Multilingual NLP can shed light on many secrets of parliamentary proceedings

Nikola Ljubešić

Jožef Stefan Institute

Ljubljana, Slovenia

January 19, 2026 | Applied NLP Tools for Digital Humanities

What this talk is about

- “Recent revolution in natural language processing” and how to properly use it for valid research
- “Semantic data processing” – research on unprecedented data sizes
- Our data are ParlaMint - transcripts (and recordings) of parliamentary sessions from 26 national and 3 regional European parliaments, 2015-2022, 8 million speeches, more than 1 billion words
- Two downstream projects - ParlaCAP (text) and ParlaSpeech (speech)

The ParlaMint project

The ParlaMint Project

CLARIN ERIC research infrastructure flagship project

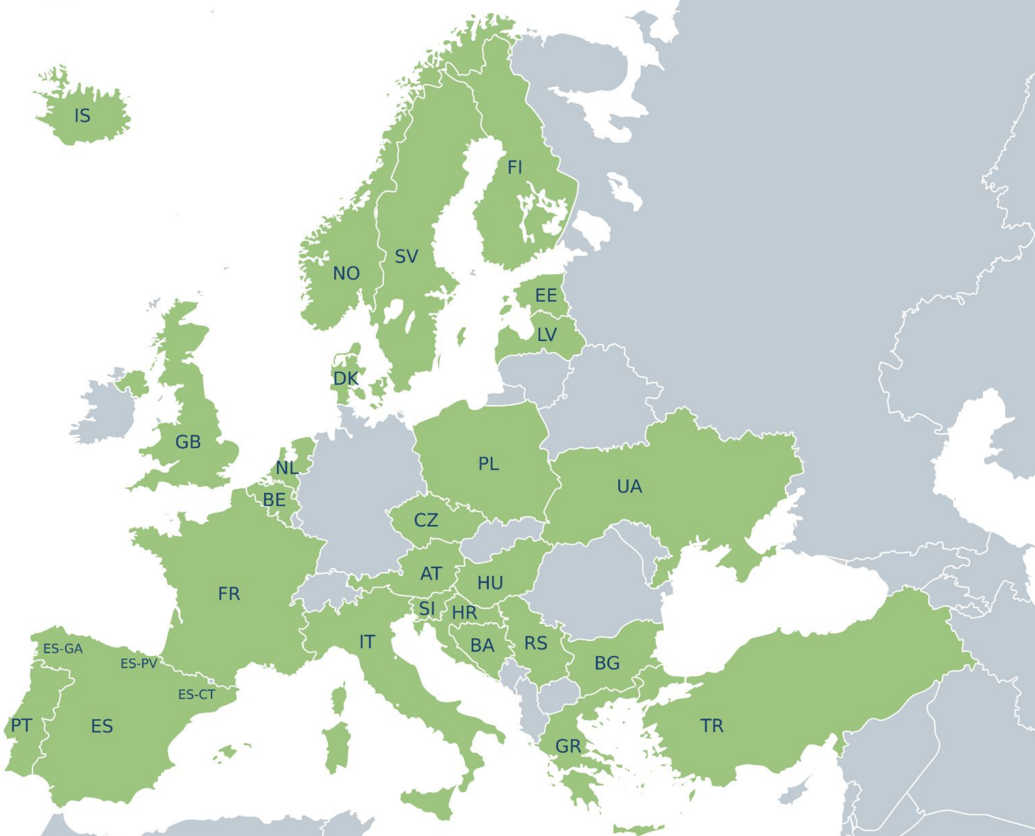
- ParlaMint I (2020–2021)
- ParlaMint II (2022-2023)

Main deliverable:

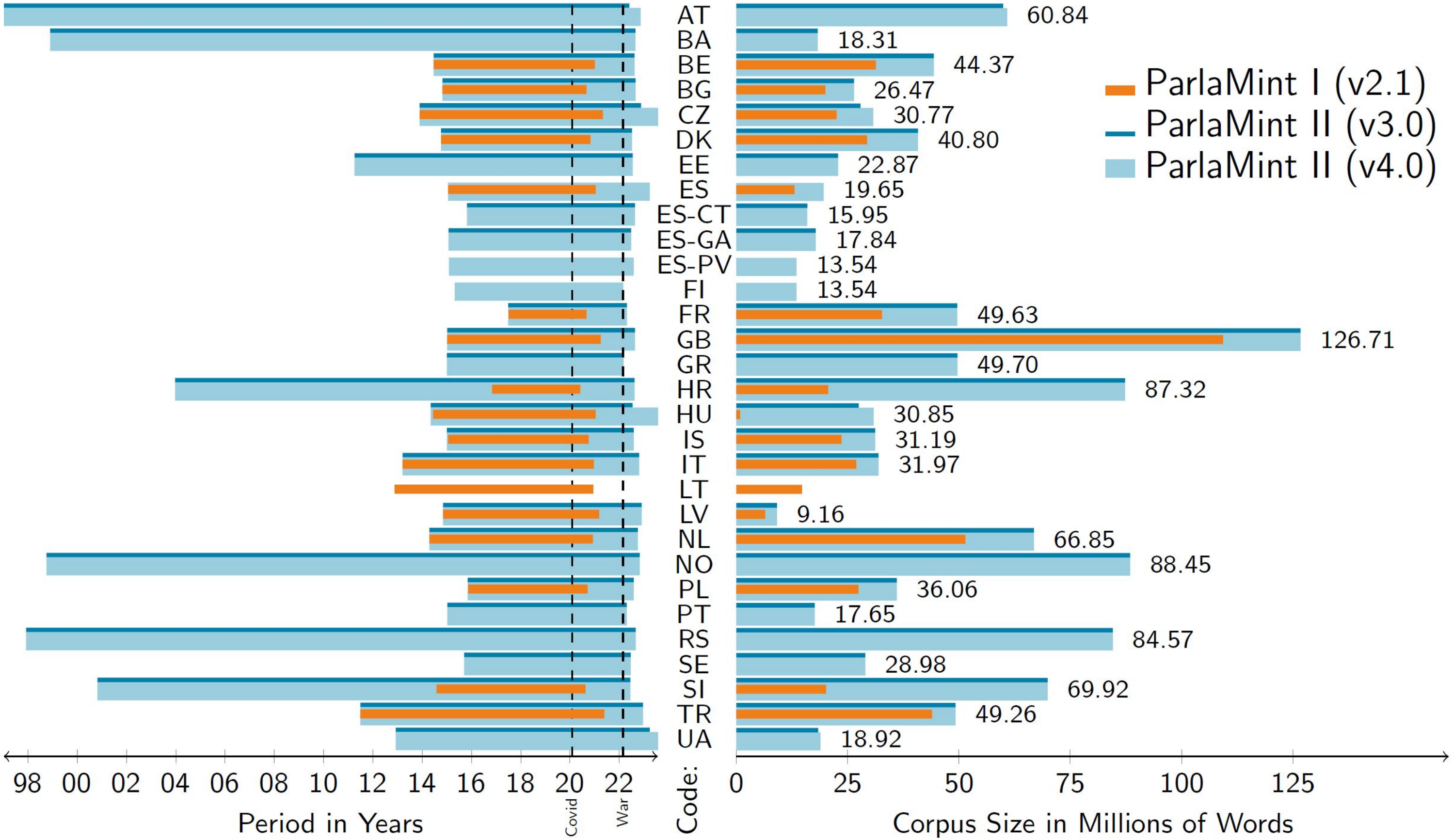
- Uniformly encoded transcriptions of speeches from European parliaments
- Rich metadata (speaker, gender, age, party, orientation, power status…)
- Linguistically annotated (part-of-speech, lemma, named entities, speeches also machine-translated into English and annotated)
- Openly available (CLARIN.SI FAIR repository and concordancer)

Geographic coverage
(26 countries and 3 autonomous regions)

Austria
Basque Country
Bosnia and Herzegovina
Belgium
Bulgaria
Catalonia
Croatia
Czech Republic
Denmark
Estonia
Finland
France
Galicia
Greece
Hungary
Iceland
Italy
Latvia
Netherlands
Norway
Poland
Portugal
Serbia
Slovenia
Spain
Sweden
Turkey
UK
Ukraine

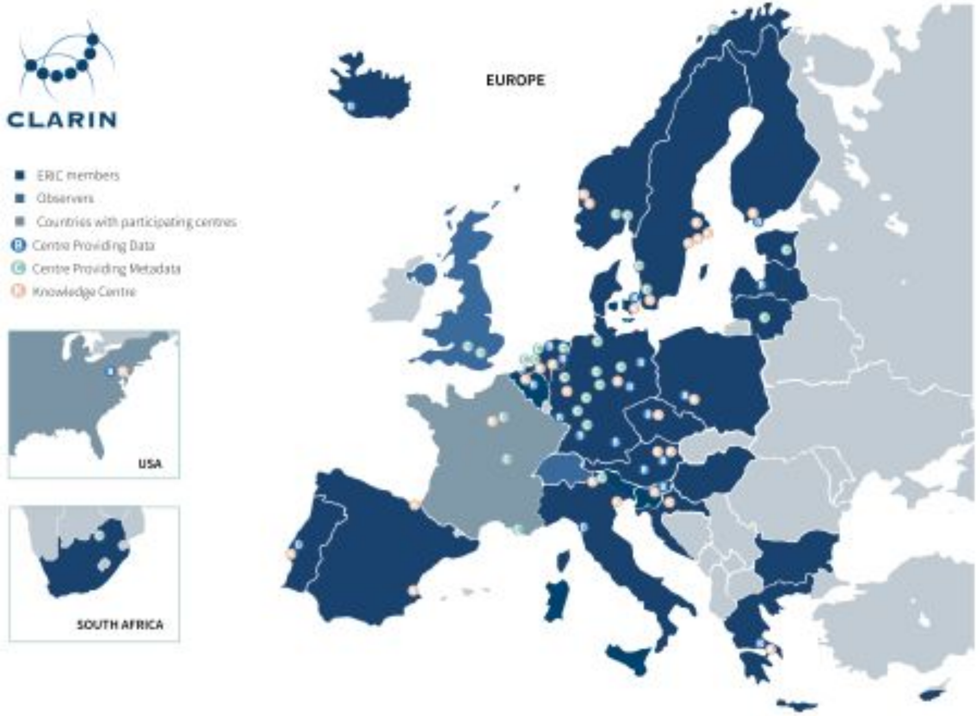


Time coverage and data size



A note on CLARIN

- CLARIN is a digital infrastructure offering data, tools and services to support research based on language resources
- A distributed network of 70 centres with 24 member countries and 2 observers



ParlaMint on the concordancer and in the repository

Data on the CLARIN.SI repository

http://hdl.handle.net/11356/2004 (text and metadata)

http://hdl.handle.net/11356/2005 (+ linguistic annotation)

http://hdl.handle.net/11356/2006 (+ machine-translated text)

Concordancer for instant search

https://www.clarin.si/ske/#dashboard?corpname=parlamint41_at

https://www.clarin.si/ske/#dashboard?corpname=parlamint41_xx_en

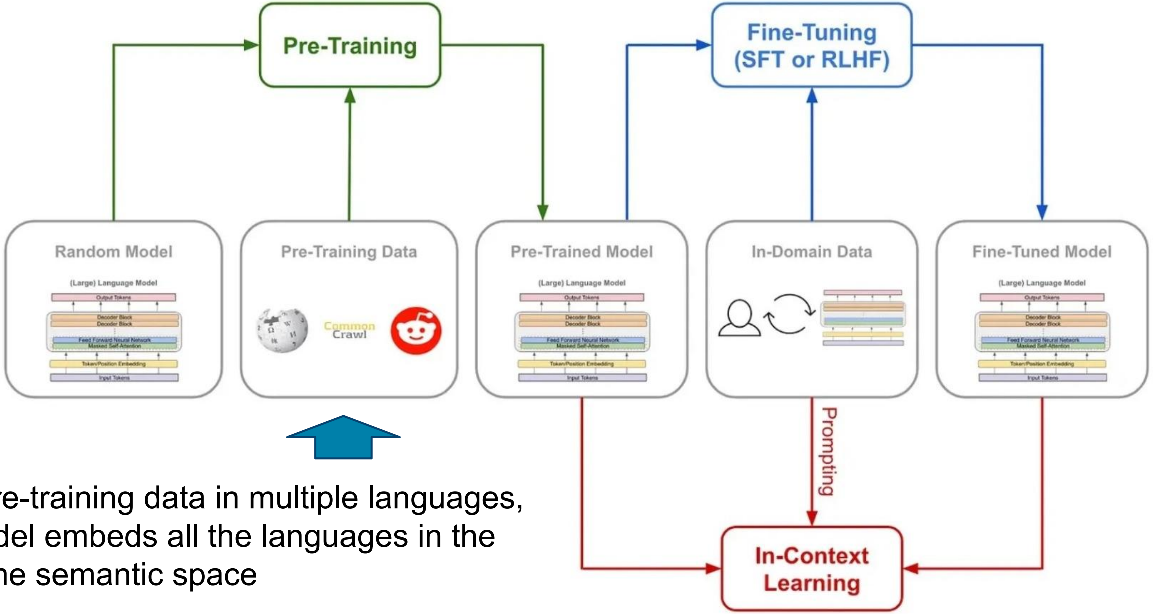
Adding (more) NLP into the mix

How to unlock the ParlaMint potential

- ParlaMint are primarily linguistic corpora, currently most useful to corpus and computational linguists
- Parliamentary data most relevant to social and political scientists, currently work on one of few parliaments due to data scarcity
- Social and political scientists less skilled in working with text
- “Text as data” paradigm - transform text into discrete values to be used in downstream analysis and modelling

Pre-trained language models

1. In-domain data in lang X
2. Model was pre-trained on langs X, Y, Z
3. Model fine-tuned on X will work on langs X, Y, Z. (zero-shot cross-lingual)



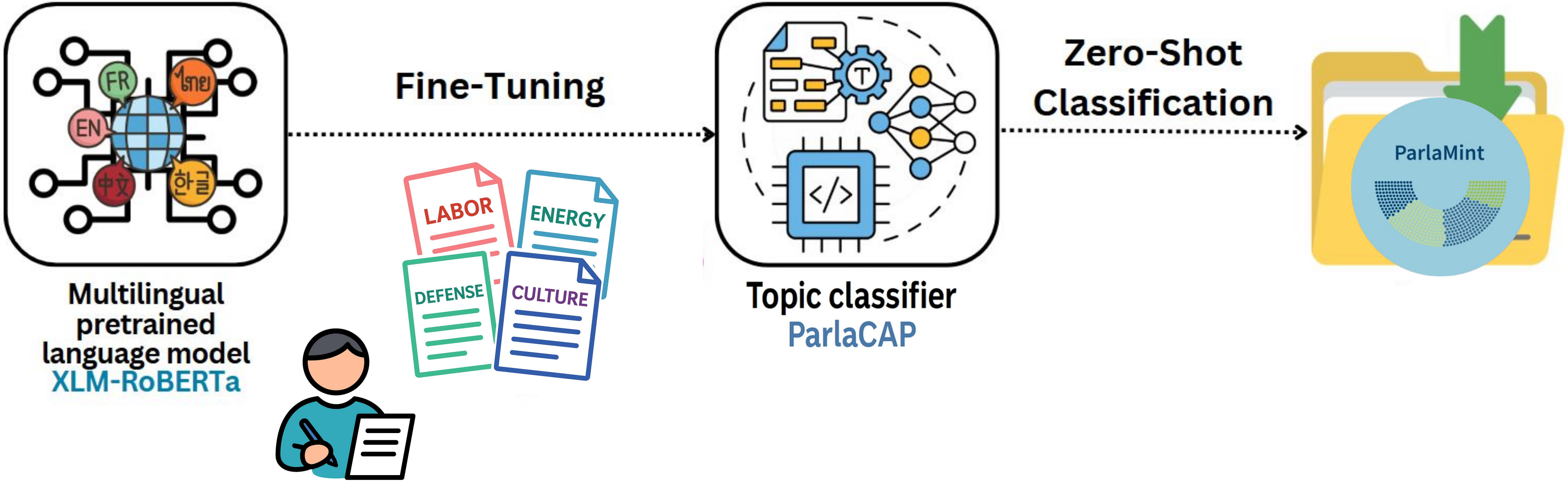
if pre-training data in multiple languages, model embeds all the languages in the same semantic space

https://medium.com/@bijit211987/the-evolution-of-language-models-pre-training-fine-tuning-and-in-context-learning-b63d4c161e49

ParlaCAP

- “Comparing agenda settings across parliaments via the ParlaMint dataset” - OSCARS Horizon Project, uptake of open science in Europe
- Cross-lingual language models to annotate more than 8 million ParlaMint speech transcripts from all 29 parliaments, 27 languages
- Annotations on
  1. Sentiment (negative, mixed negative, neutral negative, neutral positive, mixed positive, positive)
  2. Topic (Comparative Agenda Project)

Fine-tuning a model to a task



Multilingual pretrained language model
XLM-RoBERTa

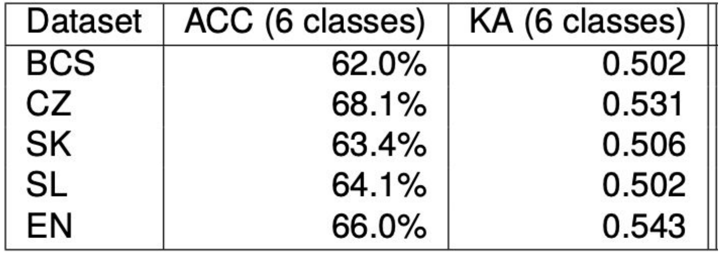
Fine-Tuning

Topic classifier
ParlaCAP

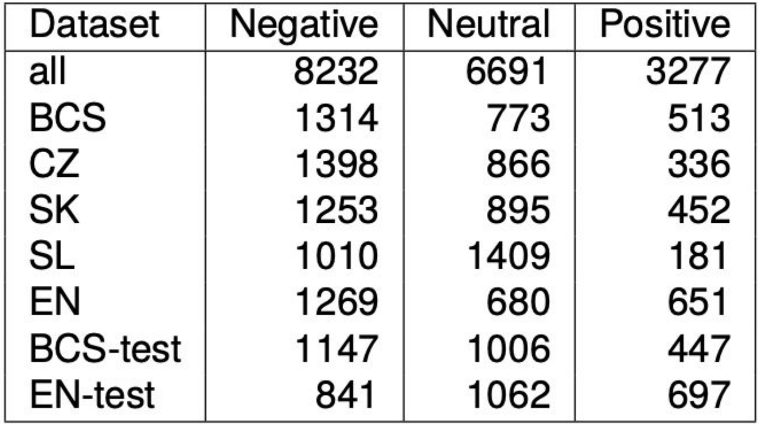
Zero-Shot Classification

ParlaSent fine-tuning dataset

- Sentiment
- Mochtak et al. (2024)
- Dataset available at http://hdl.handle.net/11356/1868



Dataset	ACC (6 classes)	KA (6 classes)
BCS	62.0%	0.502
CZ	68.1%	0.531
SK	63.4%	0.506
SL	64.1%	0.502
EN	66.0%	0.543

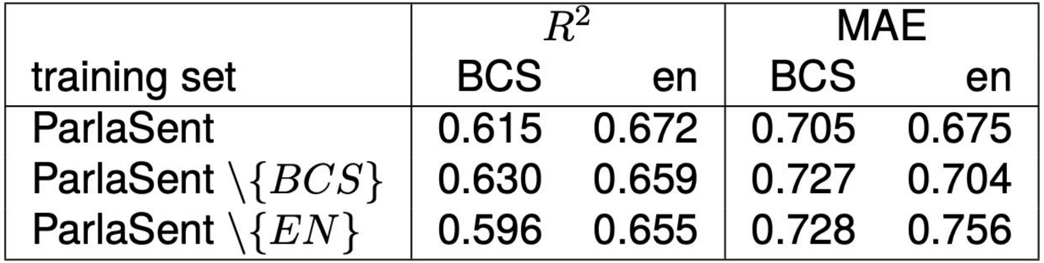


Dataset	Negative	Neutral	Positive
all	8232	6691	3277
BCS	1314	773	513
CZ	1398	866	336
SK	1253	895	452
SL	1010	1409	181
EN	1269	680	651
BCS-test	1147	1006	447
EN-test	841	1062	697

Table 2: Distribution of the three-class labels across datasets.

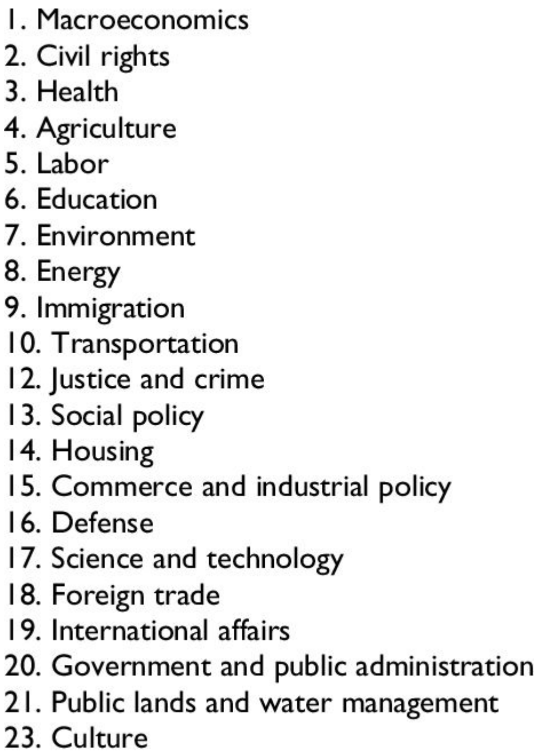
ParlaSent model and its multilingual capacity

- Measure performance on Bosnian-Croatian-Serbian and English test
- Fine-tuning on 1. all ParlaSent and 2. with specific language removed
- R^2 – higher is better (0-1), MAE – lower is better (0-5)
- Strong cross-linguality regardless of language



	R^2		MAE	
training set	BCS	en	BCS	en
ParlaSent	0.615	0.672	0.705	0.675
ParlaSent $\backslash\{BCS\}$	0.630	0.659	0.727	0.704
ParlaSent $\backslash\{EN\}$	0.596	0.655	0.728	0.756

The CAP in ParlaCAP



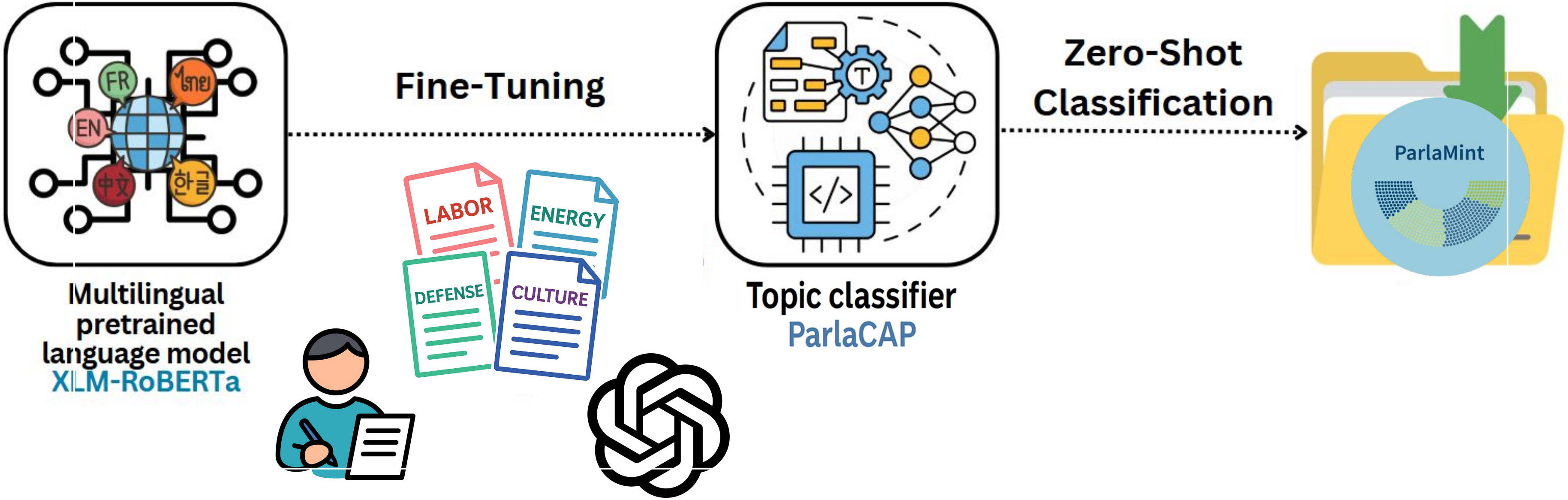
1. Macroeconomics
2. Civil rights
3. Health
4. Agriculture
5. Labor
6. Education
7. Environment
8. Energy
9. Immigration
10. Transportation
12. Justice and crime
13. Social policy
14. Housing
15. Commerce and industrial policy
16. Defense
17. Science and technology
18. Foreign trade
19. International affairs
20. Government and public administration
21. Public lands and water management
23. Culture

Comparative Agendas Project

https://www.comparativeagendas.net

Fine-tuning a model on LLM output

Teacher (GPT-4o) – student (XLM-R) setup



Measuring human vs. LLM performance (news topics)

Kuzman et al. (2025)

Comparing human performance with LLM performance via the “triangle trick”

Comparable agreement between two humans and human and machine

Conclusion - machine performance at least on human level

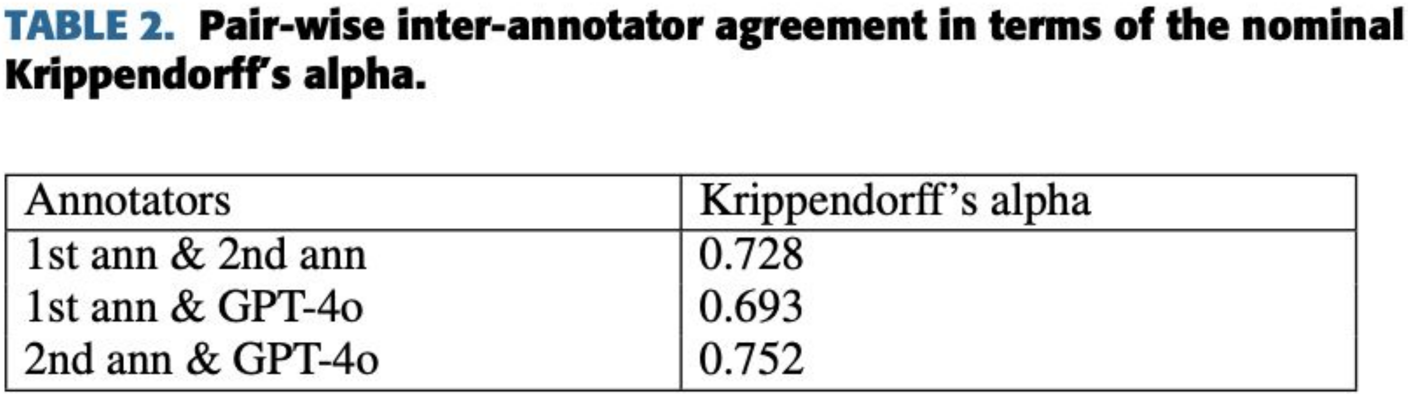


TABLE 2. Pair-wise inter-annotator agreement in terms of the nominal Krippendorff's alpha.

Annotators	Krippendorff’s alpha
1st ann & 2nd ann	0.728
1st ann & GPT-4o	0.693
2nd ann & GPT-4o	0.752

Measuring XLM-R vs. LLM performance (news topics)

XLM-R fine-tuned on LLM output vs. LLM itself

With enough data annotated by LLMs, smaller local models get to the level of performance of much much larger (and non-local) LLMs

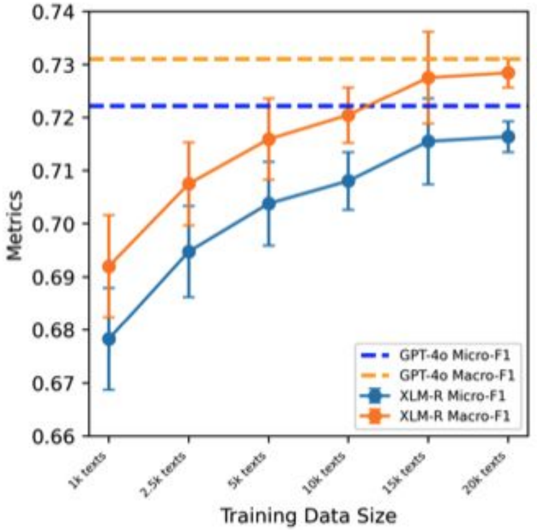


FIGURE 4. Performance in micro-F1 and macro-F1 scores of the XLM-RoBERTa (XLM-R) model fine-tuned on various sizes of training data, compared to the zero-shot GPT-4o performance as the upper limit. The scores are averaged across five iterations of fine-tuning and evaluation, each using different random sample of a specified size, drawn from the training dataset.

Why not use the LLM itself?

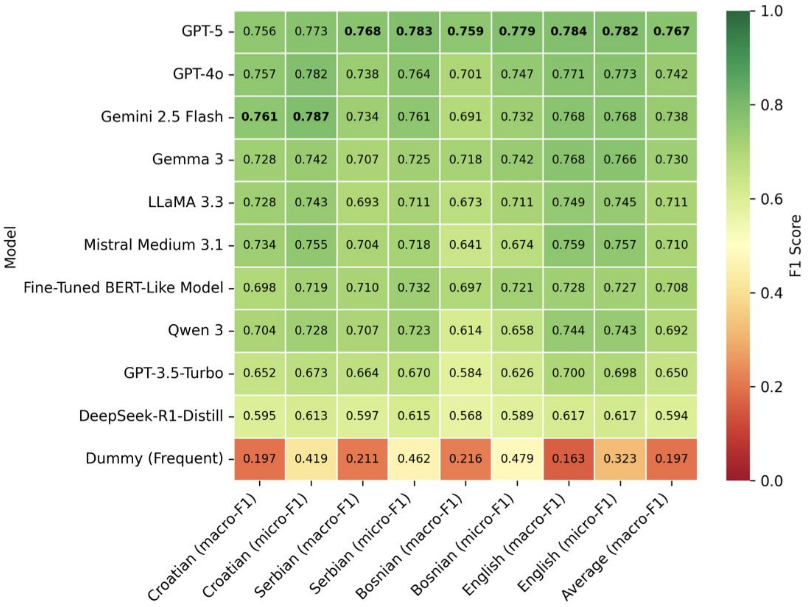
1. Costs
2. Speed
3. Reproducibility / Consistency / Availability

Our current position is that smaller local models are still the way to go in enriching data for downstream research

Will LLMs outperform the teacher-student setup?

Kuzman Pungeršek et al. (2025)

On sentiment, a series of open and closed models in a zero-shot setup are better than our model already



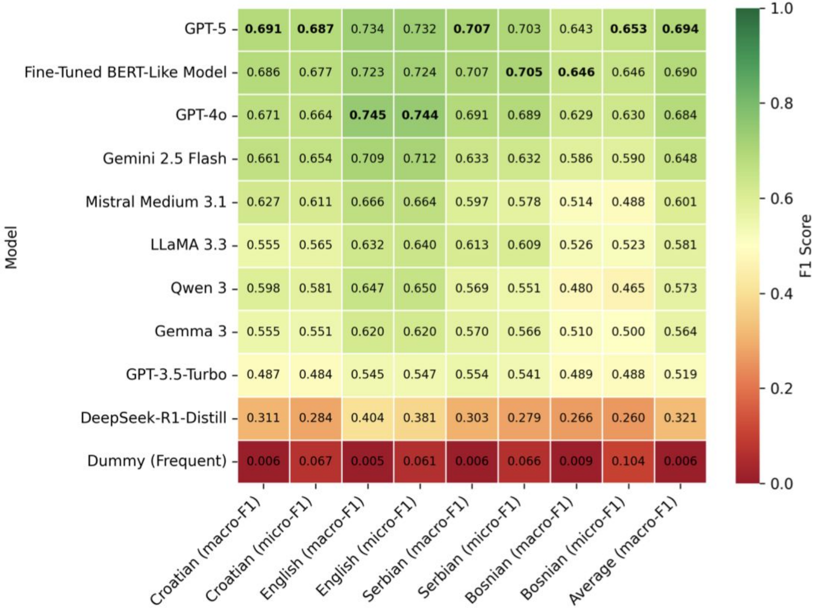
Model	Croatian (macro-F1)	Croatian (micro-F1)	Serbian (macro-F1)	Serbian (micro-F1)	Bosnian (macro-F1)	Bosnian (micro-F1)	English (macro-F1)	English (micro-F1)	Average (macro-F1)
GPT-5	0.756	0.773	**0.768**	**0.783**	**0.759**	**0.779**	**0.784**	**0.782**	**0.767**
GPT-4o	0.757	0.782	0.738	0.764	0.701	0.747	0.771	0.773	0.742
Gemini 2.5 Flash	**0.761**	**0.787**	0.734	0.761	0.691	0.732	0.768	0.768	0.738
Gemma 3	0.728	0.742	0.707	0.725	0.718	0.742	0.768	0.766	0.730
LLaMA 3.3	0.728	0.743	0.693	0.711	0.673	0.711	0.749	0.745	0.711
Mistral Medium 3.1	0.734	0.755	0.704	0.718	0.641	0.674	0.759	0.757	0.710
Fine-Tuned BERT-Like Model	0.698	0.719	0.710	0.732	0.697	0.721	0.728	0.727	0.708
Qwen 3	0.704	0.728	0.707	0.723	0.614	0.658	0.744	0.743	0.692
GPT-3.5-Turbo	0.652	0.673	0.664	0.670	0.584	0.626	0.700	0.698	0.650
DeepSeek-R1-Distill	0.595	0.613	0.597	0.615	0.568	0.589	0.617	0.617	0.594
Dummy (Frequent)	0.197	0.419	0.211	0.462	0.216	0.479	0.163	0.323	0.197

(a) Sentiment classification.

Will LLMs outperform the teacher-student setup?

Kuzman Pungeršek et al. (2025)

Larger models are also coming for the more complex parliamentary topic classification task



Model	Croatian (macro-F1)	Croatian (micro-F1)	English (macro-F1)	English (micro-F1)	Serbian (macro-F1)	Serbian (micro-F1)	Bosnian (macro-F1)	Bosnian (micro-F1)	Average (macro-F1)
GPT-5	**0.691**	**0.687**	0.734	0.732	**0.707**	0.703	0.643	**0.653**	**0.694**
Fine-Tuned BERT-Like Model	0.686	0.677	0.723	0.724	0.707	**0.705**	**0.646**	0.646	0.690
GPT-4o	0.671	0.664	**0.745**	**0.744**	0.691	0.689	0.629	0.630	0.684
Gemini 2.5 Flash	0.661	0.654	0.709	0.712	0.633	0.632	0.586	0.590	0.648
Mistral Medium 3.1	0.627	0.611	0.666	0.664	0.597	0.578	0.514	0.488	0.601
LLaMA 3.3	0.555	0.565	0.632	0.640	0.613	0.609	0.526	0.523	0.581
Qwen 3	0.598	0.581	0.647	0.650	0.569	0.551	0.480	0.465	0.573
Gemma 3	0.555	0.551	0.620	0.620	0.570	0.566	0.510	0.500	0.564
GPT-3.5-Turbo	0.487	0.484	0.545	0.547	0.554	0.541	0.489	0.488	0.519
DeepSeek-R1-Distill	0.311	0.284	0.404	0.381	0.303	0.279	0.266	0.260	0.321
Dummy (Frequent)	0.006	0.067	0.005	0.061	0.006	0.066	0.009	0.104	0.006

(d) Parliamentary topic classification.

What we got from all that NLP

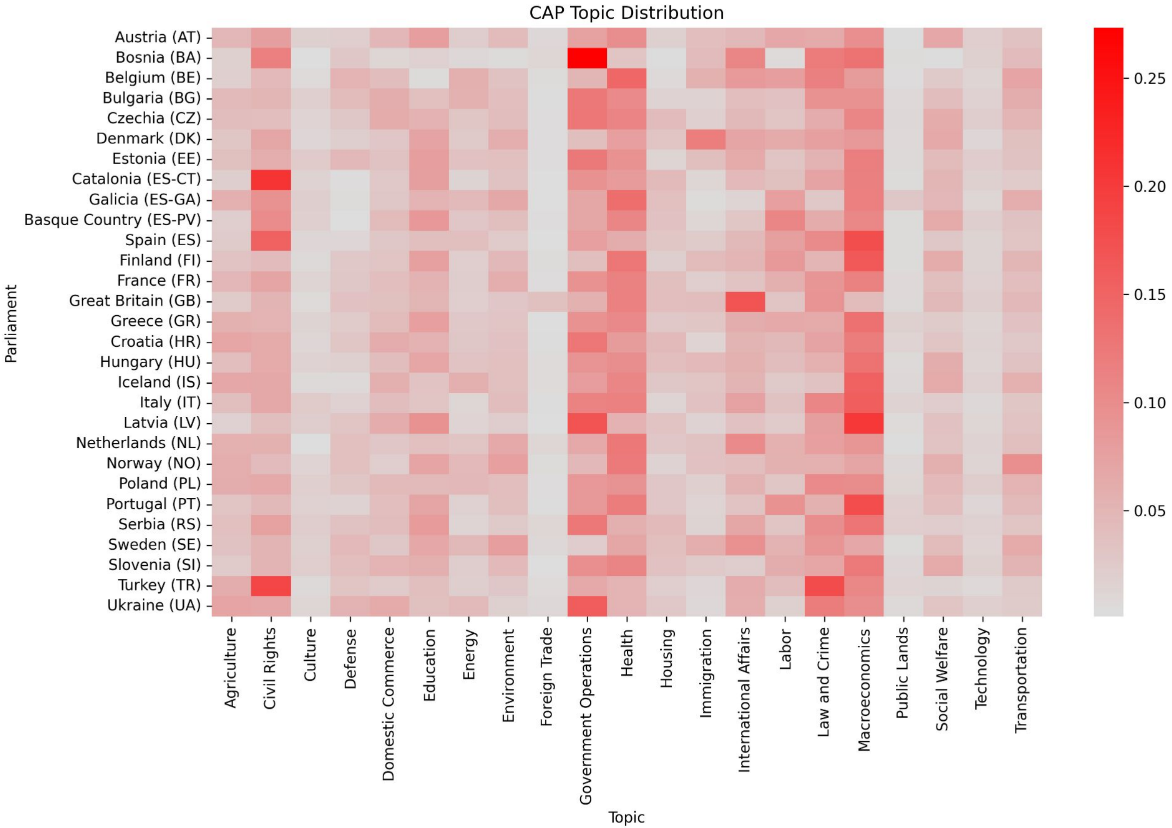
Dataset at the CROSSDA repository

CROSSDA - Croatian node of CESSDA
(Consortium of Social Science Data Archives)

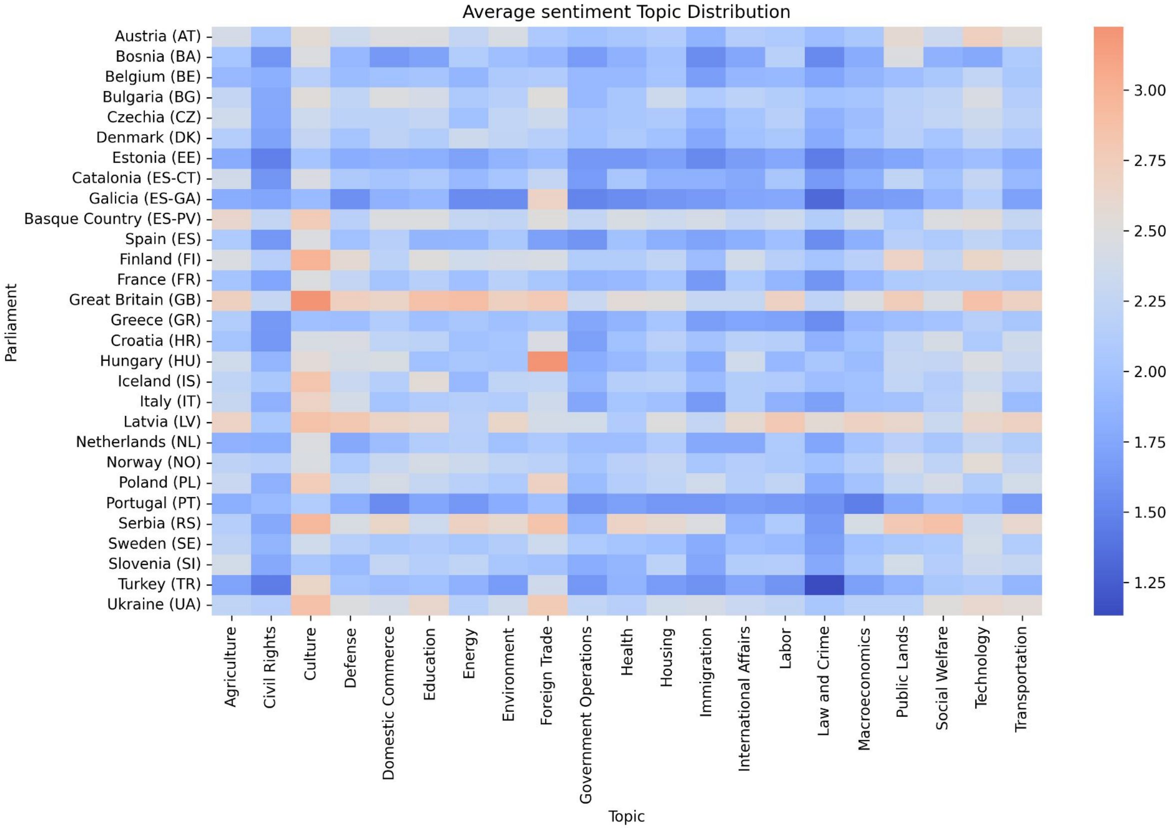
https://doi.org/10.23669/1ZTELP

We are working on an API to simplify data access (“give me all the speeches of female MPs talking about defense in a 1. positive and 2. negative way”)

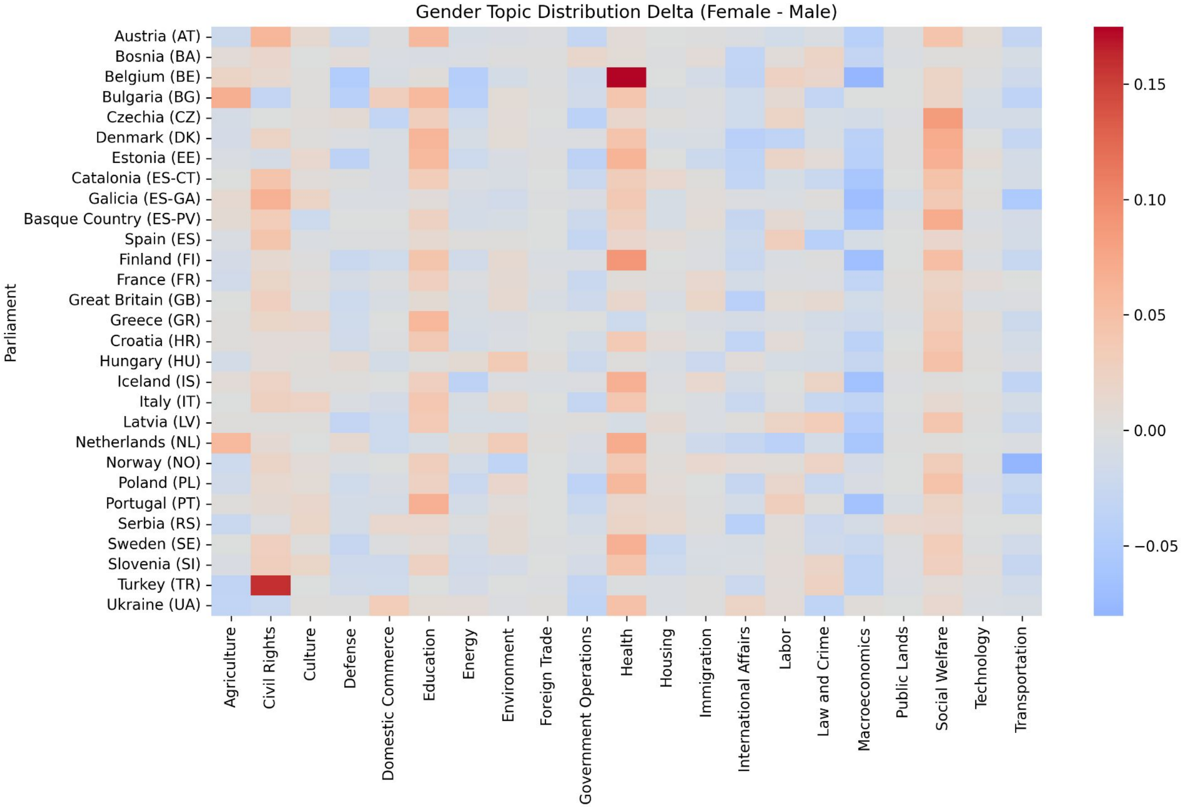
First ParlaCAP insights – topic



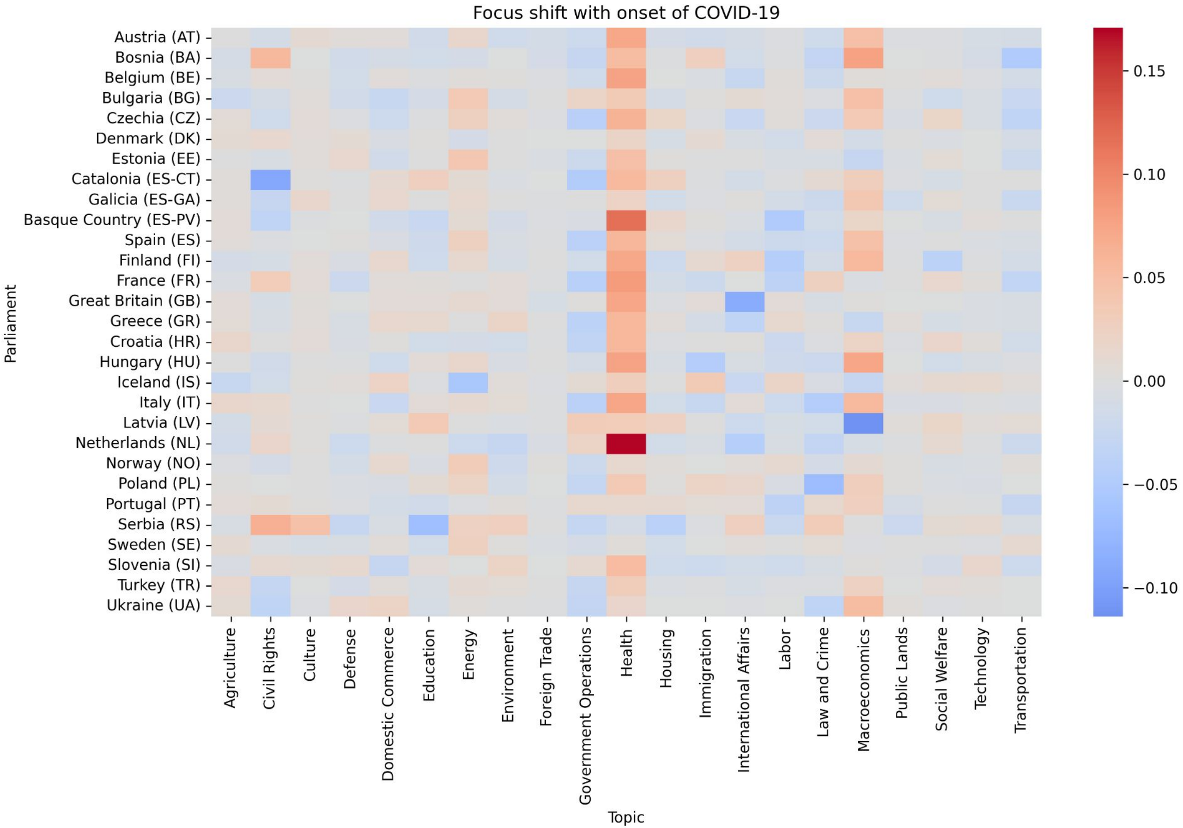
First ParlaCAP insights – sentiment



First ParlaCAP insights – gender



First ParlaCAP insights – COVID



ParlaCAP tutorials

https://clarinsi.github.io/parlacap/

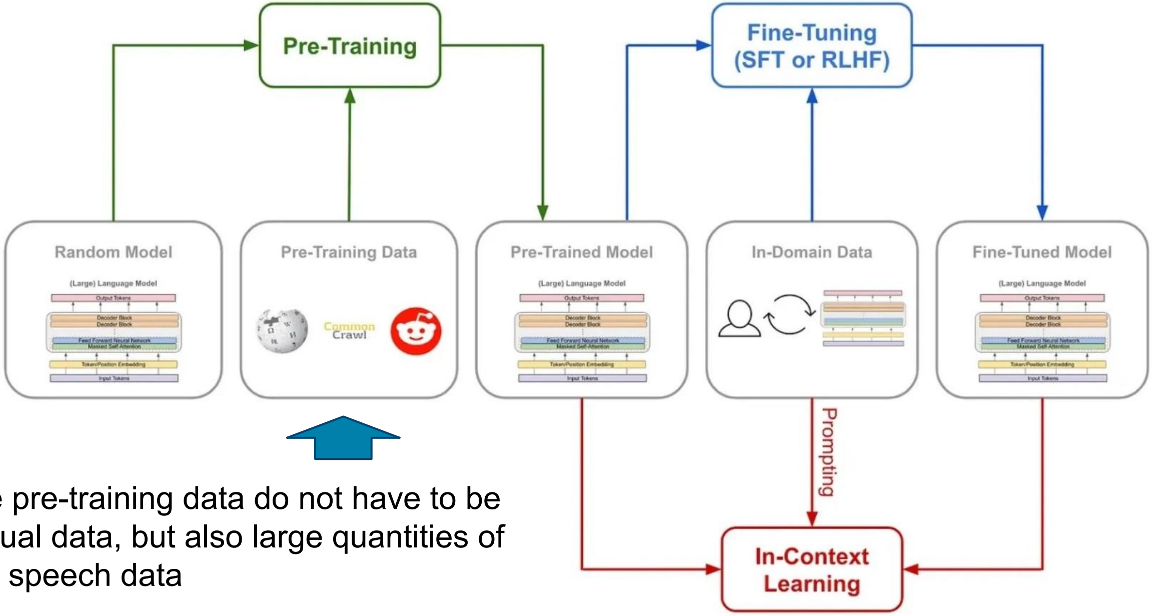
https://github.com/clarinsi/ParlaCAP-Analysis-Tutorials/

Tutorial in R being finalised

If time permits – speech data

Pre-trained language models on speech data

With small amount of labeled data, the model can learn how to identify various phenomena in speech (transcription, emotion, background sounds)



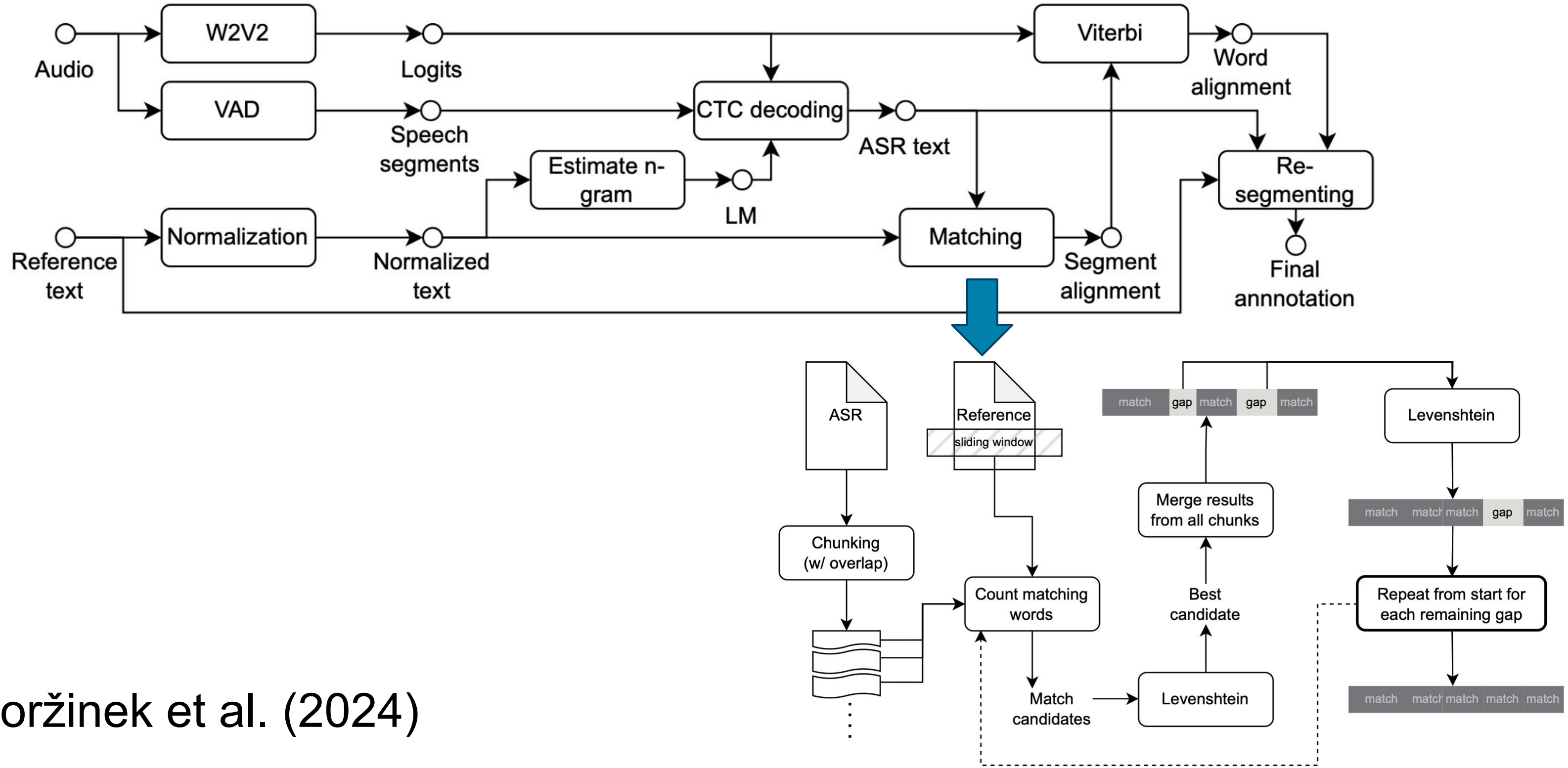
The pre-training data do not have to be textual data, but also large quantities of raw speech data

https://medium.com/@bijit211987/the-evolution-of-language-models-pre-training-fine-tuning-and-in-context-learning-b63d4c161e49

ParlaSpeech

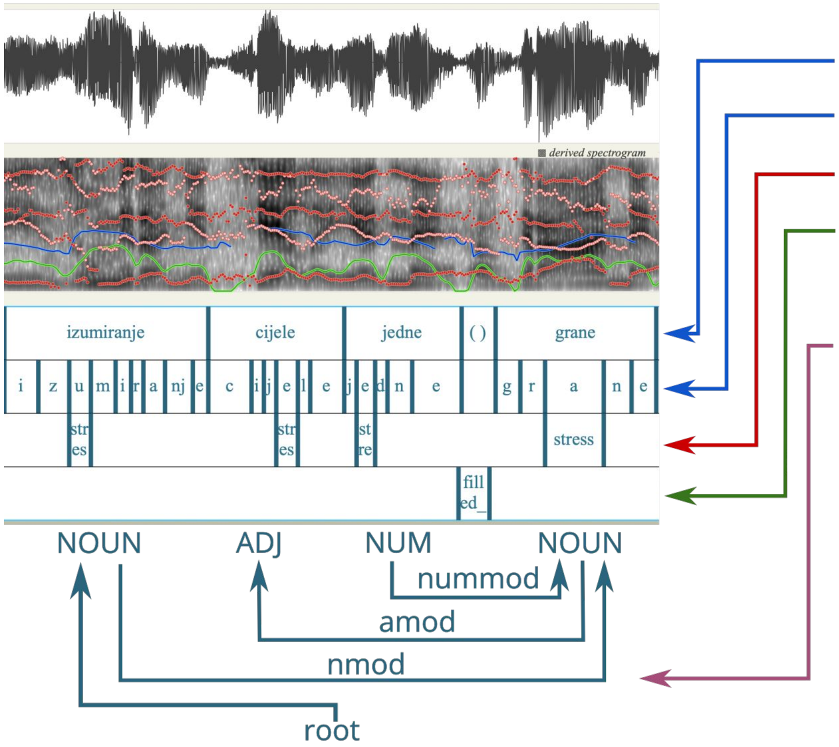
- Task inside ParlaMint, growing into a separate project
- Aligning public domain! speech data with transcripts of the parliament
- Currently aligned are Croatian, Serbian, Polish, Czech with amount of data between 1000 and 3000 hours per language, 6k all together
- Easy? No.
  - Recordings are published independently of texts with spotty metadata
  - Not all recordings are released, not everything is transcribed
  - Order of transcripts and recordings is not identical
- https://clarinsi.github.io/parlaspeech/

ParlaSpeech alignment procedure



Koržinek et al. (2024)

ParlaSpeech v3.0



Word alignment

Grapheme alignment

Primary (word) stress

Filled pauses

Sentiment: "Negative"

Linguistic annotations

Rich metadata:

Name: "Ružica Vukovac"
Gender: "Female"
Birth year: "1975"
Party: "MOST"
Status: "Opposition"
...

Interaction of acoustic variables and sentiment

Porupski and Ljubešić (2026, to be submitted)

Higher pitch, intensity, speech rate, are all predictors of negative parliamentary speech.

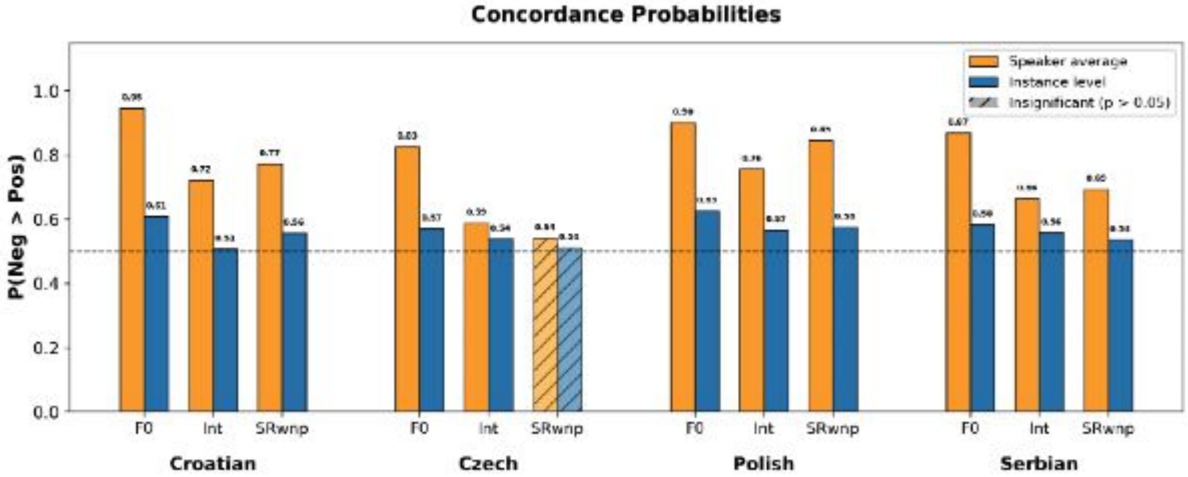


Figure 1: Concordance probability P(Neg > Pos) for pitch (F0), intensity (Int), and speech rate words/s with no pauses (SRwnp) across four languages. Speaker-averaged and utterance-level results are shown. Statistically non-significant results are indicated with hashes.

To wrap up…

- New research opportunities from advances in NLP
- Significantly larger and more diverse data at a lower cost
- Models work on multiple modalities, across languages / domains
- Limitations!, so evaluation / validation is highly advisable
- ParlaMint a rich unexplored dataset, we have just scratched the surface
- Currently we are revisiting old questions
- Collaboration with domain experts on new questions and theories
- The data and tutorials are out there, please, help yourselves!

LLMs

https://www.clarin.eu/parlamint

https://huggingface.co/classla

https://www.clarin.si/repository/xmlui/

https://nljubesi.github.io