TEXTBITE

Segmentation of Logical Units in Text

Bc. Martin Kostelník

Supervisor: Ing. Karel Beneš

What is TextBite?

A deep learning tool for extraction of articles, news, dictionary entries, book segments, etc. from historical documents.

TextBite enhances search capabilites in digitized documents used by librarians and scientists.

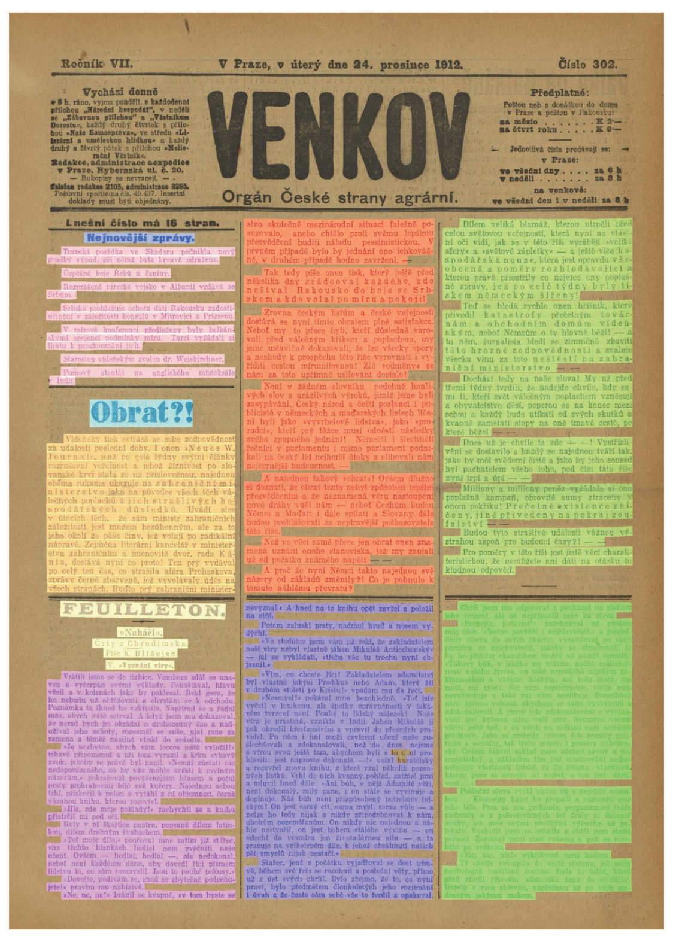
01 / Data

A custom dataset was created and labeled based upon data from Czech Digital Library.

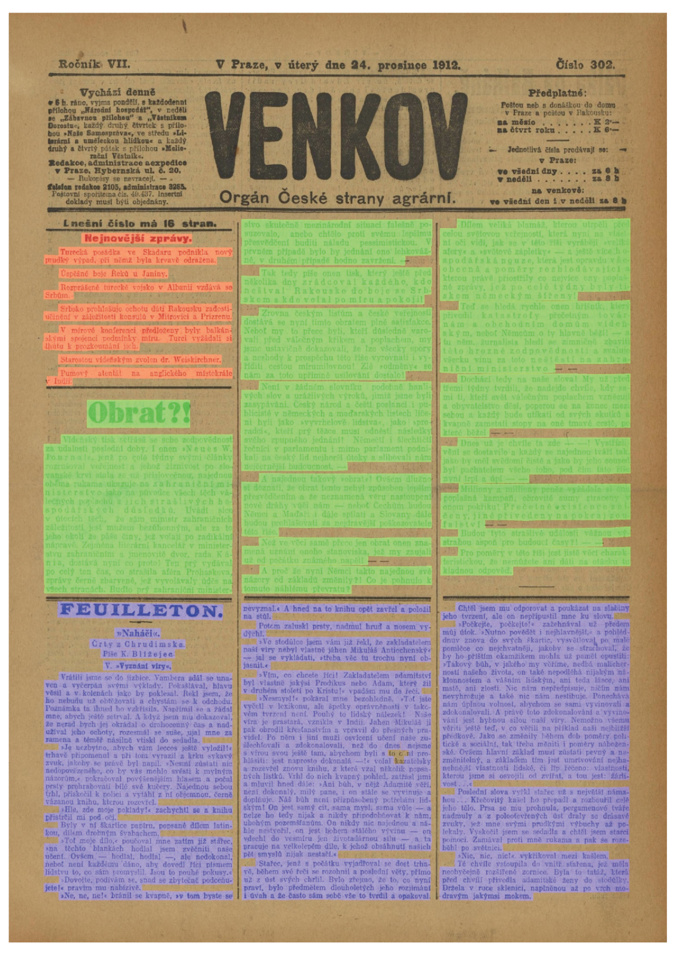
	# Pages		
	Books	Dictionaries	Periodicals
Train	169	690	3005
Validation	15	30	45
Test	15	30	45

02 / GCN Effect

GCN joins regions detected by the YOLOv8 together.



YOLOv8



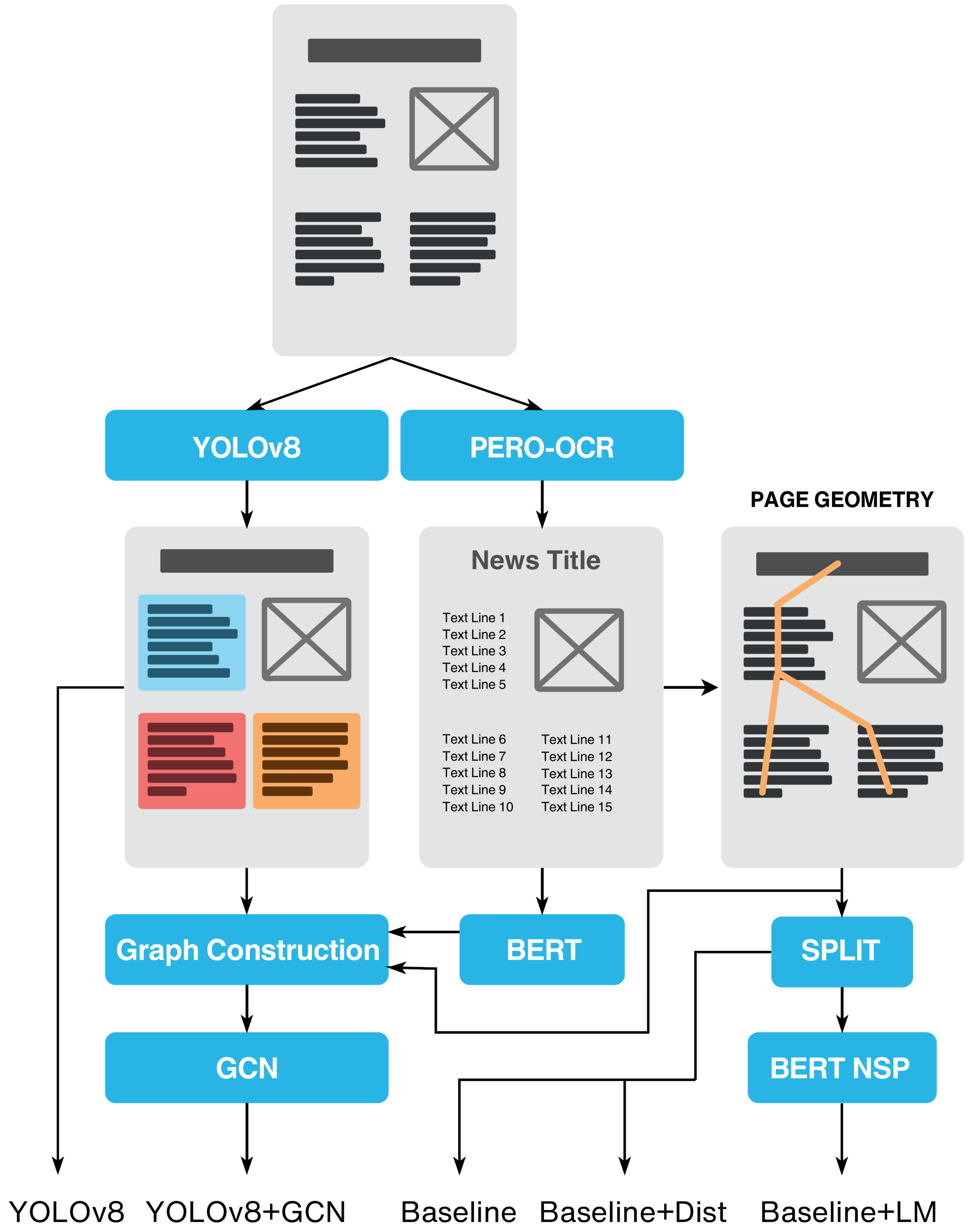
GCN

03 / Results

Method	V-Measure Books	Dictionaries	Periodical	Time [s]
Baseline	19.72	40.11	52.20	0.14
Baseline+Dist	56.56	62.68	78.61	0.14
Baseline+LM	61.09	82.83	65.35	1.66
YOLOv8	65.33	93.36	83.93	0.16
YOLOv8+GCN	73.59	95.17	89.32	0.66

On NVIDIA RTX A6000

04 / Pipeline



YOLOv8

PERO-OCR

PAGE GEOMETRY

Graph Construction

BERT

SPLIT

GCN

BERT NSP

YOLOv8

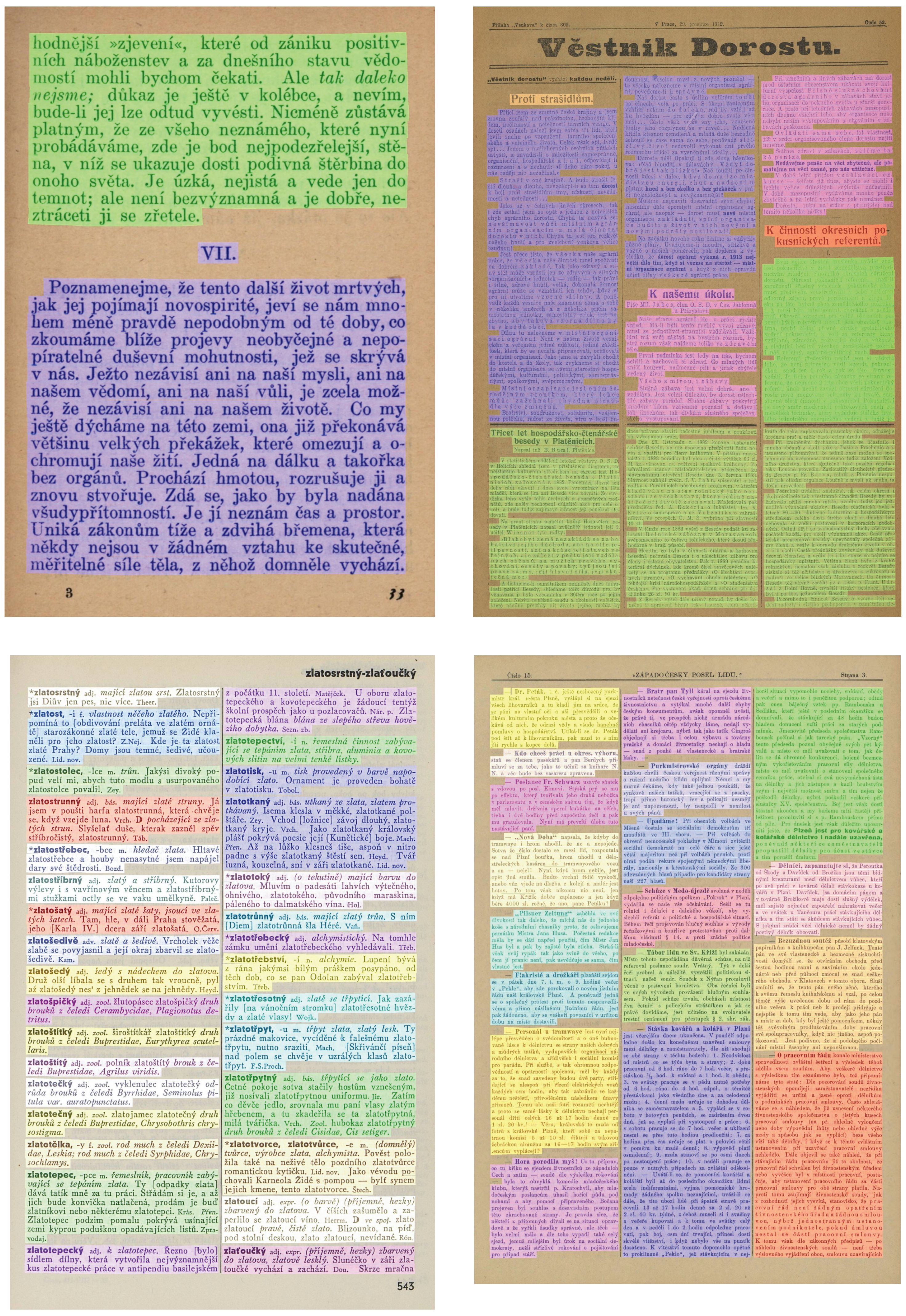
YOLOv8+GCN

Baseline

Baseline+Dist

Baseline+LM

05 / Final Output



BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2024