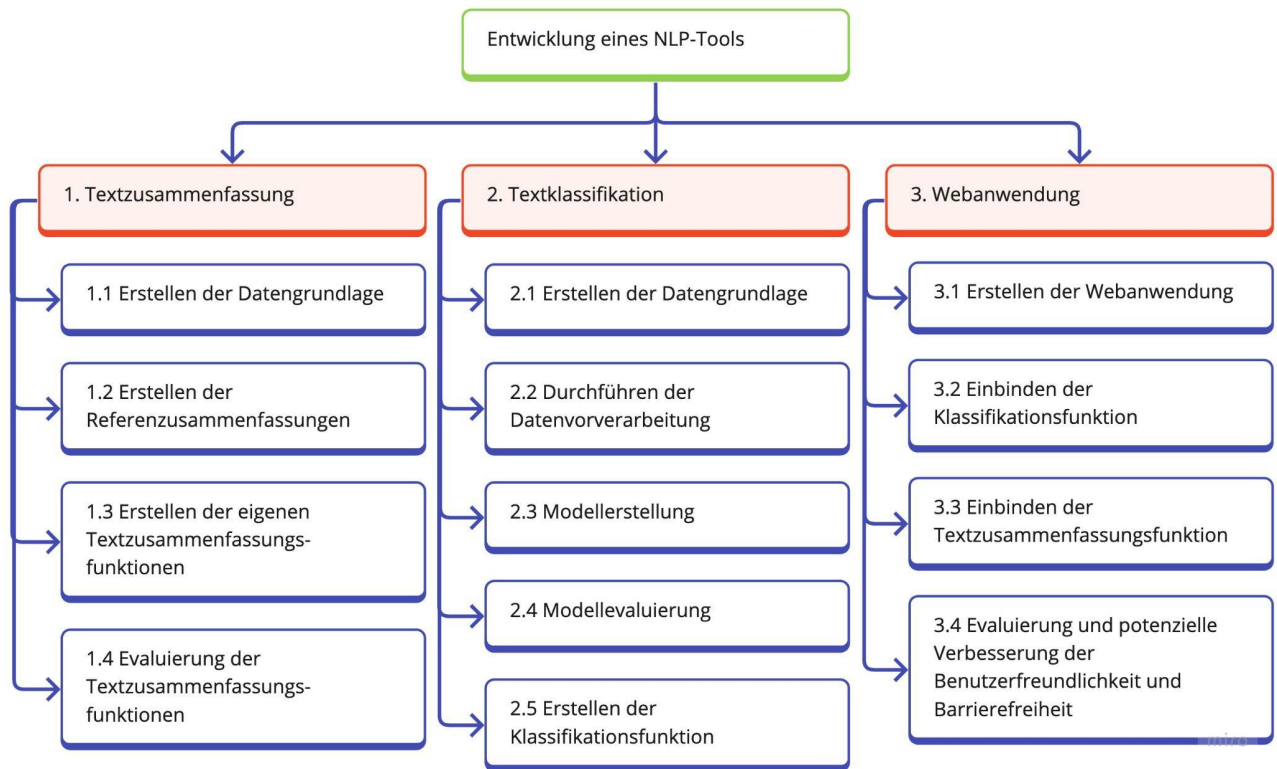


Detaillierter Projektstrukturplan

Gruppe 5: Amina Uicker-Darwish (8278962), Aymane Bouguern (1552312), Jan Rüdts (1737304), Lukas Bonn (5856761)



Beschreibung der Arbeitspakete:

Textzusammenfassung:

Arbeitspaket 1.1

PSP-Code: 1.1	Bezeichnung AP: Erstellen der Datengrundlage	Verantwortung AP: Gesamtes Team
Aufgabenstellung:	<ul style="list-style-type: none"> Durchführen von Recherchen, um eine solide Datengrundlage für die extractive Zusammenfassung zu schaffen Auswahl geeigneter Datensätze, die für die Zusammenfassung verwendet werden sollen Speichern der Texte in einer Liste 	
Zieldefinition	Es liegt eine Liste mit Texten vor, welche für die extractive	

(Ergebnisbeschreibung):	Zusammenfassung genutzt werden können	
Voraussetzungen (Vorgänger AP):	-	
Geplanter Aufwand (FTE)	76	
Hilfsmittel:	<ul style="list-style-type: none"> Suchmaschinen für Datensätze (Kaggle, Google Dataset Search, etc.) 	
Risiken:	<ul style="list-style-type: none"> Unzureichende Verfügbarkeit oder Qualität der Daten Begrenzter Zugriff auf bestimmte Datenquellen 	
Schnittstelle zu anderen AP:	<ul style="list-style-type: none"> AP 1.2 (Übergeben der Datengrundlage) AP 1.3 (Übergeben der Datengrundlage) 	
Abnahmebedingungen:	Eine umfangreiche und qualitativ hochwertige Datengrundlage wurde erstellt und zur weiteren Verarbeitung gespeichert	
Starttermin: 08.05.2023		Endtermin: 26.05.2023

Arbeitspaket 1.2

PSP-Code: 1.2	Bezeichnung AP: Erstellen der Referenzzusammenfassungen	Verantwortung AP: Amina Uicker-Darwish
Aufgabenstellung:	<ul style="list-style-type: none"> Erstellen von Referenzzusammenfassungen, welche als Grundlage für die spätere Evaluierung dienen Verwendung bestehender Tools zur Erstellung der Referenzzusammenfassungen (<i>sumy summarizer</i>) Zufällige Auswahl der Kompressionsrate und des Algorithmus (TextRank-, LSA- und LexRank-Algorithmus) für jede Zeile des Datensatzes, um dadurch Referenzzusammenfassungen mit minimalem Bias in Bezug auf das verwendete Tool zu erhalten Erstellen eines Dataframes mit den Originaltexten und den Referenzzusammenfassung als Grundlage für die weitere Nutzung 	
Zieldefinition (Ergebnisbeschreibung):	Es liegt für jeden Text der Datengrundlage eine dazugehörige Referenzzusammenfassung vor, welche durch bestehende Tools, welche auf verschiedenen Algorithmen beruhen, erstellt wurde. Die Referenzzusammenfassungen haben dabei keine fixe Kompressionsrate im Vergleich zu dem Originaltext	

Voraussetzungen (Vorgänger AP):	AP 1.1
Geplanter Aufwand (FTE)	14
Hilfsmittel:	<ul style="list-style-type: none"> Dokumentation der Zusammenfassungstools zur Erstellung der Referenzzusammenfassungen (<i>sumy summarizer</i>)
Risiken:	<ul style="list-style-type: none"> Herausforderungen bei der Erstellung der Referenzzusammenfassungen
Schnittstelle zu anderen AP:	AP 1.1 (Nutzung der Datengrundlage) AP 1.4 (Übergeben des Dataframes mit den Referenzzusammenfassungen und den Originaltexten)
Abnahmebedingungen:	Die Referenzzusammenfassungen erstellt sein und mit den dazugehörigen Originaltexten in einem Dataframe gespeichert sein. Es darf keine fixe Kompressionsrate für die Erstellung der Referenzzusammenfassungen genutzt worden sein. Es müssen verschiedene Algorithmen im Rahmen der Erstellung verwendet worden sein.
<div> <div>Starttermin: 27.05.2023</div> <div>Endtermin: 09.06.2023</div> </div>	

Arbeitspaket 1.3

PSP-Code: 1.3	Bezeichnung AP: Erstellen der eigenen Text-Zusammenfassungsfunktionen	Verantwortung AP: Aymane Bouguern Amina Uicker-Darwish
Aufgabenstellung:	<ul style="list-style-type: none"> Erstellen eigener Text-Zusammenfassungsfunktionen unter Verwendung von TextRank und LSA Vorverarbeitung und Bereinigung der Daten vor der Zusammenfassung 	
Zieldefinition (Ergebnisbeschreibung):	Es liegen zwei Zusammenfassungsfunktionen vor, welche TextRank respektive LSA verwenden	
Voraussetzungen (Vorgänger AP):	AP 1.1	
Geplanter Aufwand (FTE)	38	
Hilfsmittel:	<ul style="list-style-type: none"> Related work im Bereich der Nutzung von TextRank- und LSA-Algorithmen 	
Risiken:	<ul style="list-style-type: none"> Herausforderungen bei der Erstellung der Zusammenfassungsfunktionen 	

Schnittstelle zu anderen AP:	<ul style="list-style-type: none"> • AP 1.1 (Nutzung des Datensatzes) • AP 1.4 (Übergeben der Zusammenfassungsfunktionen) • AP 3.3 (Übergeben der "besseren" Zusammenfassungsfunktion)
Abnahmebedingungen:	<ul style="list-style-type: none"> • Eigene Text-Zusammenfassungsfunktionen müssen für TextRank bzw. LSA entwickelt und implementiert worden sein • Eingebene Texte müssen im Zuge der Funktionen vorverarbeitet und bereinigt werden, um dadurch die Zusammenfassung zu ermöglichen
Starttermin: 27.05.2023	Endtermin: 14.06.2023

Arbeitspaket 1.4

PSP-Code: 1.4	Bezeichnung AP: Evaluierung der Textzusammenfassungsfunktionen	Verantwortung AP: Aymane Bouguern Amina Uicker-Darwish
Aufgabenstellung:	<ul style="list-style-type: none"> • Erstellen von Zusammenfassungen unter Zuhilfenahme der erstellten Zusammenfassungsfunktionen für die Texte der Datengrundlage, wobei die Kompressionsrate verwendet wird, welche im Rahmen der Erstellung der Referenzzusammenfassungen genutzt wurde • Vergleichen der erstellten Zusammenfassungen mit den Referenzzusammenfassungen durch den Rouge-1 Score • Festlegen der geeigneten Zusammenfassungsfunktion 	
Zieldefinition (Ergebnisbeschreibung):	<ul style="list-style-type: none"> • Es liegt für jede Zusammenfassungsfunktion ein Rouge-1 Score vor • Die geeignete Zusammenfassungsfunktion mit dem höchsten Rouge-1 Score wurde ermittelt 	
Voraussetzungen (Vorgänger AP):	<ul style="list-style-type: none"> • AP 1.2 • AP 1.3 	
Geplanter Aufwand (FTE)	14	
Hilfsmittel:	-	
Risiken:	<ul style="list-style-type: none"> • Herausforderungen bei der Implementierung der Evaluierungsfunktion, Abweichungen zwischen den selbst erstellten und den Referenzzusammenfassungen 	

Schnittstelle zu anderen AP:	<ul style="list-style-type: none"> • AP 1.2 (Entgegennehmen der Referenzzusammenfassungen und Originaltexte) • AP 1.3 (Entgegennehmen der Zusammenfassungsfunktionen)
Abnahmebedingungen:	Für die Zusammenfassungsfunktionen wurde jeweils ein Rouge-1 Score berechnet und die zu verwendende Funktion wurde ermittelt
Starttermin: 15.06.2023	Endtermin: 21.06.2023

Textklassifikation:

Arbeitspaket 2.1

PSP-Code: 2.1	Bezeichnung AP: Erstellen der Datengrundlage	Verantwortung AP: Gesamtes Projektteam
Aufgabenstellung:	<ul style="list-style-type: none"> • Suche nach geeigneten Datensätzen mit über 500 Texten aus den folgenden fünf Oberkategorien: News-Artikel, klassische Literatur, politische Reden, Blogartikel, Juristische Texte • Laden der Texte zu den einzelnen Kategorien in einer Python-Datei (von Huggingface, aus einer Csv-Datei, etc.) und Speichern in jeweils einer Liste 	
Zieldefinition (Ergebnisbeschreibung):	Fünf Listen in einer Python-Datei mit mind. 500 Texten für die Oberkategorien News-Artikel, klassische Literatur, politische Reden, Blogartikel, Juristische Texte	
Voraussetzungen (Vorgänger AP):	-	
Geplanter Aufwand (FTE)	76	
Hilfsmittel:	- Suchmaschinen für Datensätze (Kaggle, Google Dataset Search, etc.)	
Risiken:	<ul style="list-style-type: none"> • Finden keiner passenden Datensätze aus den 5 Oberkategorien • Unzureichende Anzahl an Texten innerhalb einzelner Datensätze 	
Schnittstelle zu anderen AP:	AP 2.2 (Übergeben der Listen mit den Texten zur Vorverarbeitung der Texte)	
Abnahmebedingungen:	Es wurden geeignete Datensätze mit über 500 Texten für die 5 Oberkategorien gefunden und die Texte wurden in einer Python-Datei eingelesen und als Listen gespeichert	

Starttermin: 08.05.2023

Endtermin: 26.05.2023

Arbeitspaket 2.2

PSP-Code: 2.2	Bezeichnung AP: Durchführen der Datenvorverarbeitung	Verantwortung AP: Lukas Bonn
Aufgabenstellung:	<ul style="list-style-type: none">• Zusammenfügen der erstellten Listen zu einem Dataframe mit den jeweiligen Klassen• Gleichverteilen der Textanzahl für die unterschiedlichen Klassen• Ersetzen der Klassen gegen numerische Repräsentationen, um diese für die Modelle verständlich darzustellen• Durchführen weiterer Vorverarbeitungsschritte für die Klassifikation mit Modellen ausgenommen dem BERT-Modell (Kleinschreibung, Tokenisierung, Lemmatisierung, etc.)• Durchführen von Vorverarbeitungsschritten, um die Daten für das BERT-Modell vorzubereiten (Tokenizer, Hinzufügen von Attention_masks, Vorbereiten der DataLoader, etc.)	
Zieldefinition (Ergebnisbeschreibung):	Vorverarbeitete Texte mit den dazugehörigen Klassen, welche für die Nutzung in den Klassifikationsmodellen einschließlich dem BERT-Modell genutzt werden können	
Voraussetzungen (Vorgänger AP):	AP 2.1	
Geplanter Aufwand (FTE)	7	
Hilfsmittel:	Related Work im Bereich der Textvorverarbeitung für die Nutzung von Texten zur Klassifikation durch das BERT-Modell und andere ML-Modelle	
Risiken:	-	
Schnittstelle zu anderen AP:	<ul style="list-style-type: none">• AP 2.1 (Entgegennehmen der Listen mit den Texten zur Vorverarbeitung der Texte)• AP 2.5 (Übergeben des Dictionarys zur Umwandlung der numerischen Klassenrepräsentationen in die Originaltexte ('Political speech', 'News', etc.))	
Abnahmebedingungen:	Die Daten wurden so vorverarbeitet, dass sie von den zu verwendenden Modell genutzt werden können, um darauf trainieren und Vorhersagen treffen zu können	

Starttermin: 27.05.2023	Endtermin: 02.06.2023
-------------------------	-----------------------

Arbeitspaket 2.3

PSP-Code: 2.3	Bezeichnung AP: Modellerstellung	Verantwortung AP: Lukas Bonn Aymane Bouguern
Aufgabenstellung:	<ul style="list-style-type: none"> • Erstellen eines SVM, eines Naive Bayes und eines k-Nearest-Neighbors-Modells • Laden des vortrainierten Modells 'bert-base-uncased' • Trainieren der Modelle auf den vorverarbeiteten Daten 	
Zieldefinition (Ergebnisbeschreibung):	Vier Modelle, welche auf den vorverarbeiteten Daten trainiert (bzw. gefinetuned) wurden	
Voraussetzungen (Vorgänger AP):	AP 2.2	
Geplanter Aufwand (FTE)	10	
Hilfsmittel:	Related Work im Bereich der Textklassifikation zur Nutzung des BERT-Modells und der Modelle SVM, Naive Bayes und k-nearest Neighbors	
Risiken:	<ul style="list-style-type: none"> - Probleme beim Laden des BERT-Modells oder bei der Erstellung der weiteren Modelle 	
Schnittstelle zu anderen AP:	AP 2.4 (Übergeben der Modelle)	
Abnahmebedingungen:	Die Modelle wurden auf den vorverarbeiteten Daten trainiert und liegen in trainierter Form vor	
Starttermin: 03.06.2023		Endtermin: 07.06.2023

Arbeitspaket 2.4

PSP-Code: 2.4	Bezeichnung AP: Modellevaluierung	Verantwortung AP: Lukas Bonn
Aufgabenstellung:	<ul style="list-style-type: none"> • Evaluieren der trainierten Modelle auf einem Testdatensatz anhand der Accuracy (Anteil der korrekt klassifizierten Texte im Vergleich zu allen klassifizierten) • Vergleich der Evaluierungsergebnisse und Wahl des optimalen Modells für die Textklassifikation • Exportieren des optimalen Modells zur späteren Nutzung 	

Zieldefinition (Ergebnisbeschreibung):	Optimales Klassifikationsmodell liegt als Pickle-File vor	
Voraussetzungen (Vorgänger AP):	AP 2.3	
Geplanter Aufwand (FTE)	3	
Hilfsmittel:	-	
Risiken:	Evaluierung auf Testdaten führt zu unzureichenden Ergebnissen in Bezug auf die Accuracy	
Schnittstelle zu anderen AP:	AP 2.3 (Entgegennehmen der Modelle) AP 2.5 (Übergeben des optimalen Modells)	
Abnahmebedingungen:	Optimales Klassifikationsmodell liegt als Pickle-File vor und besitzt eine Accuracy von mindestens 95%	
Starttermin: 09.06.2023		Endtermin: 11.06.2023

Arbeitspaket 2.5

PSP-Code: 2.5	Bezeichnung AP: Erstellen der Klassifikationsfunktion	Verantwortung AP: Lukas Bonn
Aufgabenstellung:	<ul style="list-style-type: none"> • Erstellen einer Funktion zur Textklassifikation • Funktion nimmt Text entgegen, verarbeitet diesen entsprechend dem optimalen Modell vor und trifft eine Vorhersage anhand des aus dem Pickle-File geladenen Klassifikationsmodells • Festgelegter Schwellwert in der Funktion bestimmt, ob die vom Modell bestimmte Klasse oder eine Fehlermeldung zurückgegeben werden soll ('Der Text konnte nicht klassifiziert werden') 	
Zieldefinition (Ergebnisbeschreibung):	Klassifikationsfunktion steht zu Verfügung und kann dazu genutzt werden eingegebene Texte anhand des optimalen Modells zu klassifizieren	
Voraussetzungen (Vorgänger AP):	AP 2.4	
Geplanter Aufwand (FTE)	4	
Hilfsmittel:	-	
Risiken:	-	
Schnittstelle zu anderen AP:	<ul style="list-style-type: none"> • AP 2.2 (Entgegennehmen des Dictionarys zur Umwandlung der numerischen 	

	<p>Klassenrepräsentationen in die Originaltexte ('Political speech', 'News', etc.))</p> <ul style="list-style-type: none"> • AP 2.4 (Entgegennehmen des optimalen Modells) • AP 2.5 (Übergeben der Klassifikationsfunktion an die Webanwendung) • AP 3.2 (Übergeben der optimalen Klassifikationsfunktion)
Abnahmebedingungen:	Klassifikationsfunktion liegt vor und kann dazu verwendet werden, Texte zu klassifizieren
Starttermin: 12.06.2023	Endtermin: 15.06.2023

Webanwendung:

Arbeitspaket 3.1

PSP-Code: 3.1	Bezeichnung AP: Erstellen der Webanwendung	Verantwortung AP: Jan Rüdts
Aufgabenstellung:	<ul style="list-style-type: none"> • Konzeption und Entwicklung einer Webanwendung, die das NLP-Modul für Textklassifikation und -zusammenfassung integriert • Integration eines Schiebereglers zur Auswahl einer Kompressionsrate zwischen 20% und 80% • Erstellen einer Möglichkeit zur Texteingabe über ein Eingabefeld oder per Dateiupload im TXT, DOC(X) oder ODT Format • Implementierung von Feldern zur Auswahl der gewünschten Funktionen (Klassifikation und/oder Zusammenfassung) • Implementierung einer benutzerfreundlichen und barrierefreien Oberfläche (z.B. Mega Zoom-Button mit 300% Vergrößerung, Kontraständerung) mit verschiedenen Funktionen zur Texteingabe und -ausgabe 	
Zieldefinition (Ergebnisbeschreibung):	Es liegt eine erfolgreich entwickelte Webanwendung vor, welche über eine benutzerfreundliche und barrierefreie Oberfläche verfügt und verschiedene Funktionen bietet (Texteingabe und Hochladen eines Textes, Wahl einer Kompressionsrate zwischen 20% und 80%, Auswahl gewünschter Ausgabeinformationen)	
Voraussetzungen (Vorgänger AP):	-	

Geplanter Aufwand (FTE)	13
Hilfsmittel:	
Risiken:	<ul style="list-style-type: none"> • Komplexität der Implementierung der verschiedenen Funktionen • Herausforderungen bei der Umsetzung der Benutzerfreundlichkeit und Barrierefreiheit
Schnittstelle zu anderen AP:	<ul style="list-style-type: none"> • AP 3.2 (Übergeben der erstellten Webanwendung) • AP 3.3 (Übergeben der erstellten Webanwendung) • AP 3.4 (Übergeben der erstellten Webanwendung)
Abnahmebedingungen:	<ul style="list-style-type: none"> • Die Webanwendung wurde erfolgreich entwickelt und enthält alle geforderten Funktionen • Die Benutzerfreundlichkeit und Barrierefreiheit der Webanwendung erfüllt die Anforderungen
Starttermin: 17.05.2023	Endtermin: 29.05.2023

Arbeitspaket 3.2

PSP-Code: 3.2	Bezeichnung AP: Einbinden der Klassifikationsfunktion	Verantwortung AP: Jan Rüdts
Aufgabenstellung:	<ul style="list-style-type: none"> • Integration der evaluierten Klassifikationsfunktion in die Webanwendung • Anpassung der Schnittstellen und Funktionen der Webanwendung zur Nutzung der Klassifikationsfunktion 	
Zieldefinition (Ergebnisbeschreibung):	Die Klassifikationsfunktion ist in die Webanwendung eingebunden und kann dazu genutzt werden, übergebene Texte in die fünf Oberkategorien zu klassifizieren	
Voraussetzungen (Vorgänger AP):	<ul style="list-style-type: none"> • AP 2.5 • AP 3.1 	
Geplanter Aufwand (FTE)	3	
Hilfsmittel:	-	
Risiken:	<ul style="list-style-type: none"> • Kompatibilitätsprobleme zwischen der Klassifikationsfunktion und der Webanwendung • Anpassungen könnten zusätzlichen Aufwand erfordern 	

Schnittstelle zu anderen AP:	<ul style="list-style-type: none"> • AP 2.5 (Entgegennehmen der optimalen Klassifikationsfunktion) • AP 3.1 (Entgegennehmen der erstellten Webanwendung)
Abnahmebedingungen:	<ul style="list-style-type: none"> • Die Klassifikationsfunktion wurde erfolgreich in die Webanwendung integriert • Die Webanwendung nutzt die Klassifikationsfunktion zur Textklassifikation
Starttermin: 16.06.2023	Endtermin: 18.06.2023

Arbeitspaket 3.3

PSP-Code: 3.3	Bezeichnung AP: Einbinden der Textzusammenfassungsfunktion	Verantwortung AP: Jan Rüdts
Aufgabenstellung:	<ul style="list-style-type: none"> • Integration der evaluierten Funktion zur Textzusammenfassung in die Webanwendung • Anpassen der Schnittstellen und Funktionen der Webanwendung zur Nutzung der Textzusammenfassungsfunktion 	
Zieldefinition (Ergebnisbeschreibung):	Die im Rahmen der Evaluierung als bessere ermittelte Textzusammenfassungsfunktion ist in die Webanwendung eingebunden und kann dazu genutzt werden, eingegebene Texte zusammenzufassen	
Voraussetzungen (Vorgänger AP):	<ul style="list-style-type: none"> • AP 1.4 • AP 3.1 	
Geplanter Aufwand (FTE)	6	
Hilfsmittel:	-	
Risiken:	<ul style="list-style-type: none"> • Kompatibilitätsprobleme zwischen der Textzusammenfassungsfunktion und der Webanwendung • Anpassungen könnten zusätzlichen Aufwand erfordern 	

Schnittstelle zu anderen AP:	<ul style="list-style-type: none"> • AP 1.3 (Entgegennehmen der “besseren” Zusammenfassungsfunktion) • AP 3.1 (Entgegennehmen der erstellten Webanwendung)
Abnahmebedingungen:	<ul style="list-style-type: none"> • Die Textzusammenfassungsfunktion wurde erfolgreich in die Webanwendung integriert • Die Webanwendung nutzt die Textzusammenfassungsfunktion zur automatischen Zusammenfassung von Texten
Starttermin: 22.06.2023	Endtermin: 27.06.2023

Arbeitspaket 3.4

PSP-Code: 3.4	Bezeichnung AP: Evaluierung und potenzielle Verbesserung der Benutzerfreundlichkeit und Barrierefreiheit	Verantwortung AP: Jan Rüdts
Aufgabenstellung:	<ul style="list-style-type: none"> • Durchführung von Tests zur Bewertung der Benutzerfreundlichkeit der Webanwendung • Überprüfung der Barrierefreiheit der Webanwendung durch den Einsatz spezialisierter Tools und Richtlinien • Auswerten der Tests zur Benutzerfreundlichkeit und Barrierefreiheit der Webanwendung • Identifizieren und Prüfen von Verbesserungspotenzialen und Optimieren der Benutzererfahrung 	
Zieldefinition (Ergebnisbeschreibung):	Es liegt eine getestete Webanwendung vor, welche potenzielle Verbesserungen in Bezug auf die Benutzerfreundlichkeit und Barrierefreiheit beinhaltet	
Voraussetzungen (Vorgänger AP):	<ul style="list-style-type: none"> • AP 3.1 	
Geplanter Aufwand (FTE)	17	
Hilfsmittel:	<ul style="list-style-type: none"> • Spezialisierte Tools zur Überprüfung der Barrierefreiheit (z.B. <i>Wave evaluation tool</i>) 	

Risiken:	-
Schnittstelle zu anderen AP:	<ul style="list-style-type: none"> • AP 3.1 (Entgegennehmen der erstellten Webanwendung)
Abnahmebedingungen:	Die Barrierefreiheit und Benutzerfreundlichkeit der Webanwendung wurde getestet und mögliche Verbesserungsvorschläge wurden geprüft und potenziell angepasst
Starttermin: 28.06.2023	Endtermin: 14.07.2023