnissen wird ersichtlich, dass die Klassifikationsqualität auch mit den feiner aufgeteilten Unterthemen nicht schlechter wird. BoW-Vektorisierung und logistische Regression lieferten für beide Problemvarianten die besten Ergebnisse entsprechend der MCC-Bewertung. Daneben wurde die Klassifikationsgüte je Thema untersucht. Tabelle 10 fasst die Ergebnisse zusammen. Zur Einschätzung der Klassifikatorgüte eignet sich aufgrund der unbalancierten und unterschiedlichen Verhältnisse von Positiv- und Negativbeispielen je Thema im Testdatensatz der MCC-Score. Die ACC-Scores sind der Vollständigkeit halber zusätzlich angegeben.

TABELLE 8: VERGLEICH VON KLASSIFIKATIONSMETHODEN ZUR ERKENNUNG DER OBERTHEMEN

Klassifikation Oberthemen		**BoW**	**TF**	**TFIDF**
K-Nearest-Neighbour	ACC	0,4367	0,7215	0,7089
	MCC	0,2187	0,6198	0,6010
	F1	0,1648	0,5013	0,4860
Logistic Regression	ACC	**0,7563**	0,6677	0,6835
	MCC	**0,6656**	0,5415	0,5596
	F1	0,5440	0,2834	0,3256
Random Forest	ACC	0,7342	0,7310	0,7215
	MCC	0,6357	0,6319	0,6185
	F1	0,4963	0,4927	0,4939
Naive Bayes	ACC	0,7373	0,7500	0,7405
	MCC	0,6374	0,6570	0,6421
	F1	0,5603	**0,5782**	0,5655

TABELLE 9: VERGLEICH VON KLASSIFIKATIONSMETHODEN ZUR ERKENNUNG DER UNTERTHEMEN

Klassifikation Unterthemen		**BoW**	**TF**	**TFIDF**
K-Nearest-Neighbour	ACC	0,3627	0,7451	0,7451
	MCC	0,2760	0,6844	0,6832
	F1	0,2432	0,5499	0,4977
Logistic Regression	ACC	**0,7892**	0,5490	0,5833
	MCC	**0,7396**	0,4480	0,4855
	F1	**0,6705**	0,2493	0,2934
Random Forest	ACC	0,7402	0,7451	0,7206
	MCC	0,6840	0,6892	0,659
	F1	0,6515	0,6538	0,6206
Naive Bayes	ACC	0,7500	0,7500	0,7402
	MCC	0,6908	0,6911	0,6788
	F1	0,6484	0,6491	0,6407

TABELLE 10: KLASSIFIKATIONSGÜTE JE THEMA

Code	Thema	ACC	MCC
1.00	Immissionen und Emissionen	0,8810	0,6908
1.01	Schall, Lärm	0,8899	0,7088
1.02	Erschütterungen	0,9435	0,6158
1.04	Luftverunreinigungen	0,9866	0,7502
1.05	Elektromagnetische Felder	0,9777	0,6245
1.06	Immissionen und Emissionen - Sonstiges	0,9777	0,4741
2.00	Naturschutz	0,9454	0,5151
2.01	Natur- und Landschaftsschutz	0,9628	0,5262
2.03	Artenschutz	0,9717	0,4117
3.00	Umweltschutz	0,9279	0,4050
3.01	Bodenschutz	0,9702	0,3181
3.02	Gewässerschutz	0,9539	0,5561
3.03	Umweltschutz - Sonstiges	0,9807	0,2295
4.00	Landwirtschaft	0,9694	0,5338
4.01	Landwirtschaft	0,9658	0,6388
6.00	Denkmalschutz	0,9956	0,8746
6.01	Bodendenkmal	0,9940	0,8333
7.00	Abfallwirtschaft	0,9911	0,5669
8.00	Brand- und Katastrophenschutz	0,9760	0,7230
8.01	Brand- und Katastrophenschutz	0,9568	0,6281
9.00	Bauzeitliche Beeinträchtigungen	0,8865	0,5757
9.01	Bauzeitliche Beeinträchtigungen	0,8690	0,5693
9.02	Bauzeitliche Beeinträchtigungen - Sonstiges	0,9628	0,2341
10.00	Anlagen Dritter	0,9258	0,4505
10.01	Infrastrukturleitungen / Sonstige Anlagen Dritter	0,9271	0,4601
10.02	(öffentliche) Ver- und Entsorgungsanlagen	0,9836	0,4134
10.03	Straßen, Wege, Zufahrten	0,9390	0,4175
11.00	Inanspruchnahme von Grundeigentum	0,9574	0,4127
11.01	Eigentumsbeeinträchtigungen / Grunderwerb	0,9509	0,5069
12.00	Sonstiges	0,8330	0,5869

6.2.3 Erkennung von Argumentationsmustern

Ziel der Erkennung von Argumentationsmustern ist, vorwiegend mehrfach eingereichte Einwendungen oder Textbausteine zu erkennen, um so eine konsistente Beantwortung zu gewährleisten.

Einige Beispiele solcher Mehrfacheinreichungen sind auch im erstellten Datensatz vorhanden. So sind beispielsweise fünf der im Datensatz enthaltenen Sammeleinwendungen, bis auf einzelne Formulierungen oder entfallene oder hinzufügte Argumente, nahezu identisch.

Um solche Duplikate aufzufinden, wird ein zu prüfendes Dokument satz- oder absatzweise segmentiert und mit bekannten Segmenten im vorhandenen Datenbestand verglichen. Aus der Betrachtung aller relevanten Segmente kann dann ermittelt werden, bei welchen Teilen des Dokumentes es sich um identische oder leicht modifizierte Abschnitte aus dem bereits bekannten Datenbestand bzw. neue zu berücksichtigende Argumente handelt.

Da es sich bei den aufzufindenden Argumentationsmustern in der Regel um wiederverwendete Textbausteine handelt, kommen die Verfahren zur syntaktischen Ähnlichkeitsanalyse aus Abschnitt 6.1.2 für den segmentweisen Vergleich in Betracht. Der Jaccard-Index zweier zu vergleichender Texte gibt den Anteil übereinstimmender Worte im Verhältnis zu allen Worten in beiden Texten an. Durch die Wahl geeigneter Schwellenwerte, lassen sich anhand des Jaccard-Index identische, nahezu identische, ähnliche und disjunkte Segmente unterscheiden. Im Unterschied zur Levenshtein-Distanz (Anzahl der Editierungsschritte), lässt sich der Jaccard-Index auch beim Vergleich längerer Textabschnitte effizient berechnen, berücksichtigt dagegen allerdings nur die verwendeten Worte, nicht deren Reihenfolge. Praktisch ist die Aussagekraft dennoch in der Regel hoch genug. Die Wortreihenfolge kann zusätzlich berücksichtigt werden, indem der Jaccard-Index nicht über einzelne Worte, sondern überlappende Wortpaare oder n-Gramme berechnet wird.

Eine Herausforderung beim segmentweisen Vergleich eines Dokumentes mit einem großen Datenbestand ist, dass der Vergleich gegen jedes Segment im Datenbestand erfolgen muss. Auf dem vorhandenen Datensatz bestehend aus ca. 5.000 Segmenten benötigt die Prüfung eines Dokumentes bereits mehrere 100 Millisekunden Sekunden zur Verarbeitung. Dieser Zeitaufwand wächst linear mit dem zu durchsuchenden Datenbestand. Für die praktische Anwendung kann der Jaccard-Index daher mittels des MinHash-Algorithmus approximiert werden, um eine skalierbare Lösung zu realisieren. Dieser ist auch Bestandteil diverser Softwarekomponenten für Suchmaschinen, beispielsweise ElasticSearch, und kann mittels dieser in eine Lösung integriert werden.

Während der Jaccard-Vergleich sehr zuverlässige Ergebnisse in Bezug auf die Erkennung lexikalisch ähnlicher Texte liefert, ist er nicht in der Lage, paraphrasierte Textbausteine zu erkennen. Dies kann durch Verfahren erreicht werden, die semantische Ähnlichkeit messen. Hierzu können die Segmente als sogenannte Documentembeddings in Vektorrepräsentationen überführt werden, die anschließend mittels Cosinusähnlichkeit verglichen werden. Geeignete Verfahren hierzu sind Durchschnittswerte über Wortvektoren der Segmente oder Kodierungen mittels neuronaler Modelle, wie USE [6] oder sBERT [8]. Auch bei diesem Ansatz besteht die Herausforderung, dass ein Abgleich gegen einen umfangreichen Datenbestand ineffizient ist. Lösungsansätze bieten hier sogenannte Approximate Nearest Neighbour Algorithmen. Dazu gehören Ansätze auf Basis von BSP-, KD-, oder Balltrees, sowie speziell für semantische Suchmaschinen entwickelte Methoden, wie FAISS oder ScaNN. Deren Integration ist allerdings aufwändiger, da sie nicht Bestandteil von Out-of-The-Box-Komponenten für Suchmaschinen sind. Das Framework Haystack integriert derartige Algorithmen, befindet sich allerdings noch in einem frühen Entwicklungsstadium.

Bei Anwendung des segmentweisen Vergleichs mittels Jaccard-Index und einem Schwellenwert von 0,5 auf dem zur Verfügung stehenden Datensatz, wurden 522 Textbausteine identifiziert, die in ähnlicher Form in 1708 Segmenten auftraten.

6.2.4 Prüfung der Vollständigkeit und Relevanz von Erwiderungen

Das Ziel dieser Teilaufgabe besteht darin, den Rücklauf des Vorhabenträgers zu prüfen. Dabei soll einerseits ermittelt werden, ob alle Argumente verarbeitet wurden und andererseits sollen die Erwiderungen des Vorhabenträgers inhaltlich plausibilisiert werden, um beispielsweise Fehlzuordnungen von Textbausteinen zu erkennen.

Die genauen Datenformate, die zwischen EBA und Vorhabenträger ausgetauscht werden, sind nicht Bestandteil des Datensatzes. Für die folgende Analyse wurde entsprechend angenommen, dass diese analog zu den vorliegenden Daten ebenfalls in einem tabellarischen Format ausgetauscht werden, das zeilenweise bereits segmentierte Argumente und im Rücklauf zusätzlich die zugehörigen Erwiderungen in einer separaten Spalte enthält.

Eine Vollständigkeitsprüfung lässt sich entsprechend mittels einfacher regelbasierter Vergleiche ausführen. Dabei wird einerseits geprüft, ob der an den Vorhabenträger übersendete Argumentkatalog weiterhin vollständig und unverändert in der Synopse enthalten ist und andererseits erfolgt zu jedem Argument eine Prüfung, ob die zugehörige Erwiderungszeile gefüllt ist.

Zur Plausibilisierung der Relevanz sind zwei Ansätze, sowie deren Kombination denkbar. Zum einen können die zur Themenklassifikation von Argumenten vorgeschlagenen Methoden ebenfalls auf die Erwiderungen trainiert bzw. angewendet werden und so das Thema der Erwiderung bestimmen. Eine Plausibilisierung ist dann durch einen Abgleich mit den bereits identifizierten Argumentthemen möglich. Bei dieser Vorgehensweise ist zu beachten, dass die Erwiderungen neben den Themen auch generische Textbausteine, wie beispielsweise „Der Aufnahme der Auflagen wird zugestimmt." enthalten, die zu jedem Argument als relevant betrachtet werden können. Eine entsprechende Annotation ist nicht Bestandteil des vorliegenden Datensatzes und müsste entsprechend ergänzt werden.

Alternativ oder ergänzend können die vorliegenden Paare aus Argument und Erwiderung auch mittels Verfahren zur syntaktischen oder semantischen Ähnlichkeitsanalyse aus Abschnitt 6.1.2 plausibilisiert werden. Hierbei kann die Ähnlichkeit zwischen einem Einwendungssegment und der zugehörigen Erwiderung in Form eines numerischen Wertes berechnet werden. Anschließend kann mittels eines Abgleiches gegenüber vordefinierten Schwellen eine Einordnung in die in Abschnitt 3.4 definierten Kategorien grün, gelb und rot vorgenommen werden.

Um für ein derartiges Verfahren ein geeignetes Ähnlichkeitsmaß sowie passende Schwellenwerte zu wählen, sind neben den bereits im Datensatz vorhandenen Paaren aus Einwendungen und Erwiderungen, die bis auf wenige Ausnahmen plausibel sind, auch Negativbeispiele - also Einwendungssegmente mit unplausibler Erwiderung - erforderlich. Diese wurden aus dem Datensatz generiert, indem zu den vorhandenen Argumenten der Einwendungen jeweils zufällige Erwiderungen mit einem abweichenden Argumentthema ausgewählt wurden.

Für diesen Datensatz bestehend aus plausiblen und unplausiblen Paaren aus Einwendungssegmenten und Erwiderungen wurden verschiedene textuelle sowie semantische Ähnlichkeitsmaße berechnet. Zur Berechnung der textuellen Ähnlichkeit wurden die Textsegmente um Stoppworte bereinigt und der Jaccard-Index berechnet. Für die semantische Ähnlichkeit wurden die Verfahren GloVe, sBert und USE verglichen.

Um einen Vergleich der Verfahren zu ermöglichen wurden Precision-Recall-Kurven für diese vier Verfahren erstellt und gegenübergestellt. Hierzu wurden für eine Vielzahl möglicher Schwellenwerte jeweils Precision (Anteil der tatsächlich plausiblen Paare von den als plausible klassifizierten Paaren) und Recall

(Anteil der plausiblen klassifizierten Paare von allen plausiblen Paaren im Datensatz) berechnet. Die Precision-Recall-Kurve stellt diese Qualitätsmaße über alle Schwellenwerte hinweg gegenüber und ermöglicht so eine Einschätzung der Güte der Verfahren. Je besser ein Verfahren geeignet ist, desto größer ist die Fläche unterhalb der Kurve. Zudem unterstützt die Kurve die Auswahl konkreter Schwellenwerte.

Anhand Abbildung 6 wird deutlich, dass die Plausibilisierung auf Basis textueller Ähnlichkeit (Jaccard-Index) die besten Ergebnisse erzielt, während die semantischen Methoden auf Basis vortrainierter Sprachmodelle vergleichsweise ungeeignet für diese Problemstellung scheinen.

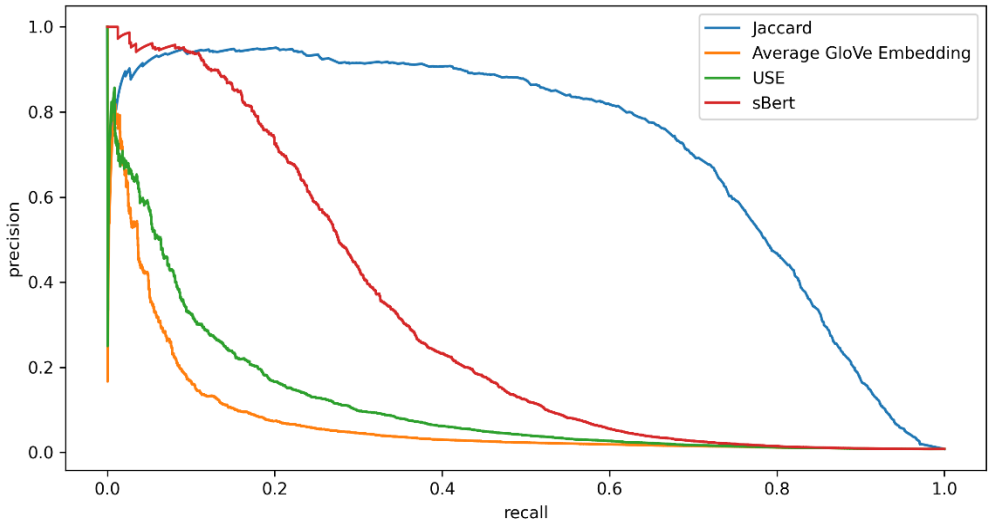


Abbildung 6: Precision-Recall-Kurven für die betrachteten Verfahren zur Plausibilisierung

Im Folgenden werden die vier vorgeschlagenen Lösungsskizzen hinsichtlich mehrerer Kriterien bewertet. Die daraus resultierende Bewertungsmatrix in Tabelle 11 fasst die Ergebnisse anhand der wesentlichsten Kriterien zusammen.

6.2.5 Kriterien

Folgende Kriterien wurden bei der Bewertung der Lösungsskizzen berücksichtigt:

Potenzieller Nutzen

Welche positiven Veränderungen gegenüber dem Status Quo sind durch die Implementierung einer Lösungsskizze grundsätzlich zu erwarten? Die Verbesserungen können quantitativer (schneller) sowie qualitativer (weniger Fehler) Natur sein. Dieses Kriterium ist vorrangig beschreibend. Es geht zwar in die Gesamtbetrachtung ein, es gibt jedoch hier keine eindeutig „besseren“ oder „schlechteren“ Werte.

Praktische Relevanz

Dieses Kriterium bildet bereits kommunizierte Einschätzungen des Auftraggebers hinsichtlich der erwarteten Nützlichkeit einer Lösung der jeweiligen Aufgabe ab. Es basiert auf Erfahrungswerten und wurde in Workshops in Erfahrung gebracht.

Mögliche Bewertungen:

- „hoch“: Eine Lösung der Aufgabe ist sehr wünschenswert, da davon ein hoher Praxisnutzen erwartet wird.
- „gering“: Eine Lösung der Aufgabe ist zwar wünschenswert, hätte aber nur begrenzten Praxisnutzen.

Realisierbarkeit mittels des vorliegenden Datensatzes

Umfang und Qualität eines Datensatzes setzen der Leistungsfähigkeit von Algorithmen des maschinellen Lernens Grenzen. Dieser Tatsache wird durch dieses Kriterium Rechnung getragen, indem eingeschätzt wird, wie gut realisierbar eine Lösungsskizze auf Grundlage der zur Verfügung stehenden Daten ist.

Mögliche Bewertungen:

- „ja“: Eine Lösung der Aufgabe auf der Basis der zur Verfügung stehenden Daten wird als ohne Einschränkungen realisierbar eingeschätzt.
- „eingeschränkt“: Eine Lösung der Aufgabe auf der Basis der zur Verfügung stehenden Daten wird als realisierbar eingeschätzt, allerdings bestehen Einschränkungen z. B. bezüglich der Ergebnisgüte, die evtl. mit einer erweiterten Datenbasis überwunden werden können.
- „nein“: Eine Lösung der Aufgabe auf der Basis der zur Verfügung stehenden Daten wird als nicht realisierbar eingeschätzt.

Erwartete Zuverlässigkeit

Welche Ergebnisgüte ist bei einer KI-Lösung im Vergleich zur manuellen Durchführung zu erwarten im Vergleich zur manuellen Durchführung?

Mögliche Bewertungen:

- „hoch“: Die zu erwartende Ergebnisgüte der maschinellen Verarbeitung wird als der (angenommenen) Güte der äquivalenten menschlichen Bearbeitung gleichwertig oder überlegen eingeschätzt.
- „manuelle Prüfung erforderlich“: Die zu erwartende Ergebnisgüte der maschinellen Verarbeitung kommt der (angenommenen) Güte der äquivalenten menschlichen Bearbeitung nahe, verbleibende Fehler sind aber möglich. – Diese sollten durch manuelle Prüfung eliminiert werden.
- „gering“: Mit der betrachteten maschinellen Verarbeitungslösung ist eine durchgehend substanziell schlechtere Ergebnisgüte als bei der äquivalenten menschlichen Bearbeitung zu erwarten.

Realisierungsrisiko

Die quantitativen Ergebnisse der dokumentierten Experimente stellen die wichtigste Grundlage der Einschätzung der Erfolgsaussichten bzw. Risiken der Umsetzung eines Lösungsvorschlags dar. Zusätzlich gehen aber auch qualitative Faktoren in diese Einschätzung ein. Z. B. versprechen neuere, häufig Deep-Learning-basierte Methoden grundsätzlich eine höhere Leistung, sind jedoch als experimenteller einzustufen und haben in der Regel hohe Anforderungen an die eingesetzte Hardware. Zudem ist die Aussagekraft von Versuchsergebnissen bei solchen Methoden erst bei größeren Datensätzen hoch. Ebenfalls ist der Reifegrad der für die Implementierung der Verfahren benötigten Softwarekomponenten zu berücksichtigen.

Unter Berücksichtigung dieser quantitativen und qualitativen Aspekte gibt dieses Kriterium eine summarische Einschätzung darüber ab, wie hoch das Risiko einer Umsetzung basierend auf den verwendeten Methoden und des vorliegenden Datensatzes ist.

Mögliche Bewertungen:

- „gering“: Bei der Umsetzung des Lösungsvorschlags sind Risiken weitgehend auszuschließen. Die quantitativen Experimentergebnisse sind überzeugend, die verwendeten Verfahren sind gut erforscht und breit in anderen Praxisanwendungen bewährt. Zur Implementierung sind ausgereifte Softwarekomponenten verfügbar.
- „mittel“: Die Aussichten für eine problemlose Realisierung sind gut, es bestehen jedoch Restrisiken, z. B. da die quantitativen Experimentergebnisse nicht ganz überzeugend ausfallen, da innovative, bisher nur im Forschungskontext ausreichend erprobte Verfahren verwendet werden, oder weil zur Implementierung nötige Softwarekomponenten nicht durchgehend ausgereift und praxiserprobt sind.
- „hoch“: Die Erfolgsaussichten einer Realisierung sind schwer zu beurteilen, da die quantitativen Experimentergebnisse nur knapp ausreichend sind, da noch sehr neue, kaum erprobte Verfahren Anwendung finden oder zur Implementierung keine geeigneten, nennenswert ausgereiften Softwarekomponenten verfügbar sind.

6.2.6 Segmentierung und Identifikation von Argumenten

Die Segmentierung und Identifikation von Argumenten ist ein mehrstufiger Prozess, der bei manueller Durchführung einen hohen Aufwand erzeugt, aufgrund dessen vergleichsweise geringer Komplexität jedoch eine geringe Fehlerrate aufweist. Dies macht diese Teilaufgabe grundsätzlich zu einem guten Kandidaten in Bezug auf die maschinelle Unterstützung. Erste Experimente auf dem vorliegenden Datensatz konnten die grundsätzliche Machbarkeit bestätigen. Wie in Abschnitt 6.2.1 dargelegt, erzielte die Kombination eines regelbasierten Segmentierers und der Nutzung der Shallow-Learning-Methode „Logistic Regression“ gute bis sehr gute Ergebnisse bei der Identifikation relevanter Segmente. Die maschinelle Unterscheidung von Argumenten und Hinweisen ist anspruchsvoller. Aufgrund des überproportionalen Vorkommens von Argumenten können allerdings auch für dieses Problem akzeptable Trefferquoten erreicht werden. So kann die maschinelle Unterstützung dieser Aufgabe die Einwendungbearbeitung potenziell beschleunigen.

Die hierfür eingesetzten Softwarekomponenten sind erprobt und die technischen Voraussetzungen an die IT-Infrastruktur sind als gering einzustufen, was die Integration der Lösung in existierende Softwarearchitekturen erleichtert.

6.2.7 Themenzuordnung von Argumenten

Bei der Themenzuordnung von Argumenten mittels Ober- und Unterthemen handelt es um zwei Mehrklassen-Klassifikationsprobleme. Der menschliche Aufwand bei manueller Durchführung des Vorgangs ist hierbei geringer als bei Segmenttrennung und Argumentidentifikation, da nur die extrahierten relevanten Segmente bearbeitet werden müssen. Die menschlichen Entscheidungen können jedoch subjektiv von der Person abhängig sein. Maschinelle Unterstützung kann bei dieser Aufgabe zu konsistenteren Ergebnissen beitragen. Zudem ist davon auszugehen, dass die Bearbeitung auf Grundlage der mit maschineller Unterstützung vorliegenden Klassifizierungsvorschläge deutlich erleichtert wird.

Im Experiment erreichte die KI-Lösung gute Genauigkeiten für die Mehrzahl der Themen. Für einige Themen ist die Anzahl der Trainingsbeispiele zu gering, um gute Ergebnisse mit Lernverfahren zu erzielen. Bei einer Verbesserung der Datenlage ist jedoch anzunehmen, dass auch diese Themen mit den vorgeschlagenen Methoden zuordenbar sind.

6.2.8 Erkennung von Argumentationsmustern

Im Gegensatz zu den beiden erstgenannten Lösungen, handelt es sich bei der Erkennung von Argumentationsmustern um ein Suchproblem, bei dem ein betrachtetes Segment mit allen anderen Segmenten in der Datenbasis verglichen wird. Hierbei soll den Sachbearbeitenden die Möglichkeit gegeben werden, nach identischen oder ähnlichen Argumentationen in früher eingereichten Einwendungen zu suchen und so Mehrfacheinreichungen und Mustereinwendungen zu erkennen. Dieser Prozess ist ohne maschinelle Unterstützung, insbesondere bei einem großen Datenbestand, kaum zuverlässig durchführbar.

Klassische Ähnlichkeitsmaße wie der Jaccard-Index ermöglichen hierbei die Suche nach lexikalisch ähnlichen Segmenten im Gesamtbestand. Modernere Deep-Learning-basierte Methoden erlauben auch den Vergleich auf semantischer Ebene (Paraphrasierungen), haben jedoch deutlich höhere Anforderungen an die eingesetzte Hardware, benötigen Trainingsbeispiele in Form von ähnlichen Satzpaaren und befinden sich noch in einem experimentelleren Stadium.

Die bei der manuellen Sichtung des vorliegenden Datensatzes identifizierten Duplikate wurden als solche markiert und konnten in Versuchen mittels Jaccard-Index aufgespürt werden.

6.2.9 Prüfung von Erwiderungen

Die Prüfung von Erwiderungen teilt sich in die Vollständigkeits- und Plausibilitätsprüfung auf. Die Betrachtung dieser Lösungsskizze basiert auf der Annahme, dass der Rücklauf des Vorhabenträgers grundsätzlich in Form von Synopsen, also einer tabellarischen Gegenüberstellung von Argumenten und Erwiderungen, vorliegt.

Unter diesen Umständen kann der Aufwand der Vollständigkeitsprüfung sowohl bei menschlicher als auch maschineller Durchführung als gering bewertet werden.

Bei der Plausibilitätsprüfung handelt es sich um einen deutlich anspruchsvolleren Vorgang, da dieser eine inhaltliche Auseinandersetzung mit den Textstellen erfordert.

Um die Leistung eines maschinellen Verfahren in diesem Prozess bewerten zu können, werden Beispiele plausibler und nicht-plausibler Argument-Erwiderungspaare benötigt. Da diese im Datensatz nicht vorlagen, mussten diese, wie in Abschnitt 6.2.4 beschrieben, konstruiert werden. Entsprechend ist die Entwicklung und Bewertung dieser Lösung mittels des vorliegenden Datensatzes nur eingeschränkt möglich.

Für eine genauere Potenzialabschätzung der maschinellen Unterstützung dieses Prozessschrittes, wäre ein Datensatz bestehend aus Beispielen praktisch anzutreffender Fehler, sowie genauere Kenntnis über die durchgeführten Prüfschritte und die Vorgehensweise der Vorhabenträger bei der Erwiderungserstellung nötig.

6.2.10 Zusammenfassung

Allen Lösungen ist gemein, dass deren Zuverlässigkeit und weitere Verbesserung über längere Zeiträume nur dann gewährleistet ist, wenn manuelle Korrekturen der Ergebnisse der eingesetzten KI-Verfahren konsequent gesammelt und periodisch als weitere Trainingsdaten zur Verfügung gestellt werden. Diesbezüglich ist positiv anzumerken, dass eine weitgehende Automatisierung dieses Prozesses durch ein entsprechendes Prozess- und Systemdesign möglich ist.

Ebenfalls anzumerken ist, dass die Ergebnisse von Shallow-Learning-Methoden durch den Einsatz von Deep-Learning potenziell gesteigert werden können. Sie stellen jedoch deutlich höhere Anforderungen

an die eingesetzte Hardware und Menge von notwendigen Trainingsbeispielen, weshalb im Zweifel eine Kosten-Nutzen-Abwägung erfolgen sollte. Zunächst erscheint eine Potenzialabschätzung anhand der vorgestellten Shallow-Learning-Methoden aussichtsreich. Mittels der daraus resultierenden Erkenntnisse und einer entsprechenden Referenzimplementierung können mögliche Verbesserungen durch den Einsatz von Deep-Learning-Verfahren geprüft werden.

TABELLE 11: ZUSAMMENFASSENDE BEWERTUNGSMATRIX

	Aufgabe 1: Segmentierung und Identifikation von Argumenten	**Aufgabe 2:** Themenzuordnung von Argumenten	**Aufgabe 3:** Erkennung von Argumentationsmustern	**Aufgabe 4:** Prüfung von Erwiderungen
Potenzieller Nutzen	schnellere Bearbeitung des Einwendungseinganges	konsistentere Bearbeitung des Einwendungseinganges	schnellere und konsistentere Bearbeitung des Einwendungseinganges	schnellere Prüfung des Rücklaufes vom Vorhabenträger
Praktische Relevanz	hoch	hoch	gering	gering
Realisierbarkeit mittels des vorliegenden Datensatzes	ja	ja	ja	eingeschränkt
Erwartete Zuverlässigkeit	manuelle Prüfung erforderlich	manuelle Prüfung erforderlich	hoch	gering
Realisierungsrisiko	mittel	mittel	gering	hoch

7 Anwendung und Potentialabschätzung

Basierend auf den Ergebnissen der in Kapitel 6 erfolgten Analyse und Bewertung der MTV-Verfahren und darauf aufsetzenden Lösungsentwürfen erfolgte im Projekt die endgültige Auswahl und Festlegung eines Anwendungsszenarios. Das ausgewählte Szenario umfasst die Segmentierung der Einwendungen, die Erkennung relevanter Segmente und Argumente sowie die thematische Zuordnung der Segmente und wird in Abschnitt 7.1 näher beschrieben und begründet.

Für dieses ausgewählte Szenario wurde sodann ein PoC-Prototyp realisiert. Hierzu wurden einerseits vorgestellte Lösungsskizzen für die genannten, ausgewählten Aufgaben in Form ausführbarer Software realisiert, andererseits wurde eine Anwenderoberfläche entworfen, die diese Funktionalitäten integriert. Technische Einzelheiten der Implementierung sind Abschnitt 7.2 zu entnehmen.

Im Ergebnis entstand ein PoC-Prototyp, der die eingesetzten Verfahren direkt aus Anwendersicht erlebbar macht und aufzeigt, was ein KI-unterstütztes Softwaresystem leisten kann und wie eine Integration in die Bearbeitungsabläufe möglich ist. Dieser PoC-Prototyp wurde mittels einer Auswahl von Einwendungen erprobt und bewertet, um so das Potenzial der vorgeschlagenen Konzepte zur KI-Unterstützung im Einwendungsmanagement für den praktischen Einsatz abzuschätzen. Die Nutzung des PoC-Prototypen und die damit in der Schlussphase des Projekts erzielten Ergebnisse sind in Abschnitt 7.3 dokumentiert.

7.1 Festlegung des Anwendungsszenarios

Für die prototypische Umsetzung war ein Anwendungsszenario zu definieren. Das Anwendungsszenario orientiert sich an den in Kapitel 3 beschriebenen Aufgaben und spannt den Rahmen, um die Güte und Performanz der jeweils bestgeeignetsten MTV-Verfahren zu untersuchen und belastbare Erkenntnisse für eine zukünftige Umsetzung im Einwendungsmanagement zu erlangen.

Eine wesentliche Rahmenbedingung bei der Festlegung des Anwendungsszenarios ist die Realisierbarkeit unter Berücksichtigung des vorliegenden Datensatzes. Die Bewertung der Lösungsentwürfe in Abschnitt 6.2 zeigt, dass die Aufgaben 1 bis 3 mit dem vorliegenden Datensatz grundsätzlich umgesetzt werden können. Bei Aufgabe 4 (Prüfung von Erwiderungen) ist die Realisierbarkeit mittels des vorliegenden Datensatzes eingeschränkt.

Weitere Entscheidungskriterien für die Festlegung des Anwendungsszenarios sind zu erwartende Erkenntnisgewinne durch die prototypische Umsetzung sowie die fachliche Relevanz der Aufgabe für den Prozess des Einwendungsmanagements im EBA.

Hinsichtlich des zu erwartenden Erkenntnisgewinns durch die Umsetzung sind die Aufgabe 1 (Segmentierung) und die Aufgabe 2 (Themenzuordnung) hoch zu bewerten. Dies gilt insbesondere für den Einsatz von Shallow-Learning-Verfahren zur Lösung von Klassifikationsproblemen, der in der Umsetzung als anspruchsvoll einzuschätzen ist. In Abgrenzung dazu erscheint die Umsetzung von Aufgabe 3 (Erkennung von Argumentationsmustern) einen geringeren Erkenntnisgewinn zu bringen. Es ist ersichtlich, dass sich Argumentationsmuster gut automatisiert erkennen lassen und dass die Aufgabe mit vergleichsweise einfachen Methoden umgesetzt werden kann. Vor dem Hintergrund scheint kein Beweis durch eine prototypische Umsetzung erforderlich zu sein.

Hinsichtlich der fachlichen Relevanz weisen insbesondere die Aufgaben 1 und 2 eine hohe Bedeutung auf und versprechen, den erforderlichen manuellen Aufwand im Bearbeitungsprozess erkennbar zu reduzieren. Hinzu kommt, dass die beiden Aufgaben den Anfang des Verarbeitungsprozesses darstellen und im PoC ohne (simulierte) Interaktion zwischen EBA und Vorhabenträger abgebildet werden können. Dem gegenüber wäre für eine prototypische Umsetzung von Aufgabe 4 (Prüfung von Erwiderungen) eine (simulierte) Interaktion zwischen EBA und Vorhabenträger erforderlich. Vor dem Hintergrund erscheint die Umsetzung von Aufgabe 4 im Rahmen der vorliegenden Studie weniger geeignet.

Zusammenfassend zeigt sich, dass die Umsetzung von Aufgabe 1 (Segmentierung) und Aufgabe 2 (Themenzuordnung) für die vorliegende Studie den höchsten Mehrwert verspricht. Inhaltlich baut die Themenzuordnung auf die Ergebnisse der Argumenterkennung auf. Die kombinierte Bearbeitung beider Aufgaben in einem PoC kann direkt aufzeigen, wie die beschriebenen MTV-Verfahren im praktischen Einsatz der Einwendungsbearbeitung wirken. Entsprechend erfolgt die Entwicklung eines PoC, in dem diese beiden Aufgaben kombiniert realisiert werden.

7.2 Entwicklung des PoC

7.2.1 Architektur

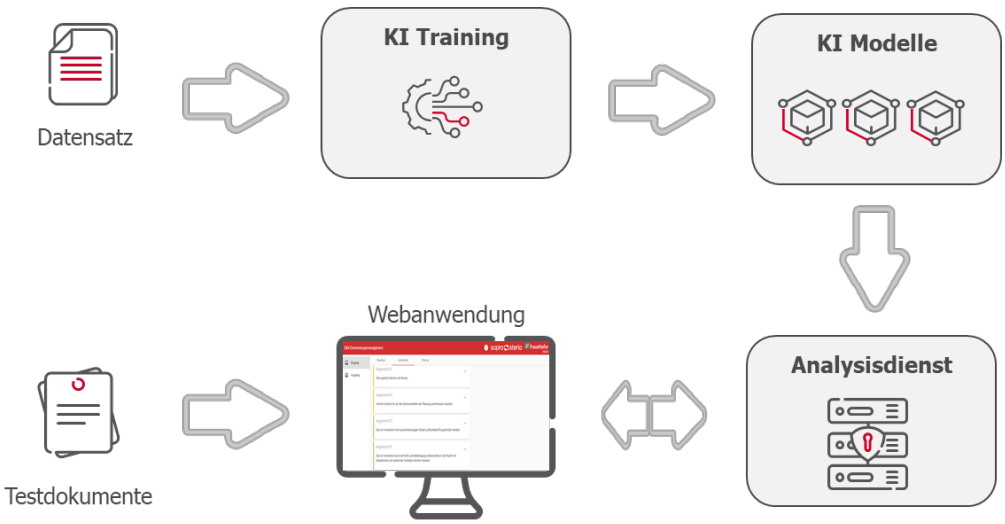


Abbildung 7: Architektur des Proof-of-Concept

Als PoC wurde ein Demonstrator entwickelt, der es mit einer beispielhaften, aber bereits Elemente der Anwendersicht berücksichtigenden Nutzerschnittstelle ermöglicht, die grundsätzliche Eignung des Lösungsansatzes für einen zukünftigen Praxiseinsatz im Rahmen des gegebenen Fachprozesses realistisch einzuschätzen. Der Demonstrator umfasst zwei wesentliche Softwarekomponenten. Dies ist zum einen eine Anwendungsoberfläche, die den Anwender durch den Bearbeitungsprozess führt, und zum anderen ein Analysedienst, der die unterstützenden Verarbeitungsdienste realisiert. Der Analysedienst verwendet KI-Modelle zur Realisierung der Klassifikationsaufgaben. Diese wurden anhand des um einige Testdokumente bereinigten Datensatzes erstellt. So ermöglicht die Verwendung dieser Testdokumente im Demonstrationsszenario eine realistische Einschätzung der Leistungsfähigkeit der Verfahren. Technisch wurden für die Modellerstellung Python-Notebooks und die Bibliotheken pandas, sklearn und imblearn verwendet. Die resultierenden Modelle wurden mittels der Bibliothek pickle als Dateien gespeichert und werden vom Analysedienst geladen.

7.2.2 Datenauswahl und Modellerstellung

Für die Vorführung des Demonstrators wurden zunächst 15 Dokumente aus dem vom Auftraggeber übermittelten Beispieldatensatz selektiert und vom Training des KI-Modells ausgeschlossen. Hierbei wurde darauf geachtet, dass eine ausgewogene Verteilung der Einwendungsquellen (privat und TöB), -formen (handschriftlich und digital) und -längen vorlag. Insgesamt wurden zehn der elf Oberthemen (siehe Tabelle 13) durch diese Auswahl abgedeckt. Um eine möglichst reale Bearbeitungssituation darzustellen, wurden auch Einwendungen mit vergleichsweise geringer Qualität in Bezug auf Grammatik und Rechtschreibung inkludiert. Hier wurde allerdings darauf geachtet, dass ausreichend Beispiele im Trainingsdatensatz verbleiben.

Mittels des verbleibenden Trainingsdatensatzes wurden die Modelle für die einzelnen Aufgaben (Relevanzerkennung, Argumentidentifikation und Themenerkennung) erstellt. Hierfür wurde entsprechend der Ergebnisse der Untersuchungen das Verfahren Logistische Regression eingesetzt.

Zur Optimierung der Modelle wurden verschiedene Vorverarbeitungstechniken angewendet und die jeweils beste Kombination gewählt. Diese Vorverarbeitungstechniken umfassen:

- Bereinigung um Sonderzeichen
- Entfernen von Stoppworten, Interpunktion und Lemmatisierung der Texte
- Maskierung von Datumsangaben und Adressen
- Reduktion des Textes auf Verben und Substantive anhand von POS-Tags

Daneben wurden verschiedene Resampling-Methoden während des Trainings eingesetzt. Resampling kann im Falle von unbalancierten Trainingsdaten die Qualität der erlernten Modelle verbessern. Folgende Techniken wurden hierfür in Betracht gezogen:

- Random Undersampling
  Bei dieser Methode werden zufällig gewählte Datensätze aus der überrepräsentierten Klasse verworfen.
- Random Oversampling
  Bei dieser Methode werden zufällig gewählte Datensätze aus der unterrepräsentierten Klasse dupliziert.
- SMOTE
  Mit dieser Technik werden während des Trainings synthetische Datenpunkte einer unterrepräsentierten Klasse erzeugt, indem bestehende Datenpunkte interpoliert werden.

Für jede Klassifikationsaufgabe wurde die jeweils erfolgreichste Kombination dieser Methoden bezüglich des erzielten F1-Score eingesetzt. Diese Kombinationen sind in Tabelle 12 zusammengefasst.

Die Themenerkennung konnte zudem durch ein mehrstufiges Modell verbessert werden. Hierzu wurden zunächst Klassifikatoren für die Themen erstellt. Anschließend wurde ein weiterer Klassifikator erstellt, der basierend auf diesem Ergebnis in Kombination mit den Klassifikationsergebnissen von Relevanz- und Argumenterkennung die finale Klassifikation berechnet. So ließ sich die Gesamtqualität und insbesondere die Precision für die Themenerkennung verbessern.

Die gelieferten Klassifikationsergebnisse lassen sich anhand von Schwellenwerten beeinflussen. Die Modelle ermitteln eine Wahrscheinlichkeit, mit der ein Datenpunkt einer Klasse zugeordnet wird und die letztendliche Klassifikation erfolgt anhand einer Wahrscheinlichkeitsschwelle. So lässt sich insbesondere die Balance zwischen *Precision* und *Recall* verschieben. Die Wahl dieser Schwellen erfolgte grundsätzlich

TABELLE 12: ERFOLGREICHSTE KOMBINATION AUS VORVERARBEITUNGS- UND RESAMPLING-VERFAHREN

Klassifikator	**Vorverarbeitung**	**Resampling**
Relevanzerkennung	Bereinigung Lemmatisierung Maskierung	SMOTE
Argumenterkennung	Bereinigung	SMOTE
Themenerkennung	Bereinigung Lemmatisierung Maskierung	Random Oversampling

so, dass der optimale *F1-Score* erzielt wird. Im Falle der Themenerkennung kann es sinnvoll sein von diesem Grundsatz abzuweichen und alternative Schwellenwerte zu verwenden. In der Anwendung sind Warnungen für unzuverlässige Themenzuordnungen vorgesehen. Entsprechend kann ein höherer Schwellenwert gewählt werden, sodass ein hoher Precision-Score (ein erkanntes Thema ist mit hoher Wahrscheinlichkeit korrekt) erzielt wird. Diese Themen können dann ohne Warnung angezeigt werden. Die Auswahl dieser Schwelle kann durch Betrachtung der Precision-Recall-Kurven und manuelle Auswahl markanter Punkte erfolgen. Abbildung 8 veranschaulicht die Wahl der Schwellenwerte anhand der Precision-Recall-Kurve für den Themenklassifikator „Erschütterungen“. Der orangene Punkt erzielt den optimalen *F1-Score* von 0,7 und einen guten *Recall* von 0,84, jedoch nur eine geringe *Precision* von 0,62. Der blaue Punkt erzielt eine Precision von 0,8, während der Recall mit 0,4 geringer ausfällt.

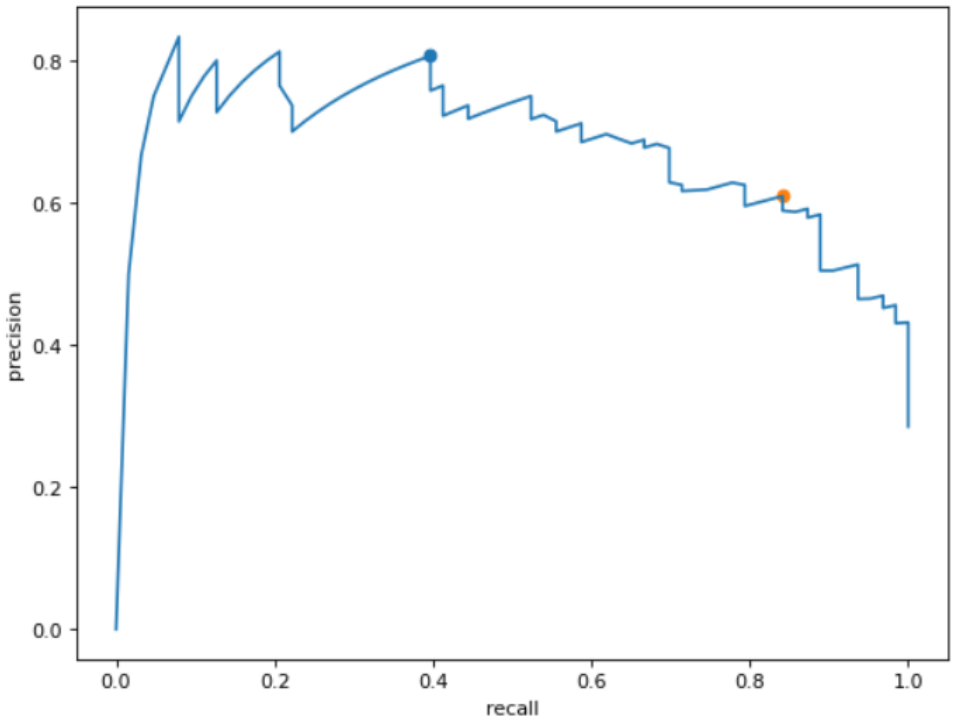


Abbildung 8: Wahl der Schwellenwerte für die Themenklassifikation "Erschütterungen" anhand der Precision-Recall-Kurve des Klassifikators

7.2.3 Anwendungsoberfläche

Die Anwenderoberfläche wurde als Single-Page-Application mittels Angular.js realisiert. Angular.js ist ein JavaScript-Framework auf Open-Source-Basis zur Realisierung von Weboberflächen.

Die realisierte Anwenderoberfläche führt durch den Bearbeitungsablauf und demonstriert dabei die KI-Unterstützung für die drei Aufgaben des Anwendungsszenarios:

- Segmentierung von Einwendungen
- Identifikation relevanter Segmente und Argumente
- Themenzuordnung von Argumenten

Aus Sicht der nutzenden Personen stellt sich der Arbeitsablauf wie folgt dar:

1. Öffnen der Startseite, Login

In der im Projektverlauf bereitgestellten Onlineversion der Anwendung ist zuerst ein Login mit den für die legitime Nutzung benötigten Zugangsdaten erforderlich (Abbildung 9).

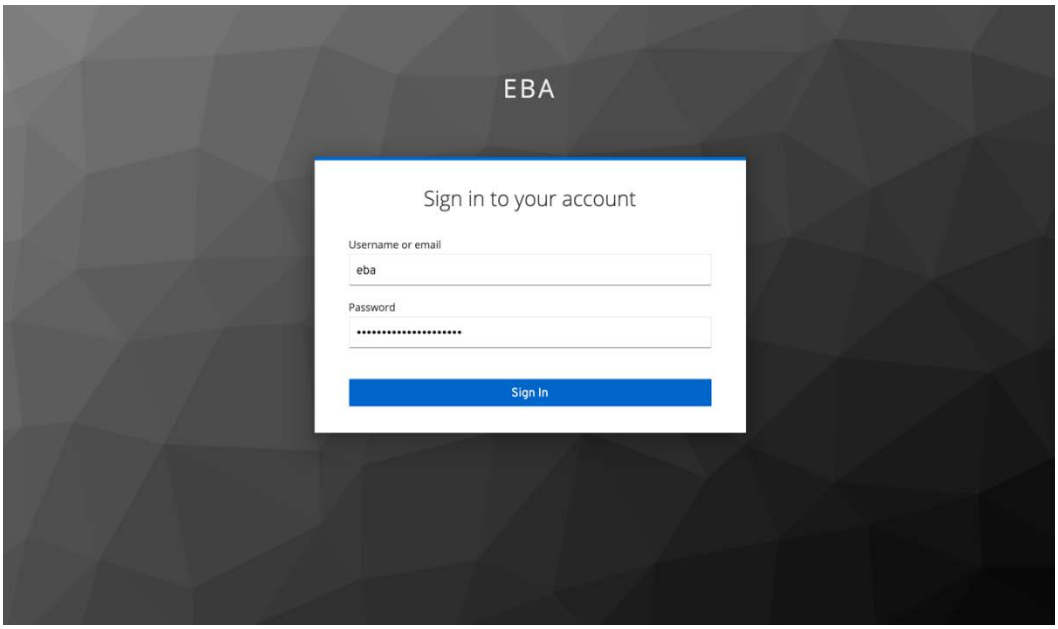


Abbildung 9: Anmeldeseite der Anwendung

2. Überblicksansicht

Hier kann ein Dokument für die weitere Bearbeitung ausgewählt werden (Abbildung 10).

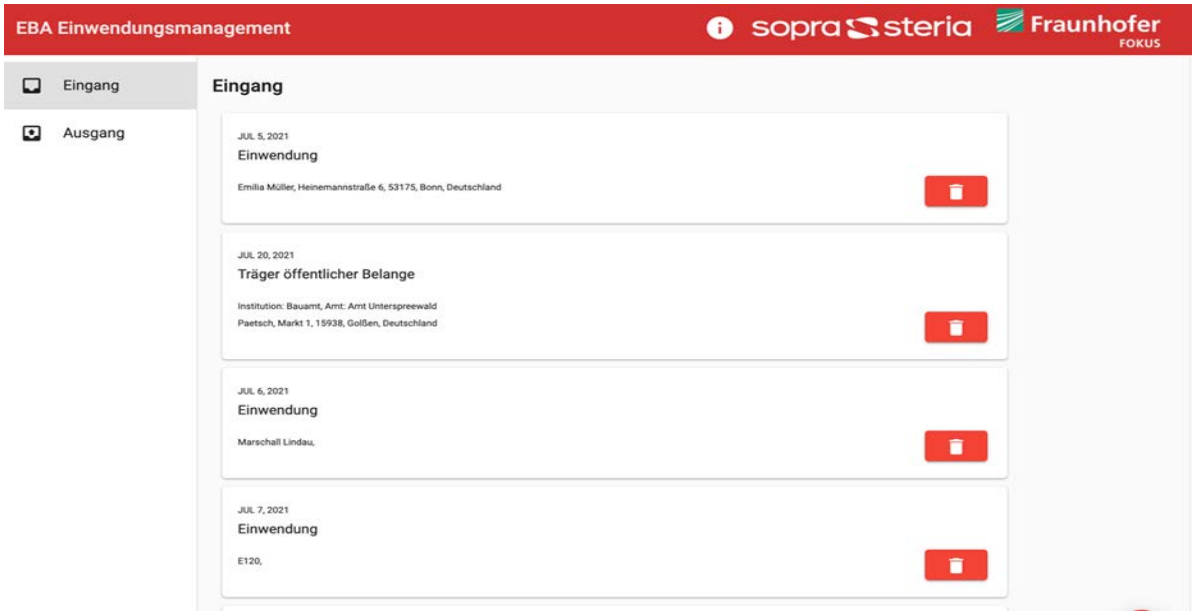


Abbildung 10: Überblicksansicht eingegangener Einwendungen

3. Detailansicht des Originaldokuments

Nach Auswahl eines Dokuments gelangt man zunächst in die Dokumentenansicht (Abbildung 11), in der das vollständige Schreiben auf der linken Seite und Felder für zugehörige Metadaten auf der rechten Seite dargestellt werden. Die Metadaten können in dieser Detailansicht manuell eingegeben oder überarbeitet werden.

Die Navigation durch den Workflow kann grundsätzlich sowohl über die obere Navigationsleiste als auch im Bereich rechts unten erfolgen.

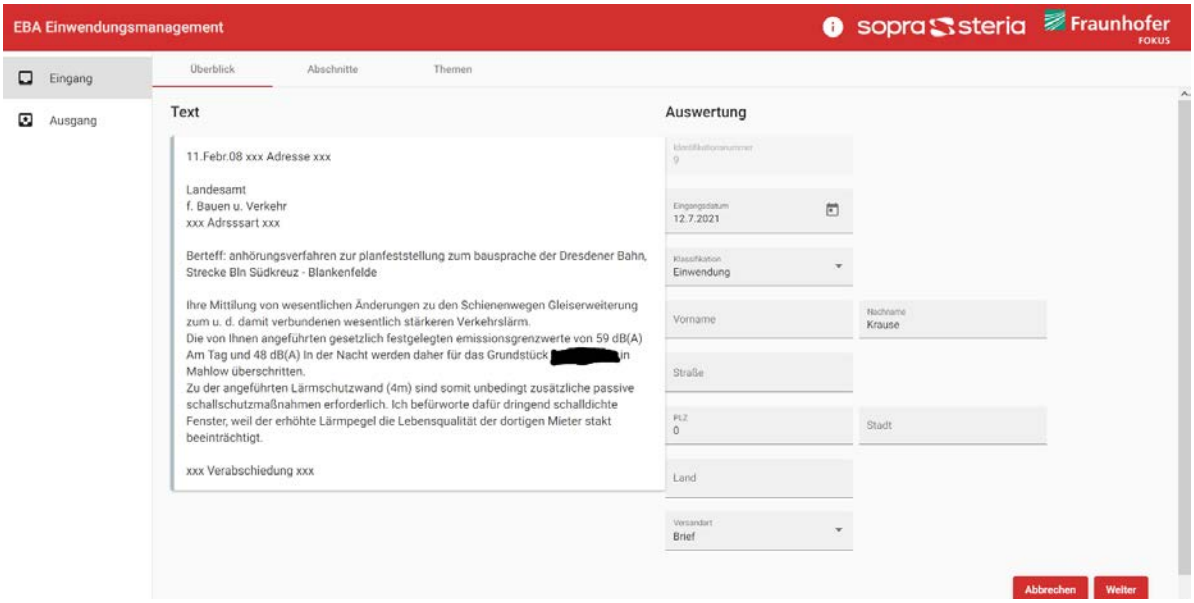


Abbildung 11: Dokumentenansicht einer Einwendung mit deren Metadaten

4. Darstellung der Textsegmente des ausgewählten Dokuments

Über die Schaltfläche “Abschnitte” in der oberen Navigationsleiste oder durch Klicken auf ”Weiter” gelangt man zum Ergebnis der Segmentierung und Segmentanalyse (Abbildung 12). Auch hier besteht die Möglichkeit, manuelle Korrekturen durchzuführen

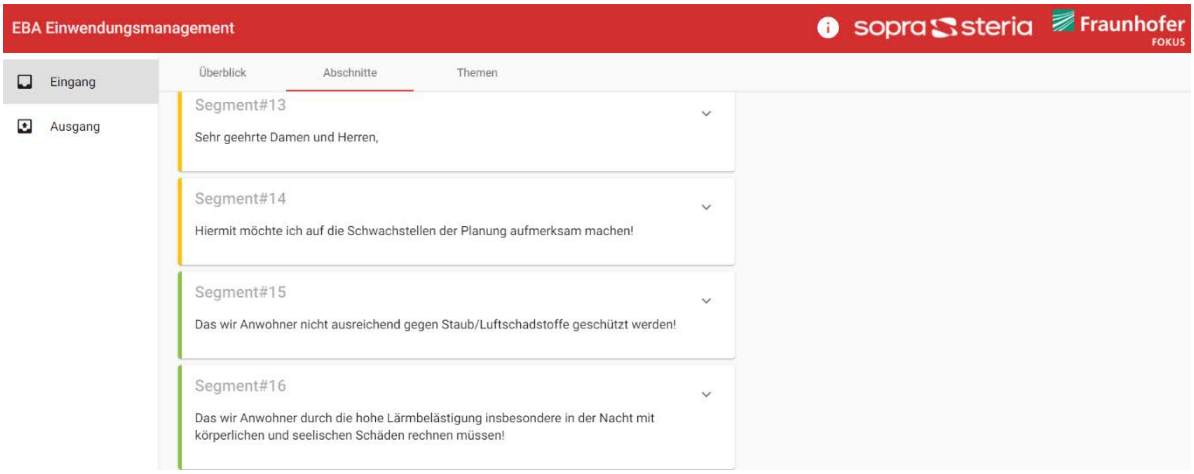


Abbildung 12: Ergebnis der Segmentierung und Segmentanalyse

5. Darstellung der ermittelten Themenzuordnungen

Beim Aufruf der nächsten Ansicht “Themen” wird die Zuordnung der Argumente zu Ober- bzw. Unterthemen ermittelt und dargestellt (Abbildung 13). Der Konfidenzwert dient hier als unterstützendes Element für den Nutzer bei der Einschätzung der Richtigkeit der KI-Auswertung. Das Hinzufügen von zusätzlichen Themen, aber auch die Korrektur der vorgeschlagenen Themen können hier über entsprechende Bedienelemente am jeweiligen Segment vorgenommen werden.

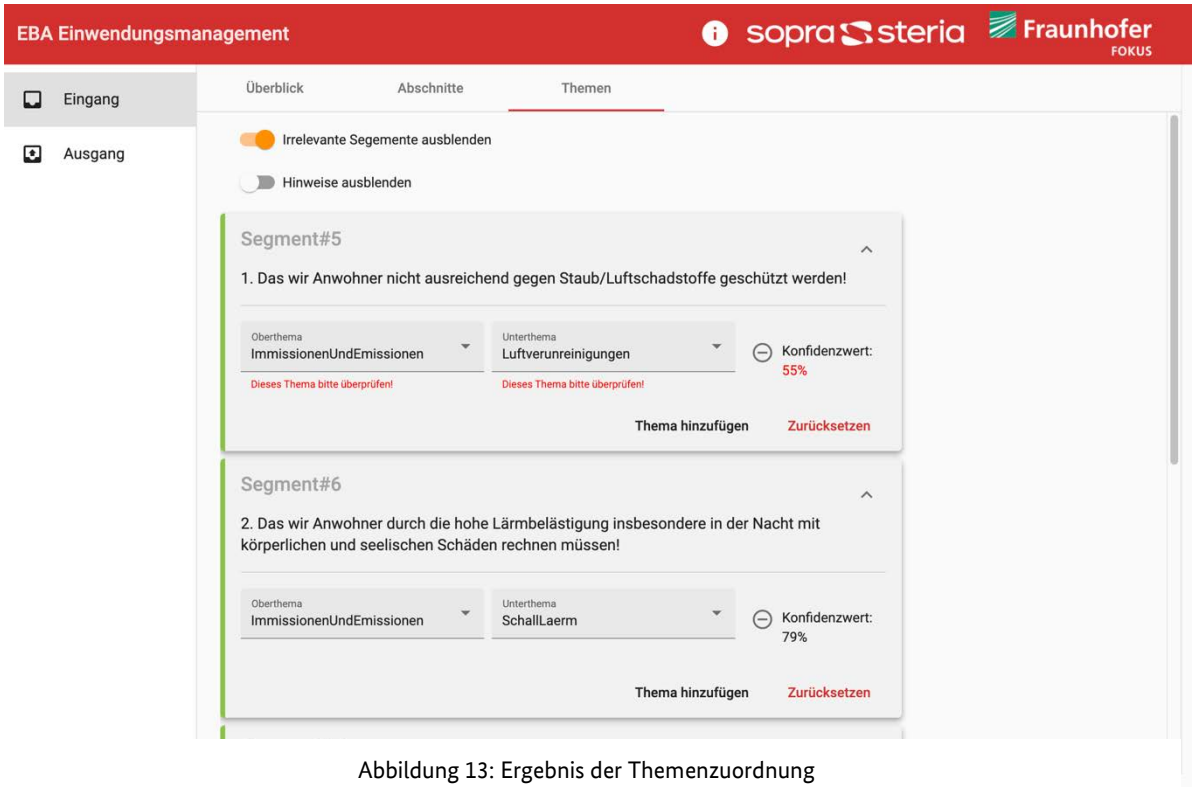


Abbildung 13: Ergebnis der Themenzuordnung

Über den Button “Bestätigen” gelangt man nach Abschluss der Bearbeitung zurück in die Überblicksansicht, und hat dort die Möglichkeit, mit der Bearbeitung weiterer Einwendungen fortzufahren.

Bereits abgeschlossene Einwendungen können im Ausgangbereich (Schaltfläche links) vor dem Absenden erneut aufgerufen, überprüft und korrigiert werden.

Die in diesem Workflow anfallenden Verarbeitungsaufgaben werden vom Analysedienst (vergleiche nächster Abschnitt) realisiert. Hierfür versendet die Anwendung zu verarbeitende Dokumente über HTTPS (also verschlüsselt und abgesichert) an den Analysedienst und empfängt auf dem gleichen Weg aufbereitete Daten von diesem.

7.2.4 Analysedienst

Der Analysedienst wurde mittels der Sprache *Python* und dem Framework FastAPI erstellt. FastAPI ist ein als Open-Source-Software veröffentlichtes Python-Framework, das die Bereitstellung von REST-Schnittstellen ermöglicht.

Für den PoC simuliert der Analysedienst zum einen ein Dokumentenmanagement, indem er die ausgewählten Testdokumente als abrufbare JSON-Datei bereitstellt. Daneben empfängt der Analysedienst für das im PoC realisierte Anwendungsszenario über eine zweite Schnittstelle ein im JSON-Format codiertes Dokument und liefert als Antwort die Liste der getrennten Segmente und zu jedem Segment die Ergebnisse der Klassifikatoren für Relevanzerkennung, Argumenterkennung und Argumentthema sowie die zugehörigen Konfidenzwerte. Hierfür greift der Dienst auf regelbasierte Logik zur Segmenttrennung zurück und verwendet vorab erstellte Modelle für die Klassifikationsaufgaben.

Um stets in sich konsistente Gesamtergebnisse zu gewährleisten, weicht die implementierte Lösung teilweise von dem zuvor skizzierten Vorschlag ab. Segmenttrennung und Relevanzerkennung wurden entsprechend der obigen Ansätze realisiert. Anschließend werden Unterthemen für die als relevant erkannten Segmente ermittelt. Falls für ein Segment mindestens ein Thema identifiziert wurde, wird automatisch das zugehörige Oberthema zugeordnet und die Kennzeichnung als Argument vorgenommen. Wird kein Unterthema identifiziert, werden die Klassifikatoren für Oberthemen ausgeführt. Erkennen diese ein Thema erfolgt eine Kennzeichnung als Argument. Verbleibende relevante Segmente werden mittels des Klassifikators für Argumenterkennung in Argumente und Hinweise differenziert.

7.2.5 Software-Bereitstellung

Die gewählte Architektur folgt dem Designparadigma moderner Webapplikationen, sodass die Bereitstellung sowohl auf lokaler Infrastruktur als auch unter Einsatz von Clouddiensten erfolgen kann. Entsprechend wurden die beiden zentralen Komponenten in Form von Webservern realisiert. So lassen sich diese zentral warten und können weitgehend unabhängig von der Hard- und Softwareausstattung der Anwenderarbeitsplätze genutzt werden.

Softwaretechnisch wurden Anwendung und Analysedienst als Containerimages realisiert. Diese Bereitstellungsform ermöglicht weitgehende Kompatibilität zu verschiedenen Infrastrukturen, da Softwareabhängigkeiten einschließlich Version und Konfiguration Bestandteil der Containerimages sind und entsprechende Containerlaufzeitumgebungen für alle gängigen Serversysteme verfügbar sind. Daneben bietet containerbasierte Software Vorteile in Bezug auf Skalierung und benötigte Hardware. Container lassen sich schneller starten und stoppen. Sie verbrauchen weniger Arbeitsspeicher als virtuelle Maschinen und nutzen somit die zugrundeliegende Hardware wirtschaftlicher.

Das ausführbare Gesamtsystem wurde mittels „Docker-Compose“ paketiert. Dabei handelt es sich um ein Tool, das es ermöglicht eine containerbasierte Applikation samt ihrer Konfiguration so zu paketieren, dass eine Bereitstellung ohne tiefere Kenntnisse der Anwendung möglich ist.

Parallel wurde bereits frühzeitig im Projektverlauf eine Onlineversion der Anwendung bereitgestellt. Hierzu wurde IT-Infrastruktur von Sopra Steria und Fraunhofer FOKUS auf Basis von OpenShift genutzt. Um dabei Anforderungen an Kommunikationssicherheit und Vertraulichkeit Rechnung zu tragen, wurden zusätzliche Sicherheitsmaßnahmen realisiert. Diese umfassen SSL-basierte Kommunikationsverschlüsselung, Nutzerauthentifizierung für die Anwendung und den Analysedienst sowie eine Komponente zur Nutzerverwaltung. Hierfür wurden die Open-Source-Softwarepakete HAProxy, Nginx und Keycloak mit geeigneten Konfigurationen eingesetzt.

7.3 Evaluierung des PoC und der damit erzielten Testergebnisse

Der Demonstrator wurde vom Projektteam getestet und bewertet sowie im Rahmen einer Ergebnispräsentation vorgestellt und dort einer ersten Bewertung durch den Auftraggeber unterzogen.

Ausführliche Tests wurden sowohl während der Entwicklung des Demonstrators als auch nach dessen Fertigstellung durchgeführt. Die während der Entwicklung kontinuierlich durchgeführten Tests hatten primär das Ziel, sicherzustellen, dass die ausgewählten Verfahren korrekt umsetzt und fehlerfrei an die Nutzerschnittstelle angebunden wurden, sowie über die Dauer der Entwicklung eventuell auftretende Regressionsfehler auszuschließen. Außerdem wurde die Eignung des Demonstrators für die spätere testweise Nutzung durch Personen aus dem realen Anwendungskontext (oder mit vergleichbarer Expertise) bereits in diesem Stadium kontinuierlich bewertet und optimiert.

Die Tests nach Abschluss der Entwicklung dienten einerseits dazu, inhaltlich die Leistungsfähigkeit der implementierten Verfahren endgültig zu validieren und zu quantifizieren. Andererseits wurden die Tests vorgenommen, um zu belegen, dass das intendierte fachliche Szenario im Demonstrator ausreichend abgebildet wird und dass dieser eine praxisnahe Bewertung der Funktionalität und Leistungsfähigkeit der untersuchten Verfahren auch durch Fachexperten und Nutzer (also ggf. auch ohne tiefere einschlägige KI- und IT-Kenntnisse) ermöglicht.

Im Folgenden sind mit dem *Proof-of-Concept* erzielten quantitativen Ergebnisse auf Basis der ausgewählten Testdokumente dokumentiert.

Der erste Verarbeitungsschritt trennt die 15 ausgewählten Testdokumente in insgesamt 243 Segmente. Davon entsprechen 197 Segmente solchen im Datensatz und 46 Segmente entsprechen Fragmenten von Segmenten des Datensatzes. Die automatisierte Trennung erfolgte also feiner als im ursprünglichen Datensatz. In diesen Fällen handelt es sich um Absenderangaben, die in den Segmenten des Datensatzes zusammengefasst wurden. Insgesamt stellt die feinere Segmenttrennung allerdings kein Hindernis für die nachfolgenden Verarbeitungsschritte dar.

Von den 243 erkannten Segmenten sind 148 relevant und bei 119 handelt es sich um Argumente. Die Häufigkeitsverteilung der Klassen in diesem Testdatensatz weicht deutlich von der des zum Training genutzten Restdatensatz ab.

Die quantitative Bewertung der Klassifikatoren erfolgt an dieser Stelle wiederum anhand der in Kapitel 6 spezifizierten und genutzten Bewertungsmetriken. Somit werden die Resultate der Modelle mit den in AP 1 durch menschliche Entscheidung vorgenommenen Bewertungen verglichen.

Die Relevanzerkennung erzielt über die Testdokumente eine *Accuracy* von 0,89, einen *F1-Score* von 0,91 und *MCC* von 0,78. Damit fallen die gemessenen Werte zwar aufgrund der abweichenden Datenzusammensetzung niedriger aus als angenommen, es werden jedoch 144 (97 %) der Segmente korrekt als relevant erkannt.

Die Argumenterkennung erzielt eine *Accuracy* von 0,85, einen *F1-Score* von 0,84 und *MCC* von 0,69 über dem Testdatensatz. Damit gelingt es 100 (85 %) der vorhandenen Argumente als solche zu erkennen.

Die Themenerkennung ist aufgrund der sehr geringen Anzahl von Beispielen für die einzelnen Themen im Testdatensatz nur bedingt quantitativ aussagekräftig. Die Berechnung von Scores ist innerhalb des Testdatensatzes nur für das Thema „Immissionen und Emissionen“ und dessen Unterthemen „Schall, Lärm“ und „Erschütterungen“ ansatzweise sinnvoll. Die folgende Tabelle fasst die quantitative Auswertung der Themenerkennung für alle Oberthemen und die Unterthemen von „Immissionen und Emissionen“ zusammen.

TABELLE 13: QUANTITATIVE AUSWERTUNG DER THEMENERKENNUNG ÜBER DEN TESTDOKUMENTEN

	Anzahl	**True Positives**	**Recall**	**Precision**	**F1**
1 Immissionen und Emissionen	63	56	0,8889	0,5833	0,7044
1.01 Schall, Lärm	30	25	0,8333	0,3086	0,4505
1.02 Erschütterungen	22	11	0,5000	0,2444	0,3284
1.03 Sekundärer Luftschall	3	1	0,3333	0,0323	0,0588
1.04 Luftverunreinigungen	1	1	1,000	0,0085	0,0169
1.05 Elektromagnetische Felder	4	3	0,7500	0,0337	0,0645
1.06 Sonstiges	3	2	0,6667	0,0541	0,1000
2 Naturschutz	0				
3 Umweltschutz	14	10	0,7143	0,0943	0,1666
4 Landwirtschaft	0				
5 Forstwirtschaft	0				
6 Denkmalschutz	8	8	1	0,0362	0,0699
7 Abfallwirtschaft	0				
8 Brand- und Katastrophenschutz	2	2	1	0,0090	0,0178
9 Bauzeitliche Beeinträchtigungen	17	11	0,6471	0,1746	0,2750
10 Anlagen Dritter	11	10	0,9090	0,1000	0,1802
11 Inanspruchnahme von Grundeigentum	4	4	1	0,0430	0,0825
12 Sonstiges	29	20	0,6897	0,2632	0,3810

Entsprechend dieser Auswertung werden für alle Themen gute bis sehr gute Recallwerte erreicht, d. h. vorhandene Themen im Datensatz werden mit hoher Wahrscheinlichkeit erkannt. Die F1-Scores sind dennoch für viele Themen sehr gering. Ursache hierfür ist eine hohe Anzahl von Fehlerkennungen. Es ist jedoch ein klarer Zusammenhang zur Menge der Trainingsdaten erkennbar. Das Thema „Immissionen und Emissionen", für das mehr als 1.000 Trainingsbeispiele im Datensatz vorliegen, erzielt einen guten F1-Score. Den zweitbesten F1-Score erzielt das Thema „Bauzeitliche Beeinträchtigungen", für das etwa 500 Trainingsbeispiele vorhanden sind. Die Scores der anderen Themen, für die im Datensatz deutlich weniger Beispiele vorliegen, sind erheblich schlechter. Ein analoger Zusammenhang zeigt sich auch für die sechs Unterthemen von „Immissionen und Emissionen". Für die Unterthemen „Schall, Lärm" und „Erschütterungen", die jeweils mehrere hundert Trainingsbeispiele umfassen, werden passable Scores erzielt, während für die anderen Unterthemen mit jeweils weniger als 100 Beispielen kaum befriedigende Ergebnisse erzielt werden. Bessere Ergebnisse wären jedoch bei Verfügbarkeit größerer Datenmengen zu erwarten.

Ziel der Entwicklung des PoC war nicht nur, die ausgewählten Verfahren zu implementieren, sondern das intendierte fachliche Szenario bereits im Demonstrator ausreichend abzubilden und eine diesem Szenario nahekommende, praxisnahe Bewertung der Funktionalität und Leistungsfähigkeit der untersuchten Verfahren auch durch Personen aus dem realen Anwendungskontext zu ermöglichen. Bereits während der Entwicklung wurde dieses Ziel durch das Projektteam kontinuierlich verfolgt und unter dieser Prämisse Optimierungsentscheidungen getroffen.

Im Anschluss an die Vorstellung während der Ergebnispräsentation erfolgte in der Diskussion eine erste Bewertung der Funktionalität des Demonstrators aus Anwendersicht durch einen mit dem Anwendungskontext vertrauten Fachexperten des Auftraggebers. Für einen Zeitraum nach der Ergebnispräsentation wurde darüber hinaus vereinbart, dass vom Auftraggeber weitere Tests mit dem Demonstrator durchgeführt werden können.

Die ersten Bewertungen weisen stark darauf hin, dass der Demonstrator eine gute Grundlage für eine hohe Aussagekraft der in der Evaluierung beschriebenen Potentialabschätzung bietet.

Die Ergebnisse der hier dokumentierten Studie wurden auf der Grundlage einer vom Auftraggeber zu Beginn der Arbeiten getroffenen Auswahl von Beispieldaten erarbeitet. Aus Aufwandsgründen sowohl für die Auswahl und Erstellung (einschließlich der Anonymisierung) als auch für die weitere Bearbeitung dieser Daten im Projekt war hier nur eine Datenbasis relativ begrenzter Größe realistisch umsetzbar. Trotz dieser begrenzten Datengrundlage konnten für die ausgewählten Aufgaben bereits überaus ermutigende Ergebnisse erzielt werden. Bei einer späteren Integration einer auf die Ergebnisse dieser Studie aufgebauten KI-Lösung in die operativen Prozesse sowie in Dokumentenmanagement- und Workflow-Systeme des Auftraggebers erscheint eine grundlegende Verbesserung der Datensituation möglich, da nunmehr Daten aus dem laufenden Betrieb heraus kontinuierlich erhoben, über die Zeit akkumuliert und wiederholt als Trainingsdaten verwendet werden können.

Den hier vorgeschlagenen maschinellen Lernverfahren ist gemeinsam, dass bei einer Vergrößerung der Trainingsdatenbasis grundsätzlich eine weitere Steigerung der Performanz zu erwarten ist. Diese lässt sich im Vorhinein kaum quantifizieren, aber da bei Datenerhebung aus dem laufenden Betrieb der Umfang der verfügbaren Trainingsdaten ganz andere Größenordnungen annehmen würde, sind spürbare Steigerungen der Performanz sehr wahrscheinlich.

Da auch langfristig ein assistiver Unterstützungsmodus anvisiert ist, fließt bei kontinuierlicher Datenerhebung ständig zusätzliches menschliches Wissen in die Daten ein, was zu einem stetigen Qualitätsanstieg des Gesamt-Datenmaterials über die Zeit führt. Längerfristig ist zudem zu erwarten, dass dieser Effekt noch gesteigert wird, indem einige Klassen von bei der manuellen Bearbeitung auftretenden Fehlern

(die bei menschlicher Entscheidung nie ganz auszuschließen sind), durch die assistive Arbeitsweise in ihrer Häufigkeit noch weiter reduziert werden. KI-Lösung und menschliche Bearbeitung könnten somit gemeinsam einen nachhaltigen Feedbackprozess zur kontinuierlichen Qualitätssteigerung bilden.

Sollten in der Zukunft durch die skizzierten Maßnahmen oder aus anderen Gründen Trainingsdaten in erheblich größerem Umfang verfügbar werden, kann zu diesem Zeitpunkt weiterhin eine Neubewertung der Eignung der Verfahren sinnvoll erscheinen. Wie in Kapitel 6 dargestellt, wären für einige der hier untersuchten Problemstellungen prinzipiell auch modernste Deep-Leaning-Ansätze sehr attraktiv. Diese konnten allerdings unter den konkreten Rahmenbedingungen des jetzigen Projekts nicht favorisiert bewertet werden. Zwar spielten hierbei auch Risiken eine Rolle, die sich aus dem weniger ausgereiften Entwicklungsstand dieser Methoden und deren hohem Bedarf an Rechenressourcen ergeben. Der vorrangige Ausschlussgrund ist jedoch der sehr erhebliche Datenbedarf dieser Verfahren, welcher den Umfang der im Projekt verfügbaren Datenbasis um Größenordnungen übersteigt. Sollten allerdings zu einem späteren Zeitpunkt in ausreichendem Umfang Trainingsdaten verfügbar sein, wäre hier folglich eine Neubewertung zweckdienlich. Da sowohl in der Erforschung neuer und optimierter Deep-Learning-Verfahren als auch bei den Produktentwicklungszyklen leistungsfähiger Multicore-CPUs und GPUs weiterhin ein enormes Fortschrittstempo festzustellen ist, könnten auch die genannten Risiken sich bereits in naher Zukunft auf ein akzeptables Maß reduzieren. Im Falle einer positiven Neubewertung könnte eine praktische Umsetzung geeigneter Deep-Learning-Verfahren somit auch für die hier betrachteten Anwendungen realistisch werden, was letztlich einen weiteren deutlichen Qualitätszuwachs erwarten ließe.

8 Zusammenfassung und Ausblick

Ziel des Forschungsprojekts war eine erste Eruierung und Evaluierung von maschinellen Lernverfahren für das Management von Einwendungen im Rahmen von Planfeststellungsverfahren. Dabei erfolgte eine Fokussierung auf vier konkret vom Auftraggeber vorgegebene Aufgaben im Kontext des Fachprozesses:

1. Identifizierung und Extraktion argumentativer Einwendungssegmente
2. Klassifizierung von Einwendungen bezüglich eines Themenkatalogs
3. Erkennung und Bewertung von Argumentationsmustern
4. Bewertung auf Vollständigkeit und thematisch plausible Zuordnung

In einem ersten Arbeitsschritt wurden die vom Auftraggeber bereitgestellten Daten aus zurückliegenden Planfeststellungsverfahren analysiert und in ein einheitliches Datenmodell überführt (AP 1). In einem zweiten Arbeitsschritt erfolgte für die vom Auftraggeber vorgegebenen Aufgaben eine Recherche grundsätzlich geeigneter Verfahren der maschinellen Textverarbeitung und die Entwicklung potenzieller Lösungsansätze. Für die Lösungsansätze wurden konkrete Experimente durchgeführt und eine Bewertung hinsichtlich der Anwendbarkeit vorgenommen (AP 2).

In einem dritten Arbeitsschritt (AP 3) wurde ein Demonstrator entwickelt, mit dem im Sinne eines PoC dargelegt wird, in welcher Form sich die Segmenterkennung und die Klassifizierung nach Themen in den Fachprozess integrieren lassen. So können in der entwickelten Web-Anwendung vorgegebene Testdokumente oder frei editierbare Texte unter Verwendung der im Rahmen der Studie vortrainierten KI-Modelle analysiert sowie die Ergebnisse der Analyse angezeigt und nachbearbeitet werden. Auf der Grundlage der Umsetzung wurde auch eine Potentialabschätzung vorgenommen.

8.1 Bewertung der Untersuchungsergebnisse

Wesentliches Ergebnis des ersten Arbeitspakets ist der **einheitlich strukturierte digitale Datensatz** der vorliegenden Einwendungs-, Erwiderungs- und Metadaten mit mehr als 5.000 Einträgen/Segmenten. Ausgangsbasis war ein vom Auftraggeber bereitgestellter Beispieldatensatz mit anonymisierten Daten aus früheren Planfeststellungsverfahren, der eingehend analysiert, harmonisiert und (größtenteils manuell) um Metadaten angereichert wurde. Das zugrundeliegende Datenmodell wurde derart konzipiert, dass es sowohl für die Analyse in AP 2 und den Demonstrator (AP 3) als auch in Zukunft für weitergehende Untersuchungen und eine mögliche Umsetzung von MTV-Verfahren im Einwendungsmanagement einsetzbar ist. Zudem liegt der Datensatz in einem einfach strukturierten Format vor, welches die manuelle Sichtung und Weiter- bzw. Wiederverwendung (z. B. in Form einer Überführung in eine Datenbank) erleichtert. Somit liegt ein wichtiger Baustein vor, der für eine zukünftige Automatisierung des Fachprozesses einen wertvollen Beitrag liefern kann.

Mit den Arbeitsergebnissen aus AP 2 liegt eine **wissenschaftlich fundierte Analyse grundsätzlich geeigneter MTV-Verfahren und konkreter Lösungsansätze** für die Automatisierung der vorgegebenen Aufgaben vor, die als Entscheidungsgrundlage für eine Priorisierung und die Konkretisierung der Anforderungen an eine mögliche Umsetzung herangezogen werden können. Dabei wurden einerseits vielversprechende Ansätze und deren Potential präsentiert, andererseits aber auch Grenzen der Verfahren aufgezeigt und ungeeignete Verfahren verworfen.

AP 3 liefert bereits auf der Grundlage der begrenzten Datenbasis eine klare und **positive Machbarkeitsentscheidung** für die Aufgaben 1 und 2 (Segmenterkennung und die Klassifizierung nach Themen), indem

geeignete Verfahren untersucht, anhand ausgewählter Testdaten validiert, in einem Demonstrator umgesetzt und mittels diesem in einer an das praktische Einsatzszenario und die Nutzersicht orientierten Umgebung erfolgreich getestet wurden.

Insgesamt ist die im Rahmen der Studie erreichte Qualität/Performanz insbesondere angesichts des begrenzten Umfangs der zur Verfügung Datenbasis sehr ermutigend. So belegen die Studienergebnisse, dass KI-basierte Lernverfahren geeignet sind, um wesentliche Aufgaben im Prozess des Einwendungsmanagements in Teilen zu automatisieren. Die Teilautomatisierung liefert Vorschläge, die Mitarbeitenden in der Anwendung zur Validierung präsentiert werden und somit zu einer deutlichen Entlastung beitragen. Die Entscheidungshoheit verbleibt dabei stets bei den Mitarbeitenden, wodurch Transparenz und Nachvollziehbarkeit gewahrt bleiben. Dabei ist eine Umsetzung auf Grundlage von Open-Source-Lösung und ohne Abhängigkeiten von externen Diensten oder Cloud-Anwendungen problemlos möglich, so dass auch Anforderungen hinsichtlich Datenschutz und Datensouveränität erfüllt werden. Darüber hinaus kann davon ausgegangen werden, dass die im Rahmen der Studie eingesetzten Verfahren in Bezug auf den zu erwartenden Ressourcenbedarf für den IT-Betrieb unkritisch sind.

8.2 Optionen zur Weiterentwicklung

Mit dem Demonstrator konnte die Problemlösungsfähigkeit dargelegt und gezeigt werden, in welcher Form sich die Segmenterkennung und die Klassifizierung nach Themen in den Fachprozess integrieren lassen. Nachfolgend sind weiterführende Überlegungen zur Realisierung des Praxispotenzials zusammengefasst.

In Bezug auf die Segmenterkennung erscheint eine Weiterentwicklung in der Form sinnvoll, dass automatisch erkannte Absätze durch Mitarbeitende manuell zusammengeführt oder unterteilt werden können (vgl. Umsetzungsskizze in Abbildung 14). Dadurch ließe sich die fachliche Anforderung nach einer flexiblen Lösung zur Unterteilung nach einzelnen Teilabschnitten und Sätzen ergonomisch umsetzen.

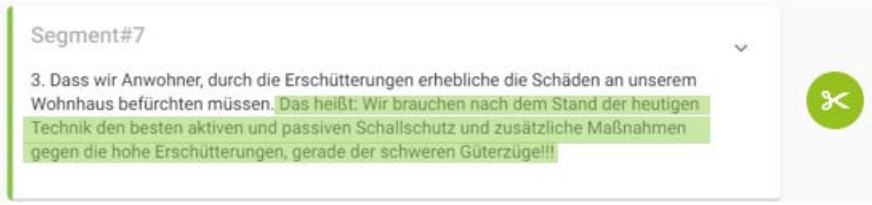


Abbildung 14: Skizzierung der möglichen Funktion zur Unterteilung von Segmenten

Unmittelbar auf dem Interaktionsprinzip des Demonstrators aufbauend ließen sich auch Verfahren zur Erkennung von Mustereinwendungen und Argumentationsmustern umsetzen und in den Prozessablauf integrieren.

In Abbildung 15 ist die mögliche Umsetzung einer Ähnlichkeitserkennung auf Dokumentenebene skizziert. Mit dieser Funktion böte sich für Mitarbeitenden die Möglichkeit, in der Übersichtsdarstellung für ein betrachtetes Element Hinweise dafür zu erhalten, wie viele der bearbeitungsrelevanten Segmente bereits in anderen Schreiben in gleicher oder ähnlicher Form vorliegen. Ein hohes Maß an Überstimmung mit anderen Schreiben kann dabei als Indikator für die Verwendung eines Musterschreibens gewertet werden. Anhand dieser Information können Schreiben sortiert und Duplikate markiert werden, die einer einheitlichen Beantwortung zugeführt werden müssen.

Analog ließe sich eine Ähnlichkeitserkennung auch auf der Ebene einzelner Segmente und Argumente umsetzen (vgl. Skizze in Abbildung 16). Vorstellbar ist, dass Mitarbeitenden bei der Markierung eines Segments Verweise auf ähnliche Argumente in anderen Schreiben angezeigt werden. Diese Information können die Mitarbeitenden prüfen, validieren und auf diesem Wege u. a. Hinweise dazu erhalten, in welchem Zusammenhang das Argument an anderer Stelle aufgeführt wird, welchem Thema dieses Argument in der Vergangenheit zugeordnet wurde, in welcher Form die Erwiderung durch den Vorhabenträger vorgenommen wurde und wie die abschließende Entscheidung ausfiel. In der Form kann das Wissensmanagement aus der Fachanwendung heraus ohne Medienbrüche unterstützt werden.

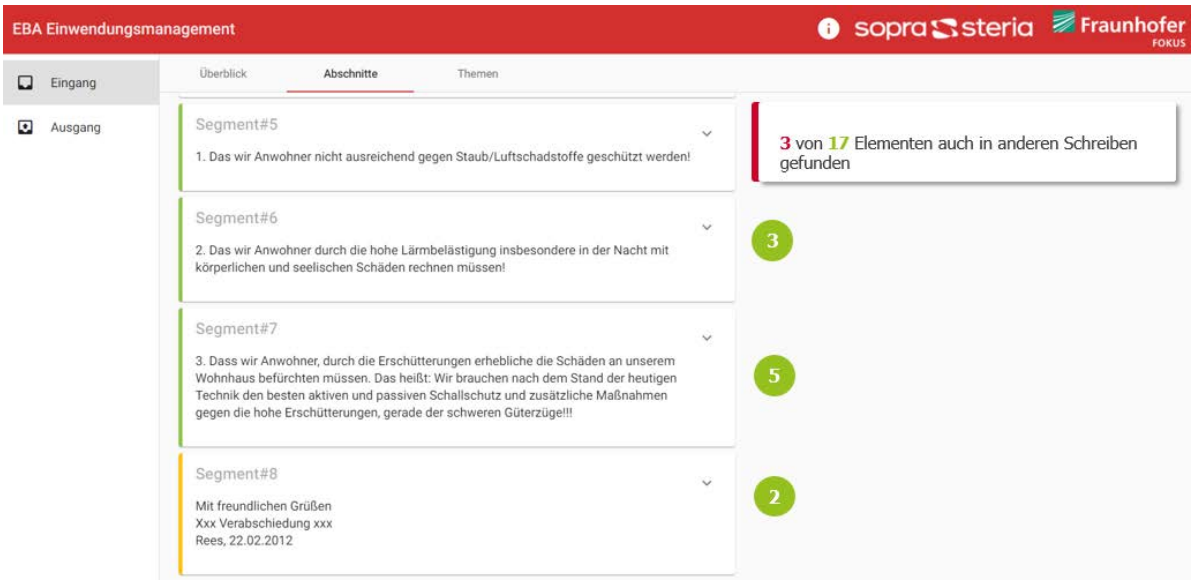


Abbildung 15: Skizzierung der möglichen Umsetzung einer Ähnlichkeitserkennung (Dokumentenebene)

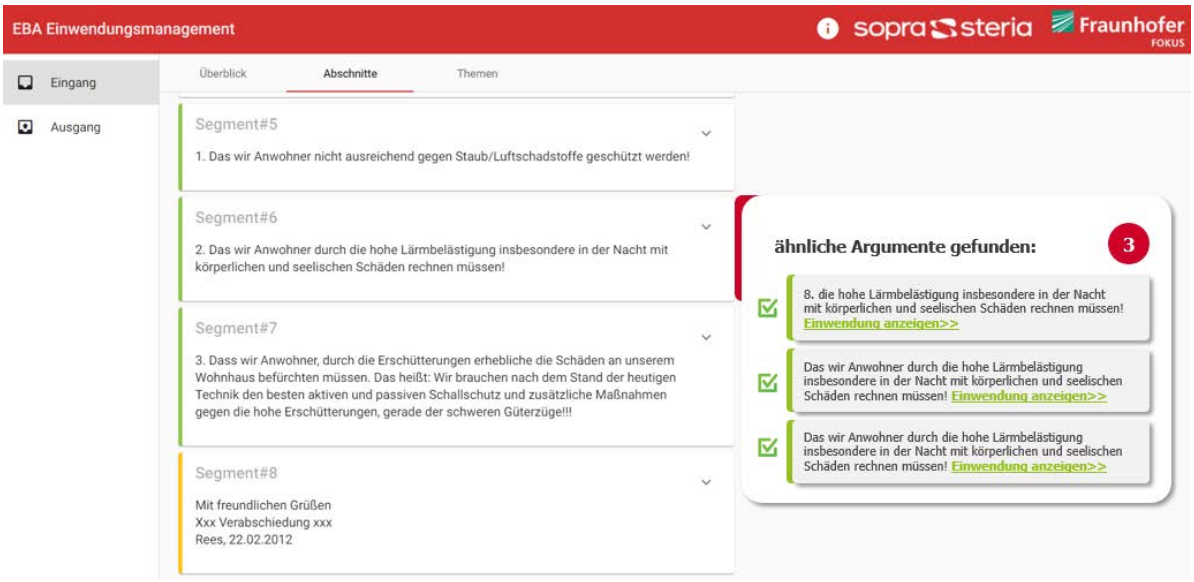


Abbildung 16: Skizzierung der möglichen Umsetzung einer Ähnlichkeitserkennung (Segmentebene)

Über die in der vorliegenden Studie vornehmlich betrachteten Aufgaben hinaus verspricht der Einsatz von MTV-Verfahren auch im erweiterten Kontext des Einwendungsmanagements hohes Potential für Automatisierung und eine effizientere Prozessdurchführung. Im Mittelpunkt stehen dabei Verfahren für die automatische Bearbeitung von Dokumenten, hier insbesondere für die Textklassifikation, die Informationsextraktion, die Prozessautomatisierung sowie Empfehlungssysteme (vergleiche Abbildung 17). Konkret erscheint eine Evaluierung der Einsatzmöglichkeiten für diese Verfahren insbesondere für die folgenden Szenarien vielversprechend:

- Informationsextraktion
  Mit Blick auf die Fülle und Heterogenität der zu verarbeitenden Daten bieten Kombinationen etablierter Verfahren hohes Potenzial in Bezug auf eine prozessspezifische Vorverarbeitung von Dokumenten. Schritte der Digitalisierung und Vereinheitlichung können dabei in einem dem Fachsystem vorgelagerten Dokumentenmanagementsystem in Form einer Digitalisierung des Posteingangs integriert werden. Ergänzend sind für ein mögliches Fachsystem Lösungen am Markt verfügbar, um verfahrensrelevante Informationen wie etwa Absender, Anschrift und adressiertes Planungsverfahren zu extrahieren und für die weitere Bearbeitung bereitzustellen.

- KI-gestütztes Content Management
  Mit Blick auf die Vielzahl von parallelen Verfahren und der damit einhergehenden Fülle von Dokumenten kommen Funktionen für den einfachen Zugriff sowie für die Sortierung und Filterung von Informationen eine hohe Bedeutung zu. Moderne Verfahren aus dem Bereich des Content Managements können diesbezüglich einen wertvollen Beitrag leisten, der auf die Zielfelder Bedienkomfort und Effizienz der Bearbeitung einzahlt.

- Gruppierung ähnlicher Segmente
  Der Einsatz von über das Content Management hinausgehende Verfahren ermöglicht es, Segmente und ganze Einwendungen auch fachverfahrensspezifisch zu gruppieren und Mitarbeitenden Funktionen anzubieten, um inhaltlich ähnliche Segmente und Dokumente einfacher analysieren und Querverweise ziehen und Analyseergebnisse validieren zu können.

- KI-gestützte Vorschläge für Entscheidungen
  Ein weiterer möglicher Punkt für den Einsatz von MTV-Verfahren betrifft die Vorbereitung der Entscheidung durch das EBA, die das Resultat der Bewertung von Erwiderung und den Ergebnissen der Anhörung darstellt. Hier ist eine (Teil-)Automatisierung in Form von Entscheidungsvorschlägen zu einzelnen Argumenten und dazugehörigen Erwiderungen vorstellbar. Zu beachten wäre hierbei, dass die jeweils aktuelle Rechtsprechung zu berücksichtigen ist, was wiederum Auswirkungen auf das (regelmäßige) Training der Modelle und die notwendige Pflege der Trainingsdaten hätte.

- KI-gestützte Erwiderungsvorschläge (ergänzendes Szenario im Systemverbund)
  Empfehlungssysteme lassen sich auch auf Seiten des Vorhabenträgers einbinden, um übermittelte Einwendungen zu analysieren und auf der Grundlage vergleichbarer Verfahren Vorschläge für die Formulierung von Erwiderungen abzuleiten. Weitere Synergien lassen sich potentiell auch dadurch generieren, dass Art und Umfang des Einsatzes von KI-Verfahren im Systemverbund geteilt und somit die Dateninterpretation und Weiterverarbeitung erleichtert werden.

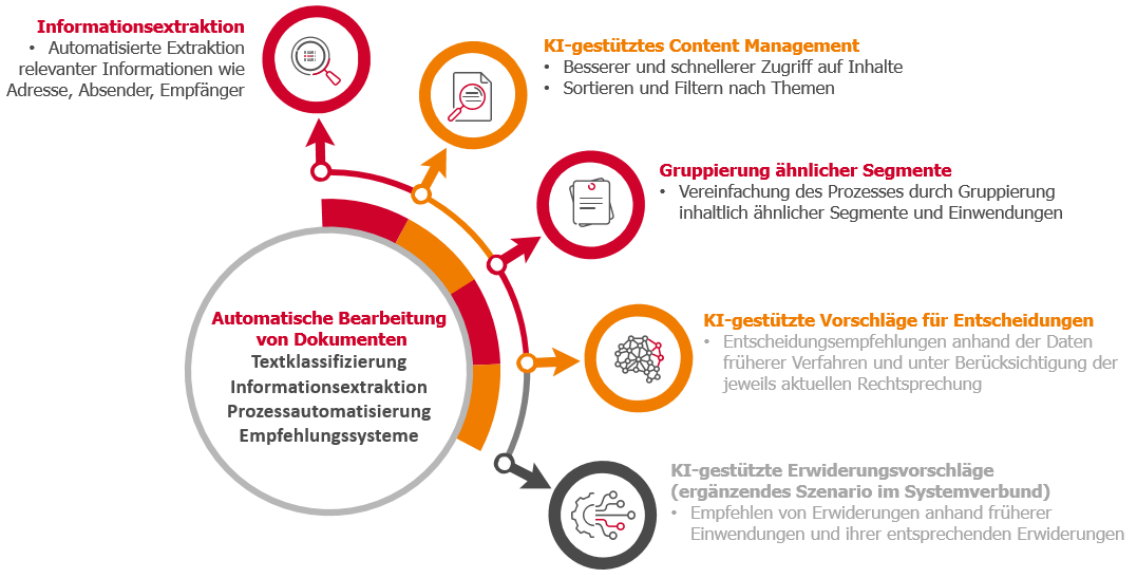


Abbildung 17: Weitere fachliche Szenarien für KI-Anwendungen im Kontext Einwendungsmanagement

8.3 Empfehlungen für die Überführung in die Praxis

Abschließend lassen sich aus der Studie Empfehlungen für eine zukünftige Überführung von MTV-Verfahren in die Praxis ableiten.

Allen im Rahmen der Studie untersuchten Verfahren gemein ist, dass mit weiteren Steigerungen der Ergebnisgüte zu rechnen ist, wenn eine deutlich vergrößerte Datenbasis zu Trainingszwecken herangezogen werden kann. Es erscheint daher empfehlenswert, **bereits vor der Umsetzung in größerem Umfang Trainingsdaten aus dem laufenden (manuellen) Betrieb heraus zu sammeln**. Mit Blick auf die zentrale Aufgabe der Klassifizierung nach Themen sollte dabei für die Datensammlung ein **einheitlicher Themenkatalog** vorgegeben werden, um Trainingsdaten gleich in einer passend strukturierten Form vorliegen zu haben und den Aufwand für eine weitere Datenaufbereitung und (manuelle) Annotation so gering wie möglich zu halten. Dies schließt nicht aus, dass der Themenkatalog im späteren Einsatz angepasst und fortgeschrieben wird, sondern bildet vielmehr das Fundament für das initiale Training der KI-Modelle.

In Bezug auf Ergonomie und Akzeptanz der Umsetzung sollte auf eine **intuitiv bedienbare Benutzerschnittstelle** geachtet werden. So sollten beispielsweise automatisch generierte Vorschläge klar erkennbar dargestellt, hinsichtlich der Ergebnisgüte kategorisiert und vom Benutzer komfortabel validier- und editierbar sein. Dies kann beispielsweise durch optische Hervorhebungen von Vorschlägen sowie das Anbieten von erweiterbaren Auswahllisten erfolgen.

Eine attraktive Möglichkeit den wünschenswerten Zugriff auf mehr Trainingsdaten auch nach der Umsetzung aus dem laufenden Betrieb heraus zu erlangen, besteht darin, die durch Mitarbeitende vorgenommenen Validierungen bzw. Anpassungen im Sinne eines **fortlaufenden Feedback-Prozesses** zu speichern. Voraussetzung dafür wäre eine Einbindung in die operativen Dokumentenmanagement- und Workflowsysteme sowie die datenschutz- und arbeitsrechtskonforme Verarbeitung der Daten. Unter Verwendung der stetig fortgeschriebenen Daten aus dem Feedback-Prozess können die **KI-Modelle periodisch nachtrainiert** werden, wodurch sich die Ergebnisgüte und somit insgesamt der Nutzen im Sinne der Effizienzsteigerung weiter steigern lassen.

Um das Potenzial insgesamt in dem skizzierten Umfang heben zu können, scheint es erforderlich, **MTV-Verfahren** bei der Konzeption von Prozessen und unterstützenden Systemen bereits initial **als integrativen Bestandteil** zu berücksichtigen und entsprechend in die IT- und Prozesslandschaft (einschließlich Dokumentenmanagementsystem) einzubetten. Bei einer fortlaufenden Datenerhebung (siehe oben) wären mittelfristig auch Trainingsdaten in einem ausreichend hohen Umfang verfügbar, so dass moderne **Deep-Learning-Verfahren** zur Anwendung kommen könnten, was wiederum einen weiteren Qualitätssprung verspricht. Auch dieser Entwicklungspfad sollte bereits in der initialen Konzeption berücksichtigt werden, um Architektur und Systeme ausreichend flexibel zu halten. Ausgehend von den Ergebnissen der Studie erscheint es möglich, KI-Lösungen auf der Basis von Open-Source-Ansätzen zu realisieren, um Abhängigkeiten von externen Diensten oder Cloud-Anwendungen zu vermeiden. Somit wären auch spezifische Anforderungen hinsichtlich Datenschutz und Datensouveränität erfüllbar.

Abbildungsverzeichnis

Tabellenverzeichnis

Quellenverzeichnis

[1] Gesmundo, Andrea, and Tanja Samardzic (2012): "Lemmatisation as a tagging task." Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2012.

[2] Porter, Martin F. (2001): "Snowball: A language for stemming algorithms."

[3] Heeman, Peter A. (1999): "POS tags and decision trees for language modeling." *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

[4] Mikolov, Tomas, et al. (2013): "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781*.

[5] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning (2014): "Glove: Global vectors for word representation." *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*.

[6] Peters, Matthew E., et al. (2018): "Deep contextualized word representations." *arXiv preprint arXiv:1802.05365*.

[7] Cer, Daniel, et al. (2018): "Universal sentence encoder." *arXiv preprint arXiv:1803.11175*.

[8] Devlin, Jacob, et al. (2018): "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805*.

[9] Reimers, Nils, and Iryna Gurevych. (2019): "Sentence-bert: Sentence embeddings using siamese bert-networks." *arXiv preprint arXiv:1908.10084*.

[10] Yang, Zhilin, et al. (2019): "Xlnet: Generalized autoregressive pretraining for language understanding." *Advances in neural information processing systems* 32.

[11] Wang, Shuai, et al. (2013): "Combination of VSM and Jaccard coefficient for external plagiarism detection." *2013 international conference on machine learning and cybernetics*. Vol. 4. IEEE, 2013.

[12] Rahutomo, Faisal, Teruaki Kitasuka, and Masayoshi Aritsugi. (2012): "Semantic cosine similarity." *The 7th International Student Conference on Advanced Science and Technology ICAST*. Vol. 4. No. 1. 2012.

[13] Levenshtein, Vladimir I. (1966): "Binary codes capable of correcting deletions, insertions, and reversals." *Soviet physics doklady*. Vol. 10. No. 8.

[14] Robertson, Stephen, and Hugo Zaragoza (2009): *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc.

[15] Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003): "Latent dirichlet allocation." *the Journal of machine Learning research* 3 (2003): 993-1022.

[16] Deerwester, Scott, et al. (1990): "Indexing by latent semantic analysis." *Journal of the American society for information science* 41.6 (1990): 391-407.

[17] Paatero, Pentti, and Unto Tapper (1994): "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values." *Environmetrics* 5.2 (1994): 111-126.

[18] Cristianini, Nello, and John Shawe-Taylor (2000): *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

[19] Guo, Gongde, et al. (2003): "KNN model-based approach in classification." *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer, Berlin, Heidelberg.

[20] Menard, Scott (2002): *Applied logistic regression analysis*. Vol. 106. Sage.

[21] Kibriya, Ashraf M., et al. (2004): "Multinomial naive bayes for text categorization revisited." *Australasian Joint Conference on Artificial Intelligence*. Springer, Berlin, Heidelberg.

[22] Breiman, Leo (2001): "Random forests." *Machine learning* 45.1 (2001): 5-32.

[23] Ramchoun, Hassan, et al. (2016): "Multilayer Perceptron: Architecture Optimization and Training." *Int. J. Interact. Multim. Artif. Intell.* 4.1 (2016): 26-30.

[24] Guo, Bao, et al. (2019): "Improving text classification with weighted word embeddings via a multi-channel TextCNN model." *Neurocomputing* 363 (2019): 366-374.

[25] Du, Changshun, and Lei Huang (2018): "Text classification research with attention-based recurrent neural networks." *International Journal of Computers Communications & Control* 13.1 (2018): 50-61.

[26] Hochreiter, Sepp, and Jürgen Schmidhuber (1997): "Long short-term memory." *Neural computation* 9.8 (1997): 1735-1780.

[27] Vaswani, Ashish, et al. (2017): "Attention is all you need." *Advances in neural information processing systems*.

[28] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. "ConceptNet 5.5: An Open Multilingual Graph of General Knowledge." In proceedings of AAAI 31.

[29] Akoglu, Haldun. "User's guide to correlation coefficients." *Turkish journal of emergency medicine* 18.3 (2018): 91-93.

Anhänge

Anhang 1: Themenkatalog

TABELLE 13: THEMENKATALOG

Code	Oberthema	Unterthema
1.00	Immissionen und Emissionen	
1.01		Schall, Lärm
1.02		Erschütterungen
1.03		Sekundärer Luftschall
1.04		Luftverunreinigungen
1.05		Elektromagnetische Felder
1.06		Immissionen und Emissionen - Sonstiges
2.00	Naturschutz	
2.01		Natur- und Landschaftsschutz
2.02		Gebietsschutz (Natura 2000)
2.03		Artenschutz
2.04		Naturschutz - Sonstiges
3.00	Umweltschutz	
3.01		Bodenschutz
3.02		Gewässerschutz
3.03		Umweltschutz - Sonstiges
4.00	Landwirtschaft	
4.01		Landwirtschaft
4.02		Landwirtschaft - Sonstiges
5.00	Forstwirtschaft	
5.01		Forstwirtschaft
5.02		Forstwirtschaft - Sonstiges

6.00	Denkmalschutz	
6.01		Bodendenkmal
6.02		Kulturdenkmal
6.03		Denkmalschutz - Sonstiges
7.00	Abfallwirtschaft	
7.01		Abfallwirtschaft
7.02		Abfallwirtschaft - Sonstiges
8.00	Brand- und Katastrophenschutz	
8.01		Brand- und Katastrophenschutz
8.02		Brand- und Katastrophenschutz - Sonstiges
9.00	Bauzeitliche Beeinträchtigungen	
9.01		Bauzeitliche Beeinträchtigungen
9.02		Bauzeitliche Beeinträchtigungen - Sonstiges
10.00	Anlagen Dritter	
10.01		Infrastrukturleitungen / Sonstige Anlagen Dritter
10.02		(öffentliche) Ver- und Entsorgungsanlagen
10.03		Straßen, Wege, Zufahrten
10.04		Anlagen Dritter - Sonstiges
11.00	Inanspruchnahme von Grundeigentum	
11.01		Eigentumsbeeinträchtigungen / Grunderwerb
11.02		Verschattung
11.03		Inanspruchnahme von Grundeigentum - Sonstiges
12.00	Sonstiges	

Anhang 2: Schlüsselworte

1.00 - IMMISSIONEN UND EMISSIONEN

Besonders überwachtes Gleis, BImSchV (Verordnung zur Durchführung des Bundes-Immissionsschutzgesetzes), BlmSchG (Bundes-Immissionsschutzgesetz), BÜG (Gesetz über die Übernahme von Staatsbürgschaften und Garantien des Freistaates Bayern), Gesundheit, Nachmessungen

1.01 - SCHALL, LÄRM

aktiver Schallschutz, Einzelimmissionsortbetrachtung, Geschwindigkeitsbeschränkung, Gesundheitsbeeinträchtigungen durch Schall, Lärm, Lärmbelastung, Lärmreduzierung, Lärmschutz, Lärmschutzmaßnahmen, Mittelungspegel/Spitzenpegel, Schall, Schallschutz, Schallschutzwand, Schallschutzwände, Umgebungslärmrichtlinie

1.02 – ERSCHÜTTERUNGEN

DIN 4150 - Erschütterungen im Bauwesen, Erschütterungen, Erschütterungsimmissionen, Erschütterungsschutzes, Erschütterungsschutzmaßnahmen, Schienenbonus

1.03 - SEKUNDÄRER LUFTSCHALL

auditive Wahrnehmung, Luftschall, Luftverunreinigungen, Schallausbreitung, Schalldämmung, Schallgeschwindigkeit, Schallwelle, Störschall, Straßenverkehrslärm

1.04 - LUFTVERUNREINIGUNGEN, FEINSTAUB

Feinstaub, Grenzwerte, Immissionswerte, Kohlenstoffdioxid, Luftqualität, Luftreinhaltung, Lunge, PM□, PM□,□, PM□□, PMIO-Belastung, Staub

1.05- ELEKTROMAGNETISCHE FELDER

Elektrifizierten, Elektromagnetische Felder, Elektrosmog, ionisierende Strahlung, Magnetfelder, Niedrigfrequenzanlagen

2.00 - NATURSCHUTZ

Gebietsschutz, Grünkonzept, Landschaftspflege, Landschaftspflegerischer, Naturschutzbehörde, Naturschutzbewegung, Naturschützer, Naturschützerin, Naturschutzgebiet, Naturschutzgesellschaft, Naturschutzgesetz, Naturschutzorganisation, Naturschutzpark, Naturschutzrechtliche, Naturpflege, Reaktorsicherheit, Vertragsnaturschutz, Vogelschutzrichtlinie

2.01 - NATUR UND LANDSCHAFTSPFLEGE

Flachwasserteiche, Landespflege, Landschaft, Maßnahmen, Rekultivierung, Renaturierung, Sanierung, Tümpel, Unterwassermatten, Wehr

2.02 - GEBIETSSCHUTZ (NATURA 2000)

Fauna-Flora-Habitat-Richtlinie, FFH-Richtlinie, Habitate 79/409/EWG, Natura 2000, Richtlinie, Richtlinie 92/43/EWG, Schutzgebieten, Vogelschutzrichtlinie

2.03 – ARTENSCHUTZ

Artenschutzprogramm, Artensterben, Artenvielfalt, Ästhetik, Aussterben, Biodiversität, Biotop, Biotopschutz, gefährdete Arten, Naturschutz, Ökotop, Rote Liste, Tierschutz

3.00 – UMWELTSCHUTZ

Altlasten, Bodenschutz, BMU– Bundesministerium für Umwelt, Erneuerbarer Rohstoff, Gewässerschutz, Kleinwasserkraft, Klima, Oberflächengewässer, Umwelt, Umweltschutz, Umweltschutzbehörde, Umweltschutzerklärung, Umweltschutzförderung, Umweltschutzkonzept, Umweltschutzmanagement, Umweltschutzmaßnahme, Umweltschutznorm, Umweltschutzorganisation, Umweltschutzrichtlinie, Umweltschutzverordnung, Umweltschutzzone, Wasser, Wassererlaubnis, wasserrechtliche, Wasserwirtschaft

3.01 – BODENSCHUTZ

Boden (Bodenkunde), Bodenart, Bodenrahmenrichtlinie, Bundes-Bodenschutz- und Altlastenverordnung, Bundes-Bodenschutzgesetz, Dünger, Grünlandwirtschaft, Hydrokultur, Landwirtschaft, ökologische Landwirtschaft, Pestizid, Pflanzenschutzmittel, Schwermetalle

3.02 – GEWÄSSERSCHUTZ

Abwasser, Aquatil, Brunnen, Deponie, Grundwasser, Grundwasserabsenkung, Gülle, Küstengewässer, LI 16 - Luft/Wasser-Wärmepumpe, Naturschutz, Oberflächengewässer, Ökosystem, Pflanzenschutzmittel, Schadstoff, Umweltverschmutzung, wassergefährdend, Wasserschutzgesetz

4.00 - LANDWIRTSCHAFT

Ackerbau, Ackerbau und Viehzucht, Agrar, Agrarchemie, Agrargeografie, Agrartechnik, Agrarwirtschaft, Agrochemie, Agrotechnik, Agrotourismus, alternative Landwirtschaft, Ausgleichsflächen, Bauernhof, biologische Landwirtschaft, Entschädigung der Landwirtschaft, Existenzgefährdung (Landwirtschaft), Feldwegenetz, Erschließung, Umwege, Gehöft, generelle Beeinträchtigung der landwirtschaftlichen Belange, Landbau, Landmaschinentechnik, Landreform, Landtourismus, Landwirtschaft, Landwirtschaftlichen, landwirtschaftlicher Betrieb, LBP-Maßnahmen (Landwirtschaft), Naturweidewirtschaft, Ökolandbau, Optimierung in Bezug auf landwirtschaftliche Nutzbarkeit, pastorale Tierhaltung auf Naturweiden, Pastoralismus, Schadstoffeinträge

5.00 - FORSTWIRTSCHAFT

BFH - Bundesforschungsanstalt für Forst- und Holzwirtschaft, BMLFUW - Bundesministeriums für Land- und Forstwirtschaft, Umwelt und Wasserwirtschaft, Forst, Forstabschätzung, Forsteinrichtung, Forsten, Forsttaxation, Forstwesen, Forstwirtschaft, Forstwirtschaftlich, Hochwald, Holzproduktion, Mischwald, Mittelwald, Niederwald, Nutzwald, Unterholz, Urwald, Waldbau, Waldwirtschaft

6.00 – DENKMALSCHUTZ

Bodendenkmale, Denkmal, Denkmalpflege, Denkmalrecht, Denkmalschutz, DSchG - Denkmalschutzgesetz, Europäische Konvention zum Schutz des archäologischen Erbes, Funde, Fundstellen, Historischen, Kulturdenkmal

6.01 – BODENDENKMAL

Archäologische, Archäologische Ausgrabungen, Archäologische Funde/Befunde, Denkmäler, Paläontologie

6.02 – KULTURDENKMAL

Baudenkmal, denkmalgeschützt, Denkmalwert, Haager Konvention zum Schutz von Kulturgut bei bewaffneten Konflikten, historisches Bauwerk, Kultur, Kulturgut, Menschheitsgeschichte, Monument, Museum, Naturdenkmal, öffentliches Interesse, Technikgeschichte, Technisches Denkmal, Zeitzeuge

7.00 – ABFALLWIRTSCHAFT

7.01 - ABFALLWIRTSCHAFT

Abfallentsorgung, Abfallvermeidung, Abwasserentsorgung, Abwasserreinigung, Kompost, Mülldeponie, Recycling, Sickerwasser, Stadtplanung, Abfall, Abfallberatung, Abfallwirtschaft, Abfallwirtschaftsamt, Abfallwirtschaftsgesetz, Kreislaufwirtschafts- und Abfallgesetz, Massenverwertungskonzept

8.00 - BRAND- UND KATASTROPHENSCHUTZ

8.1 - BRAND- UND Katastrophenschutz

Bevölkerungsschutz, Brand, Brandschutz, Brandschutzablauf, Brandschutzanforderung, Brandschutzanlage, Brandschutzanstrich, Brandschutzaufklärung, Brandschutzbeauftragte, Brandschutzbedarf, Brandschutzbegehung, Brandschutzbereich, Brandschutzbericht, Brandschutzprävention, Brandschutzrichtlinie, Brandschutzsachverständige, Brandschutzübung, Brandschutzverantwortliche, Brandschutzverordnung, Brandschutzwand, Feuerwehr (n), Gefahrguttransporte, Hochwasserschutz, Katastrophenschutz, Katastrophenschutzabkommen, Katastrophenschutzbehörde, Katastrophenschutzeinheit, Katastrophenschützer, Katastrophenschutzhelfer, Katastrophenschutzorganisation, Katastrophenschutzübung, Katastrophenschutzzug, Sicherheits- und Rettungskonzept, Sicherheitskonzept , Unfälle

9.00 - BAUZEITLICHE BEEINTRÄCHTIGUNGEN

9.01 - BAUZEITLICHE BEEINTRÄCHTIGUNGEN

Baudurchführung, Baulärm, Baulärms, Baumaschinen, Bauphase, Baustelleneinrichtungsflächen/ Baustraßen, Baustraßen, Bauzeit (z. B. Nachtarbeit), Bauzeitenplan, Bauzeitliche Beeinträchtigungen, Beeinträchtigungen durch Bauverkehr (Lärm, Abgase, Erschütterungen, Straßenverkehr), Entschädigung für Beeinträchtigungen während der Bauzeit, Gleisparallelen, Lautstärke, Materialtransport, Materialtransporte, Mobile, Schäden, Schäden durch Baumaßnahmen (Risse, Setzungen, etc.), Beweissicherung, Uhr, Wohnhäusern, Zustand

10.00 - ANLAGEN DRITTER

Entsorgungsanlagen, Infrastrukturleitungen und Sonstige Anlagen Dritter, Sicherung und Verlegung der Leitungen und Anlagen, Straßen, Wege, Zufahrten, Versorgungsanlagen

10.01 INFRASTRUKTURLEITUNGEN/SONSTIGE ANLAGEN DRITTER

Gewerbegebiet, Neubaustrecke, Plangebiet

10.02 (ÖFFENTLICHE) VER- UND ENTSORGUNGSANLAGEN

Endlager, Kläranlage, Kläranlagen, Kraftwerk, Müllhalde, Müllhalden, Wasserwerk

10.03 - STRAßEN, WEGE, ZUFAHRTEN

Auffahrt, Straße, Straßen, Straßenverkehr, Weg, Wege, Zufahrt, Zufahrten

11.00 - INANSPRUCHNAHME VON GRUNDEIGENTUM

Absicherung, Enteignung, Grunderwerb, Grundstücke, Liegenschaften, Verschattung, VwVfG (Verwaltungsverfahrensgesetz), Wert

11.01 - EIGENTUMSBEEINTRÄCHTIGUNGEN/GRUNDERWERB

VwVfG (Verwaltungsverfahrensgesetz), Sinkt, Marktwert, Minderung, Verlieren, Verlust, Eigentum, Eigentumsbeeinträchtigungen

11.02 - VERSCHATTUNG

Dunkel, kein Sonnenlicht, Licht, Verdunkeln, Verschatten