Eastern European **model** and **dataset** collection.
Language **similarity** and **distillation** effects transferability across languages.

EENLP: Cross-lingual Eastern European NLP Index

CONTRIBUTIONS

Dataset and model index

- We present a broad index of existing Eastern European language resources published as a GitHub repository.

Benchmark datasets

- We provide hand-crafted cross-lingual datasets for five different semantic tasks, compiled by processing data from various sources into the same format.

Baselines

- We perform several experiments with the existing multilingual models on our datasets to define performance baselines.

EENLP Index on GitHub

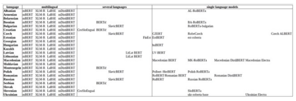
- https://github.com/altsoph/EENLP
- 90+ datasets, 45+ models

CONCLUSIONS

- Multilingual pre-training is crucial for cross-lingual transfer learning.
- Distillation decreases the ability for cross-lingual transfer learning.
- Similarity of languages matters for the cross-lingual transfer learning quality.
- XLM-RoBERTa generally dominates over the rest of the tested models.

Alexey Tikhonov, Alex Malkhasov, Andrey Manoshin, George Dima, Réka Cserháti, Md.Sadek Hossain Asif, **Matt Sárdi**

- Languages
  - Albanian, Armenian, Belarusian, Bosnian, Bulgarian, Croatian, Czech, Estonian, Hungarian, Latvian, Lithuanian, Macedonian, Moldavian, Polish, Romanian, Russian, Serbian, Slovak, Slovenian, and Ukrainian.
- Links
  - https://github.com/altsoph/EENLP/blob/main/docs/datasets.md
  - https://github.com/altsoph/EENLP/blob/main/docs/models.md
  - https://github.com/altsoph/EENLP/tree/main/build_benchmarks
- Tasks
  - news categorisation
  - paraphrase detection
  - news sentiment detection
  - tweet sentiment detection
  - NLI (Natural Language Inference)
- Models
  - BERT
  - mBERT
  - mDistilBERT
  - XLMR



4.1. News categorisation task

4.2. Paraphrase detection task

4.3. News sentiment detection task

4.4. Twitter sentiment detection task

4.5. NLI task

7. Bibliographical References

Take a picture to **download** the **full paper**