



Duale Hochschule Baden-Württemberg Mannheim

Projektrealisierung

Entwicklung eines Tools zur Textzusammenfassung und –klassifikation

Studiengang Wirtschaftsinformatik

Studienrichtung Data Science

Verfasser(in):	J. Brenzinger, L. Steinert, P. Breucker, C. Rech
Matrikelnummern:	1960679, 2617416, 5800129, 8028907
Kurs:	WWI20DSA
Bearbeitungszeitraum:	08.05.2023 – 24.05.2023

Inhaltsverzeichnis

1	Projektauftrag	1
1.1	Ausgangssituation	1
1.2	Projektziel	1
1.3	Strategische Einordnung	2
1.4	Projekthalt	3
1.5	Projektmeilensteine	3
1.6	Risiken und Gegenmaßnahmen	4
1.7	Risiken bei Nichtdurchführung des Projekts	4
2	Lastenheft	5
2.1	Vorbemerkungen	5
2.2	Projekteinführung	5
2.3	Beschreibung der Ausgangssituation (Ist-Situation)	6
2.4	Beschreibung der Zielsetzung (Soll-Zustand)	7
2.5	Anforderungen an die Projektdurchführung	8
2.5.1	Funktionale Anforderungen	9
2.5.2	Nichtfunktionale Anforderungen	9
2.6	Definition von Schnittstellen und Zuständigkeiten	10
3	Pflichtenheft	11
3.1	Vorbemerkung	11
3.2	Ausgangssituation und Zielsetzung	11
3.3	Schnittstelle und Beteiligte	12
3.4	Rahmenbedingungen	12
3.5	Dekomposition des Gesamtsystems	13
3.6	Funktionale Anforderungen	14
3.7	Nicht-funktionale Anforderungen	16
3.8	Qualitätsüberprüfung	17
3.8.1	Abnahmekriterien	17
3.8.2	Vorgehen zur Ausgangsprüfung:	17
3.9	Lieferumfang	18
	Literaturverzeichnis	19

1 Projektauftrag

Projektsteckbrief

Projektname:	Textzusammenfassung und Textklassifikation
Auftraggeber:	i.A. Michael Lang, Enzo Hilzinger
Projektmitglieder:	Pascal Breucker, Joshua Brenzinger Constantin Rech, Luis Steinert
Priorität und Tragweite	Hoch
Projektdauer	von 09.05.2023 bis 27.07.2023
Erstellungsdatum	09.05.2023

1.1 Ausgangssituation

Ein Meinungsforschungsinstitut beabsichtigt, verschiedene Quellen zu analysieren, um Stimmungsbilder zu bestimmen und auf dieser Grundlage aussagekräftige Dashboards zu entwickeln. Um diesen Prozess effizienter zu gestalten und die Analyse der Textdaten zu erleichtern, soll ein Tool zur Textzusammenfassung und -klassifikation entwickelt werden. Das Tool wird erwartet, dass es eine automatisierte Verarbeitung und Bewertung der Textquellen ermöglicht, um fundierte Entscheidungen auf Basis der Analyse zu treffen.

1.2 Projektziel

Das Ziel dieses Projektes besteht darin, ein Tool zur Textzusammenfassung und -klassifikation unter Berücksichtigung folgender Rahmenbedingungen zu entwickeln: Das Tool soll einen gegebenen Text entsprechend einer gewünschten Kompressionsrate zusammenfassen. Die Kompressionsrate entspricht der Reduzierung der Anzahl der Worte. Die Kompressionsrate soll anpassbar sein, um den Benutzerpräferenzen gerecht zu werden. Die Nachrichtenartikel sollen in die häufigsten Resorts kategorisiert werden (Foreign/Culture/Business/Washington/Metro/Science/National/Learning/RealEstate/ Politics/ Sports/...). Darüber hinaus *nicht nur Nachrichtenartikel!*

soll die Verwendung eines mehrstufigen Ansatzes, einschließlich Datenauswahl, Vorverarbeitung und Modelltraining, wobei geeignete Algorithmen ausgewählt und angepasst werden. Bewertung des erstellten Tools mithilfe geeigneter Metriken wie Confusionmatrix, Accuracy, Precision, Recall, Matthews-Correlationscoefficient (MCC), F1-Score bzw. dem Classificationreport der Scikit-learn-Bibliothek. Ein transparenter Entwicklungsprozess durch regelmäßige Commits auf GitHub muss gegeben sein. Im Rahmen des Projekts müssen die im Lastenheft aufgeführten Leistungen erbracht und im Pflichtenheft dargestellt werden.

) Aktual?

1.3 Strategische Einordnung

Die Entwicklung dieses Tools ist strategisch wichtig, da es dem Meinungsforschungsinstitut ermöglicht, die Analyse- und Auswertungsprozesse zu optimieren. Durch die Automatisierung der Textverarbeitung kann das Institut schneller und effizienter Stimmungsbilder generieren, was zu fundierten Entscheidungen und besseren Einblicken in Meinungen und Trends führt. Das einheitliche Zusammenfassen von Texten ermöglicht zudem eine Kostenersparnis durch Aufwandsreduktion manueller Arbeit und ermöglicht darüber hinaus die Erstellung von Datengrundlagen zukünftiger Analysen.

1.4 Projektinhalt

	Inhalt
In Scope	<p>Textbox zur Eingabe eines Freitexts</p> <p>Datei Upload verschiedener Dateitypen (.txt, .pdf und .doc)</p> <p>Verarbeitung von deutsch- und englischsprachigen Texten</p> <p>Individuelle Anpassung der Kompressionsrate</p> <p>Klassifikation des Eingabetexts anhand definierter Oberkategorien</p> <p>Durchführung einer Stimmungsanalyse des Eingabetexts</p> <p>Download der Ausgabe in verschiedenen Dateitypen (.txt, .pdf und .doc)</p> <p>Ausgabe der Ergebnisse als Freitext</p> <p>Analyse der Anforderungen und Spezifikationen</p> <p>Entwicklung erforderlicher Algorithmen und Modelle</p> <p>Durchführung von Tests und Qualitätskontrollen</p> <p>Dokumentation und Bereitstellung von Schulungsmaterial</p> <p>Datenbereinigung und -aufbereitung</p>
Out of Scope	<p>Technischer Support nach Implementierung</p> <p>Erhebung der Textdaten aus verschiedenen Quellen</p> <p>Employer Branding</p> <p>Persistente Speicherung der Eingabe- und Ausgabetexte oder der Klassifikationsergebnisse</p> <p>Entwicklung von Visualisierungstools für die Stimmungsbilder</p> <p>Durchführung von Meinungsumfragen</p>

1.5 Projektmeilensteine

Datum	Inhalt
17.05.23	GitHub Repository anlegen
24.05.23	<p>Projektauftrag</p> <p>Lasten-/Pflichtenheft</p> <p>Projektstrukturplan (grob)</p>
22.06.23	<p>detaillierter Projektstrukturplan</p> <p>Netzplan/Gantt-Chart und Folien für Präsentation</p>
27.07.23	<p>Projektabschlussbericht incl. Code + Kommentare</p> <p>Lessons Learned, Reflexion und Präsentation</p>

ohne groß
gefasst

1.6 Risiken und Gegenmaßnahmen

Risiko	Gegenmaßnahme
Unvorhergesehene technische Schwierigkeiten bei der Entwicklung	regelmäßige Überprüfung des Entwicklungsfortschritts. Frühzeitige Identifizierung potenzieller Herausforderungen, um die technischen Schwierigkeiten zu überwinden.
Unzureichende Datenqualität oder -verfügbarkeit für die Stimmungsanalyse	Implementierung von Datenvalidierungs- und Bereinigungsmechanismen, um sicherzustellen, dass die Daten für die Analyse geeignet sind. Ggf. Zusammenarbeit mit <u>Datenanbietern</u> , um den Zugang zu qualitativ hochwertigen Datenquellen sicherzustellen.
Mangelnde Akzeptanz oder Nutzung des Tools durch die Benutzer	Einbeziehung der Benutzer von Anfang an in den Entwicklungsprozess, um deren Bedürfnisse und Anforderungen zu verstehen. Regelmäßige Kommunikation und Schulung der Benutzer, um sicherzustellen, dass sie das Tool effektiv nutzen können.

Maßnahmen
? realisiert?

sicherlich

✓

1.7 Risiken bei Nichtdurchführung des Projekts

Neben dem Verlust der Möglichkeit, automatisch Stimmungsbilder aus verschiedenen Quellen zu analysieren und darauf aufbauend fundierte Entscheidungen zu treffen, erhöht der manuelle Aufwand die Fehleranfälligkeit bei der Verarbeitung großer Textmengen. Außerdem kann es zu Verzögerungen bei der Generierung von aussagekräftigen Stimmungsberichten und Dashboards kommen.

2 Lastenheft

2.1 Vorbemerkungen

Eine allgemeine Definition des Lastenhefts ist durch die Normenreihe DIN 69901 des Deutschen Instituts für Normung gegeben. Nach der DIN 69901 ist ein Lastenheft ein „vom Auftraggeber festgelegte Gesamtheit der Forderungen an die Lieferungen und Leistungen eines Auftragnehmers innerhalb eines (Projekt-)Auftrags“. Der Ersteller ist der Auftraggeber. Im Lastenheft werden die Anforderungen des Kunden oder Auftraggebers an das Projekt beschrieben. Dabei sollten alle wichtigen Informationen wie Funktionalitäten, Leistungen, Qualitätsanforderungen, Liefertermine, Funktionale und Nichtfunktionale Anforderungen, sowie Budgetvorgaben aufgenommen werden. Auch die Beschreibung der Ausgangslage zum Projekt (Ist-Situation, Ziel und Zweck, Geltungsbereich etc.) sind Teil dieses Dokuments (vgl. Aichele et al., 2014, S.138 ff).

Zweck: Dieses Lastenheft enthält die Zusammenstellung aller Anforderungen des Auftraggebers hinsichtlich des Liefer- und Leistungsumfangs. Es beschreibt die fachlichen Basisanforderungen aus Anwendersicht, einschließlich aller Randbedingungen bzw. Rahmenbedingungen an das Projekt der Textzusammenfassung und -klassifikation.

Referenzen: Die folgende Gliederung des Lastenheftes basiert auf den Beschreibungen: Aichele et al., 2014, sowie Peter-Johann, 2023.

2.2 Projekteinführung

Das vorliegende Projekt dient der Entwicklung eines Tools zur Textzusammenfassung und -klassifikation, das in verschiedenen Bereichen eines Meinungsforschungsinstituts Anwendung finden wird. Als Hauptanwendungsfall soll das Tool im Rahmen einer Analyse von Stimmungsbildern verwendet werden, um die Zusammenfassung von großen Mengen an Texten zu erleichtern und die gleichzeitige Kategorisierung dieser Texte zu ermöglichen. Die Entwicklung soll hierbei die folgenden Spezifikationen einschließen: Die inhaltliche

Zusammenfassung des Eingabetextes soll anhand einer bestimmten, änderbaren Kompressionsrate stattfinden, um den spezifischen Anforderungen der verschiedenen Bereiche zu entsprechen. Die Kompressionsrate soll in **Prozent** angegeben werden und entspricht einer Reduktion der Anzahl der Worte des Eingabetextes. Darüber hinaus soll eine Textklassifikation des Eingabetextes in Oberkategorien stattfinden. Diese Kategorien können beispielsweise sein: **Foreign/Culture/Business/Washington/Metro/Science/National/Learning/RealEstate/ Politics/ Sports**. Eine Eingrenzung in der Anzahl der möglichen Kategorisierungen ist nicht gegeben. Die Entwicklung soll unter Verwendung eines mehrstufigen Ansatzes stattfinden, welcher die Schritte der Datenauswahl, Vorverarbeitung und dem Modelltraining einschließt, wobei eine geeignete Auswahl vortrainierter Modelle möglich ist. Um die Güte des Modells zu bestimmen, soll das erstellte Tools mithilfe verschiedener, geeigneter Metriken bewertet werden. Der Entwicklungsprozess ist **transparent darzulegen**. Um den Entwicklungsprozess transparent und nachvollziehbar zu gestalten werden weiterhin die folgenden Meilensteine definiert, anhand der das Projekt strukturiert werden soll. Diese Meilensteine bilden die Rahmenbedingungen des Projekts:

09.05.2023: Kick-Off des Projektes zwischen den Auftraggebern und dem Projektteam. Es werden die relevanten Meilensteine und Deadlines des Projekts definiert.

24.05.2023: Abgabe des Projektauftrages, Lasten- und Pflichtenheftes, sowie des Projektstrukturplans und Vorstellung dieser im Rahmen einer Präsentation an die Auftraggeber. Es wird der genaue Projektumfang, sowie die spezifischen Anforderungen besprochen.

22.06.2023: Abgabe des detaillierten Projektstrukturplans, eines Netzplans / Gantt-Charts und der Präsentation

27.07.2023: Projektabschlussbericht, Lessons Learned, Reflexion, Folien für Präsentation, finales Coding inkl. Kommentaren

S. v.

2.3 Beschreibung der Ausgangssituation (Ist-Situation)

Innerhalb des Meinungsforschungsinstituts werden in verschiedenen Bereichen regelmäßig große Mengen an Texten aus verschiedenen Dateitypen manuell analysiert, inhaltlich zusammengefasst und klassifiziert. Es besteht daher für verschiedene Anwendungsbereiche

der allgemeine Bedarf nach einer eigenstehende Lösung, welche die automatisierte Textzusammenfassung und -kategorisierung ausgehend von einem Eingabetext ermöglicht. Bei der Zusammenfassung wird eine kürzere Version eines Dokuments oder eines Artikels erstellt, die alle wichtigen Informationen enthält. Bei der Einordnung des Eingabetexts in Oberkategorien, dementsprechend der Klassifikation des Eingabetexts, wird dem Text eine Bezeichnung oder Klasse zugeordnet, die diesen beschreibt. Die Analyse und Kategorisierung der Texte wird in den Bereichen des Instituts momentan manuell durchgeführt. Zum einen bedeutet dies einen hohen Zeitaufwand. Zum anderen ist es dadurch nicht möglich, eine bereichsübergreifend gleichbleibende Qualität sicherzustellen. Die Anwendungsbereiche umfassen **deutsch-** und **englischsprachige** Texte. Da verschiedene Dateitypen im Regelbetrieb verarbeitet werden müssen, muss diese Lösung darüber hinaus einen Dateiupload und die Einleseprozesse ermöglichen. Innerhalb des Instituts existiert bislang keine vergleichbare Lösung; das zu erbringende Produkt ist somit von Grund auf neu zu entwickeln. Deshalb ist die künftige Software im Rahmen der Entwicklung an keine institutsinterne Infrastruktur der IT oder verwendete Technologien anzupassen oder in einem spezifischen Entwicklungsumfeld zu entwickeln. Abseits der in diesem Lastenheft benannten technischen Anforderungen sind keine weiteren Einschränkungen vorgegeben. Die finale Web-Anwendung muss weiterhin keinem Employer Branding des Meinungsforschungsinstituts entsprechen.

Die Anwender des Systems können in die folgenden Gruppen unterteilt werden:

Endanwender: Diese nehmen Zugriff auf das Produkt, um eine inhaltliche Zusammenfassung und Klassifikation, sowie Stimmungsanalyse eines Eingabetexts zu erhalten.

Entwickler: Diese nehmen Zugriff auf das Produkt, um die Login-Daten des Tools anzupassen.

2.4 Beschreibung der Zielsetzung (Soll-Zustand)

Unter einem Ziel wird ein angestrebter Zustand verstanden, welcher durch zu erarbeitende Maßnahmen oder Lösungen erreicht werden soll. Über die Projektziele wird der zu erreichende Soll-Zustand konkretisiert und definiert.

Die Zielstellung der Anwender ist die vereinfachte, automatisierte Reduktion eines Textdokuments (bspw. Text-Dateien, PDF-Dateien, Word-Dokumente), sowie eines Freitexts, welcher in das Produkt zu kopieren oder übernehmen ist. Die Zusammenfassung von Texten

kann in extraktive und abstraktive Verfahren unterteilt werden. Bei extraktivem Vorgehen handelt es sich um eine Selektion von vorkommenden Sätzen/Abschnitten, denen eine besondere Wichtigkeit zugeordnet wird. Bei abstraktiven Methoden wird eine Zusammenfassung durch Paraphrasieren der Hauptmerkmale erzeugt. Das Produkt soll, ausgehen von dem Eingabetext, einen neuen Ausgabetext erzeugen, der die wichtigsten und relevantesten Informationen enthält und somit eine abstraktive Zusammenfassung durchführen. Es sollen sowohl englischsprachige als auch deutschsprachige Texte verarbeitet werden.

Der Zugriff der Endanwender auf das Produkt soll über eine Web-Anwendung stattfinden, welche gehostet wird. Die Web-Anwendung soll die folgenden Funktionalitäten besitzen:

- Einen Dateiupload für verschiedene Dateitypen
- Eine Eingabemöglichkeit für Freitext
- Anpassbare Kompressionsrate für den Eingabetext in Prozent soll möglich sein
- Ausgabe des Ergebnisses der Zusammenfassung, sowie der zugeordneten Kategorie des Eingabetexts
- Infobox zu Logik hinter Kompression und Kategorisierung

→ Handl?

Die Anwender können die Kompressionsrate der Zusammenfassung auf der Webanwendung anpassen. Die Anwender können die Ausgabe in verschiedenen Dateitypen abspeichern (.txt, .pdf und .doc). Ein Nichtziel stellt die Ausgabe verschiedener bzw. mehrerer zugeordneter Kategorien im Rahmen der Kategorisierung des Eingabetexts dar. Die tatsächliche Reduktion der Anzahl der Worte der Zusammenfassung kann um bis zu 10% von der angegebenen Kompressionsrate abweichen. Eine persistente Speicherung der Eingabe- oder Ausgabetexte, sowie der Klassifizierungen soll nicht stattfinden.

→ Change-Handling?

2.5 Anforderungen an die Projektdurchführung

Ausgehend von der Beschreibung des Soll-Zustands der Lösung sind folgend die konkreten Detailziele, die funktionalen und nichtfunktionalen Anforderungen, bestimmt. Diese werden nach dem Schema "REQ000" dokumentiert. Anforderungen beginnend mit der Ziffern 1 = Funktionale Anforderungen; Ziffer 3 = Nichtfunktionale Anforderungen

2.5.1 Funktionale Anforderungen

REQ101 Das Produkt ermöglicht die Zusammenfassung von Texten

REQ102 Das Produkt ermöglicht die Eingabe von Freitexten, sowie einen Dateiupload von verschiedenen Dateitypen (.txt, .pdf und .doc)

REQ103 Bei der Textkompression des Produkts handelt es sich um eine abstraktive Testzusammenfassung

↳ wenn nicht abstraktiv?

REQ104 Die Benutzeroberfläche ermöglicht eine individuelle Festlegung der Kompressionsrate zwischen 20 und 80 Prozent.

REQ105 Das Ergebnis der Zusammenfassung wird in der Webapplikation angezeigt

REQ106 Die Benutzeroberfläche ermöglicht die Speicherung des Ergebnisses als Word, Text oder PDF-Datei → wenn nicht 3?

REQ107 Das System unterstützt die Verarbeitung von deutschen und englischen Texten.

REQ108 Im Kontext von News Artikeln wird zwischen den Kategorien Sport, Kultur, Gesellschaft, Politik, Wissenschaft, Wirtschaft (...) unterschieden

was ist mit "Sonstige"?

REQ109 Die Oberkategorie des Eingabetextes wird angezeigt im UI der Webapplikation

REQ110 Es ist möglich, ausschließlich Dateien der gegebenen Kategorien hochzuladen

REQ111 Das Produkt führt eine Stimmungsanalyse des Eingabetexts durch

2.5.2 Nichtfunktionale Anforderungen

REQ301 Zugang zur Web-Applikation ist nur bestimmten Nutzergruppen möglich.

REQ302 Die Web-Applikation muss keinem Employer Branding entsprechen.

REQ303 Die Web-Applikation sollte eine Benutzerfreundlichkeit aufweisen und ein intuitives Benutzerinterface bieten. → Definition?

REQ304 Das System muss bei Bedarf verfügbar sein. SLA?

REQ305 Die Webapplikation sollte auf verschiedenen Browsern und Endgeräten (Desktop, Tablet, Mobil) gut funktionieren und ein responsives Design aufweisen.

REQ306 Die Qualität der Ergebnisse soll bewertet werden.

2.6 Definition von Schnittstellen und Zuständigkeiten

Die Kommunikation mit den Auftraggebern findet über einen Microsoft Teams Kanal statt. Abseits der definierten Meilensteine sind wöchentliche Austauschtermine geplant. Die Auftragnehmer unterrichten die Auftraggeber in den definierten wöchentlichen Sitzungen über die aktuelle Entwicklung des Projektes und etwaiger Hindernisse oder Schwierigkeiten. Überdies erhalten die im Auftrag des Meinungsforschungsinstitut agierenden Auftraggeber Zugang zum System während der Entwicklung über Github, was vom Auftragnehmer bereitgestellt und gepflegt werden soll.

3 Pflichtenheft

3.1 Vorbemerkung

Basierend auf dem Lastenheft erfolgt anschließend die Beschreibung der Umsetzung der Anforderung im Zuge der Erstellung des Pflichtenheftes. Nach der DIN 69901 enthält ein Pflichtenheft „vom Auftragnehmer erarbeitete Realisierungsvorgaben auf der Basis des vom Auftraggeber vorgegebenen Lastenhefts“. Juristisch stellt das Pflichtenheft die im Lastenheft **definierten Spezifikation** in feingranulare fachlicher Art und Weise dar. Dagegen betrachtet die IT das Pflichtenheft als Beschreibung der technischen Lösung zu den im Lastenheft definierten Anforderungen. Aus der Sicht des Auftragnehmers stellt das Pflichtenheft demnach die formelle und detaillierte Antwort auf die Anforderungen des Auftraggebers dar, die zuvor im Lastenheft beschrieben wurden. Die zu erbringenden Ergebnisse des Auftragnehmers werden dadurch in erforderliche Tätigkeiten (Pflichten) umgesetzt (vgl. Aichele et al., 2014, S.173 ff).

3.2 Ausgangssituation und Zielsetzung

Die Ausgangssituation besteht darin, dass das Meinungsforschungsinstitut regelmäßig große Mengen an Texten analysiert, diese zusammengefasst und klassifiziert werden müssen. Das Ziel ist daher die Entwicklung eines Tools, das die Effizienz und Genauigkeit dieses Prozesses verbessert, indem es automatische Textzusammenfassung und -klassifikation ermöglicht.

Basierend auf dem Lastenheft des Meinungsforschungsinstitutes widmen sich die nachstehenden Kapitel der Beschreibung jeglicher Anforderungen und wie diese funktional und nicht-funktional umgesetzt werden sollen. Abgrenzend zu erwähnen sei die Tatsache, dass wie im Lastenheft beschrieben der Projektauftrag eine komplette Neuentwicklung bedeutet. Es bestehen keine Systeme oder Produkte bisher in der Nutzung.

3.3 Schnittstelle und Beteiligte

Beteiligt an diesem Projekt sind zum einen die Auftraggeber Enzo Hilzinger und Michael Lang, sowie die Gruppe 1 als Entwickler und Auftragnehmer. Der fiktive Usecase eines Meinungsforschungsinstitut geht dabei zurück auf die von den Auftraggebern geforderte Entwicklung des Produktes im Rahmen der Vorlesung "Projektrealisierung".

3.4 Rahmenbedingungen

Im Rahmen dieses Projekts „Entwicklung eines Tools zur Textzusammenfassung und -klassifikation“ sind bestimmte Rahmenbedingungen festzulegen, um eine reibungslose Umsetzung zu gewährleisten. Diese Rahmenbedingungen beinhalten die Bearbeitungszeit sowie die geplanten Betriebs- und Arbeitszeiten. Im Folgenden sind die spezifischen Regelungen für diese Aspekte detailliert beschrieben:

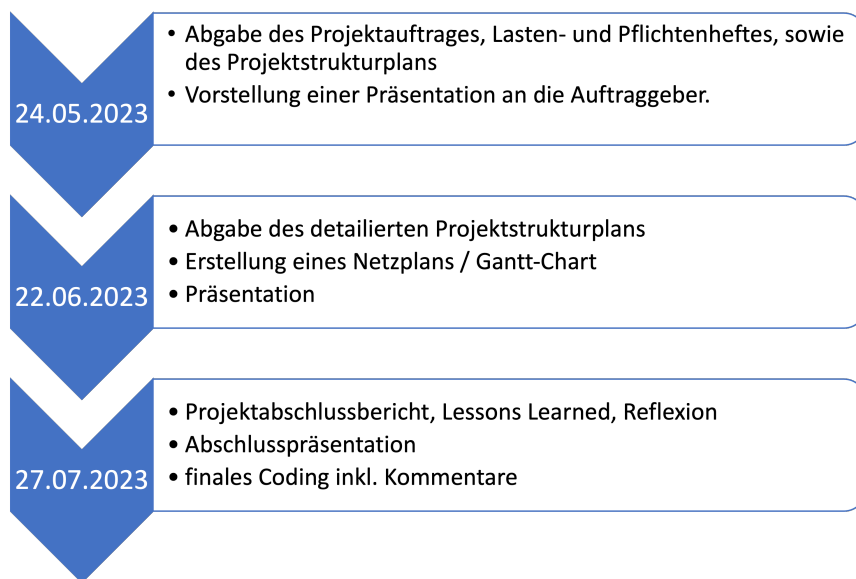


Abbildung 3.1: Projektmeilensteine. Quelle: Eigene Darstellung.

Bearbeitungszeit:

Die Bearbeitungszeit für das Projekt erstreckt sich von dem 24.05 bis zum 27.07. Dieser Zeitraum umfasst die Durchführung des Projekts parallel zu den regulären Vorlesungen der vier beteiligten Studenten. Es wird erwartet, dass die Studenten während dieses Zeitraums

ausreichend Zeit für das Projekt aufbringen und entsprechende Arbeitsstunden investieren.

Geplante Betriebszeiten:

Das entwickelte Tool zur Textzusammenfassung und -klassifikation soll grundsätzlich rund um die Uhr verfügbar sein, um den Nutzern eine maximale Flexibilität zu bieten. Es wird angestrebt, das Tool auf einer geeigneten Serverinfrastruktur mithilfe von Streamlit zu hosten, um **die Verfügbarkeit** zu gewährleisten.

Geplante Arbeitszeiten:

Die geplanten Arbeitszeiten für die Studenten während des Projekts umfassen die Zeiten außerhalb der Vorlesungen und individuellen Verpflichtungen. Es wird erwartet, dass die Studenten in Absprache miteinander und unter Berücksichtigung ihrer individuellen Verfügbarkeit regelmäßige Arbeitszeiten festlegen, um das Projekt voranzutreiben. Es ist ratsam, regelmäßige Treffen oder virtuelle Konferenzen zu vereinbaren, um den Fortschritt zu besprechen, Fragen zu klären und gemeinsame Entscheidungen zu treffen.

Es ist wichtig zu beachten, dass die oben genannten Rahmenbedingungen als allgemeine Leitlinien dienen und bei Bedarf angepasst werden können. Es wird empfohlen, eine klare Kommunikation und Zusammenarbeit zwischen den Studenten sicherzustellen, um die Arbeitszeiten bestmöglich zu koordinieren und das Projekt erfolgreich abzuschließen.

3.5 Dekomposition des Gesamtsystems

Das Gesamtsystem wird nachstehend in Hauptkomponenten zusammenfassend unterteilt:

Benutzerschnittstelle: Eine benutzerfreundliche Oberfläche, die den Benutzern die Eingabe von Texten, die Steuerung des Zusammenfassungs- und Klassifikationsprozesses sowie den Zugriff auf die Ergebnisse ermöglicht. Bereitgestellt wird dies per Webapplikation, welche mittels Streamlit entwickelt wird.

Externe Deepl API: Aufgrund der besseren Performanz der Modelle auf englischsprachigen Texten werden deutschsprachige Eingabetexte für die weitere Verarbeitung übersetzt.

Textkompression: Eine leistungsfähige Textanalyse-Engine, die in der Lage ist, Texte zu analysieren, Schlüsselinformationen zu extrahieren und automatische Zusammenfassungen zu generieren.

Klassifikationsmodul: Ein Modul, das die Texte basierend auf vordefinierten Kategorien oder benutzerdefinierten Klassifikationen einordnet.

Stimmungsanalyse: Eine weitere Klassifikation des Eingabetexts in die Kategorien positiv, neutral und negativ, um ein Stimmungsbild zu erhalten.

Datenablage: Eine persistente Speicherung der Eingabe- oder Ausgabetexte, sowie der Kategorisierung findet nicht statt.

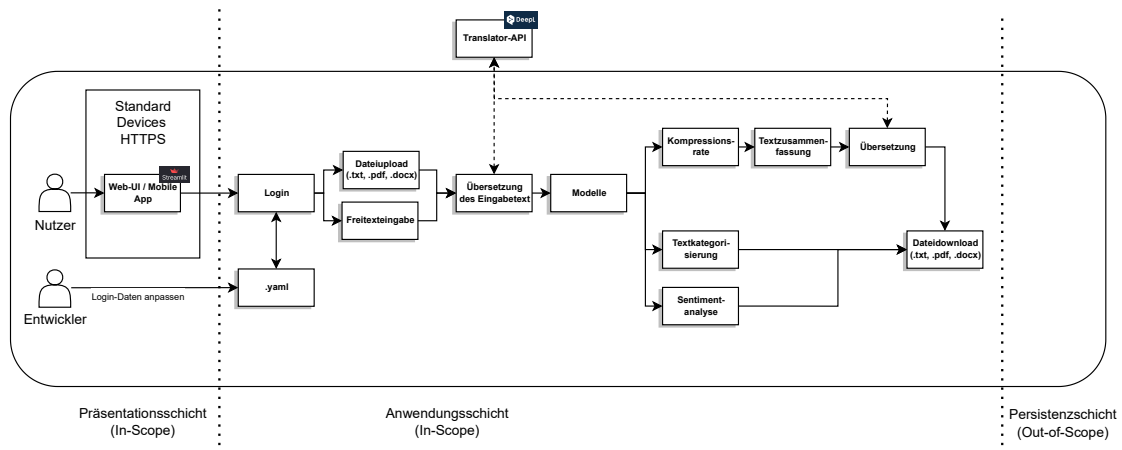


Abbildung 3.2: Vorläufige Systemskizze der geplanten Funktionalitäten. Quelle: Eigene Darstellung.

Eine feingranulare Aufgliederung des Gesamtsystems wird im Projektstrukturplan dargestellt.

3.6 Funktionale Anforderungen

REQ101 Das Produkt ermöglicht die Zusammenfassung von Texten

Umsetzung: Verwendung eines Transformer-basierten Modells zur automatischen Textzusammenfassung in der Webapplikation inclusive Data Pipeline für notwendiges Preprocessing und Vorbereiten der Daten auf den Usecase. Dabei werden die gängigen NLP Methoden zum Trainieren eines Transformers für Textzusammenfassung verwendet. *→ welche Methoden?*

REQ102 Das Produkt ermöglicht die Eingabe von Freitexten, sowie einen Dateilupload von verschiedenen Dateitypen (.txt, .pdf und .doc)

Umsetzung: Implementierung durch Streamlit-Input-Widgets zur Erfassung von Textdaten wie auch dem Dateilupload von einer Dateigröße bis 200mb.

REQ103 Bei der Textkompression des Produkts handelt es sich um eine abstraktive Textzusammenfassung

Umsetzung: Anzeige von generierter Zusammenfassung auf Benutzeroberfläche. Wie in REQ101 beschrieben wird ein Transformer Modell, speziell für die abstraktive Zusammenfassung gewählt und finetuning vorgenommen.

REQ104 Die Benutzeroberfläche ermöglicht eine individuelle Festlegung der Kompressionsrate zwischen 20 und 80 Prozent.

Umsetzung: Streamlit ermöglicht per Inputfeld die Auswahl des Wertes anhand eines verschiebbaren Reglers. ✓ *Schrittbarkeit?*

REQ105 Das Ergebnis der Zusammenfassung wird in der Webapplikation angezeigt.

Umsetzung: Anzeige von generierter Zusammenfassung auf Benutzeroberfläche. Output von Kompressionsmodell wird direkt dargestellt und nicht gespeichert.

REQ106 Die Benutzeroberfläche ermöglicht die Speicherung der Ergebnisse als Word, Text oder PDF-Datei.

Umsetzung: Einbindung in Layoutvorlage für den jeweiligen Dateityp und Download unter Zuhilfenahme etwaiger vordefinierter Python Libraries.

REQ107 Das System unterstützt die Verarbeitung von deutschen und englischen Texten.

Umsetzung: Diese, über den Projektscope hinaus angeforderte Funktionalität wird mit einer Anbindung, an die DeepL API realisiert. Das Tool bietet eine schon präzise Übersetzungsfunktion, welche jedoch noch auf den Anwendungsfall angepasst werden muss.

REQ108 Im Kontext von News Artikeln wird zwischen den Kategorien Sport, Kultur, Gesellschaft, Politik, Wissenschaft, Wirtschaft (...) unterschieden *Welche?*

Umsetzung: Anhand der definierten Evaluationsmetriken werden state-of-the-art Modelle verglichen, um die beste Performance für den hier gewählten Usecase zu identifizieren. Die Evaluierung erfolgt auf einem Testdatensplit von einem Datensatz, der für diesen Usecase erstellt wurde.

REQ109 Die Oberkategorie des Eingabetextes wird angezeigt im UI der Webapplikation

Umsetzung: Ergebnisse von Klassifikationsmodell werden auf der Benutzeroberfläche angezeigt

REQ110 Es ist möglich, ausschließlich Dateien der gegebenen Kategorien hochzuladen

Umsetzung: Der Textklassifikator erkennt keine anderen, als die vorgegebenen Klassifizierungen.

REQ111 Das Produkt führt eine Stimmungsanalyse des Eingabetextes durch. **Umsetzung:** Die Stimmungsanalyse wird ebenfalls mit einem Transformer Modell entwickelt, umgesetzt und angewendet.



3.7 Nicht-funktionale Anforderungen

REQ301 Zugang zur Web-Applikation ist nur bestimmten Nutzergruppen möglich

Umsetzung: Login über Benutzername und Passwort, welche mithilfe einer .yaml Datei gespeichert werden *nicht!*

REQ302 Die Web-Applikation muss keinem Employer Branding entsprechen

Umsetzung: Nutzung von Streamlit-komponenten mit der Möglichkeit grafische Anpassungen vorzunehmen

REQ303 Die Web-Applikation sollte eine Benutzerfreundlichkeit aufweisen und ein intuitives Benutzerinterface bieten.

Umsetzung: Die Anwendung wird nach den Grundsätzen zeitgemäßer Benutzeroberflächen bzw. Webdesigns erstellt.

REQ304 Das System muss bei Bedarf verfügbar sein.

Umsetzung: Hosting per Streamlit Cloud und Zugriff über öffentliche URL

REQ305 Die Webapplikation sollte auf verschiedenen Browsern und Endgeräten (Desktop, Tablet, Mobil) gut funktionieren und ein responsives Design aufweisen.

Umsetzung: Streamlit ermöglicht das Verwenden der Applikation auf allen gängigen Endgeräten. *→ Accessibility?*

REQ306 Die Qualität der Ergebnisse soll bewertet werden.

Umsetzung: Mithilfe der beschriebenen Evaluierungsmetriken soll von den Modellen die Güte bestimmt werden.

Spalte?

3.8 Qualitätsüberprüfung

Die Qualität des entwickelten Tools zur Textzusammenfassung und -klassifikation spielt eine entscheidende Rolle für den Erfolg des Projekts. Um sicherzustellen, dass die definierten Qualitätsanforderungen erfüllt werden, werden klare Abnahmekriterien und ein strukturiertes Vorgehen zur Ausgangsprüfung festgelegt. Im Folgenden werden diese Aspekte im Detail beschrieben.

3.8.1 Abnahmekriterien

Die Abnahmekriterien dienen als Richtlinien, anhand derer die Qualität des Tools bewertet wird. Die spezifischen Abnahmekriterien werden in Absprache mit den Projektbeteiligten festgelegt und sollten quantifizierbare und messbare Merkmale umfassen. Beispiele für Abnahmekriterien könnten sein:

Funktionalität: Das Tool erfüllt die definierten funktionalen Anforderungen, einschließlich der Textzusammenfassung und -klassifikation von Inhalten.

Genauigkeit: Das Tool erzielt eine hohe Genauigkeit bei der Zusammenfassung und Klassifikation von Texten.

Benutzerfreundlichkeit: Das Tool ist intuitiv und einfach zu bedienen.

Leistung: Das Tool zeigt eine angemessene Reaktionsgeschwindigkeit und Skalierbarkeit.

Stabilität: Das Tool läuft stabil und weist keine schwerwiegenden Fehler/Abstürze auf.

→ KPIs / Kaskaden Wech

3.8.2 Vorgehen zur Ausgangsprüfung:

Die Ausgangsprüfung dient dazu, die Qualität des entwickelten Tools vor der endgültigen Abnahme zu überprüfen. Das Vorgehen zur Ausgangsprüfung umfasst typischerweise die folgenden Schritte:

Testplanung: Es wird ein detaillierter Testplan erstellt, der die verschiedenen Testfälle, Szenarien und Daten abdeckt, um eine umfassende Überprüfung der Funktionalität, Genauigkeit, Benutzerfreundlichkeit, Leistung und Stabilität des Tools zu gewährleisten.

Durchführung von Tests: Die definierten Testfälle werden systematisch und methodisch durchgeführt. Dies umfasst sowohl manuelle Tests als auch automatisierte Tests, um sicherzustellen, dass alle Aspekte des Tools gründlich überprüft werden.

Fehlerbehebung: Wenn während der Tests Fehler oder Abweichungen von den definierten Abnahmekriterien festgestellt werden, werden diese dokumentiert und an das Entwicklungsteam gemeldet. Das Team ist dann dafür verantwortlich, die Fehler zu beheben und das Tool entsprechend zu optimieren.

Review und Abnahme: Nach Abschluss der Tests wird das Tool einer umfassenden Bewertung unterzogen, bei der die Einhaltung der Abnahmekriterien und die Erfüllung der definierten Qualitätsanforderungen bewertet werden. Wenn das Tool die festgelegten Kriterien erfüllt, erfolgt die formelle Abnahme. Die Abnahme des Projektes erfolgt zusammen mit den Auftraggebern, nachdem die Vollständigkeit des Systems zur Kenntnis genommen wurde und dieses für die Verwendung einsatzbereit ist.

3.9 Lieferumfang

Der Lieferumfang umfasst das vollständige entwickelte System, einschließlich der Benutzeroberfläche, der Textanalyse-Engine, des Klassifikationsmodells und der Datenbank. Darüber hinaus sollten auch eine umfassende Dokumentation, Installationsanweisungen und erforderliche Schulungsunterlagen bereitgestellt werden. Die ausführliche Aufschlüsselung der Projektentwicklung inklusive der Teilschritte erfolgt im Projektstrukturplan.

Literaturverzeichnis

- Aichele, C., Schönberger, M., Aichele, C., & Schönberger, M. (2014). Grundlagen des Projektmanagements. *IT-Projektmanagement: Effiziente Einführung in das Management von Projekten*, 3–16.
- Peter-Johann, H. (2023). *Lastenheft und Pflichtenheft - Peterjohann Consulting*. Verfügbar 20. Mai 2023 unter https://www.peterjohann-consulting.de/lastenheft-und-pflichtenheft/#211_einige_typische_fehler_und_fallstricke