

Dokumentation

Intelligence Engineering: Document Analyzer

Sarah Lemke: 1111111

Benedikt Lerch: 1111111

Simon von Oppenkowski: 1111111

Kristian Sabados: 2945629

Abgabe: 4. November 2025

Inhaltsverzeichnis

| | | |
|----------|--------------------------------------|-----------|
| 1 | Projektbeschreibung | 1 |
| 2 | User Story Cards und Epics | 2 |
| 3 | Eventstorming | 5 |
| 4 | Pipelineplanung - Architektur | 8 |
| A | Anhang | 10 |

1 Projektbeschreibung

Im Projekt Document Analyzer sollen Rechnungen verschiedener, on digital oder gescannt, maschinell ausgelesen und einem der Hauptkonten des Buchungswesens zugeordnet werden können. Semi-überwachte maschinelle Lernparadigma (Semi-Supervised Learning) sowie moderne Techniken aus dem Bereich Computer Vision und weiteren modernen Verfahren werden genutzt um ein Modell zu kreieren, welches Schlüsselinformationen wie Rechnungsnummer, Betreff und verwendungszweck, Datum, Betrag und Absender aus Rechnungen extrahieren kann um dieses Ziel zu erreichen. Dafür werden Trainingsdaten benötigt, bei denen diese Informationen bereits korrekt annotiert sind. Das Modell lernt also aus Beispielen, welche Textstellen zu welchen Informationskategorien gehören. Dieses Vorgehen ermöglicht eine gezielte Leistungsbewertung anhand bekannter Zielwerte (Labels) und unterstützt eine iterative Verbesserung der Extraktionsgenauigkeit. Das semi-überwachte Lernen ist daher ideal geeignet, um eine hohe Präzision und Nachvollziehbarkeit in der Dokumentenanalyse sicherzustellen.

2 User Story Cards und Epics

Key/Value Dictionary erstellen

| Prio | Name | Points |
|------|--------------------------------|--------|
| 1 | Key/Value Dictionary erstellen | 30h |

Als EntwicklerIn

möchte ich aus einem eigens zu diesem Zweck generierten Datensatz die wichtigsten und gängigsten Key-Value Paare für Start der Software extrahieren
um um einen repräsentativen Start und Default für die Software bereitzustellen

| | |
|---------------|----------------|
| Risiko | [Beschreibung] |
| Points (Post) | |

Vorausgesetzt ein solcher Datensatz konnte generiert oder beschaffen werden
wenn die Software Dokumente per OCR ausliest, **dann** wird ein festgeleger Prozentsatz der Dokumente anhand bestehender Beispiele korrekt zugeordnet

Training des Modells

| Prio | Name | Points |
|------|---------------------|--------|
| 1 | Trainng des Modells | /h |

Als EntwicklerIn

möchte ich aus den festgelegten Trainingsdaten ein Modell trainieren, welches Rechnungen inhaltlich den korrekten Buchungskonten zuweisen kann
um das Bearbeiten von Rechnungen zu beschleunigen und Fehler durch weniger anfallende manuelle und monotone Arbeit zu reduzieren

| | |
|---------------|----------------|
| Risiko | [Beschreibung] |
| Points (Post) | 3 |

Vorausgesetzt das Modell erhält repräsentative Daten,
wenn die Daten korrekt gesplittet wurden, **dann** können erste korrekte Klassifizierungen neuer Rechnungen durch das Training vorgenommen werden.

Modell evaluieren

| Prio | Name | Points |
|------|-------------------|--------|
| 2 | Modell evaluieren | 32h |

Als EntwicklerIn

möchte ich das trainierte Modell mit Testdaten evaluieren

um festzustellen, wie genau es neuen Rechnungen die korrekte Hauptkonten zuordnen kann

| | |
|---------------|----------------|
| Risiko | [Beschreibung] |
| Points (Post) | 3 |

Vorausgesetzt das Modell ist trainiert und einsatzfähig,

wenn eine Menge von mindestens 100 Testrechnungen verarbeitet wird, **dann**, werden erste korrekte Rechnungsklassifikationen erstellt und ausgegeben.

Modell integrieren

| Prio | Name | Points |
|------|--------------------|--------|
| 3 | Modell integrieren | 48h |

Als EntwicklerIn

möchte ich das trainierte Modell in die Document Analyzer Pipeline integrieren

um es Benutzern beim Upload automatisch die extrahierten Daten zeigen lassen zu können

| | |
|---------------|----------------|
| Risiko | [Beschreibung] |
| Points (Post) | 3 |

Vorausgesetzt das Modell ist erfolgreich trainiert und getestet,

wenn der Benutzer Rechnungen hochlädt, **dann**, wird das Modell automatisch aufgerufen und liefert JSON-Ausgaben mit Schlüssel/Wert-Paaren welcher als Output in im UI angezeigt wird. Cache wird nach Bestätigung DSGVO-konform gelöscht.

Modell überwachen und verbessern

| Prio | Name | Points |
|------|----------------------------------|--------|
| 4 | Modell überwachen und verbessern | 32h |

Als EntwicklerIn

möchte ich die Integrität meines Modells auf die bestmögliche Performance verbessern

um die Fehlerrate zu minimieren und User zuverlässige Ergebnisse zu präsentieren

| | |
|---------------|----------------|
| Risiko | [Beschreibung] |
| Points (Post) | 3 |

Vorausgesetzt das Modell läuft und ist stabil,

wenn permanent Rechnungen zur Klassifikation hochgeladen werden, **dann**, werden Hyperparameter so verändert, dass F1-Score und Accuracy bei mindestens 98 Prozent liegen.

3 Eventstorming

Auf der folgenden Seite eine schematische Darstellung der Ergebnisse

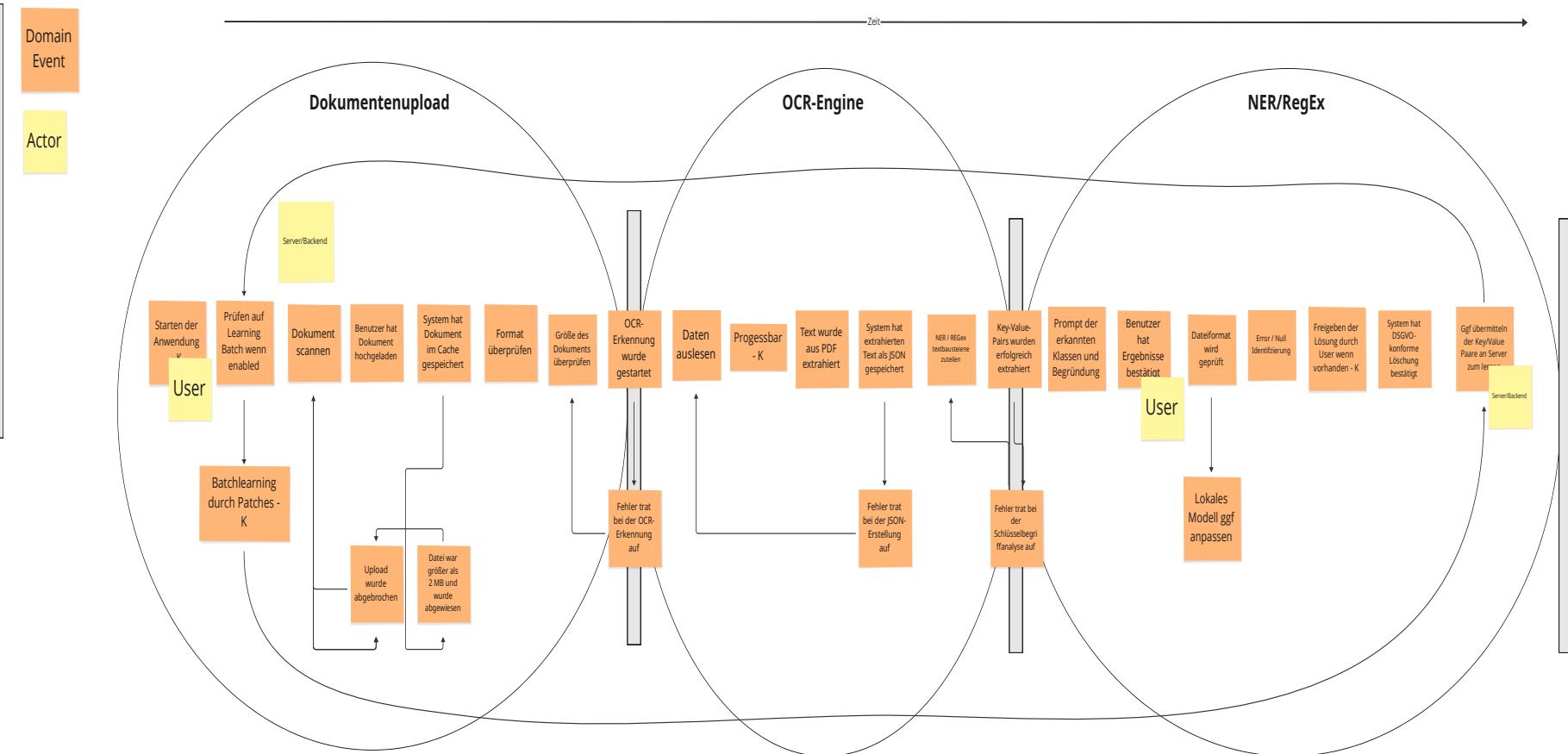


Abbildung 1: Event-Storming Übersicht

| Modell | Implementierungsaufwand | | Verständlichkeit | | Deployment Aufwand | | Einfachheitsscore |
|----------|-------------------------|--------------------------|--------------------|---------------------|--------------------|---------------|-------------------|
| | Weit verbreitet | Geprüfte Implementierung | Feature Importance | Einfach zu debuggen | Inference Zeit | Trainingszeit | |
| Modell 1 | | | | | | | |
| Modell 2 | | | | | | | |
| Modell 3 | | | | | | | |

4 Pipelineplanung - Architektur

Auf der folgenden Seite eine schematische Darstellung der Architektur und Pipeline der Software

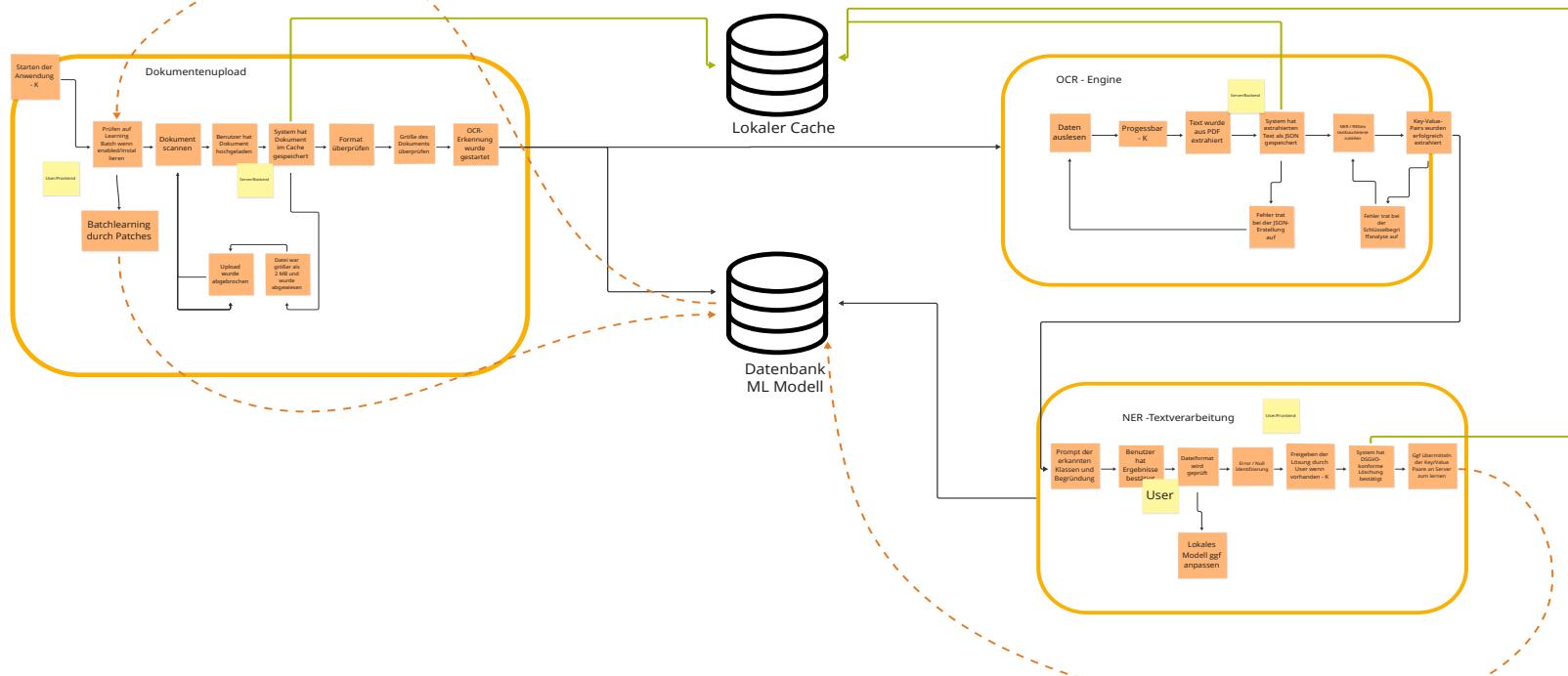


Abbildung 2: Schematische Darstellung der Architektur und Pipeline der Software

A Anhang