

CHRONICLING GERMANY

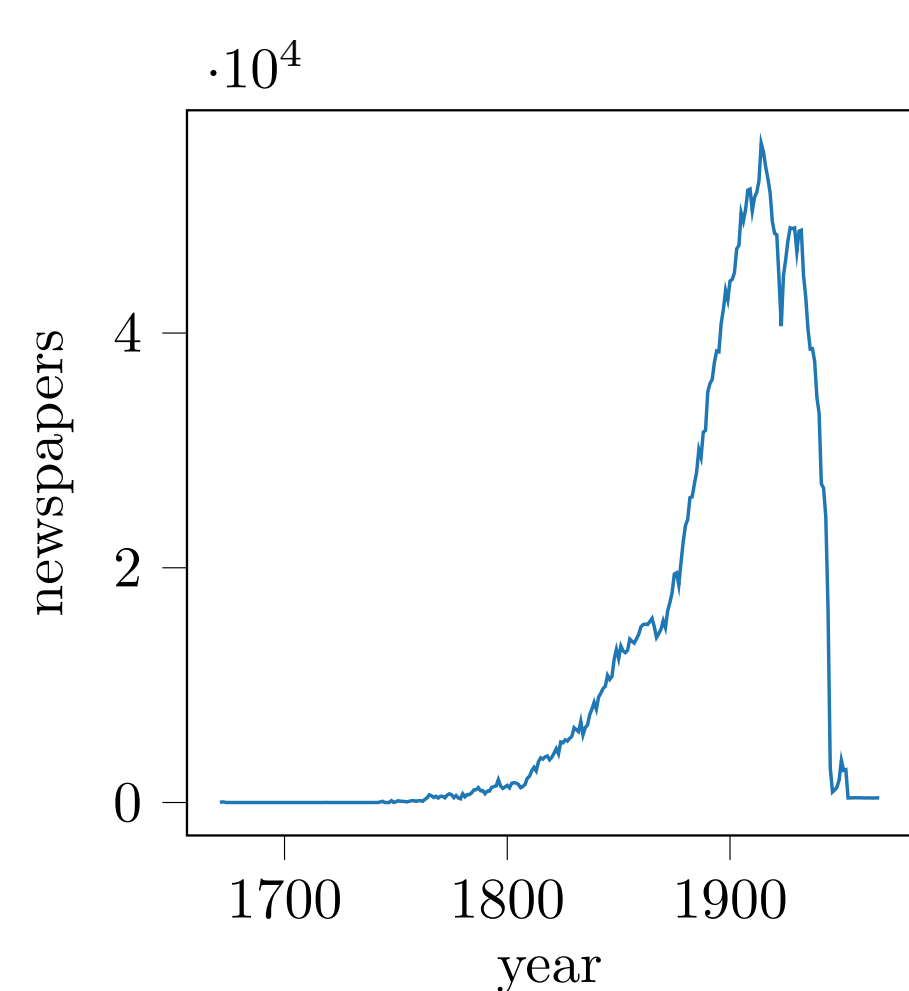
C. Schultze¹, K. Kuebart²

¹ High-Performance Computing and Analytics Lab and ² Institut für Geschichtswissenschaft, Universität Bonn

Einleitung

Zeitungsportale der Deutschen Digitalen Bibliothek und der Länder bieten Forschenden OCR-Text (Optical Character Recognition). Leider ist das Layout oft fehlerhaft erkannt worden, was die Nutzbarkeit dieser reichhaltigen Quellen einschränkt. Viele Analysemethoden der digitalen Textanalyse, wie die Kontextanalyse, Topic Modelling, Sentimentanalyse und große Teile des Natural Language Processing sind darauf angewiesen, dass der Text in korrekt erkannter Lesereihenfolge vorliegt und können daher an solchen Daten nur eingeschränkt eingesetzt werden. Wir stellen unseren Ansatz vor, Layout- und Textinformationen aus den Seitenbildern neu zu generieren. Zu diesem Zweck haben wir einen Trainingsdatensatz und eine darauf trainierte Pipeline geschaffen, die mehrere eigens angepasste KI-Methoden miteinander verbindet. In einem technischen Aufsatz, der in einem einschlägigen Computer Science Journal veröffentlicht wurde (DMLR 2, 2025), stellen wir diese Pipeline vor und erproben sie auf unserem Datensatz. Wir evaluieren die Hauptkomponenten der Pipeline auch im Vergleich zu anderen Modellen. Sowohl unser Datensatz als auch der Code unserer Pipeline sind online frei verfügbar. Diese Arbeit bildet einen Ausgangspunkt für zukünftige Forschung im Bereich der Digital History und der Verarbeitung historischer deutschsprachiger Zeitungen. Sie ermöglicht artikelbasierte und lesereihenfolgeabhängige quantitative Forschungen an digitalisierten Zeitungen.

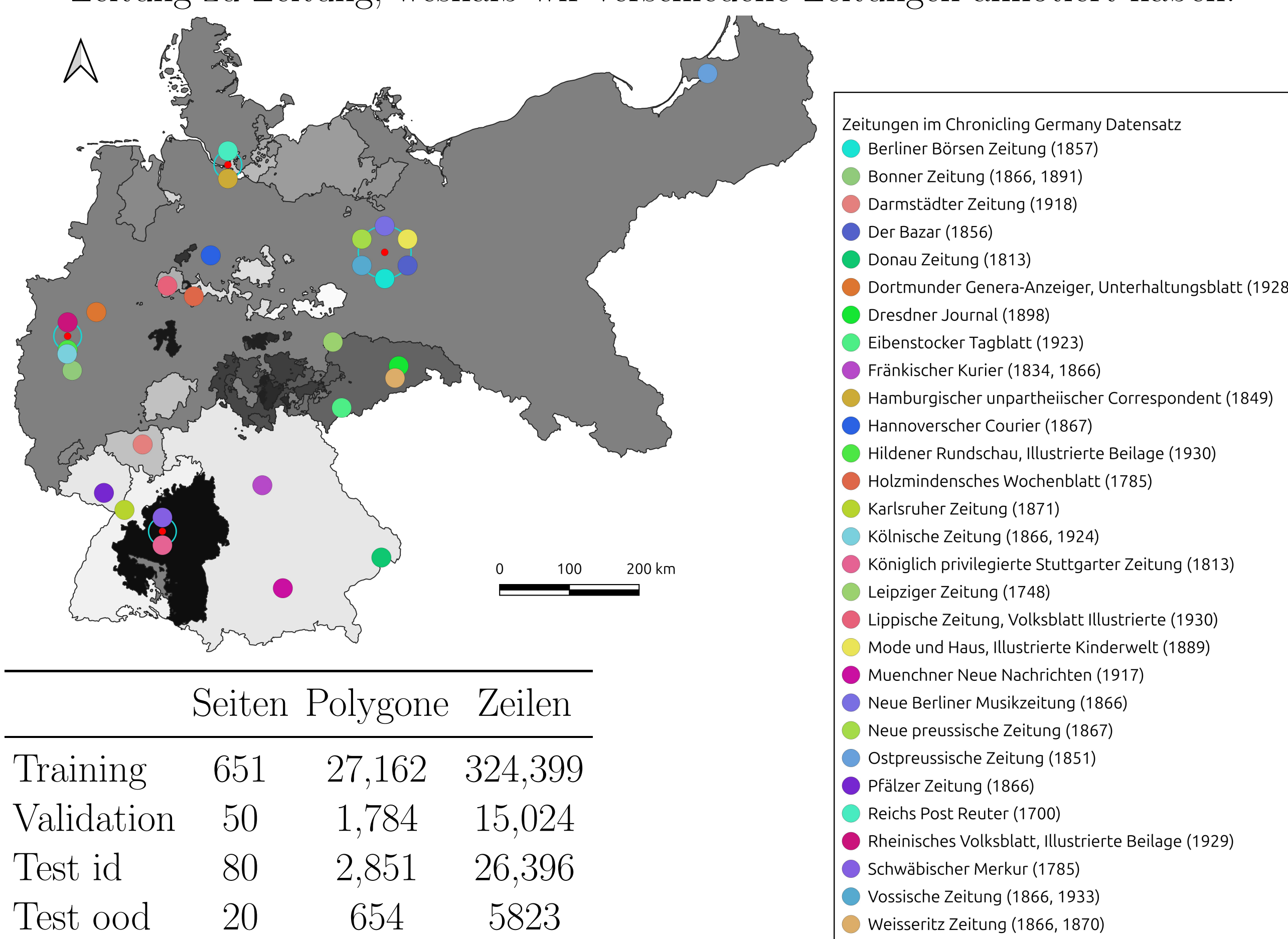
Historische Zeitungen



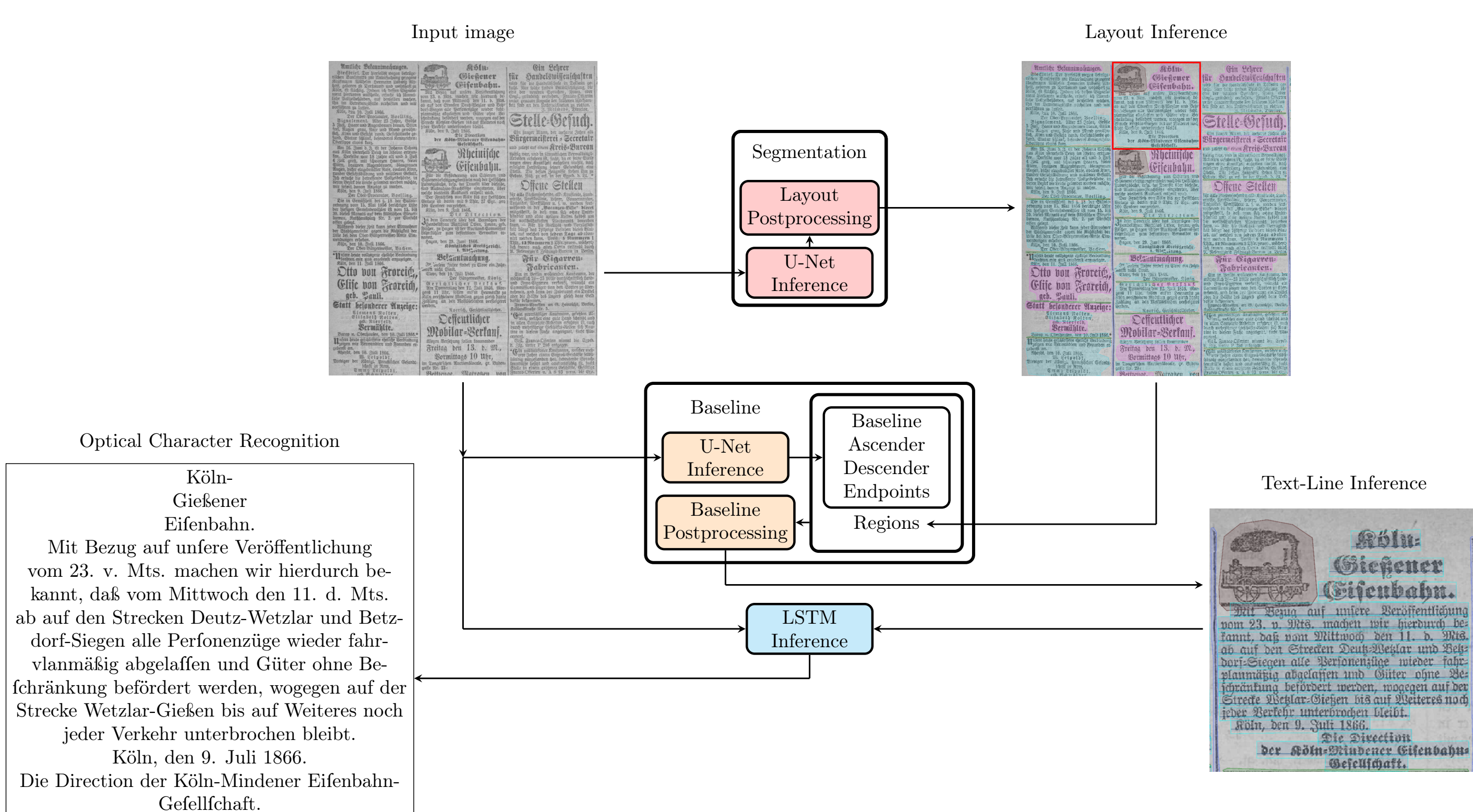
Links: Anzahl der im Deutschen Zeitungsportal der DDB (www.deutsche-digitale-bibliothek.de/newspaper) verfügbaren Zeitungsausgaben pro Jahr. Daten von Januar 2024. Mitte: Deckblatt des *Hannoverschen Couriers* vom 17. Juli 1867 aus dem Zeitungsportal der *Niedersächsischen Landesbibliothek*. Die Polygone markieren je eine Zeile des TXT-Volltexts. Rechts: Unsere Ground Truth Annotation derselben Seite. Header, Caption, Überschriften und Text sind in unterschiedlichen Farben markiert. Wir stellen XML zur Verfügung, mit Polygon-, Zeilen-, Text- und Metadaten.

Der Datensatz

Layout und Schriftart verändern sich über die Zeit und unterscheiden sich von Zeitung zu Zeitung, weshalb wir verschiedene Zeitungen annotiert haben:



Pipeline-Übersicht



Die gesamte Pipeline: Layouterkennung, Baseline-Erkennung und OCR verwenden jeweils separate Netzwerke. Die Daten werden als XML und csv ausgegeben und können so flexibel weiterverwendet werden.

Ergebnisse

Unsere auf Pero und Kraken aufbauende Texterkennung übertreffen - an unserem Datensatz bemessen - bisherige Fraktur-OCR Implementationen. Auch unsere Layouterkennung überzeugt. Wir haben unsere Pipeline an einem separaten Testdatensatz evaluiert. Die Tabellen unten zeigen die Ergebnisse für den *in distribution* (*id*) und den *out of distribution* (*ood*) Test-Datensatz. Gezeigt werden die Genauigkeit (F1-Score) für die Regions-Tags *paragraph* und *heading* sowie für Baselines innerhalb der Textregionen (links) und die OCR-Ergebnisse (rechts). Die Prozentwerte der Genauigkeit basieren auf der Levenshtein-Distanz pro Zeichen. Daneben stehen die Anteile an *perfekter*, also fehlerfreier Zeilen.

	Genauigkeit		Genauigkeit [%] Perfekt [%]			
	id	ood	id	ood	id	ood
paragraph	0,95	0,90	OCR only	97,7	98,1	70,4 58,4
heading	0,69	0,50				
baseline		0,92	Full Pipeline	93,5	92,0	55,7 43,3

Anwendungen

Unsere Pipeline ist darauf ausgelegt, im großen Stil Zeitungssseiten in ein strukturiertes Datenformat zu transkribieren. Diese Daten können mithilfe von Topic Modelling, spezialisierten LLMs oder anderen Such- oder Zuordnungsmethoden für eine Vielzahl von Forschungsfragen zu jedem in Zeitungen behandelten Thema verarbeitet und gefiltert und unter Anwendung von NLP- und *machine learning* -Techniken analysiert werden. Neben Fragestellungen der Medien- und Diskursgeschichte bieten sich auch solche der Wirtschafts- und Sozialgeschichte, aber auch der politischen Geschichte an.

Unser nächster Schritt ist es, alle digitalisierten Ausgaben der Kölnische Zeitung von 1803-1945 (etwa 425.000 Seiten) mit der Pipeline zu verarbeiten. Die Transkription resultiert in etwa 27 Millionen Text Blöcken. Wir werden mithilfe von Topic Modelling themenspezifische Korpora erstellen. Einen Aufsatz, der die Eignung verschiedener Topic Modelle zu diesem Zweck vergleicht, werden wir im DHNB - Tagungsband 2025 publizieren.

Verweise

Projekt - Paper: <https://data.mlr.press/assets/pdf/v02-16.pdf>, Code: <https://github.com/Digital-History-Bonn/Chronicling-Germany-Code>, Data: <https://gitlab.uni-bonn.de/digital-history/Chronicling-Germany-Dataset>. Topic Modeling - Paper (ab Oktober/November): <https://journals.uio.no/dhnbpub/index>
Weitere Verweise: Pero OCR: <https://arxiv.org/abs/2102.11838> Kraken: <https://kraken.re>