EMBEDDIA project

Cross-Lingual Embeddings for Less-Represented Languages in European News Media

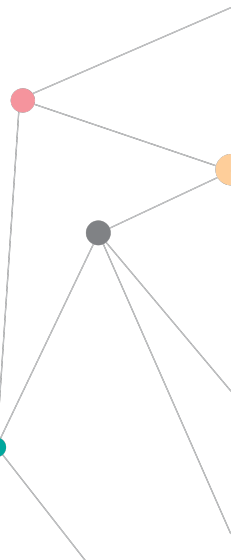
Senja Pollak (Jozef Stefan Institute), project coordinator

November 8, 2021

This project has received funding from European Union's Horizon 2020 research and innovation programme under grant agreement No 825153

Project overview

- **Cross-lingual embeddings** and **deep neural networks** enabling less-represented languages to benefit from resources and tools of well-resourced languages (English)
- Focus on morphologically-rich, **less-represented languages** in European news media: Estonian, Latvian, Lithuanian, Slovenian, Croatian and Finnish
- Applications for the **news media industry**: cross-lingual solutions for **news** and **user-generated content** analysis and **news generation**



€3M
from H2020
EU funding

10
partners

6
countries

3 years
duration
2019-2021

NEWS MEDIA APPLICATIONS

Comment analysis
hate speech filtering, trolling detection, opinion mining

News analysis
topic analysis, news linking, viewpoint & sentiment detection, summarisation, visualisation

News generation
text generation from structured data, personalised dynamic content generation, creative headlines

Consortium

- **Interdisciplinary**: media studies, natural language processing and machine learning
- **Intersectoral:**

Academic partners:

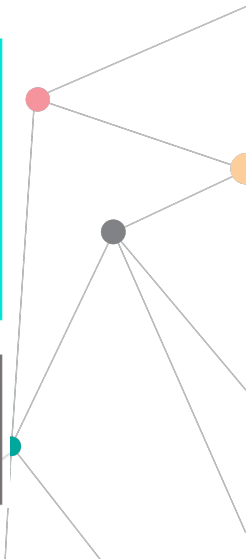
Jožef Stefan Institute (SI)
University of Ljubljana (SI)
Queen Mary Univ. of London (UK)
University of Helsinki (FI)
University of La Rochelle (FR)
University of Edinburgh (UK)

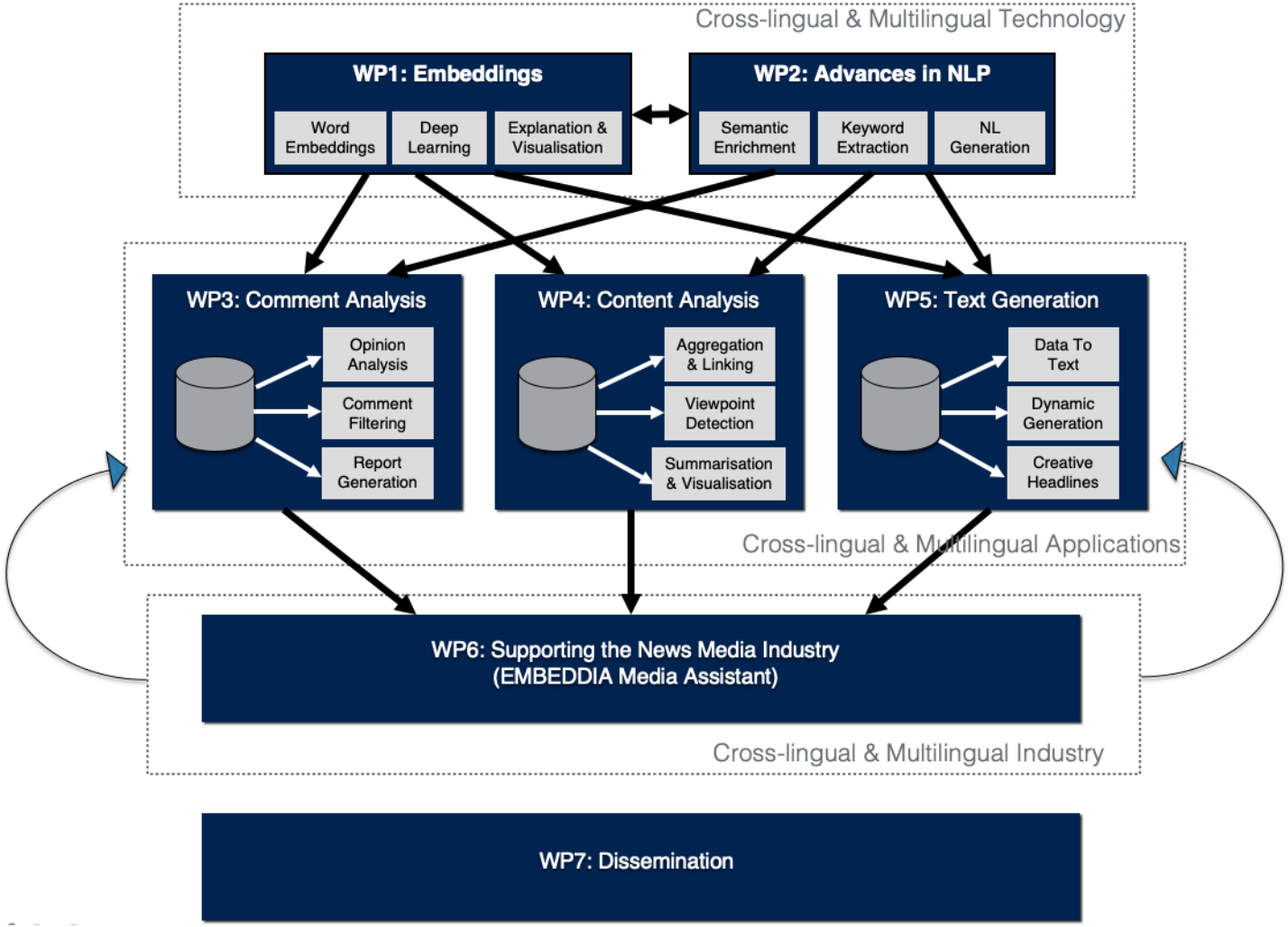
News media industry partners:

Trikoder (Styria media group) (CRO)
Ekspress Meedia (EE)
Finnish News Agency STT (FI)

Text mining industry partner SME:

TEXTA OÜ (EE)




Cross-lingual & Multilingual Technology
WP1: Embeddings
Word Embeddings
Deep Learning
Explanation & Visualisation
WP2: Advances in NLP
Semantic Enrichment
Keyword Extraction
NL Generation
WP3: Comment Analysis
Opinion Analysis
Comment Filtering
Report Generation
WP4: Content Analysis
Aggregation & Linking
Viewpoint Detection
Summarisation & Visualisation
WP5: Text Generation
Data To Text
Dynamic Generation
Creative Headlines
Cross-lingual & Multilingual Applications
WP6: Supporting the News Media Industry
(EMBEDDIA Media Assistant)
Cross-lingual & Multilingual Industry
WP7: Dissemination

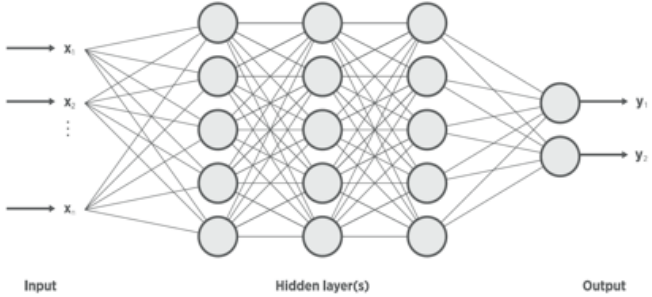
Deep neural networks

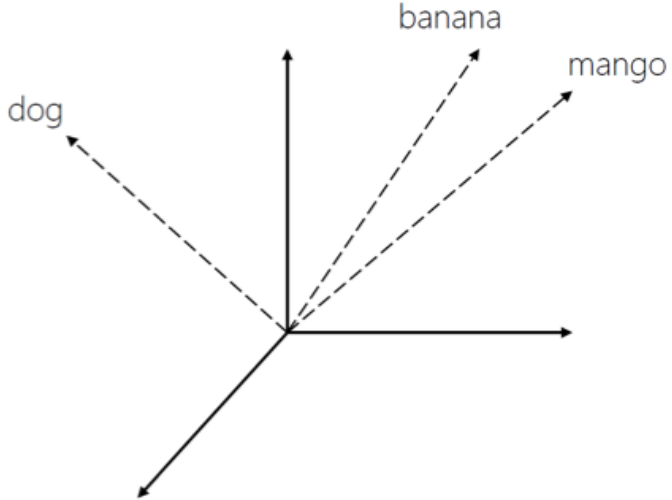
Currently the **most successful approach** to most natural language understanding tasks: machine translation, summarization, questions & answers, text generation, speech recognition and synthesis

Build knowledge representation automatically

Require **text as numeric input**
numeric input shall preserve
similarity and relations between words

Solution: **text embeddings**



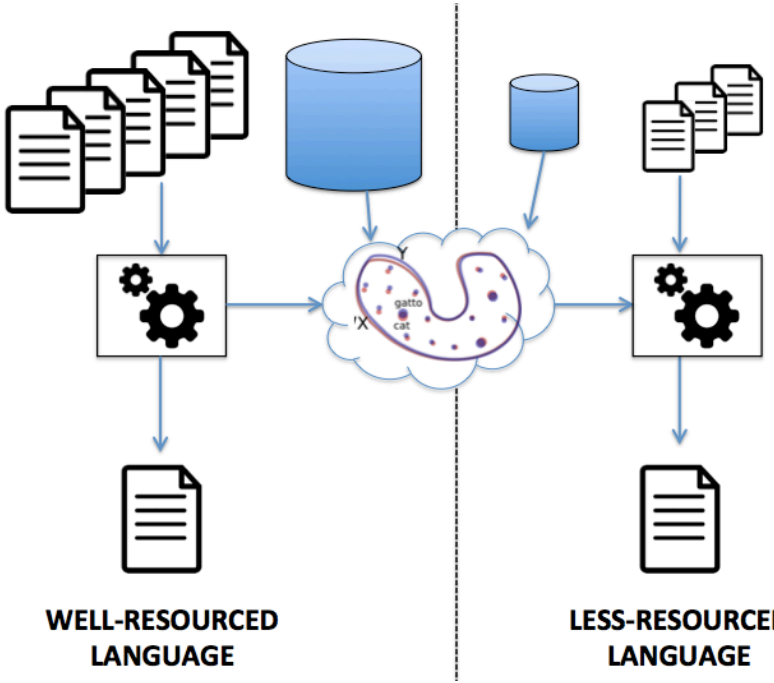


Contextual embeddings

- Simple, **static embeddings**, like word2vec, produce the same vector for a word like “cold” irrespective of its meaning and context
- Recent embeddings take the **context** into account (ELMo, transformer-based BERT and var., XLNet, XLM, T5…)
- Multilingual embeddings (mBERT, XLM-R,..)

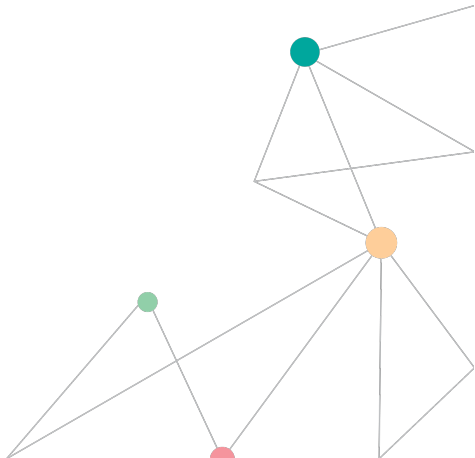
Cross-Lingual Model Transfer based on Embeddings

- Transfer of tools trained on mono-lingual resources



Selected results

- Language technology advances
- News media applications



Language technologies advances

Release of new corpora (CLARIN)

- **News corpora:**
  - 24sata news archive in *Croatian*
  - Ekspress Meedia News Archive in *Estonian* and *Russian*
  - Latvian Delfi Article Archive in *Latvian* and *Russian*
  - STT News Archive in *Finnish*
  - **keyword extraction data splits** (EE, CRO, LAT, RU)
  - **news sentiment dataset** (CRO)
- **Comments corpora:**
  - 24sata comments archive in *Croatian*
  - Ekspress Meedia comments archive in *Estonian* and *Russian*
  - Latvian Delfi comments archive in *Latvian* and *Russian*
  - **offensive/moderated content datasets**

Pollak et al. (2021): EMBEDDIA Tools, Datasets and Challenges: Resources and Hackathon Contributions. EACL EMBEDDIA Hackashop

Language technologies advances

Release of **new models** for less-resourced EU languages

- *Monolingual*: Slovene SloBERTa, Estonian Est-RoBERTa, ELMo models
- *Trilingual*: CroSloEngual BERT, FinEst BERT, LitLat BERT

Novel **evaluation tasks** and **datasets**:

- SemEval 2020 CoSimLex (Armendariz et al., 2020)

Sentence 1: The **population** of India is actually bigger than most **people** expect.

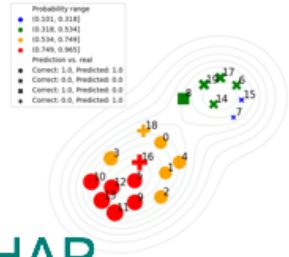
Sentence 2: The **population** of bison became a lot smaller when **people** settled in the valley.

- Cross-lingual analogies (Ulčar et al. 2020)
- Translation of SuperGLUE to Slovene
- SlavNER (Piskorski et al., BSNLP 2021) Slovene

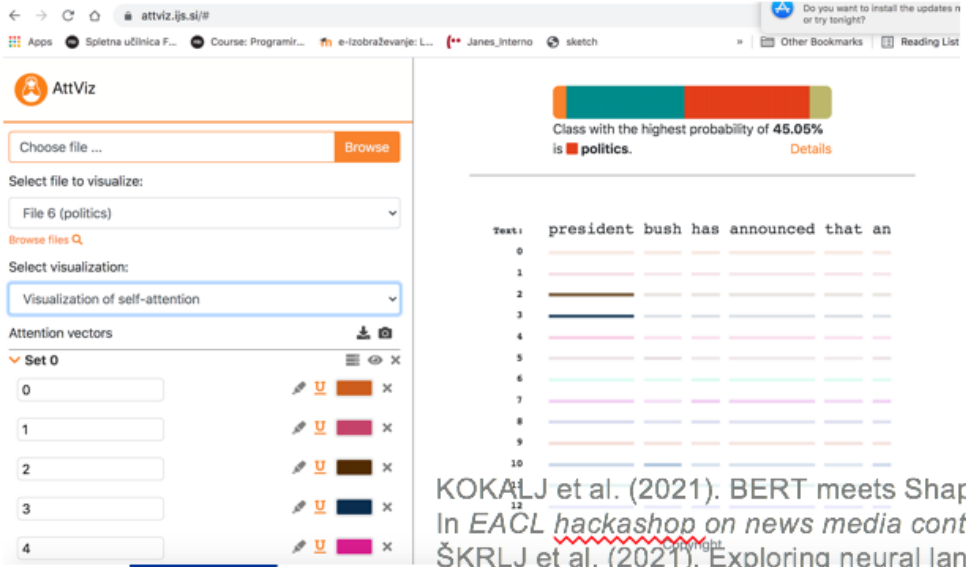
Systematic evaluation of ELMo and BERT on less-resourced EU languages

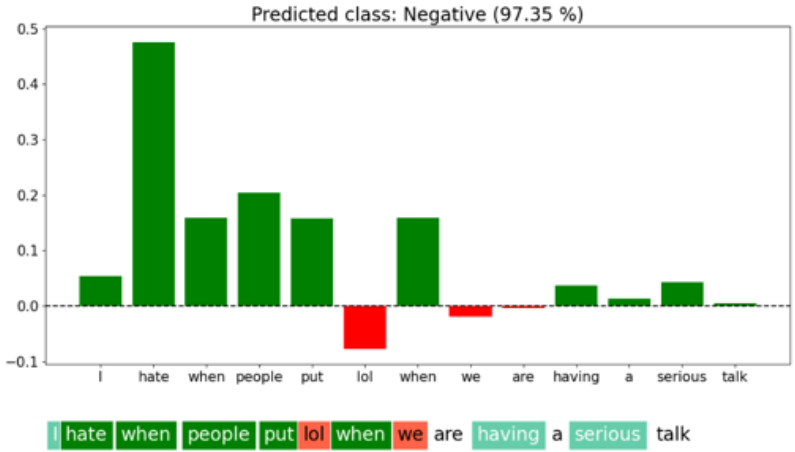
Ulčar, M., et al. (2020). Multilingual culture-independent word analogy datasets. LREC 2020.
Armendariz et al. (2020). SemEval-2020 task 3 : graded word similarity in context.
Ulčar et al. 2021, https://arxiv.org/abs/2107.10614
Piskorski et al. 2021 https://aclanthology.org/2021.bsnlp-1.15/

Advances in deep learning and interpretability



- New **reliability estimates** for deep neural networks on text
- **Explanations** and **visualizations** for BERT models with SHAP
- Attention visualisation toolkit (self-attention) https://attviz.ijs.si/





KOKALJ et al. (2021). BERT meets Shapley : extending SHAP explanations to transformer-based classifiers. In *EACL hackashop on news media content analysis and automated report generation.*

ŠKRLJ et al. (2021). Exploring neural language models via analysis of local and global self-attention spaces. EACL hackashop

MIOK et al. (2021). Bayesian BERT for trustful hate speech detection. In *ICML 2020 : Workshop on Uncertainty & Robustness in Deep Learning.*

Advances in NLP technologies

Named entity recognition and linking

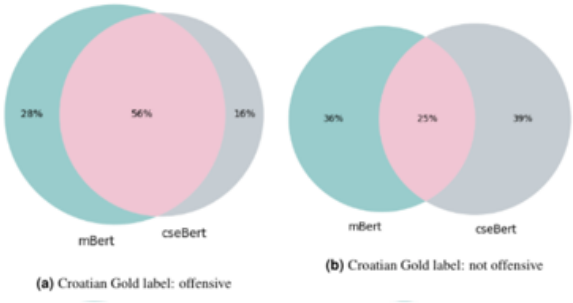
Creation of different BERT-based NER systems (Boros et al.,2020; Cabrera-Diego et al., 2021a), Improving NEL using a multilingual approach (Linhares Ponteset al., 2020)

Keyword extraction methods

- Unsupervised (RaKun by Škrlj et al., Proc. of SLSP 2019),
- Supervised TNT-KID (Martinc et al.), BERT + BiLSTM-CRF (Koloski et al., 2021)

User comments moderation

- **Monolingual comment moderation** models trained on media partners dataset (Shekhar et al., 2021, JLCL)
- **Cross-lingual** experiments **mBERT and cseBERT** in Pelicon et al. (2021, PeerJ comp. sci.)
  - models allow for cross-lingual transfer
  - various training regimes (zero-shot, few-shot intermediate training, different language combinations, various data sizes in source and target languages)
- Topic-aware models (Zosa et al., 2021 RANLP)



(a) Croatian Gold label: offensive

(b) Croatian Gold label: not offensive

Language	mBERT		cseBERT	
	TGT	LOO→TGT@10%	TGT	LOO→TGT@10%
Croatian	61.30	†**66.82**	61.04	†**70.91**
Slovenian	64.68	†**68.22**	69.52	†72.63
English	**72.40**	72.17	63.51	†**77.11**
German	**59.97**	53.20	**43.36**	39.64
Arabic	63.82	†**76.07**	48.84	**57.42**

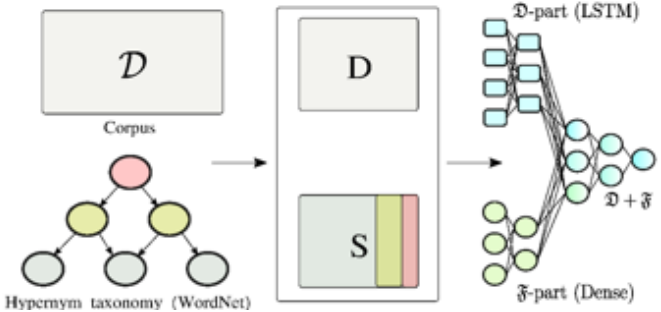
Cross-lingual sentiment analysis with intermediate training

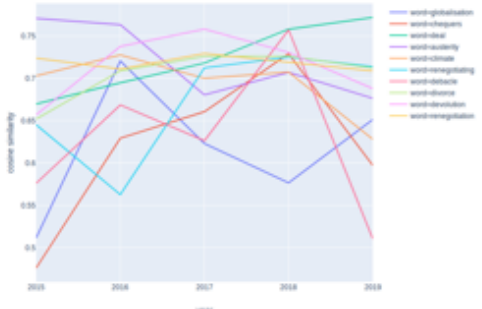
- Zero-Shot Learning for Cross-Lingual News Sentiment Classification tested on Slo->Cro (Pelicon et al., 2021b)
- An **intermediate training step** to enrich the BERT language model representations with sentiment information.
- This approach improved the model performance on the Slovenian news as well as on the Croatian news in a zero-shot setting.

- **Cross-lingual twitter sentiment classification** (Robnik Šikonja et al., Slovenscina2.0)

Various other tasks for news media

- **News linking** methods
- **Diachronic** news **analysis** allowing cross-lingual analysis of **viewpoints**
- Novel **text summarization and visualization** methods
- Headline generation methods





Conclusions

- New EMBEDDIA technologies
  - Progress in **cross-lingual approaches**, for **less-resourced** languages and for **news media** domain
  - basic NLP research and applied research in news media
  - addressing **real** news industry **needs**
  - Offering **real** cross-lingual **solutions** for **less-resourced** languages

www.embeddia.eu

@embeddiaproject

Selected publications

- Armendariz et al. (2020). SemEval-2020 task 3 : graded word similarity in context. Proceedings of the 14th International Workshop on Semantic Evaluation, pages 36–49. https://aclanthology.org/2020.semeval-1.3.pdf
- Boros, E., et al. (2021b). Event detection with entity markers. Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021,doi:10.1007/978-3-030-72240-1_20
- Cabrera-Diego, L. A., Moreno, J. G., & Doucet, A. (2021b). Using a Frustratingly Easy Domainand Tagset Adaptation for Creating Slavic Named Entity Recognition Systems. In Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing.
- Kokalj et al. (2021). BERT meets Shapley: Extending SHAP Explanations toTransformer-based Classifiers. Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation. https://aclanthology.org/2021.hackashop-1.3.pdf
- Linhares Pontes, E., Doucet, A., & Moreno, J. G. (2020). Linking named entities across languages using multilingual word embeddings. InJointed Conference on Digital Libraries (JCDL) 2020
- Martinc et al. 2021. TNT-KID : transformer-based neural tagger for keyword identification. Natural language engineering, ISSN 1469-8110, 2021, 40 str., doi: 10.1017/S1351324921000127.

Selected publications

- Miok et al. 2020. Prediction Uncertainty Estimation for Hate Speech Classification. ICML 2020 : Workshop on Uncertainty & Robustness in Deep Learning, July 17, 2020. http://www.gatsby.ucl.ac.uk/~balaji/udl2020/accepted-papers/UDL2020-paper-058.pdf.
- Pelicon et al. 2021. Investigating cross-lingual training for offensive language detection. PeerJ computer science, doi: 10.7717/peerj-cs.559.
- Pelicon et al. 2021b. Zero-shot learning for cross-lingual news sentiment classification. Applied sciences, vol. 10, no. 17, doi: 10.3390/app10175993.
- Piskorski et al. (2021). Slav-NER: the 3rd Cross-lingual Challenge on Recognition, Normalization, Classification, and Linking of Named Entities across Slavic languages. Proceedings of the 8th BSNLP Workshop on Balto-Slavic Natural Language Processing, pages 122–133 https://aclanthology.org/2021.bsnlp-1.15.pdf
- Pollak et al. (2021): EMBEDDIA Tools, Datasets and Challenges: Resources and Hackathon Contributions. Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation. https://aclanthology.org/2021.hackashop-1.14/
- Robnik Šikonja et al. 2021. Cross-lingual transfer of sentiment classifiers. Slovenščina 2.0. doi: 10.4312/slo2.0.2021.1.1-25. Shekhar et al. 2020. https://jlcl.org/content/2-allissues/1-heft1-2020/jlcl_2020-1_3.pdf

Selected publications

- Škrlj et al. 2019. RaKUn: rank-based keyword extraction via unsupervised learning and meta vertex aggregation. In Proc. of. Statistical language and speech processing : 7th international conference, SLSP 2019 Ljubljana, Slovenia, October 14-16, 2019 : proceedings, https://link.springer.com/chapter/10.1007%2F978-3-030-31372-2_26
- Škrlj et al. 2021. Exploring Neural Language Models via Analysis of Local and Global Self-Attention Spaces. https://aclanthology.org/2021.hackashop-1.11/
- Ulčar, M., et al. (2020). Multilingual culture-independent word analogy datasets. LREC 2020. https://aclanthology.org/2020.lrec-1.501/Ulčar et al. (2021). Evaluation of contextual embeddings on less-resourced languages. https://arxiv.org/abs/2107.10614
- Zosa et al. 2021. Not All Comments are Equal: Insights into Comment Moderation from a Topic-Aware Model. Accepted to RANLP. https://arxiv.org/abs/2109.10033