EMBEDDIA

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

Research and Innovation Action
Call: H2020-ICT-2018-1
Call topic: ICT-29-2018 A multilingual Next generation Internet
Project start: 1 January 2019 Project duration: 36 months

D4.1: Datasets, benchmarks and evaluation metrics for cross-lingual content analysis (T4.4)

Executive summary

This report details the dataset, benchmarks and evaluation metrics related cross-lingual content analysis tasks of WP4 of the EMBEDDIA project, covering three main research tasks: multilingual news linking, news summarisation and visualisation, and sentiment and viewpoints analysis. The deliverable first describes the datasets exported from the media partners' news archives. Next, the datasets constructed for specific WP4 tasks are presented, followed by the description of other available datasets, which are either public benchmarks or other resources that can be shared within the consortium. The report concludes with the description of evaluation methods and metrics for the three WP4 tasks.

Partner in charge: TEXTA

Project co-funded by the European Commission within Horizon 2020 **Dissemination Level**		
PU	Public	PU
PP	Restricted to other programme participants (including the Commission Services)	–
RE	Restricted to a group specified by the Consortium (including the Commission Services)	–
CO	Confidential, only for members of the Consortium (including the Commission Services)	–

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

Deliverable Information

Document administrative information	
Project acronym:	**EMBEDDIA**
Project number:	**825153**
Deliverable number:	**D4.1**
Deliverable full title:	**Datasets, benchmarks and evaluation metrics for cross-lingual content analysis**
Deliverable short title:	**Datasets for cross-lingual content analysis**
Document identifier:	**EMBEDDIA-D41-DatasetsForCrosslingualContentAnalysis-T44-submitted**
Lead partner short name:	**TEXTA**
Report version:	**submitted**
Report submission date:	**30/09/2019**
Dissemination level:	**PU**
Nature:	**R = Report**
Lead author(s):	**Kristiina Vaik (TEXTA)**
Co-author(s):	**Dragana Miljkovic (JSI), Senja Pollak (JSI), Marko Pranjić (STY/TRI), Elaine Zosa (UH-CS), Matej Martinc (JSI), Elvys Linhares Pontes (ULR)**
Status:	**_ draft, _ final, X submitted**

The EMBEDDIA Consortium partner responsible for this deliverable has addressed all comments received. Changes to this document are detailed in the change log table below.

Change log

Date	Version number	Author/Editor	Summary of changes made
28/05/2019	v1.0	Kristiina Vaik (TEXTA), Dragana Miljkovic (JSI)	Draft report
01/07/2019	v1.1	Kristiina Vaik (TEXTA)	Second draft report
17/07/2019	v1.2	Kristiina Vaik (TEXTA)	From MSWord to Overleaf
22/07/2019	v1.3	Marko Pranjić (STY/TRI)	Styria dataset description
25/07/2019	v1.4	Senja Pollak (JSI)	General comments and structure change
14/08/2019	v1.5	Senja Pollak (JSI)	Changes and comments
21/08/2019	v1.6	Kristiina Vaik (TEXTA)	Updates on media partner dataset section
22/08/2019	v1.7	Matej Martinc (JSI)	Proofreading and comments
30/08/2019	v1.8	Matej Martinc (JSI)	Editing and adding content in Section 2.3.3
30/08/2019	v1.9	Senja Pollak (JSI)	Editing
02/09/2019	v1.10	Kristiina Vaik (TEXTA)	Updates on media partner sections, editing
07/09/2019	v1.11	Marko Robnik-Šikonja (UL)	Internal review
10/09/2019	v1.12	Matthew Purver (QMUL)	Internal review
12/09/2019	v1.13	Senja Pollak and Nada Lavrač (JSI)	Structural changes implementing the reviewers' suggestions
13/09/2019	v1.14	Kristiina Vaik (TEXTA)	Consolidation based on internal reviews
21/09/2019	v1.15	Nada Lavrač (JSI)	Quality control
25/09/2019	v1.16	Kristiina Vaik (TEXTA)	Added ExM final information
29/09/2019	final	Elvys Linhares Pontes (ULR), Dragana Miljkovic (JSI), Senja Pollak (JSI)	Final corrections addressing the quality control comments
30/09/2019	submitted	Tina Anžič (JSI)	Report submitted

Table of Contents

1. Introduction 6
2. Media partners' datasets 6
   2.1 Archive datasets of EMBEDDIA media partners 7
      2.1.1 Styria Media Group archive datasets: Vecernji list and 24sata 7
      2.1.2 Ekspress Meedia archive datasets 8
      2.1.3 Finnish News Agency archive datasets 10
   2.2 WP4 task specific datasets of EMBEDDIA media partners 11
      2.2.1 Linked news articles dataset for task T4.1 11
      2.2.2 News similarity triplets dataset for task T4.1 11
      2.2.3 News of interest from neighbouring countries (NNC) for task T4.1 12
      2.2.4 Sentiment annotation datasets for task T4.3 12
3. Other datasets and benchmarks 13
   3.1 Real-time multilingual news linking (task T4.1) 13
   3.2 Cross-lingual news summarisation and visualisation (task T4.2) 17
   3.3 Cross-lingual identification of viewpoints and sentiment in news reporting (task T4.3) 17
4. Evaluation methods and metrics 22
   4.1 Evaluation methods and metrics for task T4.1 22
   4.2 Evaluation methods and metrics for task T4.2 23
   4.3 Evaluation methods and metrics for task T4.3 24
5. Conclusions and further work 25
Appendix A - Metadata of media partners datasets 30
Appendix B - Dataset samples 32
Appendix C - Annotation procedure for the Sentiment Analysis task (T4.3) 33
   Annotation process 33
   Timeline 35
   Current status of annotations organised by STY/Trikoder 36
   Annotation guidelines 37

List of abbreviations

DoA	Description of Action
EC	European Commission
ExM	Ekspress Meedia
GA	Grant Agreement
JSI	Jožef Stefan Institute
STT	Finnish News Agency, *Suomen Tietotoimisto*
STY	Styria Media Services
T	Task
TRI	Trikoder
WP	Work Package
POS	Part of Speech
CLTS	Cross-Language Text Summarisation

1 Introduction

This report is a result of the initial activities performed within WP4 of the EMBEDDIA project. The overall objective of WP4, named *Cross-lingual content analysis*, is to facilitate the analysis of news content across different languages, thus empowering news media consumers, researchers and news media professionals. The current language barriers and overflow of information prevent these groups from detecting and consuming all the relevant information, particularly across different languages, and from analysing and reflecting on the differences in news reporting. WP4 will provide real-time linking of relevant texts with informative summaries, visualizations of content, and analysis of the viewpoints and sentiment of articles from different sources, addressing content in different languages.

WP4 of the EMBEDDIA project consists of four tasks. This deliverable reports the first activities performed within Task T4.4, which is concerned with resource gathering, benchmarking and evaluation, using the gathered datasets in English, Slovene, Croatian, Estonian, Lithuanian, Latvian, Russian, Finnish, and Swedish language. The gathered resources will be used in the three main research tasks of WP4:

- T4.1: Real-time multilingual news linking;
- T4.2: Cross-lingual news summarisation and visualisation;
- T4.3: Cross-lingual identification of viewpoints and sentiment in news reporting.

During the course of the project, the texts available in the gathered news corpora (described in Section 2 of this document) will be segmented, tokenized, part-of-speech (POS) tagged, and enriched with annotations from Tasks T2.1, T2.2, and other task specific annotations. The tools developed in WP4 will be evaluated on a selection of benchmarks (described in Section 3) and using the relevant methods and metrics (described in Section 4), as well as in terms of user experience, assessed in WP6. The results of these activities will be reported in the follow-up deliverable D4.8 of Task T4.4.

The report is organised as follows. We start with a description of the resources provided by the EMBEDDIA media partners in Section 2, followed by presenting other public resources of interest for the three research tasks of WP4 in Section 3. In Section 4 we describe the evaluation methods and metrics, currently defined for the three research tasks within WP4. We present conclusions about the reported datasets and evaluation methods in Section 5, where we also outline the plans for further work.

2 Media partners' datasets

The EMBEDDIA consortium includes three news media partners: Ekspress Meedia (ExM) from Tallinn, Estonia; Suomen Tietotoimisto (STT) - the Finnish News Agency from Helsinki, Finland; and Styria Media Services (STY) from Zagreb, Croatia, and later Trikoder (TRI)[1]. In this section we introduce the datasets of news articles provided by EMBEDDIA media partners (STY/TRI, ExM, and STT) to be used in WP4 tasks. In Section 2.1 we describe the news article from media partners' archives, which were made available to EMBEDDIA, while in Section 2.2 we introduce task-specific datasets provided by EMBEDDIA media partners that have been specifically gathered/annotated for the purposes of WP4 tasks.

[1] In EMBEDDIA, STY has been recently replaced by Trikoder, a collaborating company from Croatia, which will provide the news and and other media content for the Croatian language from Styria Media Group.

2.1 Archive datasets of EMBEDDIA media partners

This section presents the news datasets provided by partners STY/TRI, ExM and STT, each presented in a separate subsection. Table 1 summarises the datasets' size, languages, date and file format.

Table 1: Overview of EMBEDDIA media partners' datasets

Partner	Size	Languages	Date	File format
STY/TRI	1 421 466	Croatian	09-07-2019	csv
ExM	1 441 112	Estonian, Russian	11-06-2019	json
STT	4 186 632	Finnish, Swedish	06-06-2019	xml

2.1.1 Styria Media Group archive datasets: Vecernji list and 24sata

Styria Media Group is one of the leading media groups in Austria, Croatia, and Slovenia. Their portfolio includes daily and weekly newspapers, magazines and book publishers, radio stations and a shareholding in a TV station. In digital format, the group operates successful news portals, marketplaces, as well as content and community portals. Styria portals *24sata* and *Vecernji list* are the leading portals in the Croatian market in terms of page visits and business results. Both *24sata* and *Vecernji list* are daily newspapers in Croatia appealing to the broad audience. *24sata* contains more emotionaly-packed, sensational content, it more readily opposes the government and makes fun of them. *Vecernji list* has slightly longer texts with more content about politics, while *24sata* contains more tragic news and show business related content.

The dataset from Styria Media Group contains news articles in Croatian from digital editions of *24sata*, *Vecernji list*, and their respective niche portals.

The 24sata subcorpora include:

- **24sata** (`www.24sata.hr`) is a daily news website with the goal of publishing the most accurate and relevant information in the shortest possible time.
- **joomboos** (`joomboos.24sata.hr`) offers readers an overview of the most interesting news from the world of YouTube, the latest news from the world of celebrities, an overview of fashion trends, and diverse and entertaining quizzes for young audiences.
- **gastro** (`gastro.24sata.hr`) is the leading foodie lifestyle community of food and cooking enthusiasts. They share a passion for exploring and trying out new delicacies for their palates.
- **miss7** (`miss7.24sata.hr`) addresses women, covering their needs and interests. The content of miss7 deals with real problems and situations women go through, and it is created by journalists who advise their readers, share with them their first-hand experiences, break taboos, and play the role of a friend they can trust.
- **missZDRAVA** (`miss7zdrava.24sata.hr`) covers the well being of the mind, body, spirit, relationships, and environment. The brand acts as a life coach by using specific content and tools to help readers get to know themselves, the laws of their psyche, body, and spirit in relation to the environment, in order to become more relaxed and fulfilled.
- **missMAMA** (`miss7mama.24sata.hr`) is a brand for mothers and pregnant women, with the most active parenting community in Croatia. It involves topics from family planning and pregnancy, through childhood until the end of elementary school.
- **autostart** (`autostart.24sata.hr`) is an automotive website in Croatia that provides the latest vehicle tests, interviews, reports, columns, and other news from the automotive industry. In addition to

video content, it provides attractive photos of the latest cars and motorcycles. It covers car and motorcycle events with useful articles.

- **Express.hr** (www.express.hr) covers a wide range of topics from various areas, and apart from the current affairs in politics and economics, it also elaborates on topics that deal with more relaxed side of life such as luxury, tourism, popular science and technology.

The Vecernji list subcorpora include:

- **Vecernji list** (www.vecernji.hr) provides daily news with an overview of the most important news and events from Croatia and the world, covering the topics on sports, fashion, culture, etc.
- **Moja Hrvatska** (mojahrvatska.vecernji.hr) brings daily news, stories, columns, and blogs about the challenges of Croatian people's lives outside the homeland.
- **Living** (living.vecernji.hr) covers the news from the world of interiors, trends in the field of architecture, horticulture, and real estate market.
- **Vojna povijest** (vojnapovijest.vecernji.hr) provides coverage on topics of military history, mostly from the Croatian War of Independence, World War I and WWII, but also from earlier periods.
- www.vecernji.ba provides a coverage of similar topics to www.vecernji.hr but focused on Bosnia and Herzegovina instead.

The data was prepared by the Data Science team of Styria Media Services – later Trikoder. The dataset is a digital archive of news portals (see above) from September 2001 until June 2019. Articles that had less than five words of content were filtered out. See Appendix A Table 6 for detailed information about the dataset's attributes. The datset can be used for research purposes by the researchers of the consortium with specific limitations and/or conditions for implementation an exploitation, where 5-10% of the data will be made available under a CC BY-NC-SA 4.0 licence. For external researchers the request to access the dataset should be submitted to Trikoder (contact person: Marko Pranjić).

2.1.2 Ekspress Meedia archive datasets

Ekspress Meedia (ExM) is the leading media group in the Baltic States, whose activities include publishing, printing services, and online media content production. ExM owns the leading online media portals in the Baltics and publishes Estonia's most widely read daily and weekly newspapers, in addition to seven out of the top ten magazines in Estonia. The group is vertically integrated with everything from content to printing and distribution.

The dataset from Ekspress Meedia contains news articles in Estonian and Russian from digital editions of www.delfi.ee domains, sections and special pages:

- **alkeemia** (alkeemia.delfi.ee) concentrates on topics such as self-development, environment, mystics, relations, spiritual world.
- **annestiil** (annestiil.delfi.ee or digileht.annestiil.delfi.ee) is a montly magazine "Anne & Stiil" covering topics on fashion, style and beauty.
- **arileht** (arileht.delfi.ee) concentrates on both local and world business news and analysis.
- **bublik** (bublik.delfi.ee) is concentrates on entertainment news in Russian.
- **catwalk** (catwalk.delfi.ee) was a previous version of *annestiil*, removed from Delfi in June 2018.
- **dekor** (dekor.delfi.ee) covers topics on furnishing and construction in Russian.
- **kalale** (digi.kalale.ee) was covering articles related to fishing, removed from Delfi in 2018.
- **kodukiri** (digi.kodukiri.ee) was a digital version of a housing magazine "Kodukiri", was removed from Delfi in 2018.

- **naisteleht** (digi.naisteleht.ee) was a digital version of a women's magazine "Naisteleht", was removed from Delfi in 2018.
- **kroonika** (kroonika.delfi.ee or digileht.kroonika.delfi.ee) is an Estonian most popular entertainment news portal and a magazine covering topics about entertainment and celebrities.
- **maakodu** (maakodu.delfi.ee or digileht.maakodu.delfi.ee) is a monthly magazine which covers topics on country-side housing, furnishing, construction, gardening.
- **maaleht** (maaleht.delfi.ee or digileht.maaleht.delfi.ee) is a weekly newspaper offering articles and news about country-side life and issues.
- **omamaitse** (omamaitse.delfi.ee or digileht.omamaitse.delfi.ee) is a monthly magazine which involves topics about food, cooking, recipes, restaurants, etc.
- **perejakodu** (perejakodu.delfi.ee or digileht.perejakodu.delfi.ee) is a monthly magazine which addresses parenting, children, pregnancy, home, etc.
- **perejalaps** (perejalaps.delfi.ee) was a former version of perejakodu.delfi.ee, was removed from Delfi in June 2018.
- **tervispluss** (tervispluss.delfi.ee or digileht.tervispluss.delfi.ee) is a monthly magazine addressing topics such as health, training, nutrition, hobbies, etc.
- **eestinaine** (eestinaine.delfi.ee or digileht.eestinaine.delfi.ee) is an Estonian longest-published women's magazine which covers a wide area of topics for women, such as love, children, home, relations, career, hobbies, travelling.
- **ekspress** (ekspress.delfi.ee) is a weekly newspaper which concentrates on investigative stories, in-depth interviews and analysis.
- **epl** (epl.delfi.ee) is a daily newspaper which covers news and features about Estonian politics, economy, culture, sports, etc.
- **erid** (erid.delfi.ee) was a sub-page for ExM special print publications. The page doesn't exist any more.
- **forte** (forte.delfi.ee) provides articles and news on technology, digital, science, history, environment, cars, etc.
- **homme** (homme.delfi.ee) was a men's portal covering topics such as cars, technology, lifestyle, relations, humour. It was removed from Delfi in September 2017.
- **ilmateade** (ilmateade.delfi.ee) covers topics on weather news, articles and forecasts.
- **ilm** (ilm.delfi.ee) is a domain name directing to *ilmateade*.
- **jana** (jana.delfi.ee) is covering topics on beauty, fashion, psychology, children, house in Russian.
- **journalist** (journalist.delfi.ee) provides user-generated stories in Russian.
- **kasulik** (kasulik.delfi.ee) covers consumer news and stories.
- **kinoveeb** (kinoveeb.delfi.ee) provides new and articles on movies, cinemas, actors, etc.
- **kodukujundaja** (kodukujundaja.delfi.ee) is a special advertising project from 2016 consisting of stories on projecting and developing a new home.
- **lemmikloom** (lemmikloom.delfi.ee) provides news and articles pets and other animals.
- **longread** (longread.delfi.ee) is a special section for long-form journalism, enriched with interactive graphics, videos and other elements.
- **moodnekodu** (moodnekodu.delfi.ee) is concentrating on topics such as modern home, furnishing, real estate, construction, renovation.

- **naistekas** (naistekas.delfi.ee) provides coverage on topics of sex, relations, health, home, horoscopes, inspiring women, women's issues, etc.
- **pogoda** (pogoda.delfi.ee) covers weather news, articles and forecasts in Russian.
- **poleznoe** (poleznoe.delfi.ee) covers consumer news and stories in Russian.
- **rahvahaal** (rahvahaal.delfi.ee) provides user-generated stories.
- **reisijuht** (reisijuht.delfi.ee) covers topics about travelling and travel offers.
- **reklaam** (reklaam.delfi.ee or reklaam.lehed.ee) is a domain directing to ExM advertising contacts and information.
- **delfi** (rus.delfi.ee or www.delfi.ee) is Estonia's most popular news portal in Russian or Estonian respectively. The content covers a wide area of topics on local and world news, opinion, economy, sports, technology, consumer news, etc.
- **sport** (sport.delfi.ee) covers all important sports news from Estonia and the world, offers comments, analysis, feature stories and sports streams.
- **superkodu** (superkodu.delfi.ee) is a special advertising project from 2017 consisting of stories on projecting and developing a new home.
- **turist** (turist.delfi.ee) is concentrating on travelling and travel offers in Russian.
- **tv** (tv.delfi.ee) is a portal which gathers all video content of delfi.ee, e.g., news clips, special shows, streams.
- **weekend** (weekend.delfi.ee) covers party galleries from Estonian nightclubs, was removed from Delfi in 2014.
- **suurespildis** (delfi.ee/news/paevauudised/suurespildis) is a special photo stories in Delfi.

The dataset was prepared by the ExM IT department. It is an archive of all publicly visible articles from Estonian and Russian news portals from year 2009 to 2019 May. See Appendix A Table 7 for detailed information about the datasets' attributes/metadata. The datasets can be used for research purposes by the researchers of the consortium without any specific limitations during the project and for research after the project. ExM plans to publicly release about 5000 items of articles (at least in two languages). For publicly available subsets CC BY-NC licence will be used. For external researchers the request to access other datasets should be submitted to ExM (contact persons[2]: Tarmo Paju and Ivar Krustok).

2.1.3 Finnish News Agency archive datasets

Finnish News Agency (Suomen Tietotoimisto, STT) is the only news agency in Finland, and the majority of the Finnish daily media subscribes to STT's services that include text, pictures, planning, tools and data.

The datasets were prepared by the STT team, with the assistance of the Language Bank of Finland (*Kielipankki*). This dataset is an archive of all articles in Finnish and Swedish sent to media outlets by STT between years 1992 to 2018. See Appendix A Table 8 for detailed information about the datasets' attributes. Both datsets are released under the CLARIN RES end-user license +NC +OTHER 1.0 licence. The Finnish dataset is downloadable[3] from the Language Bank of Finland for research purposes. To get access to the Swedish archive external researchers should contact STT directly (contact person: Salla Salmela) or via the Kielipankki application form.

[2] The main contact for data usage information was Mait Tafenau, former IT director, who left ExM in the end of August 2019.

[3] http://urn.fi/urn:nbn:fi:lb-2018121004

2.2 WP4 task specific datasets of EMBEDDIA media partners

The tasks of WP4 address multilingual news linking (T4.1), news summarisation and visualisation (T4.2), as well as sentiment and viewpoints analysis (T4.3). This section describes the datasets constructed for some of the WP4 tasks, i.e. the first three datasets for task T4.1 and the last dataset for task T.3. Summary information on the datasets is provided in Table 2.

Table 2: Overview of WP4 task specific news articles datsets.

Corpus	Task	Size	Partner	Date or Version
Linked news	T4.1	266 293	STY/TRI	10-07-2019
News similarity	T4.1	5261	STY/TRI	31-07-2019
News of interest	T4.1	21	ExM	1.0
Sentiment annotation	T4.3	approx. 1000–2000	STY/TRI/all	in progress

2.2.1 Linked news articles dataset for task T4.1

The STY linked news dataset contains a list of articles that link to other articles. Linked news refer to articles that function as a background material for the main article. These stories are usually strongly related to the article (same person, event, background, or topic). The linked news dataset was created by collecting the links from the articles. One article can link zero or more articles from the same site.

The dataset, containing information about linked news for 266 293 articles, was prepared by the Data Science team of Styria Media Services – later Trikoder. It comprises of all available articles until July 2019 that have at least one other related article. The dataset is provided as a csv and contains three attributes: article, site and related_articles. The related_articles attribute gets automatically assigned list of article IDs from the same site which were validated by journalists. This dataset is open for project partner research purposes only, discussions about potential public release are in progress.

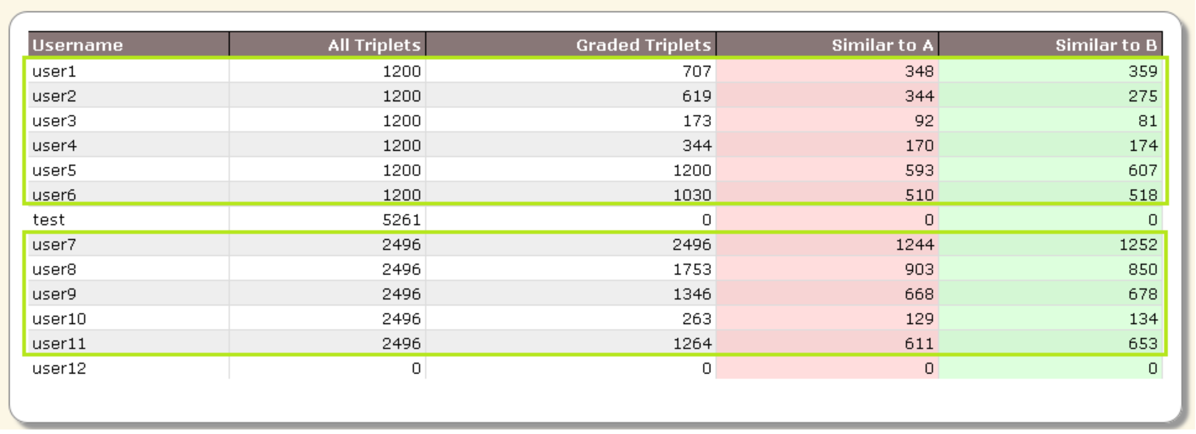
2.2.2 News similarity triplets dataset for task T4.1

Trikoder is preparing a dataset from Croatian articles of Styria Media Group, which will serve to measure topic similarity. The goal of this dataset is to understand how humans perceive topic similarity and to use this information in building and evaluating the models used for various news tasks by STY/Trikoder (e.g., recommender of similar articles, model for personalisation, model for finding related images (from text), tag recommendation).

The news similarity triplets dataset is a collection consisting of three similar articles combined with human similarity annotation. The question to the annotators was: “Given the input article, which of the other two articles (second or third) is more similar to the first one in terms of topic?”

The news similarity triplets are constructed in such a way that all three articles are similar. Three different types of triplet datasets were created. The articles of these triplet datasets originate from `www.24sata.hr` which have been published after January 2017. Firstly, 2000 article triplets were chosen such that all three articles belong to the same section or subsection of the news portal. Another 2000 triplets were chosen such that among five most similar articles (based on doc2vec cosine distance) to the query article, two were selected randomly to create a triplet. The last 1261 triplets were selected from linked articles such that all three articles are linked. If there were more than three linked articles, three were selected on random.

Figure 1: Number of articles annotated by the annotators working on the Article Triplets Similarity task.



Username	All Triplets	Graded Triplets	Similar to A	Similar to B
user1	1200	707	348	359
user2	1200	619	344	275
user3	1200	173	92	81
user4	1200	344	170	174
user5	1200	1200	593	607
user6	1200	1030	510	518
test	5261	0	0	0
user7	2496	2496	1244	1252
user8	2496	1753	903	850
user9	2496	1346	668	678
user10	2496	263	129	134
user11	2496	1264	611	653
user12	0	0	0	0

The annotators shall provide a ground truth for their similarity. The illustration of current progress (as of 24th of August 2019) is provided in Table 1, which shows that two annotators already finished the annotation task, and others are also making a good progress. The annotators were exposed to training prior to starting the task.

The dataset was prepared by the Data Science team of STY – later Trikoder. The articles will be publicly available with the shuffled data, and the annotations for these articles will be available. For non-shuffled version of articles, we are still discussing if they can be made available.

2.2.3 News of interest from neighbouring countries (NNC) for task T4.1

This dataset contains Latvian articles from the Delfi portal which have gained popularity among Estonian readers. Delfi is a major internet portal in Estonia, Latvia, and Lithuania providing daily news, ranging from gardening to politics. It ranks as one of the most popular websites among the Baltic users. The aim of this is to automatically detect other relevant articles/stories in other Baltic states which may be of interest to Estonian readers.

The articles of this dataset originate from beginning of year 2018 to June 2019. For creating this dataset a small scraper was implemented. The final articles were selected manually based on their relevance (number of clicks for the same story generated at `delfi.ee` and the domain knowledge of a managing editor). The information scraped from an article consists of the URL, the language, the article title, the main body of text, the article summary, titles of related articles as suggested by the host site. The dataset is stored in JSON format where one record is a two-element list which contains Latvian and Estonian scraped article's information.

Data collection was done by Tarmo Paju from ExM. It can be used for research purposes by the researchers of the consortium without any specific limitations during the project and for research after the project.

2.2.4 Sentiment annotation datasets for task T4.3

For testing cross-lingual sentiment analysis on news, STY/TRI organised annotation of 1000-2000 news written in Croatian, and the datasets for other project languages can be created by media partners using the same procedure. The article specification and annotation procedure are provided in Appendix C.

The datasets have the same structure as the SentiNews dataset[4], described in the study of Bučar et al. (2018). The details of this dataset are further described at the beginning of Section 3.3. At the moment of writing this deliverable, the annotation organised by STY/Trikoder is in progress.

For task T4.2 at the moment of writing this deliverable, no specific datasets have been created by media partners, but the sentiment datasets have been identified as a good candidate for initial experiments in visualisation, allowing for developing tools for visualisation based on differences in metadata (i.e. sentiment).

3 Other datasets and benchmarks

This section describes publicly available resources and the resources gathered by the academic partners of the EMBEDDIA project. The included resources are grouped into subsections corresponding to the tasks in WP4.

3.1 Real-time multilingual news linking (task T4.1)

An overview of the datasets identified as useful for T4.1 is presented in Table 3. This task aims to develop tools for linking news stories across languages based on their topics and contents. In collaboration with WP1 and WP2, different events, entities, and keywords will be obtained which can be used for linking stories across languages.

Table 3: Datasets of interest to real-time multilingual news linking

ID number	Dataset	Public availability	Project languages
T4.1.-1	One Week of Global News Feeds provided by Kaggle	Yes	all apart from Lithuanian and Latvian
T4.1.-2	News Aggregator Dataset provided by Kaggle	Yes	all apart from Lithuanian, Latvian
T4.1.-3	TREC 2018/2019 News Track (Washington post)	Yes	English
T4.1.-4	Dataset of Slovene & English crawled news	Only for project partners	Slovene, English
T4.1.-5	Wikipedia	Yes	all
T4.1.-6	CLEF datasets	Yes	Finnish
T4.1.-7	YLE dataset	Yes	Finnish, Swedish

T4.1 will use information from the semantic enrichment (especially multilingual event detection) and cross lingual word-embeddings based topic modelling. The methodology for ranking and grouping such links will be developed, so that users will be offered a representative and reliable selection of links. Moreover, event emergence (where did the news first appear) will be identified and the news spreading will be analysed. This task will also produce the linked dataset for analysis of different viewpoints and sentiment annotation (T4.3), as well as for comment analysis (WP3). Special emphasis will be on the efficiency of the developed methods for the task of real-time news linking.

Each dataset is described seperately below.

One Week of Global News Feeds provided by Kaggle (ID number T4.1-1)

[4] SentiNews dataset is available at https://www.clarin.si/repository/xmlui/handle/11356/1110

This news dataset[5] captures a snapshot of one week of most of the new news content published online and includes approx. 1.3 million articles generated by over 18 000 news sources worldwide in 20 languages during a one week period in August 2017 (Thursday 24th to Wednesday 30th). This dataset was prepared by Rohit Kulkarni. The sources include news sites, government agencies, tech journals, blogs, and Wikipedia updates. The data has been collected by RSS feeds and by crawling other large news aggregators.

News Aggregator Dataset provided by Kaggle (ID number T4.1-2)

The dataset contains headlines, URLs, and categories for 422 937 news stories collected by a web aggregator between March 10th 2014 and August 10th 2014.

News categories in this dataset include business, science and technology, entertainment, and health. News articles that refer to the same news item (e.g., several articles about a recently released employment statistics) are categorized together.

TREC 2018 and 2019 News Track (Washington Post) (ID number T4.1-3)

We included two editions of the data from the TREC News Track workshop, first organised in 2018 (Huang et al., 2018). The TREC 2019 News Track workshop has set two goals: background linking and entity ranking. We describe only the first task, as it corresponds to the project's needs. The goal of the background linking task is to retrieve other news articles that provide important context or background information, by using a given news story. The motivation is to help readers understand or learn more about the story or main issues in the current article using the best possible sources.

For example, news websites nearly always link to related articles in a sidebar, at the end of an article, from within the text of the article, or all three. The participants of this task need to look at a particular case for linking: given that the user is reading a specific article (the query article), recommend articles that this person should read next that are the most useful for providing context and background for the query article.

The TREC Washington Post corpus used in the TREC 2019 News Track is identical to the 2018 workshop and contains 608 180 news articles and blog posts from January 2012 through August 2017. The articles are stored in a JSON format.

In TREC, the results are judged by NIST (National Institute of Standards & Technology, agency of the U.S. Commerce Department) assessors on the following scale:

- The linked document provides little or no useful background information.
- The linked document provides some useful background or contextual information that would help the user understand the broader story context of the query article.
- The document provides significantly useful background.
- The document provides essential useful background.
- The document MUST appear in the sidebar otherwise critical context is missing.

Altogether, 100 topics were developed on the Washington Post collection for the purpose of evaluating the background linking approaches of TREC 2018 and 2019 News Track participants. The background linking systems are required to find articles related to these topics. For TREC 2018 News Track, a gold standard test set was developed, containing manually assigned relevance scores for links between 50 topics and 823 articles from the Washington Post corpus. This test set can be used for the evaluation of new background linking approaches, since it enables comparison between the approaches developed in the scope of T4.1 and the approaches of the TREC 2018 News Track participants. A very similar gold standard test set for the TREC 2019 News Track, which will allow comparison with the approaches of

[5] This dataset is available at `https://www.kaggle.com/therohk/global-news-week`

the TREC 2019 News Track participants, is expected to be developed and made publicly available in the near future. The same evaluation criteria could be used for building cross-lingual datasets. For using the datasets, researchers have to sign the NDA. For EMBEDDIA consortium the NDA was signed by JSI.

Dataset of Slovene and English crawled news (ID number T4.1-4)

The tools for creation of topical datasets supporting individuals in the creation and analysis of collections of Slovene and English media presented in Kralj Novak et al. (2015) were used to create the datasets described below:

The Slovenian dataset contains 879 049 articles published online by various Slovenian news providers in the period from 17th of February 2014 to 24th of April 2019. The English dataset contains 59 560 949 articles published online by 175 world-wide English-language news sites in the period from 20th of October 2011 to 27th of August 2019. The news acquisition is ongoing for both datasets, hence the final date can be extended. A custom made web crawler (Sluban & Grčar, 2013) is used to download the articles, parse the content and remove the boilerplate. The data is collected, parsed, and stored in the XML format. It is then indexed in the ElasticSearch engine and can be exported into the JSON format via a Python script. The Web interface for data exploration, allowing for full-text search based on the ElasticSearch engine is accessible for Slovene at `http://annotate.ijs.si/senttweet/` and for English at `http://simpol.ijs.si/Home/NewsSearch/`.

This dataset can be shared by the researchers of the consortium for the purposes of T4.1 in order to create new topic specific datasets for Slovene and English (e.g., we have already created a new dataset containing just articles on the topic of Brexit). The exported data can be linked with datasets on the same topic created from media partners' datasets. As the datasets contain the information about news sources and automatically assigned sentiment labels, they can also be explored in the scope of task T4.3.

Wikipedia (ID number T4.1-5)

Wikipedia is the free online encyclopedia written by volunteer contributors and available in more than 200 languages. A Wikipedia dump for each language is publicly available from Wikimedia[6]. Alternatively, document-aligned Wikipedias for several language pairs are also available[7]. We have opted to make use of the cross-language link property on each Wikipedia page in order to link pages from different languages that are about the same topic. This enables us to link pages across several languages (not just two).

Document-aligned Wikipedia has been used in existing cross-language information retrieval (CLIR) models for training and evaluation (Litschko et al., 2018; Josifoski et al., 2019; Balikas et al., 2018). Another advantage of this dataset for our task is that it links documents according to a broader theme and does not rely on keywords and named entities. This will enable us to develop and test models that find connections across documents that go beyond matching metadata.

CLEF Ad-hoc Retrieval Datasets (ID number T4.1-6)

CLEF (Conference and Labs of the Evaluation Forum) is an organisation that organises tasks in CLIR. They have made datasets for some of the tasks from previous years available for download[8]. These datasets are especially important because they are the only ones identified that are suitable for ad-hoc retrieval and are standard datasets used by existing CLIR systems for evaluation and benchmarking. For statistics and description of these datasets from 2001 to 2003 for Finnish, see Litschko et al. (2018).

[6] `https://dumps.wikimedia.org/`

[7] `https://linguatools.org/tools/corpora/wikipedia-comparable-corpora/`

[8] See `http://direct.dei.unipd.it/`

Datasets such as CLEF and Wikipedia (see above) will also be useful when we want to submit papers to venues such as the European Conference on Information Retrieval (ECIR) where they recommend that systems are evaluated and benchmarked on standard publicly-available datasets[9].

YLE dataset (ID number T4.1-7)

YLE is the national broadcasting company of Finland. They have a datasets of news articles in Finnish and Swedish that have been written separately from January 2011 (January 2012 for Swedish) to December 2018. This dataset is publicly available through the Finnish Language Bank[10]. The dataset is annotated with named entities which can be used to link articles along with other metadata (such as publication date).

Event Registry[11]

EventRegistry[12] is a media intelligence platform, enabling users to gather and analyze current and past news content. There are several services offered within their Text Analytics tool:

- Semantic annotation: input text can be automatically semantically annotated by recognizing mentioned people, locations, organisations and recognizable things.
- Categorize text: the content can be categorized into a set of predefined categories. The categorization will return up to 5 categories together with their scores (relatedness of the category to your text). Currently it is possible to categorize into two category taxonomies:
  - DMOZ taxonomy, which is based on the DMOZ website. The categorization can be used only to categorize documents in the English language.
  - News taxonomy trained on 8 general news content categories (Business, Politics, Technology, Arts and Entertainment, Sports, Health, Science, and Environment).
- Semantic similarity: when observing two documents, it can be difficult to determine how semantically similar or different they are. By simply looking at the words appearing in the document, one can miss the fact that the same things can be expressed with different words or that the same word can have different meanings. If the content is in different languages, the problem is even harder. The available API endpoint allows the comparison of two documents, potentially in different languages, and compute the semantic similarity between them. Computation of the similarity will ignore the common words that don't convey meaning as well as take into account the fact that different words can have the same meaning (e.g. refugee = migrant, or VW = Volkswagen).
- Named entity recognition: the list of people, organisations locations, dates and currencies are detected in the input text.
- Sentiment analysis: detects the sentiment expressed in the provided input text. The API endpoint is able to extract sentiment for English language only.
- Extract article information given the URL: given the article URL, extract the available information about the article from the page (article title, body, source name, article links, etc).
- Detect language: detect the language of the input text.

The EventRegistry covers news in several project languages, including Slovene, Croatian, Finnish, Swedish, Russian, English, but it does not contain three EMBEDDIA languages: Estonian, Latvian and Lithuanian. As EventRegistry is a payable service, we tested the free version to identify the links of topical articles (selected topic of Brexit) and collected a small subset of articles in Slovenian and

[9] ECIR guidelines: https://irsg.bcs.org/proceedings/ECIR_Draft_Guidelines.pdf

[10] See https://www.kielipankki.fi/corpora/

[11] Since Event Registry is a tool and not a dataset, we did not assign it an ID number.

[12] See https://eventregistry.org/

Croatian regarding Brexit to check the dataset quality. At the time of writing this report, we haven't yet decided to pay for this service to gather data for the needs of EMBEDDIA.

3.2 Cross-lingual news summarisation and visualisation (task T4.2)

There are not many datasets for summarisation for languages other than English. We are primarily interested in news summarisation and therefore we focus on this domain. As a testing language pair for cross-lingual model transfer we selected English (resource-rich language) and Slovene (less-resourced language). We plan to add other project languages later. The intention of task T4.2 is to take a trained state-of-the-art news summarisation model for English, apply our cross-lingual transfer learning approach to that model, and use the resulting model on the target language (Slovene, in our case).

For this reason, we took a collection of news articles in Slovene (Bučar et al., 2018). This collection is the same as the sentiment analysis corpus SentiNews (see Section 3.3), containing Slovene news with political, business, economic, and financial content published between 1st of September 2007 and 31th of December 2013. The news were retrieved from five frequently read Slovene web media resources (*24ur*, *Dnevnik*, *Finance*, *RtvSlo*, and *Žurnal*).

In the extracted dataset, every news instance in a tab separated values format (tsv) includes the ID, title, summary, and the content of the news. The collection contains 217 532 documents (Finance: 110 841, Dnevnik: 47 684, Žurnal: 39 886, Rtvslo: 10 450, and 24ur: 8 671).

Regarding the Cross-Language Text Summarisation (CLTS) task, most works analyzed the generation of Chinese-to-English and English-to-Chinese cross-lingual summaries (Wan et al. (2010); Yao et al. (2015a,b); J. Zhang et al. (2016); Wan et al. (2018)). Unfortunately, there are few available datasets for CLTS. The MultiLing Pilot 2011 dataset (Giannakopoulos et al. (2011)) is derived from publicly available WikiNews English texts. This dataset is composed of 10 topics, each topic having 10 source texts and 3 reference summaries. Each reference summary contains a maximum of 250 words. Native speakers translated this dataset into Arabic, Czech, French, Greek, Hebrew and Hindi languages. Despite the absence of the languages of the EMBBEDIA project, this dataset can provide an initial analysis of cross-lingual approaches.

For visualization, no specific public datasets have been identified for this deliverable, but the visualisation will support all other tasks, and therefore all the datasets listed in this deliverable are relevant. The bilingual dataset with ID number T4.3-5 has been identified as a good candidate for creating specific datasets for cross-lingual visualisation.

3.3 Cross-lingual identification of viewpoints and sentiment in news reporting (task T4.3)

In T4.3, we plan to conduct an extensive study on how different viewpoints on a number of subjects and topics can be automatically identified in news articles from different sources and how these viewpoints change and develop through time. We also want to analyze the news from the perspective of the perceived sentiment in various news content. We plan to design tools for embeddings-based exploratory analysis of news content, as well as levaraging the explainability of classification models in predictive analysis.

For news sentiment analysis, manually annotated training sets in at least one language are required for training machine learning models for the tasks at hand, while test sets in multiple languages are needed to support the cross-lingual news sentiment analysis. We should point out that while several benchmark sentiment analysis datasets already exist for English (e.g., the IMDB dataset of movie reviews (Maas et al., 2011), the Stanford Sentiment Treebank (SST) dataset of phrases (Socher et al., 2013), the Yelp review dataset (X. Zhang et al., 2015), etc. none of them contains news articles. Generally, the news articles tend to be more objective than the above mentioned benchmark datasets of reviews, and the general sentiment datasets are not necessarily appropriate for news sentiment analysis model training and testing. Therefore, here we only focus on sentiment annotated datasets containing news articles that are far more scarce than the general sentiment analysis datasets.

For viewpoint analysis, we will consider analysing concept level changes across time (e.g., using dynamic word embeddings), differences in news reporting across different news sources, or identifying specific aspects (e.g., detecting subjectivity, fake news, fact checking). For exploratory tasks, cross-lingual embeddings can support analysis beyond single language, while in predictive tasks, training sets in at least one language are required for training machine learning models, and test sets in multiple languages allow for testing the transferability of the implemented approaches across different languages.

In this deliverable, we focus on datasets supporting viewpoints analysis in news articles, therefore the datasets with annotated viewpoints which do not contain news articles (e.g., a tweet dataset with annotated viewpoints used in the SemEval-2016 Task 6: Detecting Stance in Tweets (Mohammad et al., 2016) are excluded.

Datasets useful for sentiment analysis and viewpoints are listed in Table 4. All the datasets are described separately below.

Table 4: Datasets for sentiment analysis and viewpoints.

ID number	Dataset	Publicly availability	Language	Annotated (YES/NO)	Annotation procedure (Manual/Lexicon-based)
T4.3.-1	SentiNews dataset	Publicly available	Slovene	YES	Manual ; 5 categories (from very negative to very positive)
T4.3.-2	SemEval 2017 Task 5 - Subtask 2 "Fine-Grained Sentiment Analysis on Financial News Headlines"	Publicly available	English	YES	Manual (sentiment); number between -1 and 1
T4.3.-3	NYSK dataset	Publicly available	English	NO	/
T4.3.-4	The New York Times Annotated Corpus	Publicly available	English	Only small subset for sentiment (but not available) categories	
T4.3.-5	Dataset of Slovene and English crawled news	Only for project partners	Slovene	YES	Automatically
T4.3.-6	Webhose's Free Online Datasets	Publicly available	English, Russian, Swedish	NO	/
T4.3.-7	Emergent	Publicly available	English	YES	Manually annotated
T4.3.-8	Sports articles for objectivity analysis dataset	Publicly available	English	Yes, with the subjectivity/objectivity labels	Manually annotated
T4.3.-9	CNN and FOX News	Publicly available	English	YES	/
T4.3.-10	Kenyan presidential and parliamentary elections	Only for project partners	English	YES	Source (local, Western)
T4.3.-11	NELA2017	Publicly available	English	YES	/
T4.3.-12	Webis Bias Flipper 2018	Publicly available	English	YES	bias
T4.3.-13	PolNeAR v.1.0.0	Publicly available	English	YES	attributions
T4.3.-14	LIAR	Publicly available	English	YES	truthfulness

SentiNews dataset (ID number T4.3-1)

SentiNews[13] (Bučar et al., 2018) is a manually sentiment annotated Slovenian news corpus. The dataset contains Slovenian 10 427 news texts from Slovenian news portals (www.24ur.com, www.dnevnik.si, www.finance.si, www.rtvslo.si, www.zurnal24.si) which were published between 1th of September 2007 and 31th of December 2013. These texts have been annotated by 2 to 6 annotators using the five-level

[13] The dataset is available at https://www.clarin.si/repository/xmlui/handle/11356/1110

Likert scale on three levels of granularity, i.e. on the document, paragraph, and sentence level. This item is publicly available and licensed under the Creative Commons - Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)[14].

Description of the annotation procedure:

- between 2 and 6 annotators independently annotated sentiments of a stratified random sample of 10 427 documents from the Slovenian news portals. The texts were annotated using the five-level Likert scale (1 – very negative, 2 – negative, 3 – neutral, 4 – positive, and 5 – very positive) on three levels of granularity, i.e. on the document, paragraph, and sentence level;
- 6 annotators, it took more than a year to manually annotate the sample;
- instructions to annotators: "Please specify the sentiment from the perspective of an average Slovenian web user. How did you feel after reading the news?"; five-level Likert scale;
- levels of granularity: document level, paragraph level, sentence level
- sentiment allocation: negative (if average of scores less or even 2.4); neutral (if average of scores is between 2.4 and 3.6); positive (average of annotated scores over 3.6)

This dataset has been identified as the most appropriate for cross-lingual sentiment analysis, and building test set for Croatian has already begun (see Appendix C).

SemEval 2017 Task 5 - Subtask 2 "Fine-Grained Sentiment Analysis on Financial News Headlines" (ID number T4.3-2)

As part of the SemEval-2017 competition, a shared task "Fine-Grained Sentiment Analysis on Financial Microblogs and News" (Cortis et al., 2017) was organised. This task contained two tracks, the first one concerning microblog messages and the second one covering news statements and headlines. In the latter track the goal was to predict the sentiment score for each of the mentioned companies/stocks. To evaluate the participating systems, a dataset containing a collection of financially relevant news headlines, which have been annotated for fine-grained sentiment, was created[15].

It is licensed under the Creative Commons Attribution-Non-Commercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) license. When using this dataset it is required to cite the SemEval 2017 task description paper for task 5.

NYSK dataset (ID number T4.3-3)

NYSK[16][17] (New York v. Strauss-Kahn) is a collection of English news articles about the case relating to allegations of sexual assault against the former IMF director Dominique Strauss-Kahn (May 2011). Documents are first obtained via a Web search using AMI EI: an integrated platform for delivering enterprise intelligence, developed by AMI Software[18] with the following query: "dsk" OR "strauss-kahn" OR "strauss-khan". Documents are filtered and presented in XML format. NYSK dataset was used to extract topic-sentiment correlation and evolution over time (Dermouche et al., 2014, 2015) but may be used for other text mining tasks (e.g., topic or sentiment analysis).

The New York Times Annotated Corpus (ID number T4.3-4)

The New York Times Annotated Corpus[19] contains articles written and published by the New York Times between 1st of January 1987 and 19th of June 2007. All in all, it contains 1.8 million articles, 650 000

[14] https://creativecommons.org/licenses/by-sa/4.0/

[15] The dataset is available at https://bitbucket.org/ssix-project/semeval-2017-task-5-subtask-2/overview

[16] This dataset is not labelled with specific sentiment labels or labels related to viewpoints.

[17] The dataset is available at: http://archive.ics.uci.edu/ml/datasets/NYSK?ref=datanews.io

[18] See https://www.bertin-it.com/en/when-the-speech-to-text-learns-the-language-of-trading-rooms/

[19] The dataset is available at https://catalog.ldc.upenn.edu/LDC2008T19

out of them being manually summarized by a staff of library scientists and 1.5 million out a them are tagged for persons, places, organisations, titles and topics.

Additionally, the dataset was used for sentiment analysis in the study "CS224N Final Project: Sentiment analysis of news articles for financial signal prediction[20]. For evaluation of their sentiment classifier, the authors manually labelled the subset containing articles from January and June 2006 with positive, neutral, and negative sentiment classes.

The entire dataset is not freely available, but one could pay the fee if the authors of the above mentioned study were prepared to share their manual annotations of the dataset.

The text in this corpus is formatted with the News Industry Text Format (NITF) developed by the International Press Telecommunications Council, an independent association of news agencies and publishers. NITF is an XML specification that provides a standardized representation for the content and structure of discrete news articles. NITF encompasses structural markup such as bylines, headlines and paragraphs. The format also provides management attributes for categorizing articles into topics, summarisation, usage restrictions, and revision histories. The goals of NITF are to answer the essential questions inherent in news articles:

- Who: who owns the copyright, who has rights to republish the article, and who the article is about.
- What: the subjects reported, the named entities inside the article, and the events it describes.
- When: when the article was written, when it was issued, and when it was revised.
- Where: where the article was written, where the events took place, and where it was delivered.
- Why: the metadata describing the newsworthiness of the article.

The dataset sample is shown in Figure 2 in Appendix B.

Dataset of Slovene and English crawled news (ID number T4.3-5)

The structure of Slovene and English news datasets is the same as of the Slovene and English crawled news described in Subsection 3.1 under the subtitle "Dataset of Slovene and English crawled news (ID number T4.1-4)".

Articles in the English dataset are labeled by lexicon-based approach and can be used for sentiment-based news analysis. The lexicon consists of a predefined set of positive and negative English words. Articles in the Slovene dataset were labeled automatically with sentiment and can be used for sentiment analysis. The sentiment classifier was trained by linear SVM from 10 000 manually annotated articles. The sentiment labels can serve for analysis of sentiments and topics, while news source infromation can serve for analysis of viewpoints.

Webhose's Free Online Datasets (ID number T4.3-6)

These datasets[21] contain popular news articles on various topics: sports, finances, travel, entertainment, etc. There are news available in three languages which are of interest for the EMBEDDIA project: English, Russian and Swedish. The dataset is not annotated and the files are in the JSON format.

Emergent (ID number T4.3-7)

The dataset[22] served a digital journalism project which dealt with rumour debunking (Ferreira & Vlachos, 2016). It contains 300 rumoured claims and 2 595 associated news articles, collected and labelled by journalists. The rumors were collected from a variety of sources such as rumour sites and social

[20] For more information, see https://nlp.stanford.edu/courses/cs224n/2011/reports/nccohen-aatreya-jameszjj.pdf

[21] The datasets are available at https://webhose.io/free-datasets/

[22] The dataset is available at https://drive.google.com/drive/folders/0BwPdBcatuO0vYTAxSnA1d09qdGM

media. Each rumor is manually classified into three classes (true, false, or unverified), according to the estimation of its veracity. For each rumor, the journalists searched for articles in which the rumors are mentioned. The associated articles were manually annotated according to its stance on the rumor into three classes:

- **for**: The article states that the rumor is true
- **against**: The article states that the rumor is false
- **observing**: The rumor is reported in the article without assessment of its veracity

Sports articles for objectivity analysis dataset (ID number T4.3-8)

The dataset[23] contains 1000 sports articles which were labelled using Amazon Mechanical Turk as being objective or subjective. The dataset was made public 9th of April 2018. The dataset contains raw texts, extracted features, and the URLs from which the articles were retrieved. Some of the features were extracted using the Stanford POS tagger and the tags are as defined in Penn Treebank Project.

CNN and FOX News (ID number T4.3-9)

The dataset[24] contains web news from CNN and FOX. The news articles originate from 1st of January 2014 to 4th of April 2014 in order to collect altogether 3630 news articles from media houses from two opposites of the political spectrum (Qian & Zhai, 2014). Each instance in the dataset contains a title, abstract and body of the article. Some of the instances contain images that were appended to the news articles. This dataset is planned for viewpoint classification models.

Kenyan presidential and parliamentary elections (ID number T4.3-10)

The corpus presented by Pollak et al. (2011)[25] contains 464 articles, (about 320 000 words) concerning Kenyan presidential and parliamentary elections, held on 27th December 2007, and the crisis following the elections. The corpus was originally collected by the IPrA Research Center, University of Antwerp, and the documents originate from six different daily newspapers in English, covering the time period from 22nd December 2007 to 29th February 2008. Articles in the corpus are classified into two classes according to the origin of the news organisation that produced the article. The articles from the US and British press, i.e. The New York Times, The Washington Post, The Independent, The Times and Post, belong to the class “Western” and articles from local Kenyan newspapers Daily Nation and The Standard are categorized as “Local”. Each class contains 232 documents. The corpus can be used for the viewpoints analysis, allowing the comparison between local (Kenyan) news and foreign (US and British) news viewpoints on the subject of Kenyan elections.

NEws LAndscape (NELA2017) (ID number T4.3-11)

NEws LAndscape (NELA2017) dataset[26] (Horne et al., 2018) contains over 136 000 political news articles from 92 news sources which were collected over 7 months in 2017. News sources were manually chosen to include well-established and mainstream sources, maliciously fake sources, satire sources, and hyper-partisan political blogs.

Additionally, for each article 130 content-based and social media engagement features were computed and can be used for bias and viewpoints classification. These features are generated according to 7 aspects of the article, namely its structure, complexity, sentiment, bias, morality, topic, and engagement.

[23] The dataset are available at `https://archive.ics.uci.edu/ml/datasets/Sports+articles+for+objectivity+analysis`

[24] The dataset are available at `https://sites.google.com/site/qianmingjie/home/datasets/cnn-and-fox-news`

[25] The dataset has not been publicly released, but one of the authors is part of the consortium and has access to the dataset.

[26] The dataset are available at `https://dataverse.harvard.edu/dataverse/nela`

Webis Bias Flipper 2018 (ID number T4.3-12)

The Webis Bias Flipper 2018 dataset[27] (Ajjour et al., 2018) contains news articles about 2781 events from allsides.com, ranging from June 1st, 2012 to February 10, 2018. For each event, a number of news portals with opposite political biases were crawled in order to retrieve headlines and the content of the articles covering the event. The dataset altogether contains 6458 news articles labelled according to their bias (left- or right-oriented news source).

PolNeAR v.1.0.0 – Political News Attribution Relations Corpus (ID number T4.3-13)

PolNeAR (Newell et al., 2018) is a corpus[28] of news articles with annotated attributions, where the attribution is understood as a citation of a statement or a description of the internal state (e.g., thoughts, intentions) of some person or group. The dataset contains 1008 articles which cover the campaigns of candidates in the US General Election on 8 Nov 2016 and only articles that contain attribution to at least one of the candidates, Donald Trump or Hillary Clinton, are included. The news articles come from 7 US national news publishers: Huffington Post, Breitbart, New York Times, Politico, Washington Post, Western Journalism and USA Today.

The publicly available dataset contains timestamp annotations and can therefore be used for the temporal bias and viewpoint analysis.

LIAR - a dataset for fake news detection (ID number T4.3-14)

The LIAR dataset[29] (Wang, 2017) contains 12 836 manually labeled short statements from POLITIFACT .COM and can be used for fact-checking and viewpoints research. Each statement is labeled by the POLITIFACT.COM editor for truthfulness, subject, context/venue, speaker, state, party, and prior history. By the number of statements and a time span of a decade, this is one of the largest publicly available corpus for fact-checking and viewpoints detection. We should point out that the majority of statements in the corpus do not originate from the news articles, but come from a variety of other contexts, such as political debates, TV ads, Facebook posts, tweets and political interviews.

4 Evaluation methods and metrics

In this section we describe evaluation methods and metrics which will be used within all WP4 tasks. We present below the methods and metrics per each WP4 task.

4.1 Evaluation methods and metrics for task T4.1

The task of cross-lingual news linking can be considered a cross-lingual document retrieval task (CLDR), where the query is given in one language and the expected results are documents in another language. In EMBEDDIA, the query can be a document (news article) in one language and the results are documents in multiple languages (including the query language). CLDR can be a 'known-search' retrieval task where there is exactly one 'correct' document for each query or an 'ad-hoc' retrieval task where there are multiple relevant documents expected for each query. We focus on the ad-hoc task because in news linking several related news articles are expected.

There are several evaluation metrics used for CLDR that we can use to measure the performance of models and compare them with state-of-the-art CLDR models (Balikas et al., 2018; Josifoski et al.,

[27] The dataset is available at https://webis.de/data/webis-bias-flipper-18.html

[28] The dataset is available at https://github.com/networkdynamics/PolNeAR

[29] The dataset is available at https://www.cs.ucsb.edu/~william/data/liar_dataset.zip

2019). The CLDR evaluation measures for the task of cross-lingual document retrieval (Josifoski et al., 2019; Balikas et al., 2018; Litschko et al., 2019, 2018) include mean average precision (MAP), mean reciprocal rank (MRR), and precision at k.

- Precision at k
  In information retrieval, a precision is the ratio of the number of relevant documents divided by the number of documents returned by the system (returned documents).

$$precision = \frac{|relevant \quad documents|}{|returned \quad documents|} \tag{1}$$

  In a ranking task, we are interested in precision at rank k. This means that only documents ranked higher than k are considered and the rest are disregarded.

- Mean average precision (MAP) is computed by averaging the precision at k for all queries.
- Mean reciprocal rank (MRR)
  In a known-search task, MRR is computed by getting the reciprocal rank of the correct response document for each query i in the query set Q, and averaging over all $|Q|$ queries.

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \tag{2}$$

To obtain reliable statistics on model performance, n-fold cross-validation will be used, where a dataset is split into n parts (folds) and each fold becomes once the test set while the rest are used as the training set. The metrics are then averaged across the n folds.

4.2 Evaluation methods and metrics for task T4.2

Evaluation of the cross-lingual news summarisation and visualization technology to be developed in task T4.2 will follow the guidelines established by the nested model of visualization design and validation (Munzner, 2009). To establish the validity of the developed technologies evaluation must be completed at four levels of abstraction: domain problem characterization, data abstraction, encoding/interaction choice, and algorithm/implementation.

At each level of abstraction, threats to the validity of the design must be mitigated. We will use distinct evaluation techniques to deal with these threats at each level.

At the lowest level of abstraction we must evaluate the validity of algorithm/implementation. The evaluation should determine if the design has been realized as an efficient algorithm and implementation. Complexity analysis and instrumentation of the system to measure time and memory statistics are common evaluation techniques at this level. These evaluations can be periodically rerun during the iterative design of the technology.

At the visual encoding and interaction level the perceptual efficiency and user experience are evaluated. The design rational and encoding choices will first be evaluated using a structured comparison with existing systems and techniques. The identified goals and tasks will be discussed with regard to the choice of visual variables for each data type in the collection of systems under investigation (Bertin, 1983; Mackinlay, 1986). Prototype designs are evaluated by user pilot studies or qualitative image analysis. Once the design is finalized and implemented a laboratory study will run to evaluate the error rate and time taken for relevant tasks.

Before creating the technology it is important to address the validity of the design at first two levels of the nested model. The domain problem characterisation should identify and involve real users of the finished technologies. Requirements gathering, structured interviews and observational research are all techniques used to evaluate if proposed designs will solve real problems faced by users. Adoption rates are a down stream measure of validity at this level.

Finally, once valid prototypes have been created and deployed, the translation between domain specific issues to general visualization techniques can be evaluated. Field studies to monitoring usage patterns, and user interviews or questioners will be used to draw conclusions about the contributions of the developed technologies.

We will also consider commonly used evaluation approaches for summarisation, such as BLEU and ROUGE. BLEU (bilingual evaluation understudy) is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another (Papineni et al., 2002). BLEU uses a modified form of precision to compare a candidate translation against multiple reference translations. It was one of the first metrics to claim a high correlation with human judgements of quality, but there is no guarantee that an increase in BLEU score is an indicator of improved translation quality (Lin & Och, 2004). ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a modification of BLEU that focuses on recall rather than precision (Lin, 2004). The metrics compare an automatically produced summary or translation against a reference or a set of references (human-produced) summary or translation. However, this metric also has some drawbacks, such as: Relative perfect scores are highly diverseand unattainable by humans, 100 % perfect scores are impossible forhigher quality datasets, etc. (Schluter, 2017).

Both BLUE and ROUGE are widely used metrics, but they have shortcomings. For this reason, we will consider them as a help to automate early evaluation, compare versions quickly, etc. We will still consider human user evaluation as a very important part in summarisation task.

4.3 Evaluation methods and metrics for task T4.3

The comparison and evaluation of the sentiment analysis methods in a multilingual setting belongs to the extrinsic evaluation of cross-lingual embeddings (see deliverable D1.1 for a detailed explanation of intrinsic and extrinsic evaluation of embeddings).

Regardless of the cross-lingual context, the choice of a sentiment analysis method highly depends on the data and the intended application. Review of Ribeiro et al. (2016) reveals different performance of several sentiment analysis methods when applied on different texts: product reviews, tweets, blogs, user comments. We are concerned with the news articles and therefore, we will focus comparison of the developed methods solely to the news texts.

We will evaluate the performance of the sentiment analysis methods developed within the EMBEDDIA project on the manually annotated datasets and compare the algorithm performance with the human judgement scores of the sentiments in texts.

The main dataset, which will be used for training of the sentiment models, is the SentiNews dataset of (Bučar et al., 2018). This dataset contains various news articles written in Slovene language. To test how the cross-lingual embedding models work for the classification of sentiments, we will also prepare media partners will provide manually annotated datasets. These datasets will contain at least 1000 news articles and will be annotated in a similar manner as the SentiNews dataset of Bučar et al. (2018) was annotated. Details on the annotation procedure are described in Appendix C

Presently, annotation procedure is taking place in Croatia, organised by STY/Trikoder. Recruitment and training of six annotators was already performed and annotation process of 2025 articles in Croatian language will finish by M10.

Since the SentiNews dataset of Bučar et al. (2018) is our starting point, we will follow the example from their study and reduce the labels from the 5-point Likert scale (very negative, negative, neutral, positive, very positive) to three categories: positive, neutral and negative.

Table 5: Confusion matrix for experiments with three classes

		Actual		
		Positive	Neutral	Negative
Predicted	Positive	a	b	c
	Neutral	d	e	f
	Negative	g	h	i

We consider that we have 3-class classification task, so we will use traditional Precision, Recall, and F1 measures for the automated classification. Each letter in Table 5 represents the number of instances which are actually in class X and predicted as class Y, where X;Y ∈ {positive;neutral;negative}. The recall (R) of a class X is the ratio of the number of elements correctly classified as X to the number of known elements in class X. Precision(P) of a class X is the ratio of the number of elements classified correctly as X to the total predicted as the class X. For example, the precision of the negative class is computed as: P(neg)=i/(c+f+i); its recall, as: R(neg)=i/(g+h+i). F1 measure is the harmonic mean between both precision and recall. In the case of negative class, F1 (neg) is calculated as F1(neg)=2*P(neg)*R(neg)/(P(neg)+R(neg)).

We will also compute the overall accuracy as: A=(a+e+i)/(a+b+c+d+e+f+g+h+i). It considers equally important the correct classification of each sentence, independently of the class and measures the method capability to predict the correct output.

When the classes are imbalanced, we will consider Macro-F1, which is a variation of F1 and is used to evaluate classification effectiveness on skewed datasets. Macro-F1 enables performance evaluation for the smaller classes. Macro-F1 values are computed by first calculating F1 values for each class separately, as described above for the negative class, and then averaging over all classes. Thus, accuracy and Macro-F1 provide complementary assessments of the classification effectiveness.

Comparison with other datasets, which have continuos scores, such as SemEval challenges will be based on cosine similarity (Ghosh et al., 2015), as orifinally. If the sentiment scores to be predicted by systems lie on a continuous scale between minimal and maximal value, cosine enables comparison of the proximity between gold standard and predicted results (conceptualized as vectors), while not requiring exact correspondence between the gold and predicted score for a given instance. E.g., an instance is a message or headline which can include several entities (companies or cashtags). The cosine similarity is calculated according to the equation:

$$cosine(G,P) = \frac{\sum_{i=1}^{n} G_i \times P_i}{\sqrt{\sum_{i=1}^{n} G_i^2} \times \sqrt{\sum_{i=1}^{n} P_i^2}}$$

where G is the vector of gold standard scores and P is the vector of corresponding scores predicted by the system.

In Cortis et al. (2017) modifications of the standard cosine approach were proposed during the competition with the reason that cosine similarity treats all predicted scores with the same weight. The details of the modified formula are available in "Section 6 Alternative Evaluation Metric" of the paper Cortis et al. (2017). We will take also the modified approaches into account.

5 Conclusions and further work

In this report we presented the resources collected in order to build the tools and evaluate the results in the tasks dedicated to cross-lingual content analysis within WP4 of the EMBEDDIA project. The deliverable describes the datasets provided by the media partners and other available resources, which are either public or can be shared within the consortium. The current collection is large, but we expect that it will be further extended during the course of the project. The report also presented the benchmarks and the evaluation methods and metrics for the WP4 tasks.

During the course of the project, the news texts will be segmented, tokenized, part-of-speech (POS) tagged, and enriched with annotation approaches developed/used in Tasks T2.1, T2.2, and other task specific annotations. The tools developed in WP4 will be evaluated on the benchmarks of Section 3 and using the methods and metrics described in Section 4, as well as in terms of user experience to be assessed in WP6. The results of these activities and additional resources will be also reported in the follow-up deliverable D4.8.

Bibliography

Ajjour, Y., Wachsmuth, H., Kiesel, D., Riehmann, P., Fan, F., Castiglia, G., ... Stein, B. (2018, November). Visualization of the topic space of argument search results in args.me. In *2018 conference on empirical methods in natural language processing (emnlp 2018) - system demonstrations* (p. 60-65). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/D18-2011

Balikas, G., Laclau, C., Redko, I., & Amini, M.-R. (2018). Cross-lingual document retrieval using regularized wasserstein distance. In *European conference on information retrieval* (pp. 398–410).

Bertin, J. (1983). *Semiology of graphics*. University of Wisconsin Press.

Bučar, J., Žnidaršič, M., & Povh, J. (2018, September). Annotated news corpora and a lexicon for sentiment analysis in slovene. *Lang. Resour. Eval.*, *52*(3), 895–919. Retrieved from https://doi.org/10.1007/s10579-018-9413-3 doi: 10.1007/s10579-018-9413-3

Cortis, K., Freitas, A., Daudert, T., Huerlimann, M., Zarrouk, M., Handschuh, S., & Davis, B. (2017, August). SemEval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)* (pp. 519–535). Vancouver, Canada: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/S17-2089 doi: 10.18653/v1/S17-2089

Dermouche, M., Kouas, L., Velcin, J., & Loudcher, S. (2015). A joint model for topic-sentiment modeling from text. In *Proceedings of the 30th annual acm symposium on applied computing* (p. 819-824). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2695664.2695726 doi: 10.1145/2695664.2695726

Dermouche, M., Velcin, J., Khouas, L., & Loudcher, S. (2014, dec). A joint model for topic-sentiment evolution over time. In *2014 ieee international conference on data mining (icdm)* (p. 773-778). Los Alamitos, CA, USA: IEEE Computer Society. Retrieved from https://doi.ieeecomputersociety.org/10.1109/ICDM.2014.82 doi: 10.1109/ICDM.2014.82

Ferreira, W., & Vlachos, A. (2016). Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 1163–1168).

Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Barnden, J., & Reyes, A. (2015, June). SemEval-2015 task 11: Sentiment analysis of figurative language in twitter. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)* (pp. 470–478). Denver, Colorado: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/S15-2080 doi: 10.18653/v1/S15-2080

Giannakopoulos, G., El-Haj, M., Favre, B., Litvak, M., Steinberger, J., & Varma, V. (2011). TAC2011 multiling pilot overview. In *4th text analysis conference TAC.*

Horne, B. D., Khedr, S., & Adali, S. (2018). Sampling the news producers: A large news and feature data set for the study of the complex media landscape. In *Twelfth international aaai conference on web and social media.*

Huang, S., Soboroff, I., & Harman, D. (2018). Trec 2018 news track. *NewsIR@ ECIR*, *2079*, 57–59.

Josifoski, M., Paskov, I. S., Paskov, H. S., Jaggi, M., & West, R. (2019). Crosslingual document embedding as reduced-rank ridge regression. In *Proceedings of the twelfth acm international conference on web search and data mining* (pp. 744–752).

Kralj Novak, P., Grčar, M., Sluban, B., & Mozetič, I. (2015). Analysis of financial news with newsstream.. Retrieved from https://arxiv.org/abs/1508.00027v2

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Sage.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *140*, 1-55.

Lin, C.-Y. (2004, July). ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out* (p. 74-81). Barcelona, Spain: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W04-1013

Lin, C.-Y., & Och, F. J. (2004, July). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd annual meeting of the association for computational linguistics (ACL-04)* (p. 605-612). Barcelona, Spain. Retrieved from https://www.aclweb.org/anthology/P04-1077 doi: 10.3115/1218955.1219032

Litschko, R., Glavaš, G., Ponzetto, S. P., & Vulić, I. (2018). Unsupervised cross-lingual information retrieval using monolingual data only. In *The 41st international acm sigir conference on research & development in information retrieval* (pp. 1253–1256).

Litschko, R., Glavaš, G., Vulic, I., & Dietz, L. (2019). Evaluating resource-lean cross-lingual embedding models in unsupervised retrieval. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval* (pp. 1109–1112).

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1* (pp. 142–150).

Mackinlay, J. (1986, April). Automating the design of graphical presentations of relational information. *ACM Trans. Graph.*, *5*(2), 110–141. Retrieved from http://doi.acm.org/10.1145/22949.22950 doi: 10.1145/22949.22950

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016). Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (semeval-2016)* (pp. 31–41).

Munzner, T. (2009, Nov). A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics*, *15*(6), 921-928. doi: 10.1109/TVCG.2009.111

Newell, E., Margolin, D., & Ruths, D. (2018, May). An attribution relations corpus for political news. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC-2018).* Miyazaki, Japan: European Languages Resources Association (ELRA). Retrieved from https://www.aclweb.org/anthology/L18-1524

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics* (p. 311-318). Philadelphia, Pennsylvania, USA: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P02-1040 doi: 10.3115/1073083.1073135

Pollak, S., Coesemans, R., Daelemans, W., & Lavrač, N. (2011). Detecting contrast patterns in newspaper articles by combining discourse analysis and text mining. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)*, *21*(4), 647–683.

Qian, M., & Zhai, C. (2014). Unsupervised feature selection for multi-view clustering on text-image web news data. In *Proceedings of the 23rd acm international conference on conference on information and knowledge management* (pp. 1963–1966).

Ribeiro, F. N., Araújo, M., Gonçalves, P., André Gonçalves, M., & Benevenuto, F. (2016). Sentibench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, *5*(1), 23. Retrieved from `https://doi.org/10.1140/epjds/s13688-016-0085-1` doi: 10.1140/epjds/s13688-016-0085-1

Schluter, N. (2017, April). The limits of automatic summarisation according to ROUGE. In *Proceedings of the 15th conference of the European chapter of the association for computational linguistics: Volume 2, short papers* (p. 41-45). Valencia, Spain: Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/E17-2007`

Sluban, B., & Grčar, M. (2013). Url tree: efficient unsupervised content extraction from streams of web documents. In *Proceedings of the 22nd acm international conference on conference on information &#38; knowledge management* (pp. 2267–2272). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/2505515.2505654` doi: 10.1145/2505515.2505654

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1631–1642).

Wan, X., Li, H., & Xiao, J. (2010). Cross-language document summarization based on machine translation quality prediction. In *Acl* (pp. 917–926).

Wan, X., Luo, F., Sun, X., Huang, S., & Yao, J.-g. (2018, Jan 17). Cross-language document summarization via extraction and ranking of multiple summaries. *Knowledge and Information Systems*. Retrieved from `https://doi.org/10.1007/s10115-018-1152-7` doi: 10.1007/s10115-018-1152-7

Wang, W. Y. (2017). " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Yao, J., Wan, X., & Xiao, J. (2015a). Compressive document summarization via sparse optimization. In *IJCAI* (pp. 1376–1382). AAAI Press.

Yao, J., Wan, X., & Xiao, J. (2015b). Phrase-based compressive cross-language summarization. In *EMNLP* (pp. 118–127).

Zhang, J., Zhou, Y., & Zong, C. (2016). Abstractive cross-language summarization via translation model enhanced predicate argument structure fusing. *IEEE/ACM Trans. Audio, Speech & Language Processing*, *24*(10), 1842–1853.

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. In *Advances in neural information processing systems* (pp. 649–657).

Appendix A - Metadata of media partners datasets

In the three tables below we outline the types and meaning of variables used in the media partners' datasets.

Table 6: Description of the Styria article dataset.

Variable	Variable type	Description
site	string	the URL of the web portal. The article can be accessed by joining site URL and article_id with a slug. For example to access the article with article_id 614684, you can get it on `www.24sata.hr/a-614684`
article_id	integer	the public id of the article on the new site
title	string	the title of the news article
content	string	the content of the news article
lead	string	a short introduction to the content
tags	string	the article tags which are set manually by the journalists with proposals from the recommender system developed for this purpose. Journalists can add new tags unknown to or missed by the recommender system. There can be zero or more tags, separated by the '|' character
section	string	the main section of the news portal where the article was posted (does not need to be set). The most frequent section is *Vijesti* (News)
subsection	string	the subsection of the section where the article was posted (does not need to be set). Each section can have multiple subsections
authors	string	the author(s) of the article, zero or more, separated by the '|' character. An article may not have the author due to a number of reasons, e.g. the author does not want to sign the article
published_from	datetime	the date when this article appeared on the portal; sometimes journalists write articles in advance, so the publish date can be much later than the *date_created*
date_created	datetime	the date when the article was originally written

Table 7: Description of the Ekspress Meedia article dataset.

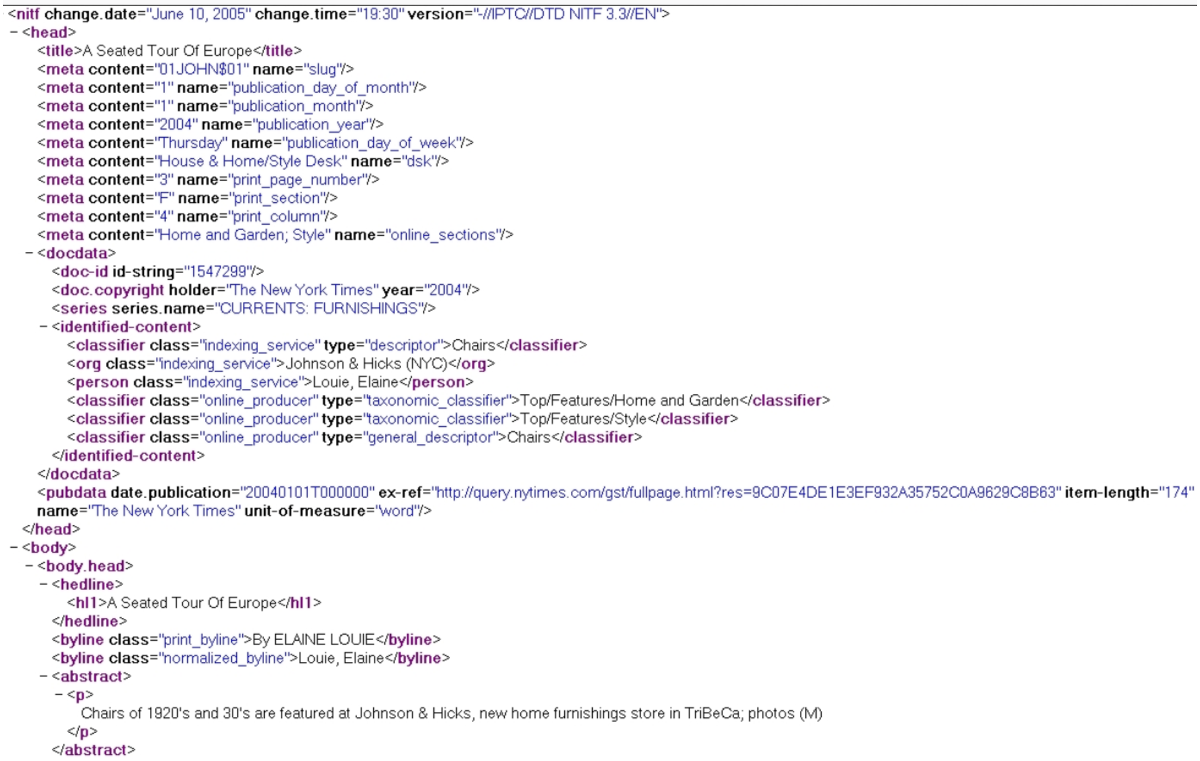
Variable	Variable type	Subvariable	Subvariable type	Description
id	integer	-	-	the ID of the article
title	string	-	-	the title of the article
lead	string	-	-	the lead of the article
url	string	-	-	the URL of the article
tags	list of dictionaries or None			each dictionary represents one tag
		domain_id	string	the ID of the domain
		id	string	the ID of the tag
		lang	string	the language of the tag
		tag	string	the tag itself
		translitted_name	string	modified version of the tag
rawBody	string	-	-	the raw text of the article
bodyText	string	-	-	clean article text (stripped from HTML)
publishDate	string	-	-	published date & time of the article
category-Primary	dictionary or empty list			Information about the category
		articleId	integer	the ID of the article
		categoryId	integer	the ID of the category
		categoryName	string	the name of the category (e.g. World)
		categoryPrimary	boolean	True if the category is primary, False if the category is not primary
		categoryUrl	string	the URL of the category
		categoryVisible	boolean	True if the category is visible online, False if the category is not visible online
		channelId	integer	the ID of the channel
		ChannelUrl	string	the URL of the channel
		directoryName	string	the name of the URL's directory
		parentId	integer	the ID of the parent category
channelLanguage	string or None	-	-	the language of the channel
categoryLanguage	integer or None	-	-	\textbf{the ID of the channel's language}
commentCount	integer	-	-	the number of comments
RelatedArticles	list of integers	-	-	a list of related articles' ID's

Table 8: Description of the STT article datasets for Finnish and Swedish.

Variable	Variable type	Variable description
title	string	caption of the image in the article
urgency	integer	refers to a priority given to the story. Default is 3, breaking news is 2 (or even 1)
contentCreated	datetime	the date of creating the content
contentModified	datetime	the date of modifying the content
altID	integer	the ID of the article
located	string	location of the certain city
subject	string	the department name (domestic, foreign, sport, etc.)
slugline	None	not in use
dateline	string	the location where the news is "dated"
description	string (can contain multiple variables)	short summary or caption of the article
headline	string	the headline of the article
creditline	string	byline author
genre	string	indicates the type of the article. For example 'Pääjuttu' refers to a head story
keyword	string	the name of the category, e.g "Onnettomuudet ja tuhot" ("Accidents and disasters")
body	string	the content of the article

Appendix B - Dataset samples

Example document for the New York Times Annotated Corpus (ID number T4.3-4) is presented below.



```
<nitf change.date="June 10, 2005" change.time="19:30" version="-//IPTC//DTD NITF 3.3//EN">
- <head>
    <title>A Seated Tour Of Europe</title>
    <meta content="01JOHN$01" name="slug"/>
    <meta content="1" name="publication_day_of_month"/>
    <meta content="1" name="publication_month"/>
    <meta content="2004" name="publication_year"/>
    <meta content="Thursday" name="publication_day_of_week"/>
    <meta content="House & Home/Style Desk" name="dsk"/>
    <meta content="3" name="print_page_number"/>
    <meta content="F" name="print_section"/>
    <meta content="4" name="print_column"/>
    <meta content="Home and Garden; Style" name="online_sections"/>
  - <docdata>
      <doc-id id-string="1547299"/>
      <doc.copyright holder="The New York Times" year="2004"/>
      <series series.name="CURRENTS: FURNISHINGS"/>
    - <identified-content>
        <classifier class="indexing_service" type="descriptor">Chairs</classifier>
        <org class="indexing_service">Johnson & Hicks (NYC)</org>
        <person class="indexing_service">Louie, Elaine</person>
        <classifier class="online_producer" type="taxonomic_classifier">Top/Features/Home and Garden</classifier>
        <classifier class="online_producer" type="taxonomic_classifier">Top/Features/Style</classifier>
        <classifier class="online_producer" type="general_descriptor">Chairs</classifier>
      </identified-content>
    </docdata>
    <pubdata date.publication="20040101T000000" ex-ref="http://query.nytimes.com/gst/fullpage.html?res=9C07E4DE1E3EF932A35752C0A9629C8B63" item-length="174"
    name="The New York Times" unit-of-measure="word"/>
  </head>
- <body>
  - <body.head>
    - <hedline>
        <hl1>A Seated Tour Of Europe</hl1>
      </hedline>
      <byline class="print_byline">By ELAINE LOUIE</byline>
      <byline class="normalized_byline">Louie, Elaine</byline>
    - <abstract>
      - <p>
          Chairs of 1920's and 30's are featured at Johnson & Hicks, new home furnishings store in TriBeCa; photos (M)
        </p>
      </abstract>
```

Figure 2: A sample of the NYT Annoted Corpus.

Appendix C - Annotation procedure for the Sentiment Analysis task (T4.3)

This section gives an overview of the annotation procedure, guidelines, timeline and the current status of annotations.

Annotation process

Media partners will annotate between 1000 and 2000 general news articles in their language, which will serve for testing the cross-lingual transfer of the models trained on manually sentiment annotated Slovenian news corpus SentiNews (Bučar et al., 2018). The number of the annotated articles depends on the media partner resources, but we have set the minimum number to 1000 articles. The articles should be of general, random topics and can be short or medium – not too long.

Below are the annotation procedure steps:

I Select preliminary candidate articles for a dataset[30].

Media partner will choose between 1500 and 2500[31] articles in the following way:

- A half of the articles should be from the period from 1st of September 2007 to 31st of December 2013 and the other half should be from the last 5 years. Media partners choose one article on every 50 or 100 articles. It is important that a media partner does not take, for example first 50 articles from each year, but to pay attention to the balanced distribution of articles throughout the whole year.

II Filter out articles using text length criterion.

JSI will check the length of texts and select at least 1000 articles of short and medium length. These texts will be then pre-processed and annotated by media partners. The annotation task will be performed on three levels:

a. Document-level annotation: media partner will annotate at least 1000 general news articles.

b. Paragraph and sentence-level annotation: media partner will annotate at least 1000 articles on a paragraph and sentence level.

III Check the quality of the data (cleaning and preprocessing).[32]

Once the dataset is identified, media partner will proceed to this step to ensure data quality. The duration of this step depends on the procedures to ensure quality of published texts that media partners implement in their daily routine.

During the project the validation datasets need to be annotated in a similar manner as the training datasets were annotated (Bučar et al., 2018), so we describe both annotation approaches for an easier overview. In Bučar's approach grammar and spelling errors were removed from the texts. In EMBEDDIA we propose a semi-automatic approach, i.e use a spell checker to detect grammar and spelling errors after which perform a manual correction of the words for which the spell checker marks as potentially incorrect.

IV Recruit native-speaker annotators.

[30] This step will be a collaboration between JSI and a media partner.

[31] This is a higher number than planned for annotation, but there will be filtering out in a next step

[32] This step is done both automatically and manually. Media partner will probably need at least 2 weeks to do both for 1500-2500 news articles. We have set the duration of this step to 21 days in the Section 1.4 Timeline to take into account any additional delays.

EMBEDDIA's approach follows the same criteria as in Bučar et al. (2018) in which the following were taken into account:

1. candidate's suitability for carrying out the task;
2. candidate's interest;
3. candidate's organisation;
4. gender and age equality.

Similarly as in the study of Bučar, EMBEDDIA's approach would also be to recruit 6 native-speaker annotators (3 women and men, aged 19– 30, from two different universities).

V Train recruiters in 2 phases:

PHASE 1:

EMBEDDIA's approach follows Bučar's approach. The first phase for the recruiters was to read basic guidelines for annotation and learn how to use the web application for annotating. Together with a referee, they annotated ten news articles on three levels, i.e. document, paragraph and sentence levels, and discussed individual instances. The process of sentiment annotation consisted of two subprocesses: *comprehension*, in which the annotator understands the content, and *sentiment judgment*, in which the annotator identifies the sentiment. Using a five-level Likert scale (Likert, 1932), the annotators specified the evoked sentiment with the use of the following instruction: "Please specify the sentiment from the perspective of an average Slovenian web user. How did you feel after reading this news?"

PHASE 2:

Bučar's approach in the second phase was for the recruiters to annotate 50 news items individually. The agreement among the annotators was then analysed. The instances with lower agreement were discussed and the issues were resolved by introducing additional annotation guidelines points.

EMBEDDIA's approach would be to do the annotation in different levels:

- **a.** document annotation level: in the second phase, each of the recruiters will annotate 25 news items individually. The agreement among the annotators will be analysed. Repeat this step in collaboration with JSI team who will calculate the agreement coefficients until agreement (Krippendorff's alpha coefficient (Krippendorff, 1980)) of 0.8 is reached.
- **b.** paragraph and sentence-level annotation: the procedure is the same as for the document level. The agreement of 0.8 (Krippendorff's alpha coefficient) within the paragraph and sentence level is much harder to achieve, so we will adjust a threshold after the Phase 1 of recruiters' training.

VI Annotation process (to be done by media partners):

In the study of Bučar et al. (2018) each annotator did not manually annotate all the items in the random sample of 10 427 news articles. Approximately 2000 documents per web medium were annotated on the three levels of granularity. Almost 9% of the articles in the sample was annotated by all the annotators, and slightly more than 70% by at least two of them. The sentiment of an instance was defined as the average of the sentiment scores given by the different annotators.

EMBEDDIA's approach would be to do the annotation in different levels:

- **a.** document-level annotation: in case of minimal number of 1000 news articles that need to be annotated each annotator will annotate 250 articles. 50 articles should be annotated by all 6 annotators. Among the rest of 950 articles, 700 articles were annotated by 1 annotator and 250 articles by 2 annotators. As a final result:
  - 50 news articles will be annotated by 6 annotators;
  - 250 news articles will be annotated by 2 annotators;

  - 700 articles will be annotated by 1 annotator.

  See Figure 3 for the distribution of 1000 news articles among the annotators. In case the number of annotated articles is higher than 1000, the above numbers will be proportionally adjusted.

b. paragraph and sentence-level annotation: same as document level.

VII Evaluation of the annotation process (to be done by JSI)

In Bučar's approach the correlation coefficients were used to determine inter-rater agreement on all three levels of granularity: Cronbach's alpha, Krippendorff's alpha, Fleiss' kappa and Kendall's coefficient of concordance (W) between the annotators, as well as the minima (min), maxima (max) and averages (avg) for the Pearson (rP) and Spearman (rS) correlation coefficients at the document, paragraph and sentence levels of granularity.

EMBEDDIA's approach would be to narrow down the estimation of inter-rater agreement and use only one correlation coefficients for any level of granularity: Krippendorff's alpha.

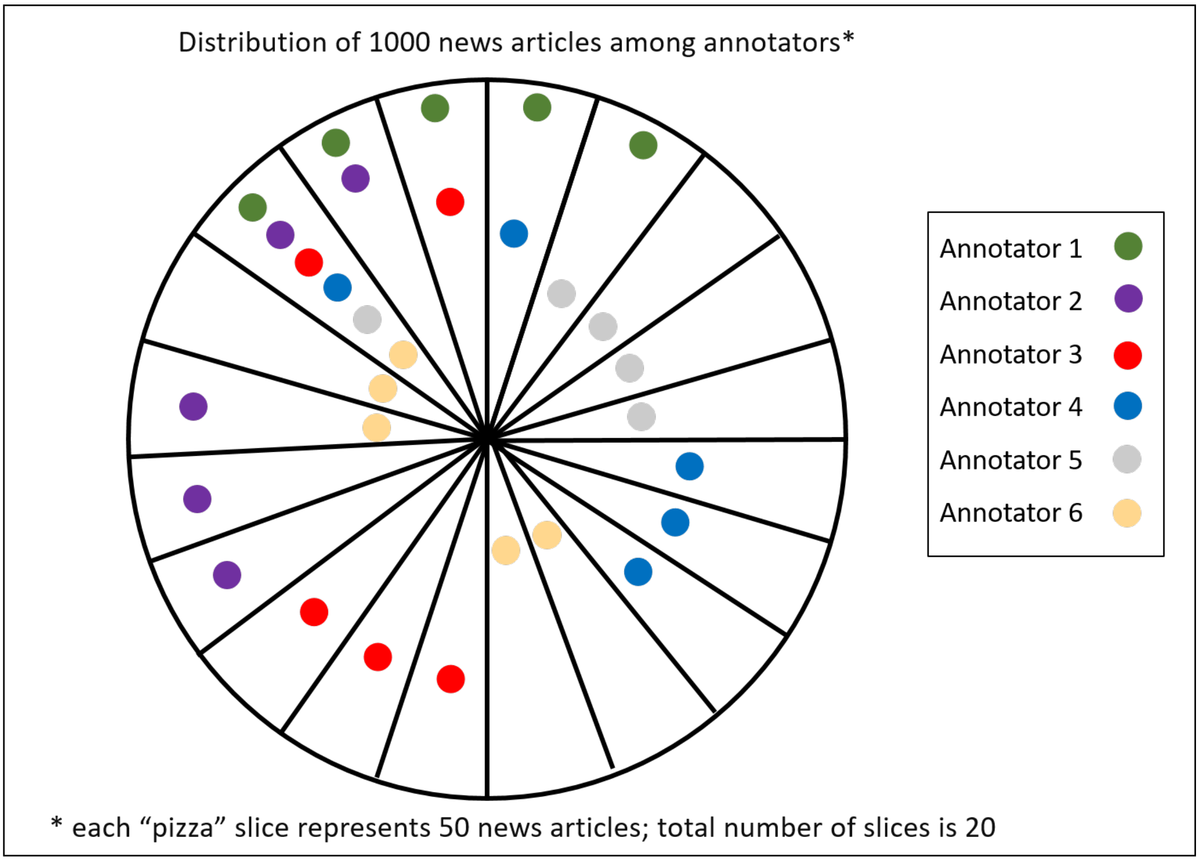

Distribution of 1000 news articles among annotators*
Annotator 1 ●
Annotator 2 ●
Annotator 3 ●
Annotator 4 ●
Annotator 5 ●
Annotator 6 ●
* each "pizza" slice represents 50 news articles; total number of slices is 20

Figure 3: Distribution of 1000 news articles among the annotators.

Timeline

The annotation process is expected to last maximum 70 working days (see Table 9), which translates to 14 weeks or 3.5 months (holidays excluded).

Table 9: Annotation timeline

Step	Institution responsible	Duration
Select preliminary candidate articles for a dataset	Media partner & JSI	5 days
Filter out articles using text length criterion	JSI	2 days
Check the quality of the data (cleaning and pre-processing)	Media partner	21 days
Recruit native-speaker annotators	Media partner	10 days
Train recruiters in 2-phases: PHASE 1	Media partner & JSI	1 day
Train recruiters in 2-phases: PHASE 2	Media partner & JSI	3 days
Annotation process	Media partner	25 days
Evaluation of the annotation process	JSI	3 days

Current status of annotations organised by STY/Trikoder

Media partner Styria Media Services (STY, Croatia) and Trikoder (TRI, Croatia) that continued the work of STY, provided 2025 articles for the purpose of Sentiment Analysis. The articles correspond to annotation specifications within the project (see Section Annotation process above), which means that the news articles are of the general and random topics; they are either short or medium – not too long; they are pre-processed (the texts are proofread for spelling and grammar errors).

Articles for the purpose of Sentiment Analysis are from the period from 12th of April 2007 to 28th June 2019. Out of 2025 articles, 1013 articles are from the period from 12th of April 2007 to 31th of December and the other 1012 are from the period 1st of January 2014 to 28th June 2019. The dataset has a balanced distribution of articles throughout the whole period.

Recruit native-speaker annotators.

Trikoder recruited 11 native-speaker annotators, 6 of them mainly for the purpose of Sentiment Analysis and the rest mainly for the purpose of Article Triplets Similarity task. The annotators were chosen since they are interested in the task; the media partner also assured gender and age balance of the annotators that comes from different faculties.

Training of the recruiters in 2 phases.

Both phases, the first and the second, were executed on 26th of July 2019 in the media partner's offices in Zagreb. In the first phase we introduced the project and the goals of the project. A referee introduced the web application for the annotation task. The annotators received the basic guidelines (see the Annotation Guidelines section of Appendix C). we went through the guidelines together and a referee explained them in more details. This was followed by annotation of 5 articles, which we annotated together on the three levels (sentence, paragraph and document level). Using a five-level Likert (1932) scale (1–very negative, 2–negative, 3–neutral, 4–positive and 5–very positive), the annotators specified the evoked sentiment with the use of the following instruction: "Did this news evoke very positive/positive/neutral/negative/very negative feelings? (Please specify the sentiment from the perspective of an average Croatian web user)". Together with a referee, we discussed about individual instances, about every single decision, annotation grade and resolved possible issues and doubts.

In the second phase the annotators (6 annotators) annotated the same 25 articles individually. Afterwards we analysed the results of annotation. The agreement (Cronbach's alpha measure) between the annotators is 0.816 which is very satisfying achievement with only 25 articles. We planned to achieve 0.8 threshold. If the annotators would not achieve the planned threshold, they would repeat the second phase until they achieve it. The instances with lower agreement were discussed and the issues were resolved.

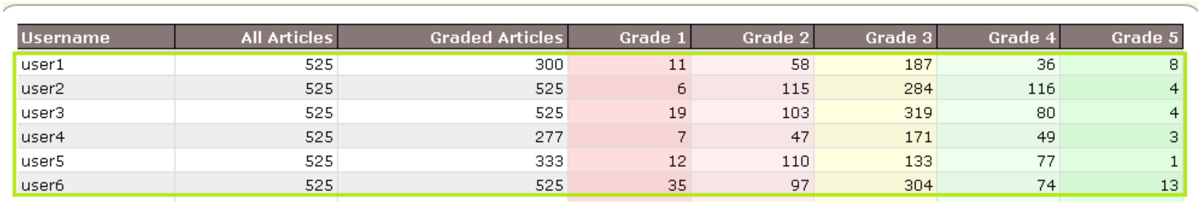
Annotation process.

Out of 2025 articles, 25 articles were annotated within the second phase, the rest 2000 are being annotated as follows. Each annotator will annotate 500 articles, 100 are supposed to be annotated by all 6 annotators. Among the rest 1900 articles, 1400 will be annotated by 1 annotator and 500 articles by 2 annotators.

Preliminary statistics for the Sentiment Analysis task.

Table 10 shows the number of articles that were annotated by each annotator for the Sentiment Analysis task (situation on 24th of August 2019). Half of annotator already finished with their task, however, the others annotated more than half of their articles and are making a good progress. Table 10 also shows that some other annotators, which are included in the Article Triplets Similarity task, also annotated some articles for the Sentiment Analysis task. This is additional data which can help researchers to improve cross-lingual embedding models and lexicon-based models.

Table 10: Number of articles annotated by each annotator for the Sentiment Analysis task.



Username	All Articles	Graded Articles	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5
user1	525	300	11	58	187	36	8
user2	525	525	6	115	284	116	4
user3	525	525	19	103	319	80	4
user4	525	277	7	47	171	49	3
user5	525	333	12	110	133	77	1
user6	525	525	35	97	304	74	13

Annotation guidelines

Below are the guidelines, as given to the annotators of the sentiment of news articles provided by STY/Trikoder.

INTRODUCTION:

In the first phase, you will obtain the basic guidelines for annotation and learned how to use the web application. Together with a referee, you will annotate 10 news on three levels, i.e. document, paragraph and sentence level, and discussed about individual instances.

The process of sentiment annotation consists of two sub-processes: comprehension, where the annotator understands the content, and sentiment judgment, where the annotator identifies the sentiment. Using the five-level Likert scale (1 – very negative, 2 – negative, 3 – neutral, 4 – positive and 5 – very positive) the annotators will specify evoked sentiment using the following instructions: "Please specify the sentiment from the perspective of an average Croatian web user. How did you feel after reading this news?"

In the second phase, along with a referee, each annotator will annotate 25 news items individually. When finished we will check the inter-rater reliability measure (Cronbach's alpha).

GUIDELINES FOR ANNOTATION:

- Web app: `http://dejan.amadej.si/sean/login.php`
- Basic guideline: Please specify the sentiment from the perspective of an average Croatian web user.
- Try to ignore your personal your personal beliefs, value norms, cultural, ethnic, religious inclinations ...

- If there are very long sentences and it is hard to determine the sentiment -> neutral sentiment (grade: 3)
- Sentences (usually long ones) with compound-complex sentences with more than one sentiment expressed, such as journalist's quotes and comments of politician's statements -> sentiment which prevails or neutral sentiment (grade: 3)
- Business results/stock markets/economic growth . . .
  - Higher than + 1,0 -> very positive (grade: 5)
  - + 0,5 to + 1,0 -> positive (grade: 4)
  - -0,5 to + 0,5 -> neutral (grade: 3)
  - - 0,5 to - 1,0 -> negative (grade: 2)
  - Lower than - 1,0 -> very negative (grade: 1)
- Gas prices/electricity/telecommunications/water/communal/costs of living. . .
  - Higher than + 1,0 -> very negative (grade: 1)
  - + 0,5 to + 1,0 -> negative (grade: 2)
  - -0,5 to + 0,5 -> neutral (grade: 3)
  - - 0,5 to - 1,0 -> positive (grade: 4)
  - Lower than - 1,0 -> very positive (grade: 5)