

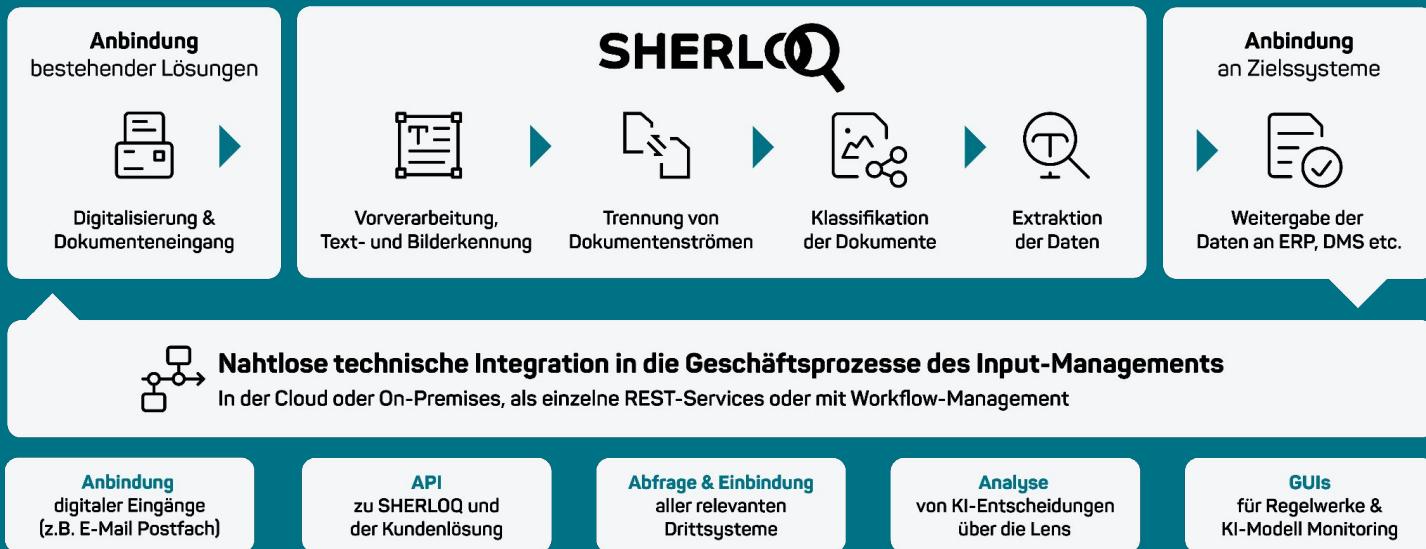


KI-Lösung für Intelligente
Dokumentenverarbeitung &
Kundenkommunikation

Hallo Java User Group!



KI-gestützte Automatisierung im Dokumentenprozess



Wir machen...

-  **Projekte** und nicht nur Technik
Beratung, Integration, ...
-  **On-Prem** wenn gewünscht
fast immer
-  **KI** eingestellt auf **Kundendaten**
train + evaluate + explain
-  **faires Lizenzmodell**
pay for success

statt...

-  komplette **Produkt-Suite** für Input-Management
-  reine **Cloud-Lösung** oder **Software-as-a-Service**
-  **AI-as-a-Service** und **general-purpose-Modelle**

Wie sehen typische SHERLOQ-Projekte aus?



Beispiel #1

Klassifikation im Posteingang eines Energieversorgers



Hilferuf

In unserer Poststelle bleiben E-Mails
und Scans xxx Tage unbearbeitet liegen.

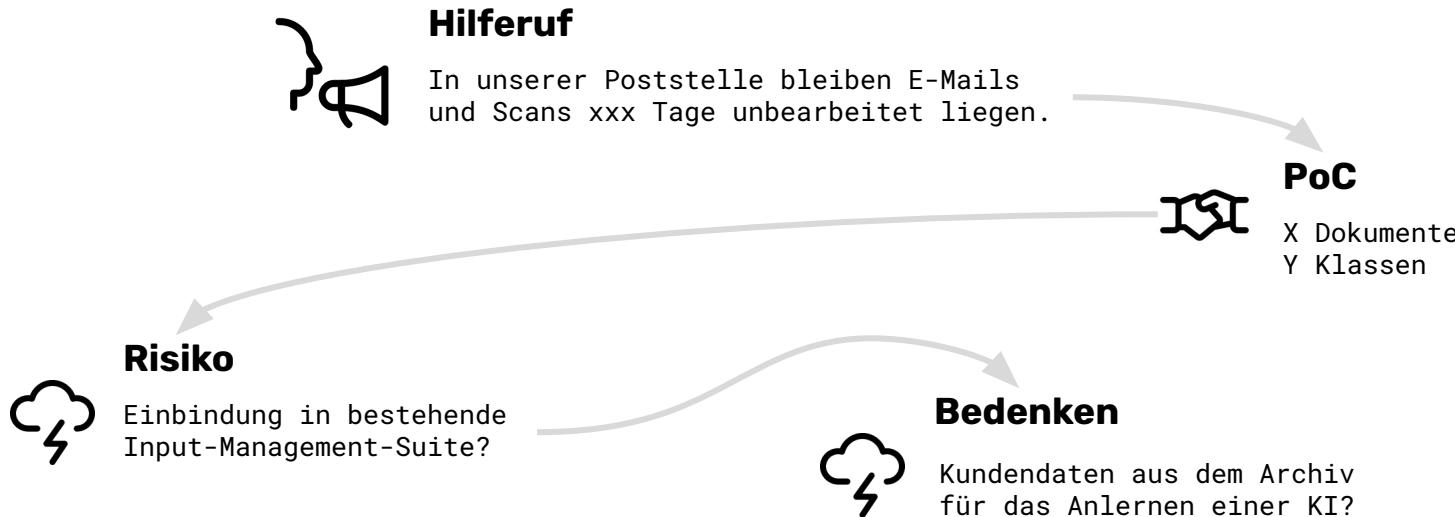


PoC

X Dokumente
Y Klassen

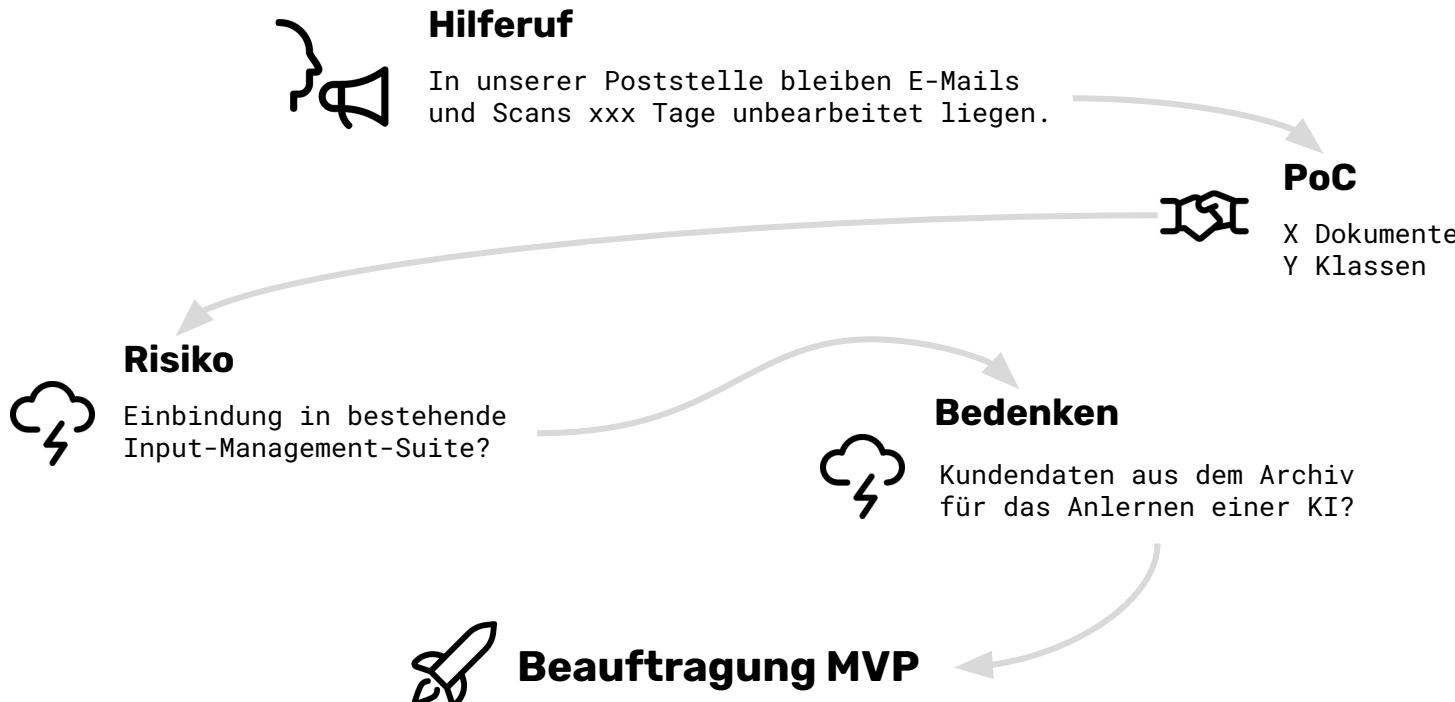
Beispiel #1

Klassifikation im Posteingang eines Energieversorgers



Beispiel #1

Klassifikation im Posteingang eines Energieversorgers



Beispiel #1

Klassifikation im Posteingang eines Energieversorgers

Ziel des MVP

Erkennung der **10 häufigsten Klassen**

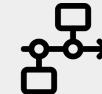


Daten



- aus dem **Archiv**
- Scans & E-Mails als **PDFs** von der Input-Management-Suite
- **500-10.000** Dokumente/Klasse und **200.000** vom ganzen Rest

Integration



- KI als **REST-Service**
- Anbindung durch Kunden-IT (MuleSoft, RabbitMQ)

Infrastruktur



- On-Prem Linux-VMs für Training, Test, Prod

2108-0229

Beispiel #2

Extraktion aus Lieferscheinen

Ziel PoC

Verprobung der Extraktion von



- Lieferdatum
- Lieferscheinnummer
- Kreditor
- ...

Daten



- **821** Seiten Scans / Fotos
- davon nur **70%** lesbar
- keine Werte / Annotationen
- "Macht damit, was ihr wollt."

Lieferschein

Kundennummer:	141213
Auftragsnummer:	4001917
Lieferschein-Nr.:	463306
Lieferdatum:	15.07.2021
Unser Zeichen:	HS
Telefon:	
Fax:	
Best./Abh. KFZ:	
MA-SK-69	

Versandart: LKW
Ab Werk

Baustelle/Objekt:

Zur Verfügung:
Bestellz. Kunde: R200044

Menge	Artikel	Pakete	Lagen	Stück	Kommt-zuschlag	Stück Pal.
6,390 qm	BP10729 CombiStabil 13/12/10 cm grau mit Microfase DIN EN1338	925	1	-	-	WERK
28,000 qm	BP10054 CombiStabil 18/18/10 cm ->>grau mit Microfase DIN EN1338	511	4	-	-	WERK

Rücklieferung: Stück Werkspaletten: Stück Europaletten:

Der Abholer ist für die Sicherung der Ladung und für die Einhaltung des zulässigen Gesamtgewichtes selbst verantwortlich. Unsere Leistung beschränkt sich auf das Binden der Zettelkette. Die Sicherung der Ladung obliegt dem Abholer.

Unterschriften:

Verladung Spediteur / Abholung Empfänger

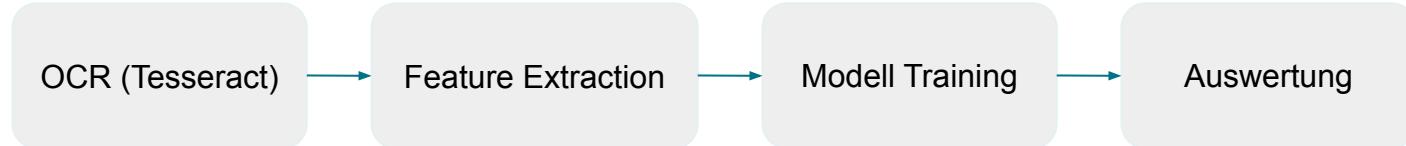
7000057

SHERLOQ

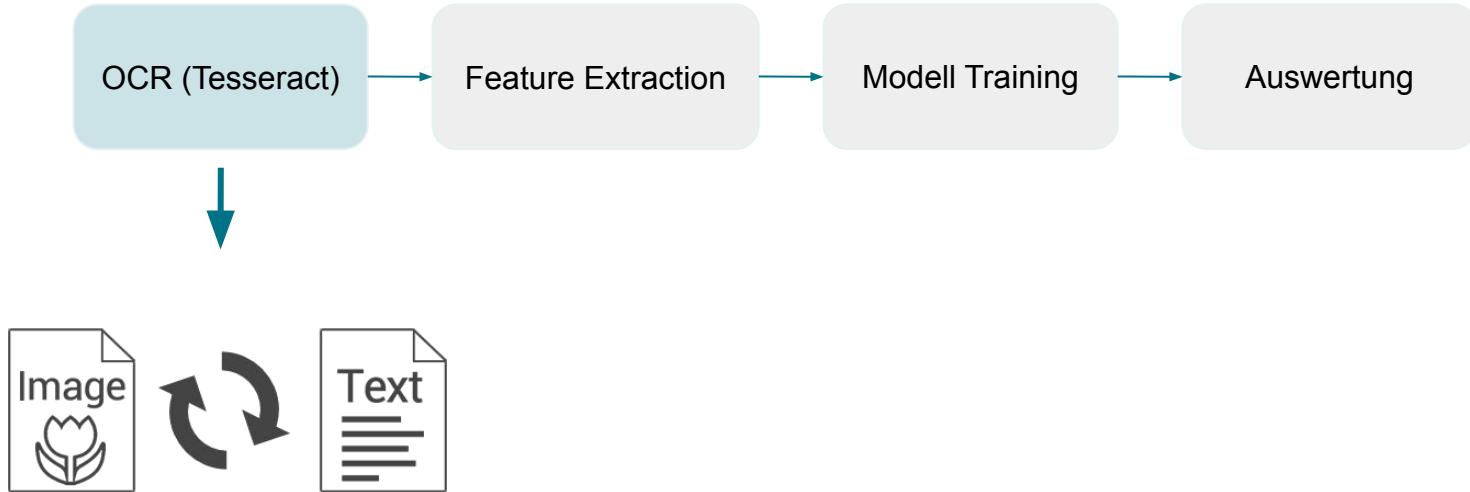
AI Überblick



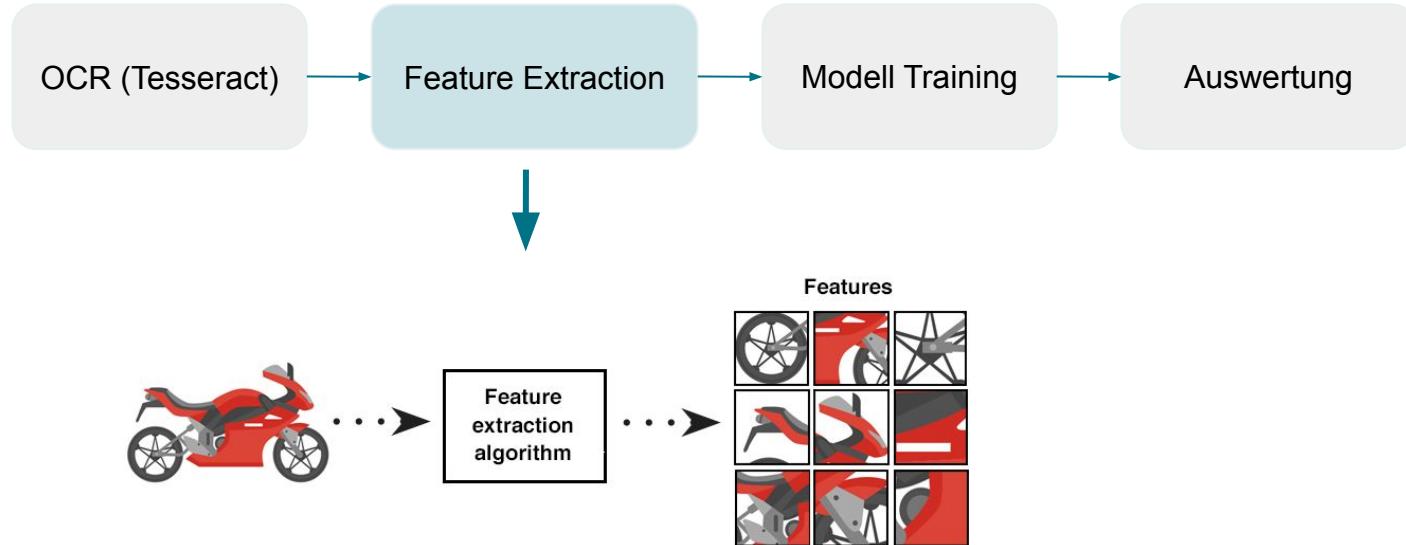
Von Kundendaten zum trainierten Modell



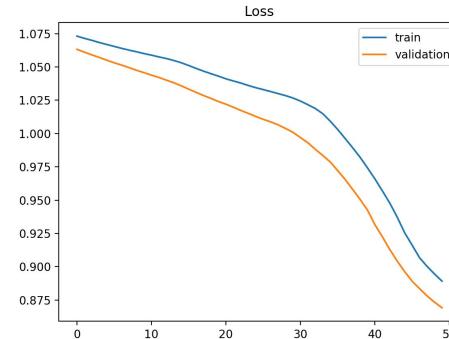
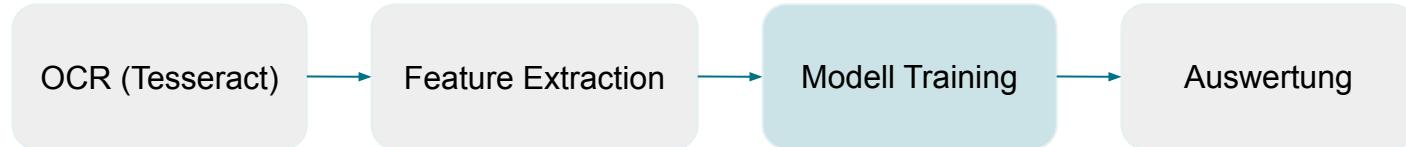
Von Kundendaten zum trainierten Modell



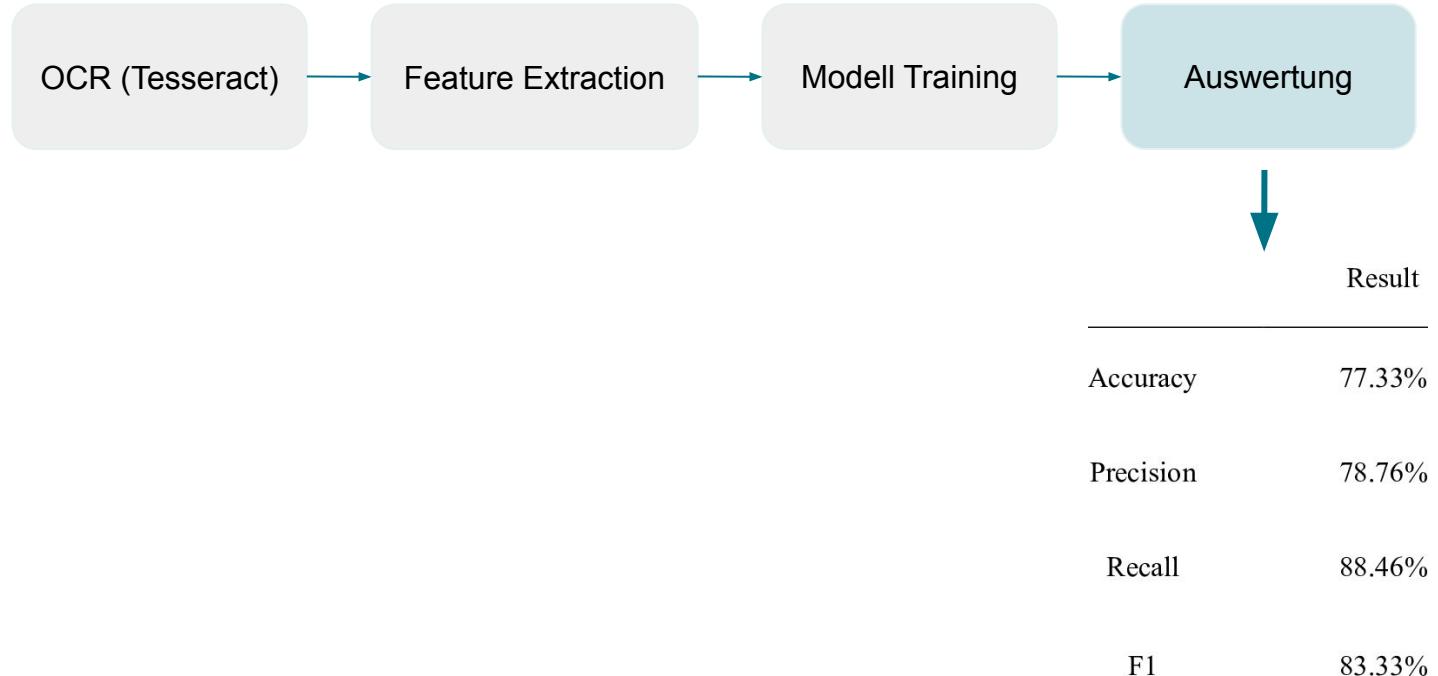
Von Kundendaten zum trainierten Modell



Von Kundendaten zum trainierten Modell



Von Kundendaten zum trainierten Modell



Anwendungsfall Klassifikation

Training auf Kundendaten (AI-Lab)

- PDFs, E-Mails
- xgboost, spacy, keras, ...

Auswertungs-Dashboard

- explainable AI mit lime
- Live-Test

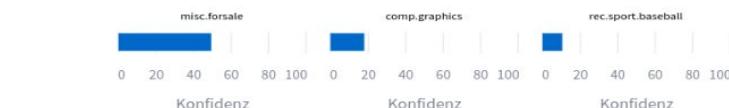
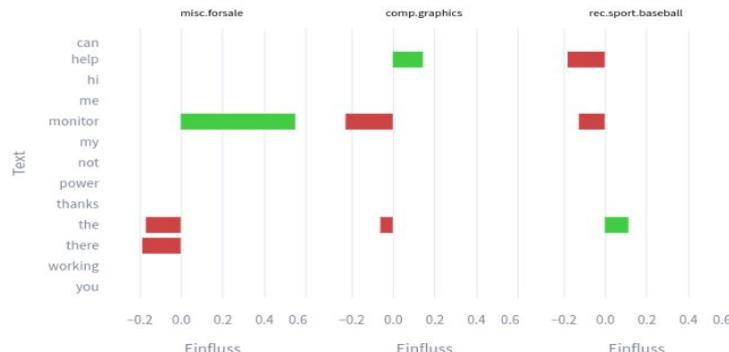
Bereitstellung als REST-Service

- optionales Regelwerk

Erklärbarkeit

Berechnen

Einfluss der relevantesten Schlagworte auf die Klassifikation.



Hervorhebung der relevantesten Schlagworte im Text

hi **there**,

my **monitor** is not working, maybe it's **the** power button.
can you help me?

thanks,
john

ML/DL für Text-Klassifikation

Entwicklung der Ansätze

Eingabe



Text als

- **Bag of Words / Wort-Statistiken**

Modelle



klassisches ML

dmlc
XGBoost

- **Entscheidungsbäume
on steroids**

 LightGBM

ML/DL für Text-Klassifikation

Entwicklung der Ansätze

Eingabe



Text als

- Bag of Words / Wort-Statistiken
- **Zeichenfolge**
- **Wortfolge**

Modelle



klassisches ML

dmlc
XGBoost

- Entscheidungsbäume
on steroids

 LightGBM

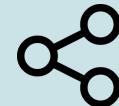
oder neuronale Netze / DL

 PyTorch

- **Faltungsnetze**
- **rekurrente Netze**

 TensorFlow

Transfer Learning



- vortrainierte
Wort-Einbettungen

ML/DL für Text-Klassifikation

Entwicklung der Ansätze

Eingabe



Text als

- Bag of Words / Wort-Statistiken
- Zeichenfolge
- Wortfolge
- **Subtokenfolge**



Hugging Face

Modelle



klassisches ML

dmlc
XGBoost

- Entscheidungsbäume *on steroids*



oder neuronale Netze / DL

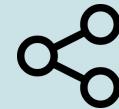
 PyTorch

- Faltungsnetze
- rekurrente Netze
- **Transformer**

 TensorFlow

 Hugging Face

Transfer Learning



- vortrainierte Wort-Einbettungen
- vortrainierte **Sprachmodelle**

 Hugging Face

Auswertung Klassifikation

Konfusionsmatrix

		Vorhersage		
		Rest	Schaden	Vertrag
Wahrheit	Rest	37	8	23
	Schaden	12	751	67
	Vertrag	21	2	113

23 Test-Dokumente aus **Rest** wurden falsch als **Vertrag** vorhergesagt.

Auswertung Klassifikation

Klassen-Metriken

		Vorhersage		
		Rest	Schaden	Vertrag
Wahrheit	Rest	37	8	23
	Schaden	12	751	67
	Vertrag	21	2	113
Precision		37/70 = 53%	751/761 = 99%	113/203 = 57%

Recall

$$37/68 = 54\%$$

54% der Schaden-Dokumente wurden richtig erkannt

$$751/830 = 90\%$$

$$113/136 = 83\%$$

57% der Vertrag-Vorhersagen stimmten

Auswertung Klassifikation

Key Performance Indicators

		Vorhersage		
		Rest	Schaden	Vertrag
Wahrheit	Rest	37	8	23
	Schaden	12	751	67
	Vertrag	21	2	113

Accuracy

"Diagonale / Alles"

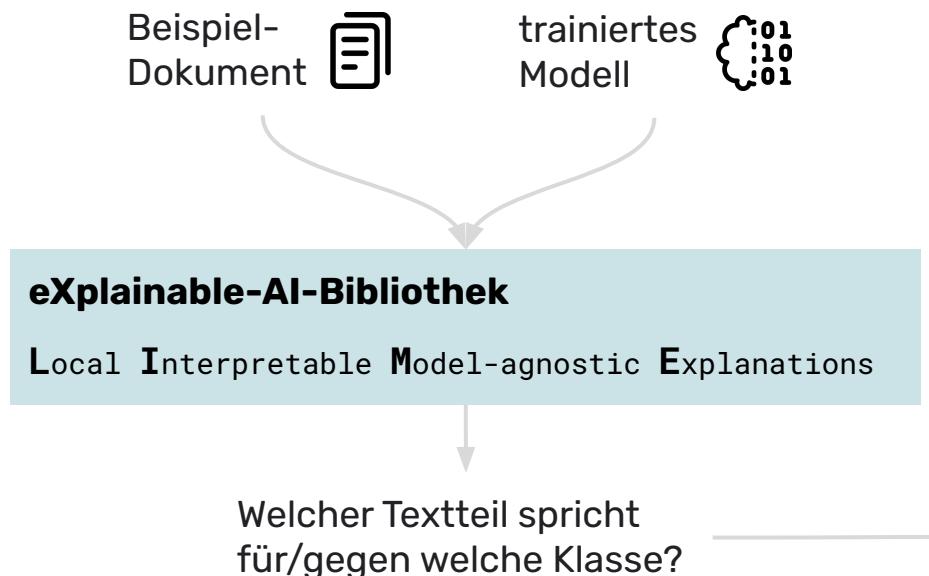
87%

Macro F1

*harmonisches Mittel von Precision und Recall,
Durchschnitt über alle Klassen*

72%

Erklärbarkeit der Klassifikation



Anwendungsfall Extraktion



spezifische Entitäten

- Lieferdatum, Belegnummer, etc.

Freiform-Dokumente

- Kandidaten-Ranking / custom-NER

Unterstützung beim Labeln

Label Studio



The screenshot shows the Label Studio interface. On the left, there's a sidebar with links: PoC-Lieferscheine, Datenbasis, Ergebnisse, Analyse, and Demo. Below that is a 'Dokument-Auswahl' section with a dropdown for 'Qualität auswählen' (set to 'gut') and a radio button for 'mittel'. A 'Test-Dokument auswählen' dropdown is set to '11'. At the bottom, there are checkboxes for 'Zeige Original', 'Zeige erkannten Text', 'Zeige Annotationen', and 'Zeige Vorhersagen', with the last one checked. A 'Zoom-Faktor' slider is set to 0.38. The main area shows a blurred document preview with the word 'TIEFBAUHANDEL' visible.

Analyse der Test-Dokumente

The screenshot shows the SHERLOQ analysis interface. It displays a scanned document titled 'LIEFERSCHEIN' with various data fields annotated. Annotations are color-coded: LS-NR (red), LS-Datum (green), Kostenstelle (cyan), Kreditor (blue), Baustelle (magenta), 'annotiert' (light blue), and 'extrahiert' (light green). The document details include: Kunden-Nr. 9376217, Bestell-Nr. 2121290, KST: 4750034950, Liefertermin: 27.07.2021, Abholung: 27.07.2021 Ganztags, Warenempfänger Firma (blurred), and POS. ARTIKEL-NR. ARTIKELBEZEICHNUNG, BESTELL MENGE, OFFENE MENGE, LIEFERMENGE ME UMRECHNUNG, 4 ROL, 0 ROL, 4 ROL. The document number 4750034950 is highlighted in red.

Extraktion Labeling



SHERLOQ

Labelstudio-Oberfläche mit selbstdefinierten Labels

- vorinstallierten Labeling-Templates
- kollaboratives Arbeiten

The screenshot shows the Labelstudio interface for a labeling task. On the left, there's a sidebar titled 'image' with a list of 10 items, each with a thumbnail and a blue double-headed arrow icon. To the right of the sidebar is a header bar with the project ID '#18', the author 'labelstudio@codecentric...', and other details. Below the header is a row of colored buttons corresponding to labels: Kreditor (green), LS-NR (purple), LS-Datum (pink), Baustelle (yellow), Kostenstelle (orange), and Kreditor-Fußzeile (light green). Underneath these labels are three checkboxes: 'unbrauchbar' (disabled), 'unsicher' (disabled), and 'schlecht_lesbar' (disabled). The main area displays a document from 'GRAVIERANSTALT' with the number '2108-0280'. The document includes sections for 'Gravuren | Schilder | Stempel | Pokale', a menu with options like 'Lasergeschrägen', and a table with columns for 'Pos', 'Menge', 'Gewicht kg', and 'Text'. At the bottom, there's a note about a 'Lieferschein Nr. 9854' and a stamp mentioning 'Trodat Printy 4913'.

Anwendungsfall Extraktion

Machine Learning Methoden:

Kandidaten-Ranking

Kandidaten mit **RegEx** extrahieren

Merkmale jedes Kandidaten
ermitteln (**Position + umliegender
Text**)

ein ML Modell anhand der
Kandidaten-Features **trainieren**

Auftragsdatum	29.07.2021	Lieferzeit	28.07.2021	Druckdatum / Zeit	28.07.2021 09:36:01
Ihre Referenz		Kontaktperson		Projektnummer	201744
Umsatzsteuer Identnummer		Mailadresse von Kontaktperson		Projekt	
Ihre Kontaktperson		AD-Mitarbeiter		Spediteur	
class	page	x	y	text	text_left
LS-Datum	1	261	884	09.07.2021	Auftragsdatum
LS-Datum	1	684	886	29.07.2021	Lieferzeit
LS-Datum	1	1137	885	28.07.2021	Druckdatum / Zeit

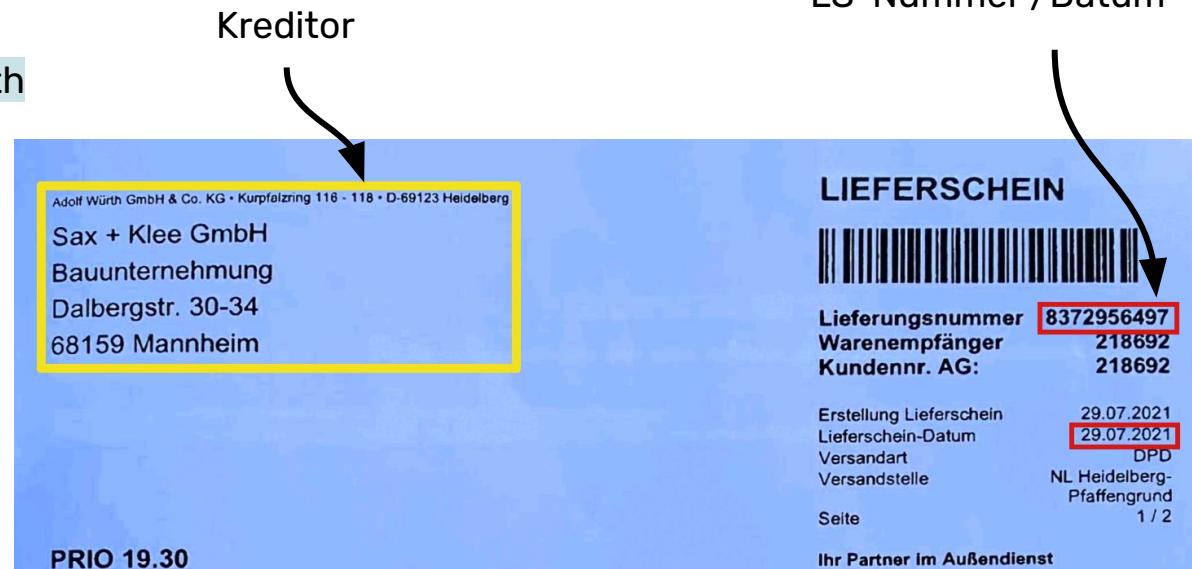
Anwendungsfall Extraktion

Machine Learning Methoden:

Custom Named Entity Recognition with

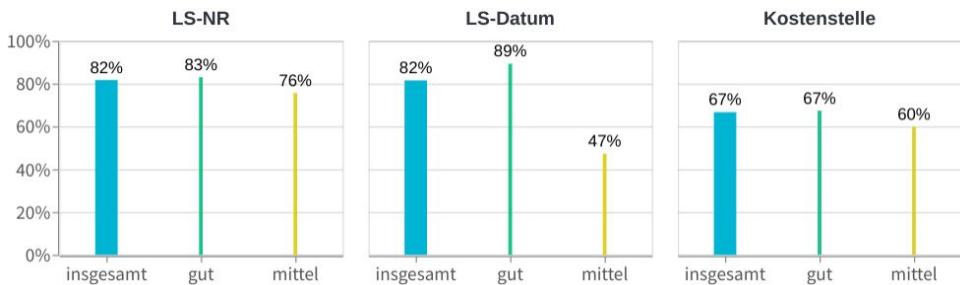
spaCy

CNN-LSTM Model Architektur

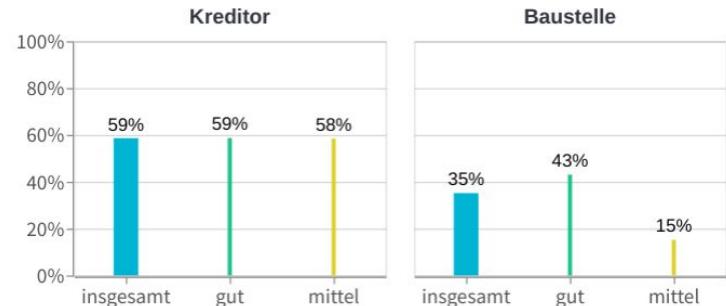


Ergebnisse der Extraktion

Kandidaten-Ranking



Custom Named Entity Recognition with Spacy



Von Daten zur KI-Lösung

Herausforderungen und Tools

Daten
annotieren



Label
Studio

Experimente
tracken



Daten
verwalten



Daten
versionieren

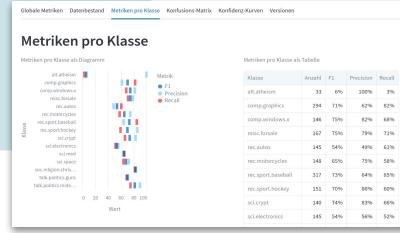
ML-Pipelines
orchestrieren



Pipelines
konfigurieren



Ergebnisse
analysieren



AI-Service
bauen

Service APIs BentoML Service API endpoints for inference.

- POST** /v1/debug/preprocess InferenceAPI[ByteIOFile --> JSON]
- POST** /v1/debug/featureize InferenceAPI[ByteIOFile --> JSON]
- POST** /v1/debug/predict InferenceAPI[ByteIOFile --> JSON]
- POST** /v1/predict InferenceAPI[JSON --> JSON]

Predict request for classification.

The request data is expected to be a single PDF file passed as base64-encoded content with content type `"application/pdf"`. The PDF is classified using:

- optional rules passed in the request metadata field under `request.meta.extra.rules`, as a list in the format explained in <https://bentoml.org>.
- If a rule applies, it overrides the AI result.

The response data contains the classification result and candidates, both in form of prediction items. For each prediction item, the following fields are available:

- `key`: rule id
- `value`: a class name in long form (up to three components separated by `/`) or `unknown`
- `selection`: set to `True` if the result is the best and to `False` for other candidates.

SHERLOQ Engineering



SHERLOQ Bausteine



Training (AI-Lab)

Training der AI Modelle

- bekannter Werkzeugkoffer
- schnelles Deployment
- skalierbar
- ressourcen-intensiv
- PoC Umgebung (On-Demand)
- KI-Checkup (Nach-Training)



Prediction

Betrieb der AI Modelle

- Individuelle AI Modelle/Preprocessing
- wiederverwendbare Integrationen (WebService, PubSub, Kofax)
- Daten Sammeln (für KI-Checkup)
- skalierbar



Zentrale Services

Kundenübergreifend

- Sammeln von Nutzungsdaten für die Abrechnung
- Erstellung von monatlichen Reports für die Abrechnung
- Lizenzierung
- ...



Rahmenbedingungen

- On-Premises / Hybrid
- Konfigurierbar
- schneller Start für Data Scientists
- Authentifizierung / Autorisierung

SHERLOQ Engineering

codecentric

codecentric Gitlab (Repositories, Container Registry)

On-Premises

Kubernetes
(microk8s)



Training (AI-Lab)

- Dagster
- MinIO
- MLflow
- JupyterLab
- LabelStudio
- Streamlit



Prediction

- AI App Gateway
- AI App (+Mock)
- Lens
- Brain
- Insights
- MinIO

ArgoCD - Keycloak - Prometheus -
Grafana - Loki - ...

ArgoCD - Keycloak - Prometheus -
Grafana - Loki - ...

Azure

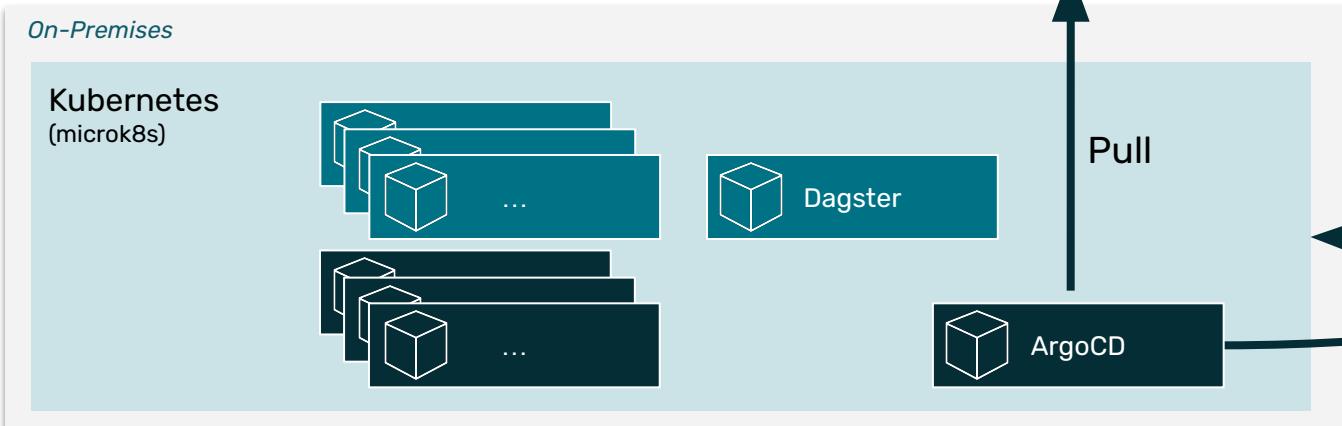
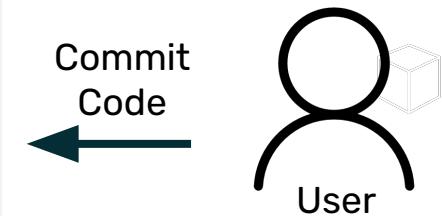


Zentrale Services

- Subscription Manager
- Lizenzservice

SHERLOQ Deployment

Am Beispiel von Dagster



SHERLCQ

a solution by @codecentric



Elvira Siegel
Data Scientist
elvira.siegel@codecentric.de



codecentric AG
Am Mittelhafen 14
48155 Münster

www.codecentric.de



Thomas Timmermann
Data Scientist
thomas.timmermann@codecentric.de



Johannes Voscout
Developer
johannes.voscout@codecentric.de



a solution by @codecentric

