# EMBEDDIA: AI Technology for the Media Industry

Senja Pollak (Project coordinator)

Jožef Stefan Institute, Slovenia

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

6th ELRC, March 31 2022

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

- **Cross-lingual embeddings** and **deep neural networks**
- **Less-represented EU languages** to benefit from resources and tools of well-resourced languages
- Focus on **morphologically-rich** languages incl. **Estonian, Croatian, Finnish**, Latvian, Slovenian
- **AI applications** for the **news media industry**

**€3M** from H2020 EU funding

**10** partners

**6** countries

**3 years** duration 2019-2022

# NEWS MEDIA APPLICATIONS

**Comment analysis**
hate speech filtering, opinion mining, fake news spreaders detection

**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:**

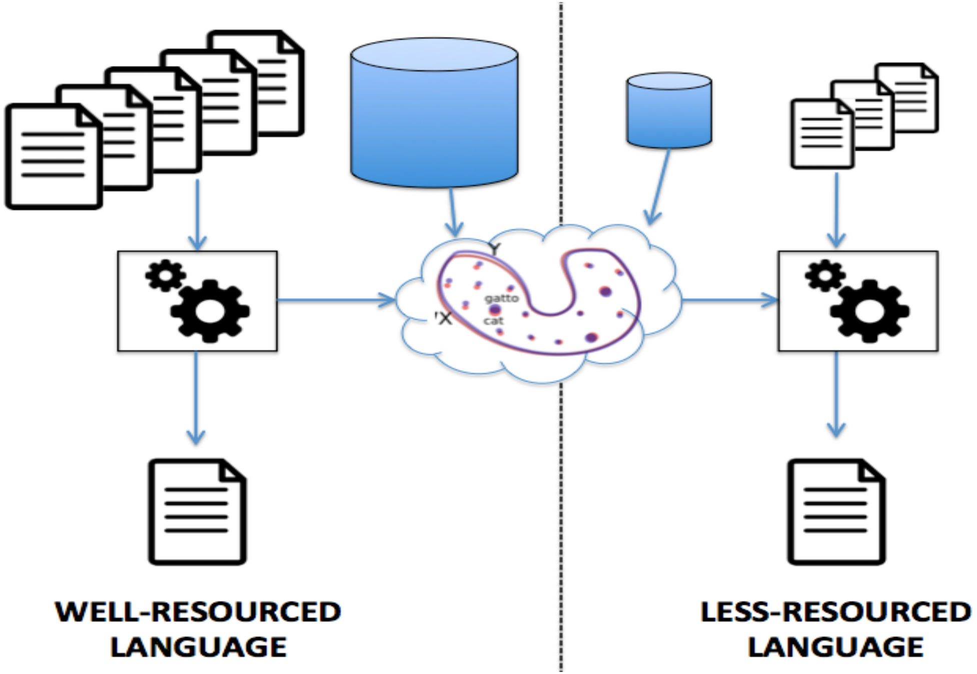
Trikoder (Styria media group) (CRO)
Ekspress Meedia (EE)
Finnish News Agency STT (FI)

**Text mining industry partner SME:**

TEXTA OÜ (EE)

# Technological background

- Embeddings, deep neural networks, transformers



WELL-RESOURCED LANGUAGE

LESS-RESOURCED LANGUAGE

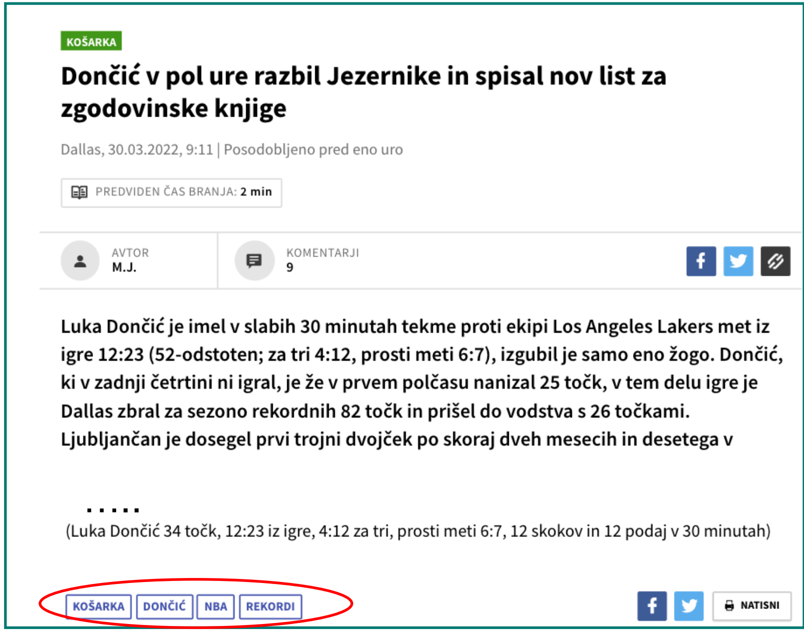
# Selected results

## Language technologies advances and applied tasks

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

# Article tagging with keywords

- Recognise keyword in articles
- Learning from collections with available keywords, adapts to specific tagging regime
- monolingual results: P@10 32%-71%
- cross-lingual methods (no training data): P@10:23-33%



KOŠARKA

**Dončić v pol ure razbil Jezernike in spisal nov list za zgodovinske knjige**

Dallas, 30.03.2022, 9:11 | Posodobljeno pred eno uro

PREDVIDEN ČAS BRANJA: **2 min**

AVTOR
M.J.

KOMENTARJI
9

**Luka Dončić je imel v slabih 30 minutah tekme proti ekipi Los Angeles Lakers met iz igre 12:23 (52-odstoten; za tri 4:12, prosti meti 6:7), izgubil je samo eno žogo. Dončić, ki v zadnji četrtini ni igral, je že v prvem polčasu nanizal 25 točk, v tem delu igre je Dallas zbral za sezono rekordnih 82 točk in prišel do vodstva s 26 točkami. Ljubljančan je dosegel prvi trojni dvojček po skoraj dveh mesecih in desetega v**

.....

(Luka Dončić 34 točk, 12:23 iz igre, 4:12 za tri, prosti meti 6:7, 12 skokov in 12 podaj v 30 minutah)

KOŠARKA DONČIĆ NBA REKORDI

NATISNI

# Sentiment Analysis

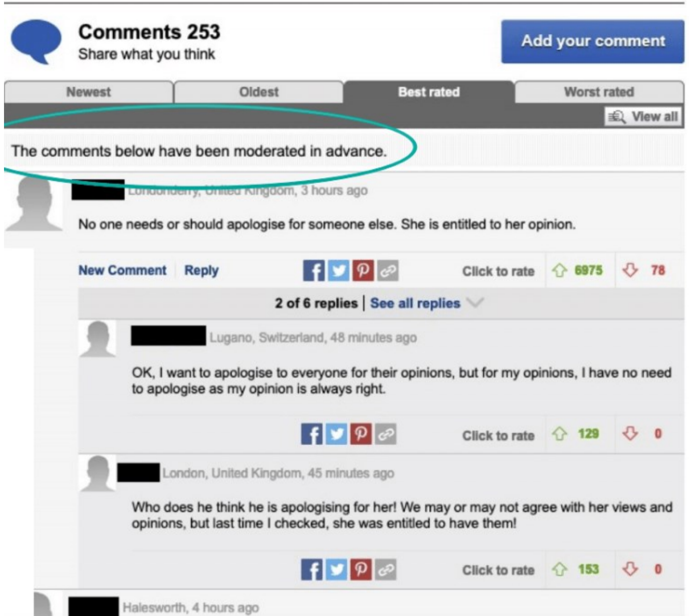
- Classifies news articles in three classes based on their sentiment: **negative, neutral, positive.**
- **Zero-Shot Learning setting:** trained on Slovenian news articles and applies it to languages other than Slovenian without training data.
- Results: **66% F1** on Slovene, where training data is available, and between **55-75% F1** when tested on other languages.

# Other selected news analysis applications

- Topic modelling
- Article retrieval and linking
    - background linking
    - related articles recommendation
    - interesting news identification
- Viewpoints analysis and analysis across time

# User generated content analysis

- **Comment moderation**
  - Monolingual, specific for media partners rules
  - Cross-lingual models for offensive speech filtering
  - Topic-aware models



Comments 253
Share what you think

The comments below have been moderated in advance.

[redacted] Londonderry, United Kingdom, 3 hours ago

No one needs or should apologise for someone else. She is entitled to her opinion.

2 of 6 replies | See all replies

[redacted] Lugano, Switzerland, 48 minutes ago

OK, I want to apologise to everyone for their opinions, but for my opinions, I have no need to apologise as my opinion is always right.

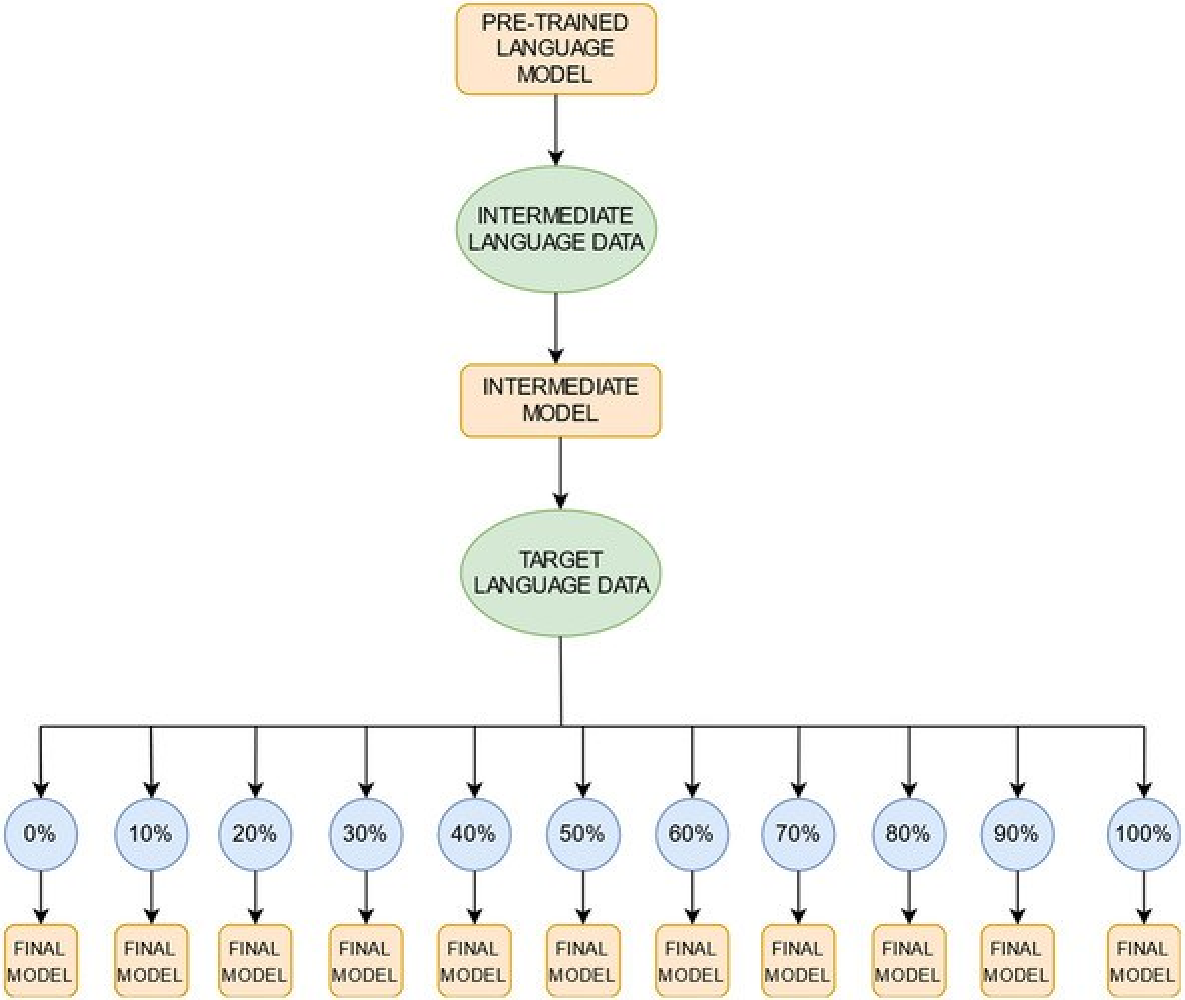
[redacted] London, United Kingdom, 45 minutes ago

Who does he think he is apologising for her! We may or may not agree with her views and opinions, but last time I checked, she was entitled to have them!

[redacted] Halesworth, 4 hours ago

# Cross-lingual offensive speech experiments

- 5 hate speech datasets: Arabic, Croatian, German, English, and Slovenian
- 2 models: mBERT, cseBERT
- Monolingual results: 71-84%



PRE-TRAINED LANGUAGE MODEL

INTERMEDIATE LANGUAGE DATA

INTERMEDIATE MODEL

TARGET LANGUAGE DATA

0%

10%

20%

30%

40%

50%

60%

70%

80%

90%

100%

FINAL MODEL

FINAL MODEL

FINAL MODEL

FINAL MODEL

FINAL MODEL

FINAL MODEL

FINAL MODEL

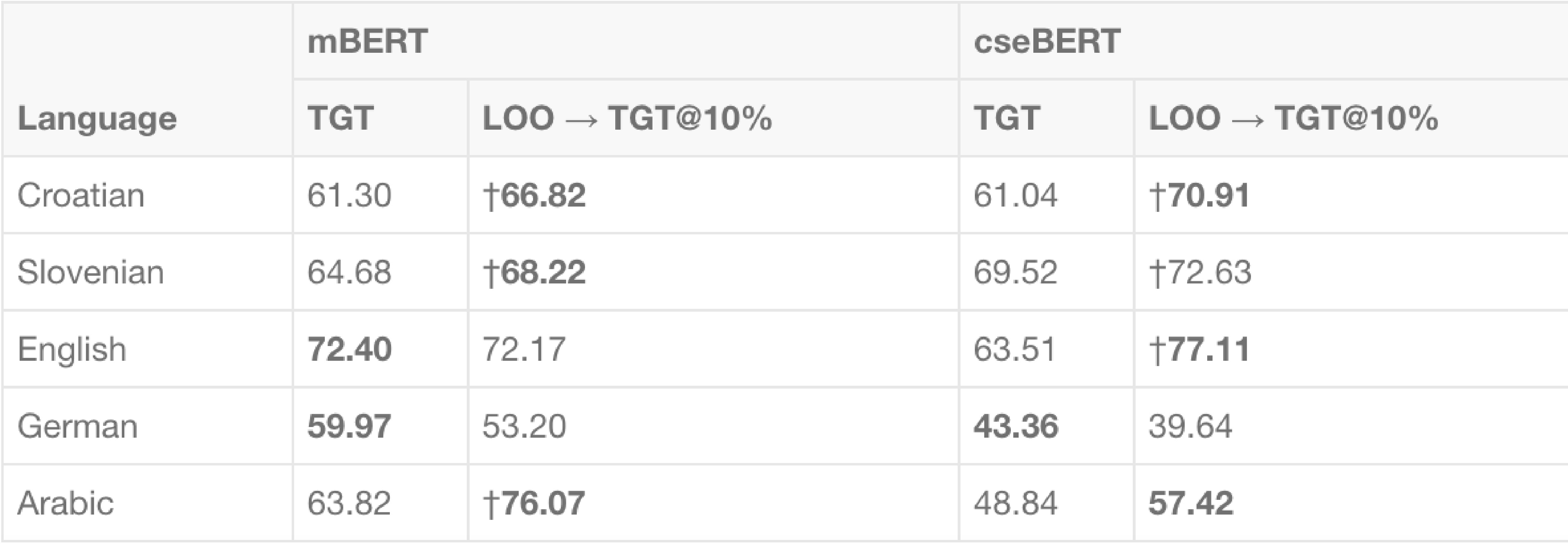
FINAL MODEL

FINAL MODEL

FINAL MODEL

FINAL MODEL

# LOO: all except target language



| | mBERT | | cseBERT | |
|---|---|---|---|---|
| **Language** | **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** |

# Other tasks on user generated content

- Multilingual sentiment analysis
  - Including zero-shot Twitter sentiment transfer
- User demographics characteristics
- Fake-news detection
  - Knowledge graph informed fake news classification via heterogeneous representation ensembles

# Text generation

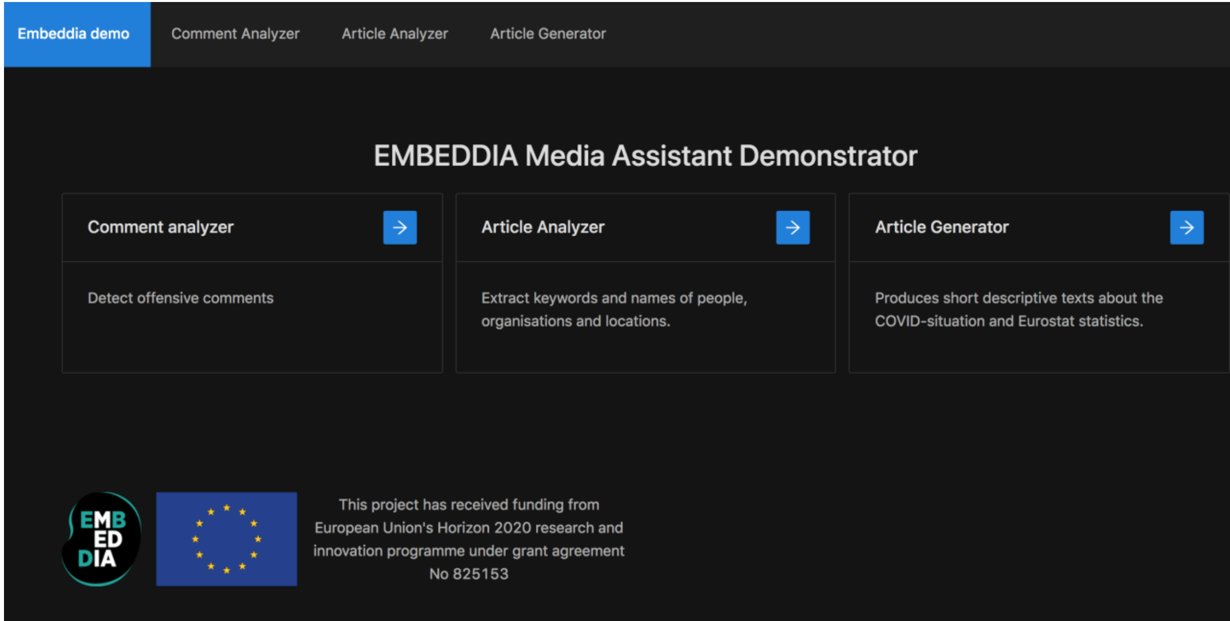
- Report generation from data
- Text summarization
- Headline generation

| TRUE | Generated |
|---|---|
| love rules : remembering my mom on mother's day | 10 rules my mom taught me |
| formula one unveils driver numbers for 2014 season | vettel takes pole as f1 number one for 2014 season |
| russian premier steps into fray on jailed tycoon | russia's premier defies putin on oil case |
| central banks move to rescue an ailing euro | central banks step up efforts to stem euro's slide |
| suit by detainee on transfer to syria finds support in jet's log | records support detainee's claim of torture in syria |
| looking anew at value of a corporate pedigree | bush's corporate past becomes a test of his values |
| sybrina fulton seeks to trademark trayvon rallying cries | trayvon martin's mom seeks to trademark his name |
| questions of death row justice for poor people in alabama | alabama's death - row system is failing its citizens |

# Making the tools available: EMBEDDIA Media assistant (code, dockers, demos)

Key results:

*keyword tagging*
*comment filtering*
*article generator*



EMBEDDIA Media Assistant Demonstrator

Comment analyzer →

Detect offensive comments

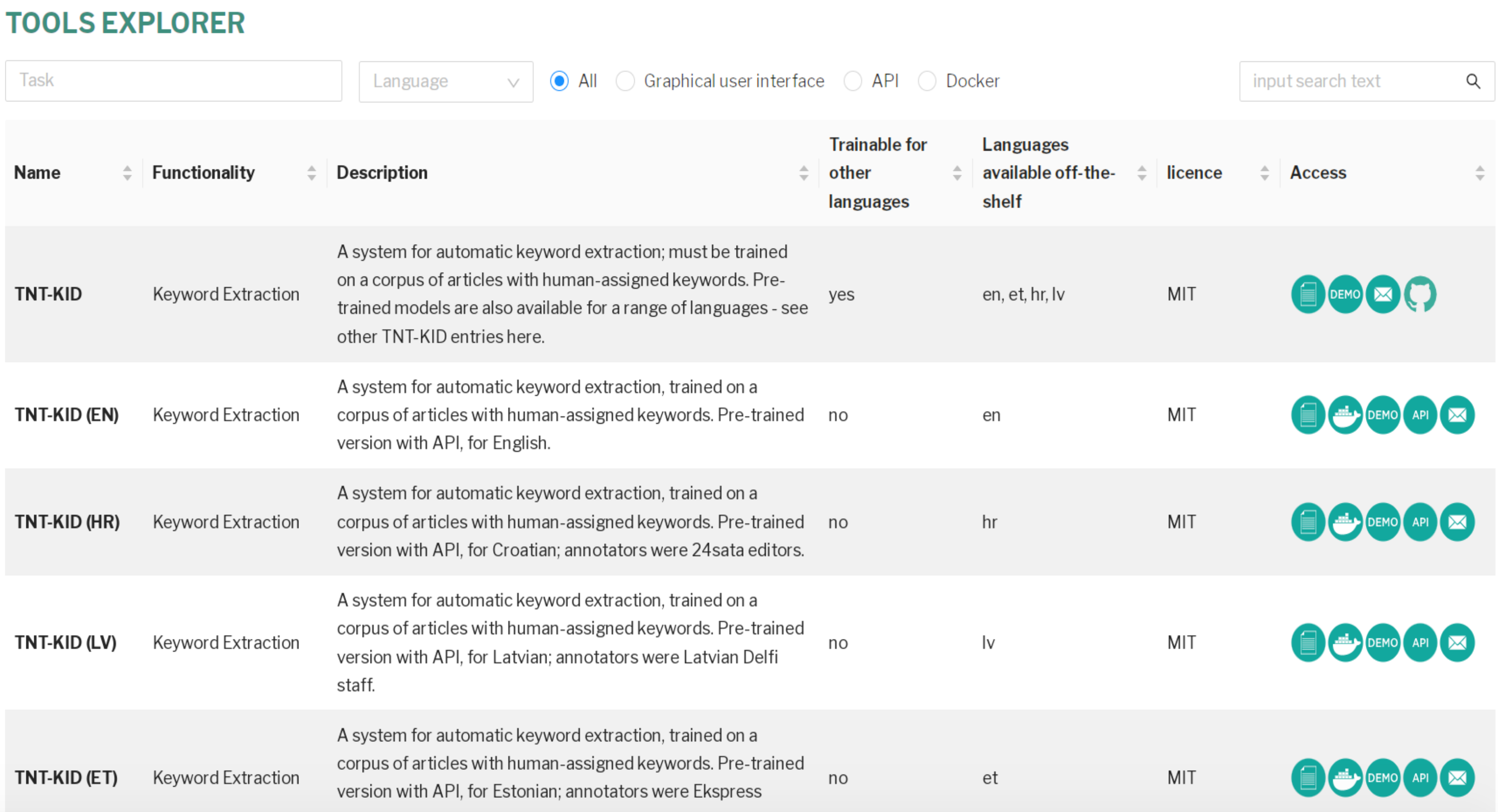
Article Analyzer →

Extract keywords and names of people, organisations and locations.

Article Generator →

Produces short descriptive texts about the COVID-situation and Eurostat statistics.

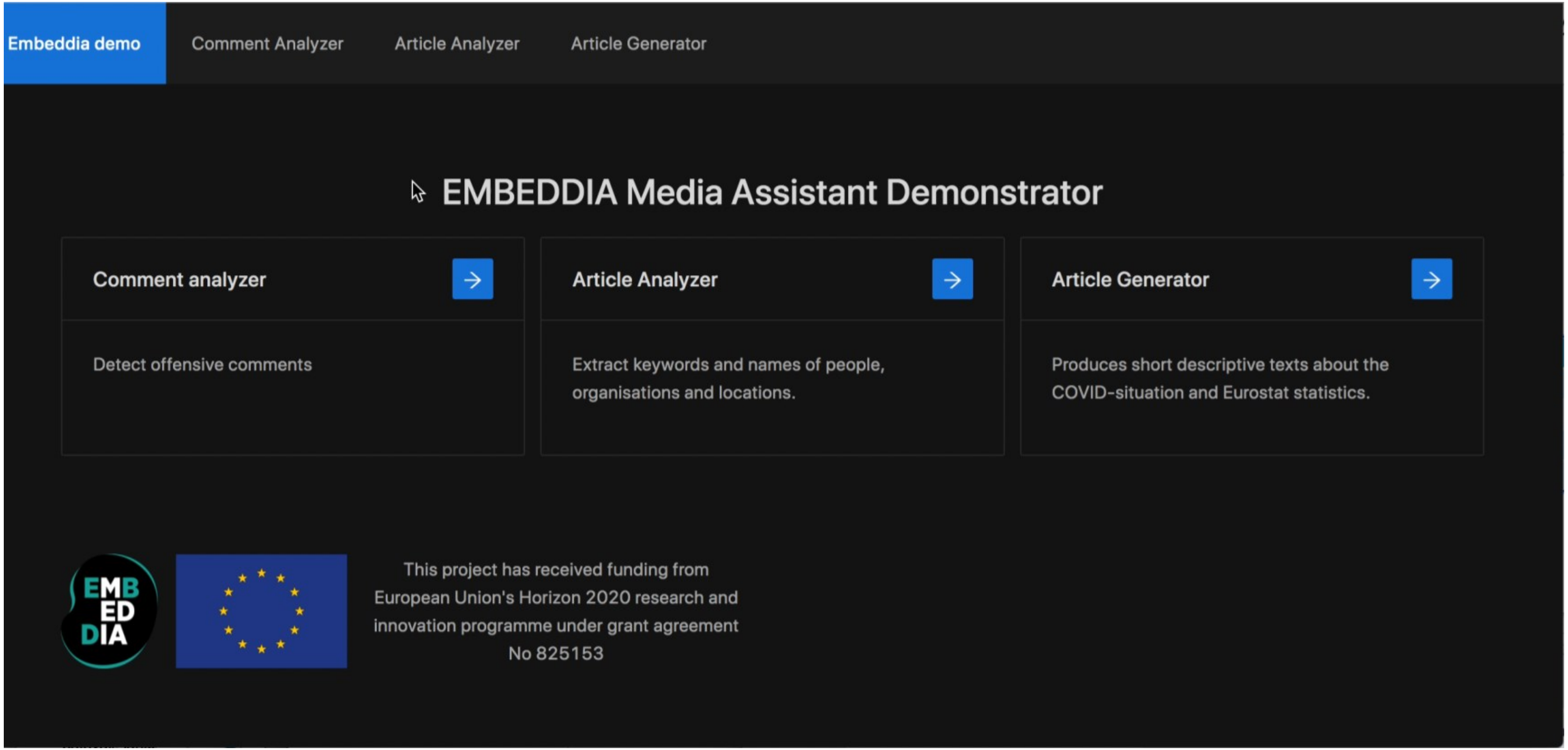
This project has received funding from European Union's Horizon 2020 research and innovation programme under grant agreement No 825153



## TOOLS EXPLORER

Task | Language ⌵ | ● All ○ Graphical user interface ○ API ○ Docker | input search text 🔍

| Name | Functionality | Description | Trainable for other languages | Languages available off-the-shelf | licence | Access |
|---|---|---|---|---|---|---|
| **TNT-KID** | Keyword Extraction | A system for automatic keyword extraction; must be trained on a corpus of articles with human-assigned keywords. Pre-trained models are also available for a range of languages - see other TNT-KID entries here. | yes | en, et, hr, lv | MIT | 📄 DEMO ✉ 🐙 |
| **TNT-KID (EN)** | Keyword Extraction | A system for automatic keyword extraction, trained on a corpus of articles with human-assigned keywords. Pre-trained version with API, for English. | no | en | MIT | 📄 🐳 DEMO API ✉ |
| **TNT-KID (HR)** | Keyword Extraction | A system for automatic keyword extraction, trained on a corpus of articles with human-assigned keywords. Pre-trained version with API, for Croatian; annotators were 24sata editors. | no | hr | MIT | 📄 🐳 DEMO API ✉ |
| **TNT-KID (LV)** | Keyword Extraction | A system for automatic keyword extraction, trained on a corpus of articles with human-assigned keywords. Pre-trained version with API, for Latvian; annotators were Latvian Delfi staff. | no | lv | MIT | 📄 🐳 DEMO API ✉ |
| **TNT-KID (ET)** | Keyword Extraction | A system for automatic keyword extraction, trained on a corpus of articles with human-assigned keywords. Pre-trained version with API, for Estonian; annotators were Ekspress | no | et | MIT | 📄 🐳 DEMO API ✉ |



EMBEDDIA Media Assistant Demonstrator

Comment analyzer

Detect offensive comments

Article Analyzer

Extract keywords and names of people, organisations and locations.

Article Generator

Produces short descriptive texts about the COVID-situation and Eurostat statistics.

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

DEMO:
https://embeddia-demo.texta.ee/

EMBEDDIA

www.embeddia.eu

@embeddiaproject