Projektrealisierung

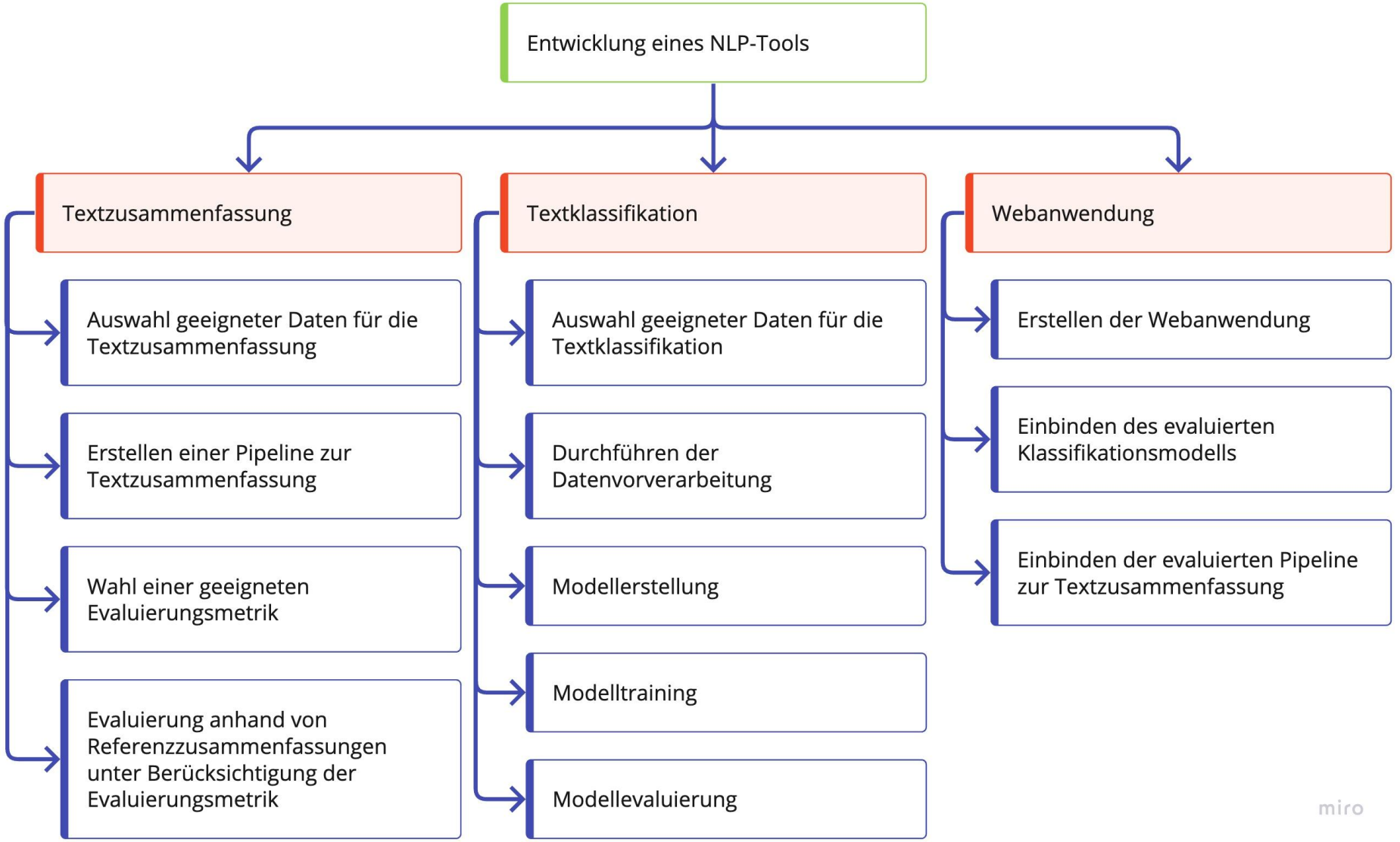
Zwischenstand – 20.06.23

Von Lukas Bonn, Amina Uicker-Darwish,
Jan Rüdt, Aymane Bouguern

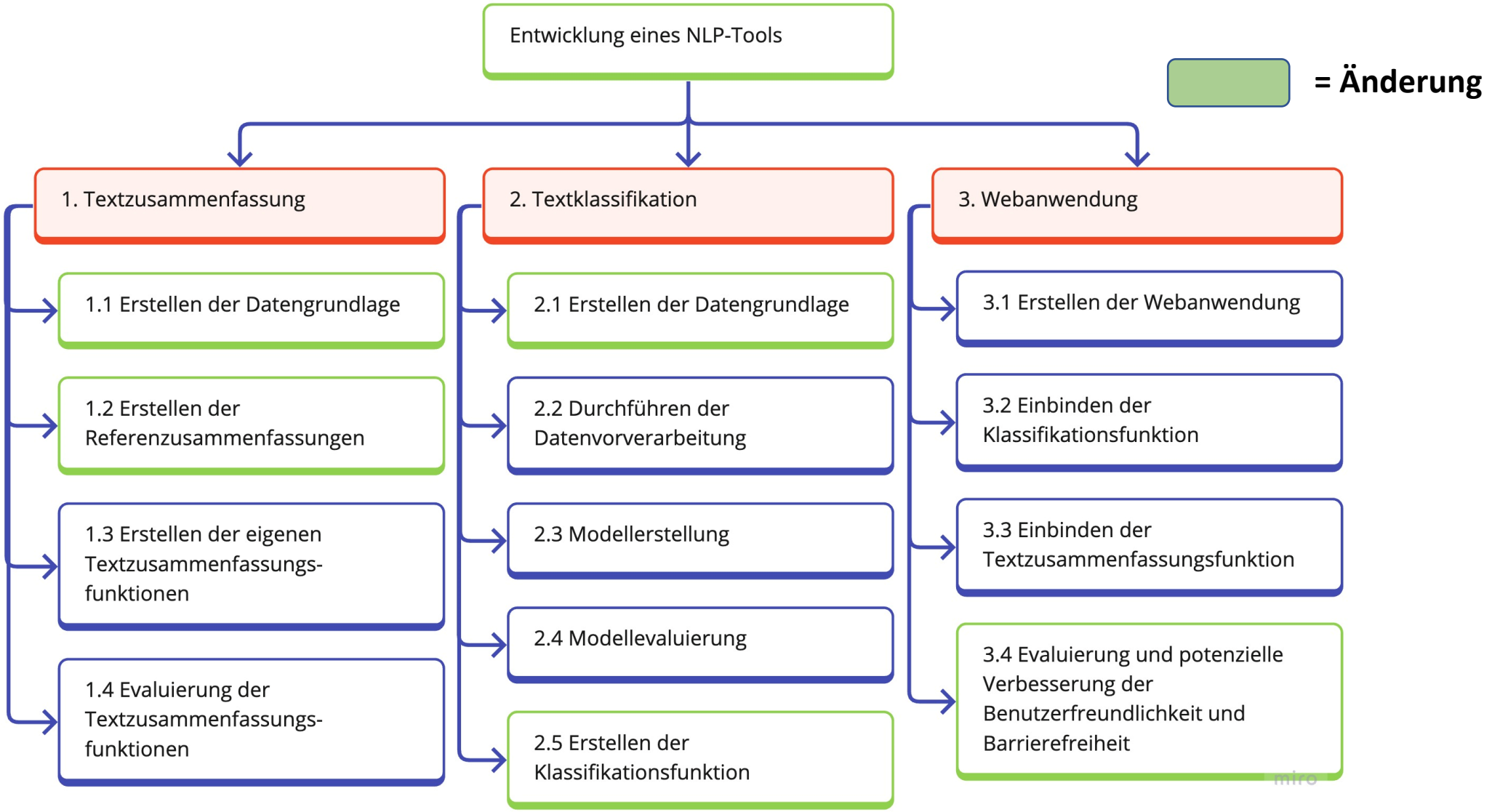
Agenda

1. Übersicht: Projektstrukturplan und Arbeitspakete
2. Entwicklungsstand der Anforderungen
    - Klassifikation
    - Zusammenfassung
    - Frontend
3. Herausforderungen
4. Ausblick

Recap - Projektstrukturplan



Aktualisierung - Projektstrukturplan



Entwicklungsstand - Zusammenfassung

Aspekte aus dem PSP

Erstellung des Datensatzes

- Texte aus 5 verschiedenen Kategorien, Texlänge je 500-800 Wörtern ✓

Referenzzusammenfassungen erstellen

- Anhand von LSA, Textrank, Lexrank (Sumy Libary), Unterschiedliche Kompressionsraten (random) ✓

Datenvorverarbeitung

- Bereinigung der Datensätze ✓
- NLP-Pipeline zur Vorverarbeitung ✓

Funktionen zur Textzusammenfassung

- 2 Extractive Ansätze: Textrank und LSA ✓

Geeignete Evaluierungsmetrik

- ROUGE-Score 1 ✓

Erfüllung der Anforderungen

Erfüllt

- Kompressionsrate 20%-80% ✓
- Englische Sprache ✓
- Zusammenfassung unterschiedlicher Textarten ✓
- Eigenständige Auswahl der Toolchain (Spacy, NLTK, Sumy) ✓

Noch offen

- Extrahierung wichtigster Informationen ✗

Entwicklungsstand - Klassifikation

Aspekte aus dem PSP

Erstellung der Datengrundlage

- 5 Kategorien: News Artikel, Literatur, Poitische Rede, Blogartikel, Juristische Texte ✓

Datenvorverarbeitung

- Vorverarbeitung der Datensätze für die Klassifikation ✓
- NLP-Pipeline (klassische Algorithmen) -> Tokenizer, Lemmatizer, .. ✓
- Datenvorverarbeitung (BERT-Modell) -> Tokenizer ✓

Modellerstellung

- BERT-Model (Transfer Learning) ✓
- Algorithmen, kein Deep Learning -> SVM, KNN, Naive Bayes (Modelltraining) ✓

Modellevaluierung

- Richtig klassifizierte Texte im Vergleich zu allen klassifizierten Texten (Accuracy) ✓

Erstellen der Klassifikationsfunktion

- Erstellen einer Funktion zur Nutzung des Modells in der Webanwendung ✓

Erfüllung der Anforderungen

Erfüllt		Noch offen
Präzise Klassifikation in Oberkategorien	✓	✗
Auswahl geeigneter Modelle und Algorithmen	✓	
Eigenständige Auswahl der Toolchain (Spacy, Scit-kit Learn)	✓	
Klassifikation englischer Texte	✓	

Entwicklungsstand - Webanwendung

Aspekte aus dem PSP

Erstellen der Webanwendung

- Flask Anwendung (Benutzerfreundliches Frontend, verschiedene Funktionalitäten -> Demo)

✓

Einbinden des evaluierten Klassifikationsmodells

- Einbindung des vorläufigen Modells (inkl. der Vorerarbeitungsfunktion, BERT)

✓

Einbinden der evaluierten Textzusammenfassungsfunktion

- Einbindung der vorläufigen Zusammenfassungsfunktion (LSA)

✗

Evaluierung des Frontends nach Benutzerfreundlichkeit

- Tools zur Evaluierung
- Erprobung durch Testuser

✗

Erfüllung der Anforderungen

Erfüllt

- Texte können in unterschiedlichen Formaten entgegengenommen werden (Textfeld, ODT, DOC(X), TXT) ✓
- Auswahl einer Kompressionsrate ✓
- Anzeigen der Resultate ✓
- Benutzerfreundlichkeit (Super Zoom, Kontraste) ✓

Noch offen

- Finale Textzusammenfassungsfunktion einbinden
- Evaluierung der Benutzerfreundlichkeit und Barrierefreiheit

Kurzdemo...

Herausforderungen

- Datenerstellung:
  - Neue Anforderungen (5 Oberkategorien)
  - 5 geeignete Datensätze finden (+ Referenzzusammenfassungen erstellen)
  - Größe der Datensätze
  - Datenvorverarbeitung
- Barrierefreies Frontend

Ausblick

Die nächsten Schritte:

1. Evaluierung der Zusammenfassung

2. Einbinden der Zusammenfassungsfunktion

3. Evaluierung der Gesamtanwendung

4. Abschlussdokumentation/präsentation, Repository