ParlaCAP – Mining the ParlaMint Treasures with Multilingual Topic and Sentiment Classification

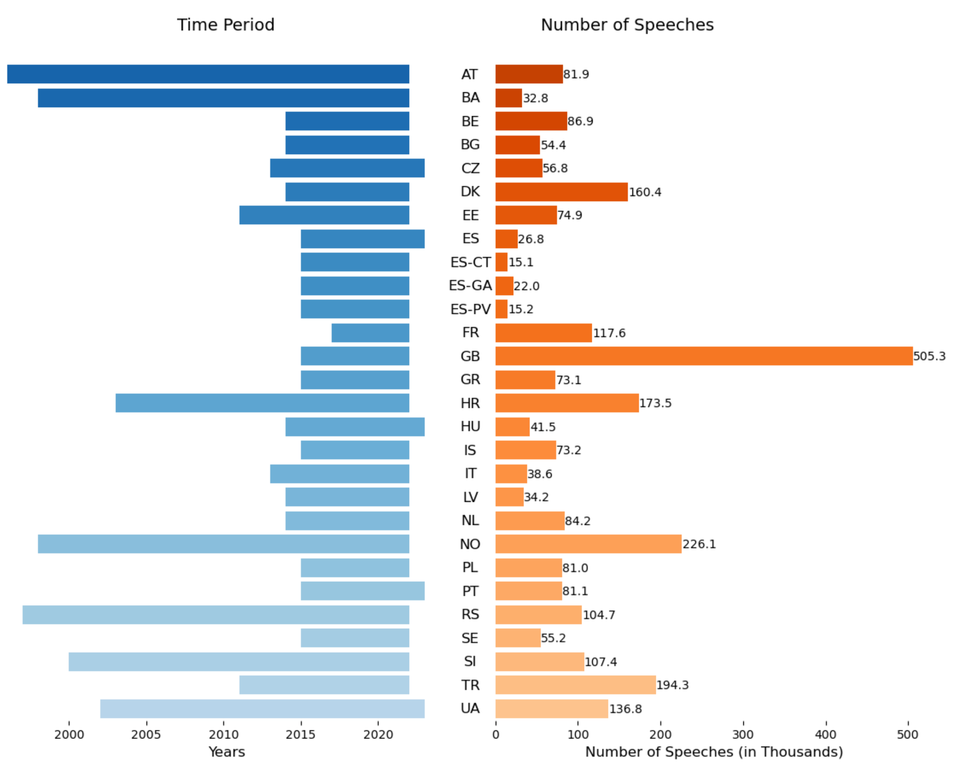
Nikola Ljubešić[1,2,3], Taja Kuzman Pungeršek[1], Peter Rupnik[1], Ivan Porupski[1], Vuk Dinić[1]

[1] Jožef Stefan Institute, Ljubljana, Slovenia
[2] Faculty of Computer and Information Science, University of Ljubljana, Slovenia
[3] Institute of Contemporary History, Ljubljana, Slovenia

Overview

- ParlaCAP is an OSCARS (`https://oscars-project.eu`) Open Science Cascading Grants project, 2025-2026
- Main challenge: analyse the content of the 7+ million speeches in the ParlaMint corpora, given in 28 parliaments, by topic and sentiment
- Follow the “text as data” paradigm known in political science and automatically label each speech by topic and sentiment

ParlaCAP Dataset



28 parliaments, 7.92M speeches

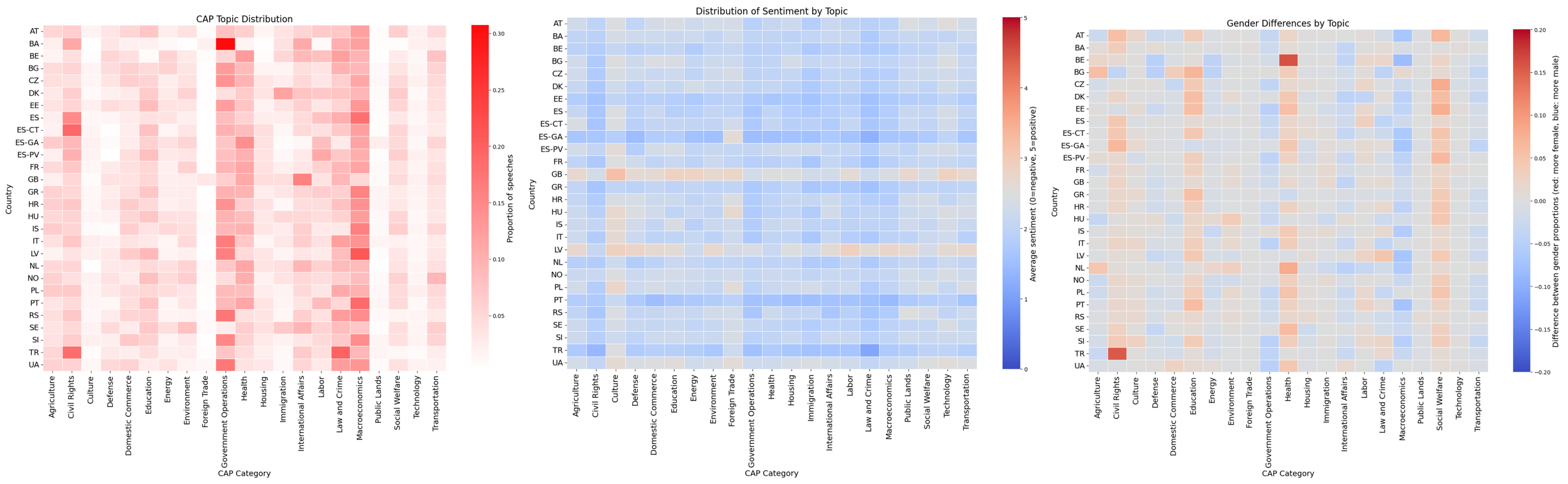
ParlaCAP Topic Classifier

- 22 topic labels from Comparative Agendas Project (CAP) schema
- LLM Teacher-Student framework: GPT-4o model (teacher) used to label 29k+ ParlaMint speeches, XLM-R-parla (student) fine-tuned on these data
- Comparable inter-annotator agreement between (human, human) and (machine, human)
- Manually-annotated test sets in English, Bosnian, Croatian and Serbian (app. 3400 instances in total)

Performance (without Mix instances - prediction confidence below 0.6):

language	macro-F1 (ParlaCAP)	macro-F1 (GPT-4o)	% of Mix labels
en	0.758	0.754	8.9
bs	0.680	0.656	11.1
hr	0.726	0.706	11.4
sr	0.743	0.743	11.1

First Analyses



Data Structure

Three TSV files per parliament:

- Speech-level TSV
  - ParlaMint metadata (speaker, party, party status, chairing...)
  - PartyFacts Party ID, V-DEM Country ID
  - Topic label
  - Aggregated sentiment label
  - Text of the speech, machine-translated text
  - ...
- Speech-level TSV without text (88% reduction in size)
- Sentence-level TSV
  - Speech ID
  - Sentiment label
  - Text of the sentence

Data released through CROSSDA (Croatian node of CESSDA)

`https://doi.org/10.23669/1ZTELP`

Links

- Parliamentary sentiment model
  `https://huggingface.co/classla/xlm-r-parlasent`
- ParlaCAP topic classifier
  `https://huggingface.co/classla/ParlaCAP-Topic-Classifier`
- Tutorials for analyzing ParlaCAP datasets with Python
  `https://github.com/clarinsi/ParlaCAP-Analysis-Tutorials`

Acknowledgement

OSCARS Open Science Clusters' Action for Research & Society

CLARIN.SI