OSCARS
Open Science Clusters' Action for Research & Society

Funded Project

ParlaCAP - Comparing agenda settings across parliaments via the ParlaMint dataset

Presenter: Petya Osenova, IICT-BAS, ORCID 0000-0002-4484-5027

Implemented by

Jožef Stefan Institute | IICT | PAN Polska Akademia Nauk | fpzg

CLARIN.SI

Funded by the European Union

What problem(s) did you plan to solve?

- **Project start:** 1 January 2025, duration – 24 months
- **Background:** ParlaMint dataset with transcripts from 28 European parliaments
- **Opportunity**: great dataset for political and other social science research
- **Requirement**: cannot work with textual data (billion+ words),
  a need for tabular data to be exploited via statistical analysis and inference

What have you done to solve the problem?

- **Challenge 1:** *Classifying speeches by sentiment and topic in 28 languages*

Solution: Fine-tuning scalable multilingual transformers on in-domain data pre-annotated with frontier LLMs (+human-coded test data)

- **Challenge 2:** *User-friendly interface for SSH researchers*

Solution: Developing an API to deliver data subsets from 8 million speeches

- **Challenge 3:** *Appropriate materials for data and model uptake*

Solution: Series of tutorials, books, presentations, posters

What are the key results achieved to date and how have you made them available to the broader community?

- **Data releases:**
  - ParlaCAP 1.0 dataset (FAIR CESSDA repo, API)
  - ParlaMint 5.0 corpus collection (FAIR CLARIN repo, CLARIN concordancer)
- **Datasets for LLM training and testing** (freely available)
  - ParlaCAP fine-tuning data (FAIR CLARIN repo)
  - ParlaCAP test data (only on request to counter LLM contamination)
- **Multilingual transformer models**
  - ParlaCAP model for topic classification (HuggingFace, FAIR CLARIN repo)
  - ParlaSent model for sentiment (HuggingFace, FAIR CLARIN repo)
- **Tutorials in Python and R, presentations, posters papers, book** (project website, Zenodo)

How will make your results sustainable over time - How will the scientific community/-ies further exploit them?

- Data and models on FAIR repositories
- API uptime ensured for 5 years, then integration into CLARIN ERIC
- Expect follow-up open-science projects on the developed data and models
- Models and data already exploited in research by third parties

Who has been doing it?

- **Project Coordinator:** Nikola Ljubešić (Jožef Stefan Institute, Slovenia)
- **Project members:** Tomaž Erjavec, Taja Kuzman, Peter Rupnik, Katja Meden, Jure Skubic, Anna Kryvenko (Slovenia), Daniela Širinić (Croatia), Petya Osenova (Bulgaria), Maciej Ogrodniczuk, Łukasz Kobyliński (Poland).

- **External collaborators:** Michal Mochtak (Radboud University), Matyáš Kopp (Institute of Formal and Applied Linguistics)