EMBEDDIA Tools, Datasets and Challenges: Resources and Hackathon Contributions

Senja Pollak
Jožef Stefan Institute
senja.pollak@ijs.si

Marko Robnik Šikonja
University of Ljubljana

Matthew Purver
Queen Mary University of London
Jožef Stefan Institute

Michele Boggia
University of Helsinki

Ravi Shekhar
Queen Mary University of London

Marko Pranjić
Trikoder d.o.o.

Salla Salmela
Suomen Tietotoimisto STT

Ivar Krustok
Tarmo Paju
Ekspress Meedia

Carl-Gustav Linden
University of Bergen

Leo Leppänen
Elaine Zosa
University of Helsinki

Matej Ulčar
University of Ljubljana

Linda Freienthal
Silver Traat
TEXTA OÜ

Luis Adrián Cabrera-Diego
University of La Rochelle, L3i

Matej Martinc
Nada Lavrač
Blaž Škrlj
Jožef Stefan Institute

Martin Žnidaršič
Andraž Pelicon
Boshko Koloski
Jožef Stefan Institute

Vid Podpečan
Janez Kranjc
Jožef Stefan Institute

Shane Sheehan
Usher Institute
University of Edinburgh

Emanuela Boros
University of La Rochelle, L3i

Jose G. Moreno
University of Toulouse, IRIT

Antoine Doucet
University of La Rochelle, L3i

Hannu Toivonen
University of Helsinki
hannu.toivonen@helsinki.fi

Abstract

This paper presents tools and data sources collected and released by the EMBEDDIA project, supported by the European Union's Horizon 2020 research and innovation program. The collected resources were offered to participants of a hackathon organized as part of the EACL Hackashop on News Media Content Analysis and Automated Report Generation in February 2021. The hackathon had six participating teams who addressed different challenges, either from the list of proposed challenges or their own news-industry-related tasks. This paper goes beyond the scope of the hackathon, as it brings together in a coherent and compact form most of the resources developed, collected and released by the EMBEDDIA project. Moreover, it constitutes a handy source for news media industry and researchers in the fields of Natural Language Processing and Social Science.

1 Introduction

News media industry is the primary provider of information for society and individuals. Since the first newspaper was published, the propagation of information has continuously changed as new technologies are adopted by the news media, and the advent of the internet has made this change faster than ever (Pentina and Tarafdar, 2014). Internet-based media (e.g., social media, forums and blogs) have made news more accessible, and dissemination more affordable, resulting in drastically increased media coverage. Social media can also help provide source information for newsrooms, as shown in e.g., disaster response tasks (Alam et al., 2018).

Suitable Natural Language Processing techniques are needed to analyze news archives and gain insight about the evolution of our society, while dealing with the constant flow of information. Relevant datasets are equally important in

Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation, pages 99–109

order to train data-driven approaches. To encourage the development and uptake of such techniques and datasets, and take on the challenges presented by the introduction of new technologies in the news media industry, the EMBEDDIA project[1] organized, in conjunction with EACL 2021, a hackathon[2] as part of the EACL Hackashop on News Media Content Analysis and Automated Report Generation[3].

For this event, held virtually in February 2021, the datasets and tools curated and implemented by the EMBEDDIA project were publicly released and made available to the participants. We also provided examples of realistic challenges faced by today's newsrooms, and offered technical support and consultancy sessions with a news media expert throughout the entire duration of the hackathon.

The contributions of this paper are structured as follows. Section 2 presents the tools released for the event. The newly gathered, publicly released EMBEDDIA datasets are reported in Section 3. Section 4 presents sample news media challenges. Section 5 outlines the projects undertaken by the teams who completed the hackathon. The hackathon outcomes are summarized in Section 6.

2 Tools

The EMBEDDIA tools and models released for the hackathon include general text processing tools like language processing frameworks and text representation models (Section 2.1), news article analysis (Section 2.2), news comment analysis (Section 2.3), and news article and headline generation (Section 2.4) tools.

These tools require different levels of technical proficiency. Language processing tools and frameworks require little to no programming skills. On the other hand, for some tasks, we provide fully functional systems that can be used out of the box but require a certain level of technical knowledge in order to be fully utilized. Moreover, some tools and text representation models require programming skills and can be employed to improve existing systems, implement new analytic tools, or to be adapted to new uses.

[1] http://embeddia.eu

[2] http://embeddia.eu/hackashop2021-call-for-hackathon-participation/

[3] http://embeddia.eu/hackashop2021/

2.1 General Text Analytics

We first present two general frameworks, requiring no programming skills: the EMBEDDIA Media Assistant, incorporating the TEXTA Toolkit that is focused exclusively on text, and the ClowdFlows toolbox, which is a general data science framework incorporating numerous NLP components. Finally, we describe BERT embeddings, a general text representation framework that includes variants of multilingual BERT models, which are typically part of programming solutions.

2.1.1 TEXTA Toolkit and EMBEDDIA Media Assistant

The TEXTA Toolkit (TTK) is an open-source software for building RESTful text analytics applications.[4] TTK can be used for:

- searching and aggregating data (using e.g. regular expressions),
- training embeddings,
- building machine learning classifiers,
- building topic-related lexicons using embeddings,
- clustering and visualizing data, and
- extracting and creating training data.

The TEXTA Toolkit is the principal ingredient of the EMBEDDIA Media Assistant (EMA), which includes the TEXTA Toolkit GUI and API, an API Wrapper with a number of APIs for news analysis, and a Demonstrator for demonstrating the APIs.

2.1.2 ClowdFlows

ClowdFlows[5] is an open-source online platform for developing and sharing data mining and machine learning workflows (Kranjc et al., 2012). It works online in modern Web browsers, without client-side installation. The user interface allows combining software components (called widgets) into functional workflows, which can be executed, stored, and shared in the cloud. The main aim of ClowdFlows is to foster sharing of workflow solutions in order to simplify the replication and adaptation of shared work. It is suitable for prototyping, demonstrating new approaches, and exposing solutions to potential users who are not proficient in programming but would like to experiment with their own datasets and different tool parameter settings.

[4] https://docs.texta.ee/

[5] https://cf3.ijs.si/

2.1.3 BERT Embeddings

CroSloEngual[6] BERT and FinEst[7] BERT (Ulčar and Robnik-Šikonja, 2020) are trilingual models, based on the BERT architecture (Devlin et al., 2019), created in the EMBEDDIA project to facilitate easy cross-lingual transfer. Both models are trained on three languages: one of them being English as a resource-rich language, CroSloEngual BERT was trained on Croatian, Slovenian, and English data, while FinEst BERT was trained on Finnish, Estonian, and English data.

The advantage of multi-lingual models over monolingual models is that they can be used for cross-lingual knowledge transfer, e.g., a model for a task for which very little data is available in a target language such as Croatian or Estonian can be trained on English (with more data available) and transferred to a less-resourced language. While massive multilingual BERT-like models are available that cover more than 100 languages (Devlin et al., 2019), a model trained on only a few languages performs significantly better on these (Ulčar and Robnik-Šikonja, 2020). The two trilingual BERT models here are effective for the languages they cover and for the cross-lingual transfer of models between these languages. The models represent words/tokens with contextually dependent vectors (word embeddings). These can be used for training many NLP tasks, e.g., fine-tuning the model for any text classification task.

2.2 News Article Analysis Tools

The majority of provided tools cover different aspects of news article analysis, processing, and generation. We present keyword extraction tools TNT-KID and RaKUn, named entity recognition approaches, tools for diachronic analysis of words, tools for topic analysis and visualization, and tools for sentiment analysis.

2.2.1 Keyword Extraction

Two tools are available for keyword extraction: TNT-KID and RaKUn.

TNT-KID[8] (Transformer-based Neural Tagger for Keyword Identification, Martinc et al., 2020) is a supervised tool for extracting keywords from news articles in several languages (English, Estonian, Croatian, and Russian). It relies on the modified Transformer architecture (Vaswani et al., 2017) and leverages language model pretraining on a domain-specific corpus. This gives competitive and robust performance while requiring only a fraction of the manually labeled data needed by the best performing supervised systems. This makes TNT-KID especially appropriate for less-resourced languages where large manually labeled datasets are scarce.

RaKUn[9] (Škrlj et al., 2019) offers unsupervised detection and exploration of keyphrases. It transforms a document collection into a network, which is pruned to keep only the most relevant nodes. The nodes are ranked, prioritizing nodes corresponding to individual keywords and paths (keyphrases comprised of multiple words). Being unsupervised, RaKUn is well suited for less-resourced languages where expensive pre-training is not possible.

2.2.2 Named Entity Recognition[10]

The Named Entity Recognition (NER) system is based on the architecture proposed by Boros et al. (2020). It consists of fine-tuned BERT with two additional Transformer blocks (Vaswani et al., 2017). We provided models capable of predicting three types of named entities (Location, Organisation and Person) for eight European languages: Croatian, Estonian, Finnish, Latvian, Lithuanian, Russian, Slovene and Swedish. These models were trained using the WikiANN corpus (Pan et al., 2017), specifically using the training, development and testing partitions provided by Rahimi et al. (2019). Regarding BERT, for Croatian and Slovene we used *CroSloEngual BERT* (Ulčar and Robnik-Šikonja, 2020); for Finnish and Estonian *FinEst BERT* (Ulčar and Robnik-Šikonja, 2020); for Russian *RuBERT* (Kuratov and Arkhipov, 2019); for Swedish *Swedish BERT* (Malmsten et al., 2020); for Latvian and Lithuanian *Multilingual BERT* (Devlin et al., 2019).

2.2.3 Diachronic News Analysis[11]

The tool for diachronic semantic shift detection (Martinc et al., 2019a) leverages the BERT contextual embeddings (Devlin et al., 2019) for generat-

[6] https://huggingface.co/EMBEDDIA/crosloengual-bert

[7] https://huggingface.co/EMBEDDIA/finest-bert

[8] https://github.com/EMBEDDIA/tnt_kid

[9] https://github.com/EMBEDDIA/RaKUn

[10] https://github.com/EMBEDDIA/stacked-ner

[11] https://github.com/EMBEDDIA/semantic_shift_detection

ing time-specific word representations. It checks whether a specific word (or phrase) in the corpus has changed across time by measuring the rate of change for time-specific relations to semantically similar words in distinct time periods. Besides measuring long-term semantic changes, the method can also be successfully used for the detection of short-term yearly semantic shifts and has even been employed in the multilingual setting.

2.2.4 Topic Analysis

We present three tools dealing with news topics: PTM, PDTM and TeMoCo. The first two use topics to link articles across languages, and the third one visualizes distributions of topics over time.

PTM[12] (Polylingual Topic Model, Mimno et al., 2009) can be used to train cross-lingual topic models and obtain cross-lingual topic vectors for news articles. These vectors can be used to link news articles across languages. An ensemble of cross-lingual topic vectors and document embeddings can outperform stand-alone methods for cross-lingual news linking (Zosa et al., 2020).[13]

PDTM[14] (Polylingual Dynamic Topic Model, Zosa and Granroth-Wilding, 2019) is an extension of the Dynamic Topic Model (Blei and Lafferty, 2006) for multiple languages. This model can track the evolution of topics over time aligned across multiple languages.

TeMoCo[15] (Temporal Topic Visualisation, Sheehan et al., 2019, 2020) visualizes changes in topic distribution and associated keywords in a document or collection of articles. The tool can investigate a single document or a corpus which has been temporally annotated (e.g., a transcript or corpus of dated articles). The user can examine an overview of a dataset, processed into time and topic segments. The changes in topic size and keywords describe patterns in the data. Clicking on the segments brings up the related news articles with keyword highlighting.

[12] https://github.com/EMBEDDIA/cross-lingual-linking

[13] https://github.com/EMBEDDIA/cross-lingual-linking

[14] https://github.com/EMBEDDIA/multilingual_dtm

[15] https://github.com/EMBEDDIA/TeMoCo

2.2.5 News Sentiment Analysis[16]

Sentiment analysis is likely the most popular NLP application in industry. Our multilingual model for news sentiment classification is based on multilingual BERT. The model was trained on the Slovenian news sentiment dataset (Bučar et al., 2018) using a two-step training approach with document and paragraph level sentiment labels (Pelicon et al., 2020). The model was tested on the document-level labels of the Croatian news sentiment dataset (Section 3.2.2) in a zero-shot setting. The model maps the input document into one of the three predefined classes: positive, negative, and neutral.

2.3 News Comment Analysis Tools

Several of the tools in the sections above can also be applied to comments. We describe the following comment-specific tools: comment moderation, bot and gender detection, and sentiment analysis tools.

2.3.1 Comment Moderation[17]

Our comment moderation tool flags inappropriate comments that should be blocked from appearing on news sites (Pelicon et al., 2021a,b). It uses multilingual BERT (Devlin et al., 2019) and the trilingual EMBEDDIA BERT models (Section 2.1.3). The models were trained on combinations of five datasets: Croatian and Estonian (see Section 3.3 and details in Shekhar et al. (2020)), Slovenian (Ljubešić et al., 2019), English (Zampieri et al., 2019), and German (Wiegand et al., 2018). For Croatian, we also provide a model to predict which rule is violated, based on the moderation policy of 24 sata, the biggest Croatian news publisher (see Section 3.3.3).

2.3.2 Bot and Gender Detection[18]

An author profiling tool for gender classification and bot detection in Spanish and English, trained on Twitter data (Martinc et al., 2019b), was developed for the PAN 2019 author profiling shared task (Rangel and Rosso, 2019). It uses a two-step approach: in the first step distinguishing between bots and humans, and in the second step determining the gender of human authors. It relies on a Logistic Regression classifier and employs a number of different word and character n-gram features.

[16] https://github.com/EMBEDDIA/crosslingual_news_sentiment

[17] https://github.com/EMBEDDIA/hackashop2021_comment_filtering

[18] https://github.com/EMBEDDIA/PAN2019

2.3.3 Sentiment Analysis[19]

The code for sentiment analysis allows training a model that classifies text into one of three sentiment categories: positive, neutral, or negative. The classifier is trained on the Twitter datasets[20] provided by Mozetič et al. (2016). The models and datasets support cross-lingual knowledge transfer from resource-rich language(s) to less-resourced languages.

2.4 News Article and Headline Generation

Two of our tools are for generating text, either news for specific topics, or creative language.

Template-Based NLG System for Automated Journalism The rule-based natural language generation system—similar in concept to Leppänen et al. (2017)—produces news texts in Finnish and English from statistical data obtained from EuroStat. The system provides the text inputs used in the NLG challenges, described in Section 4.3. Access to the tool is provided through an API.[21]

Creative Language Generation We provide a framework[22] to help in generation of creative language using an evolutionary algorithm (Alnajjar and Toivonen, 2020).

3 Datasets

For the purposes of the hackashop, the EMBEDDIA media partners released their news archives, the majority of which are now being made publicly available for use after the project.

3.1 General EMBEDDIA News Datasets

Four publicly available datasets released by the EMBEDDIA project are described below.

3.1.1 Ekspress Meedia News Archive (in Estonian and Russian)

Ekspress Meedia belongs to the Ekspress Meedia Group, one of the largest media groups in the Baltics. The dataset is an archive of articles from the Ekspress Meedia news site from 2009–2019, containing over 1.4M articles, mostly in the Estonian (1,115,120 articles) with some in the Russian language (325,952 articles). Keywords (tags) are included for articles after 2015. The dataset is publicly available in the CLARIN repository.[23]

3.1.2 Latvian Delfi Article Archive (in Latvian and Russian)

Latvian Delfi belongs to Ekspress Meedia Group. This dataset is an archive of articles from the Delfi news site from 2015-2019, containing over 180,000 articles (c. 50% in Latvian and 50% in Russian language). Keywords (tags) for articles are included. The dataset is publicly available in CLARIN.[24]

3.1.3 24sata News Archive (in Croatian)

24sata is the biggest Croatian news publisher, owned by the Styria Media Group. The 24sata news portal consists of a daily news portal and several smaller portals covering news on specific topics, such as automotive news, health, culinary content, and lifestyle advice. The dataset contains over 650,000 articles in Croatian between 2007–2019, as well as assigned tags. The dataset is publicly available in CLARIN.[25]

3.1.4 STT News Archive (in Finnish)

The Finnish corpus (STT, 2019) contains newswire articles in Finnish sent to media outlets by the Finnish News Agency (STT) between 1992–2018. The corpus includes about 2.8 million items in total. The news articles are categorized by department (domestic, foreign, economy, politics, culture, entertainment and sports), as well as by metadata (IPTC subject categories or keywords and location data). The dataset is publicly available via CLARIN,[26] as is a parsed version of the corpus in CoNLL-U format (STT et al., 2020).[27]

3.2 Task-specific News Datasets

For the purposes of the hackashop, a set of task-specific datasets were also gathered.

3.2.1 Keyword Extraction Datasplits

For the keyword extraction challenge, we created train and test data splits, given as article IDs from datasets in Section 3.1. The number of articles for Estonian, Latvian, Russian and Croatian (see Koloski et al. (2021a) for details) are:

[19] https://github.com/EMBEDDIA/cross-lingual-classification-of-tweet-sentiment

[20] http://hdl.handle.net/11356/1054

[21] http://newseye-wp5.cs.helsinki.fi:4220/documentation/

[22] https://github.com/EMBEDDIA/evolutionary-algorithm-for-NLG

[23] http://hdl.handle.net/11356/1408

[24] http://hdl.handle.net/11356/1409

[25] http://hdl.handle.net/11356/1410

[26] http://urn.fi/urn:nbn:fi:lb-2019041501

[27] http://urn.fi/urn:nbn:fi:lb-2020031201

- Croatian: 32,223 train, 3,582 test;
- Estonian: 10,750 train, 7,747 test;
- Russian: 13,831 train, 11,475 test;
- Latvian: 13,133 train, 11,641 test.

The data is publicly available in CLARIN.[28]

3.2.2 News Sentiment Annotated Dataset

We selected a subset of 2,025 news articles from the Croatian 24sata dataset (see Section 3.1.3 and Pelicon et al., 2020). Several annotators annotated the articles on a five-point Likert-scale from 1 (most negative sentiment) to 5 (most positive). The final sentiment label of an article was then based on the average of the scores given by the different annotators: negative if average was less than or equal to 2.4, neutral if between 2.4 and 3.6, or positive if greater than or equal to 3.6. The dataset is publicly available in CLARIN.[29]

3.2.3 Estonian-Latvian Interesting News Pairs

For the purposes of the challenge on finding interesting news from neighbouring countries (see Section 4.1.2 and Koloski et al., 2021b) an Estonian journalist gathered 21 news articles from Latvia that would be of interest for Estonians, paired with 21 corresponding Estonian articles.[30]

3.2.4 Corpus of Computer-Generated Statistical News Texts

This corpus, consisting of a total 188 news texts produced by the rule-based natural language generation system described in Section 2.4, is provided to allow for easier offline development of solutions to the NLG challenges. The corpus contains news texts in both Finnish and English,[31] discussing consumer prices as well as health care spending and funding on the national level within the EU.

3.3 News Comments Datasets

Three news comment datasets have been made publicly available. To ensure privacy, user IDs in all news comment datasets in this section have been obfuscated, so they no longer correspond to the original IDs on the publishers' systems. User IDs for moderated comments have been removed.

[28] http://hdl.handle.net/11356/1403
[29] http://hdl.handle.net/11356/1342
[30] https://github.com/EMBEDDIA/interesting-cross-border-news-discovery
[31] https://github.com/EMBEDDIA/embeddia-nlg-output-corpus

3.3.1 Ekspress Meedia Comment Archive (in Estonian and Russian)

This dataset is an archive of reader comments on the Ekspress Meedia news site from 2009–2019, containing approximately 31M comments, mostly in Estonian language, with some in Russian. The dataset is publicly available in CLARIN.[32]

3.3.2 Latvian Delfi Comment Archive (in Latvian and Russian)

The dataset of Latvian Delfi, which belongs to Ekspress Meedia Group, is an archive of reader comments from the Delfi news site from 2014–2019, containing approximately 12M comments, mostly in Latvian language, with some in Russian. The dataset is publicly available in CLARIN.[33]

3.3.3 24sata Comment Archive (in Croatian)

In this archive, there are over 20M user comments from 2007–2019, written mostly in Croatian. All comments were gathered from 24sata, the biggest Croatian news publisher, owned by Styria Media Group. Each comment is given with the ID of the news article where it was posted and with multi-label moderation information corresponding to the rules of 24sata's moderation policy (see Shekhar et al., 2020). The dataset is publicly available in CLARIN.[34]

3.4 Other News Datasets

EventRegistry (Leban et al., 2014), which is a news intelligence platform aiming to empower organizations to keep track of world events and analyze their impact, provided free access to their data for hackathon participants.

Datasets relevant to the hackathon have also been made available for academic use by the Finnish broadcasting company Yle in Finnish[35] and in Swedish[36].

4 Challenges

Sample news media challenge addressed in the EMBEDDIA project come from three different areas: news analysis, news comments analysis, and article and headline generation.

[32] http://hdl.handle.net/11356/1401
[33] http://hdl.handle.net/11356/1407
[34] http://hdl.handle.net/11356/1399
[35] https://korp.csc.fi/download/YLE/fi/2011-2018-src/
[36] https://korp.csc.fi/download/YLE/sv/2012-2018-src/

4.1 News Analysis Challenges

4.1.1 Keyword Extraction

The EMBEDDIA datasets from Ekspress Meedia, Latvian Delfi and 24sata contain articles together with keywords assigned by journalists (see Section 3.2.1). The project has produced several state-of-the-art approaches for automatic keyword extraction on these datasets (see Section 2.2.1). The challenge consists of providing alternative methods to achieve the most accurate keyword extraction and compare with our results.

4.1.2 Identifying Interesting News from Neighbouring Countries

Journalists are very interested in identifying stories from cross-border countries, that attract a large number of readers and are "special". A journalist at Ekspress Meedia in Estonia gave the example of selecting news from Latvia that would be of interest to Estonian readers. Example topics include: drunk Estonians in Latvia, a person in Latvia living in a boat, stories from Latvia about topics that also interest Estonians (for example, alcohol taxes, newsworthy actions that take place near the border, certain public figures). At the moment it is easy to detect all the news from Latvia with the mentions of words "Estonia" or "Estonians", but the challenge is to identify a larger number of topics, e.g. scandals, deaths, gossip that might be somehow connected to Estonia, and news and stories that Estonians relate to (for example, when similar things have happened in Estonia or similar news has been popular there). Given the collection of news from two different countries (e.g. Estonia, Latvia, see Section 3.1), the task is to identify these special interesting news stories; 21 manually identified examples were provided (see Section 3.2.3).

4.1.3 Diachronic News Article Analysis

Media houses with large news articles collections are interested in analysing the reporting on certain topics to investigate changes over time. This can not only help them understand their reporting, but also help journalists to discover specific aspects related to these concepts.

An example from a news media professional from Estonia is as follows: "the doping affairs in sports regularly appear and for example for one of our skiers, a few years ago, we have already reported on a potential doping affair, but did not analyse it in depth. Few years later it has turned out that the sportsman was indeed involved in a doping affair. Having a better overview of doping related persons and topics over time, would be interesting for us." An even more straightforward application is the monitoring of politicians and parties; controversial topics are also of interest, as they can show general changes in society towards them.

Each of the media partners provided some people/parties/concepts of their interest. Examples are reported in Appendix A.

4.2 News Comments Analysis

4.2.1 Comment Moderation

The EMBEDDIA datasets from Ekspress Meedia and 24sata contain comments with metadata showing the ones blocked by the moderators (see Section 3.3). In the case of the 24sata dataset, specific moderation policies exist with a list of reasons for blocking, and the metadata also shows which of the reasons applied. The policies are applied by humans, though, and therefore the metadata will reflect the way moderators actually behave, including making mistakes and showing biases. During the EMBEDDIA project, we have developed and evaluated multiple automatic filtering approaches on these datasets, which can be used off-the-shelf or can be re-trained or modified (see Section 2.3.1). The hackathon participants were invited to propose alternative comment filtering methods, to improve over the existing approaches, or apply them to other datasets; to use them to investigate how human moderators actually behave; and/or to investigate how to analyse, understand or use the outputs.

4.2.2 Comment Summarization

Each of the comment datasets available contains about 10 years of data. The EMEBDDIA project has developed and evaluated a range of classifiers that can detect useful information in comments and comment-like text (including sentiment, topic, author information etc; see Section 2.3). The participants were invited to use these and other methods to extract meaningful information from comment threads and develop new ways of presenting this information in a way that could be useful to a journalist or analyst. Example approaches given were summarizing topics, views and opinions; and detecting and summarizing constructive or positive comments, as an antidote to the negative comments so often focused on in NLP.

4.3 Natural Language Generation

4.3.1 Improving the Fluency of Automatically Generated Articles

Despite recent strides in neural natural language generation (NLG) methods, neural NLG methods are still prone to producing text that is not grounded in the input data. As such errors are catastrophic in news industry applications, most news generation systems continue to employ rule-based NLG methods. Such methods, however, lack to adequately handle the variety and fluency of expression. One potential solution would be to combine neural post-processing with a rule-based NLG system. In this challenge, participants are provided with black box access to a rule-based NLG system that produces statistical news articles. A corpus of the produced news articles is also provided.[37] The challenge is to use automated post-processing methods to improve the fluency and grammaticality of the system's output without changing the meaning of the text.

The system is multilingual (English and Finnish), and optimally the proposed solutions should be language-independent, taking advantage of e.g., multilingual word embeddings. At the same time, we also welcome monolingual solutions.

4.3.2 Headline Generation

Headlines play an important role in news text, not only summarizing the most important information in the underlying news text, but also presenting it in a light that is likely to entice the reader to engage with the larger text. In this challenge, the participants are invited to create headlines for automatically generated articles (see Section 4.3.1).

5 Hackathon Contributions

Six teams with 24 members in total participated in the hackathon during 1–19 February 2021. The challenges described in Section 4 were offered to the teams as examples of interesting problems in the area of news media analysis and generation. The teams had, however, the freedom to choose and formulate their own aims for the hackathon. Likewise, they were offered the data, tools and models described above.

The hackathon was organized online, with three joint events to kick off the activities, to meet and talk about the ongoing work halfway, and to wrap up the work at the end. Ample support on tools, models, data and challenges was provided by the EMBEDDIA experts via several channels.

The six teams all picked up different challenges and set themselves specific goals. Reports from five teams are included in these proceedings.

Three teams worked on news content analysis:

- One team developed a COVID-19 news dashboard to visualise sentiment in pandemic-related news. The dashboard uses a multilingual BERT model to analyze news headlines in different languages across Europe (Robertson et al., 2021).
- Methods for cross-border news discovery were developed by another team using multilingual topic models. Their tool discovers Latvian news that could interest Estonian readers (Koloski et al., 2021b).
- A third team used sentiment and viewpoint analysis to study attitudes related to LGBTIQ+ in Slovenian news. Their results suggest that political affiliation of media outlets can affect sentiment towards and framing of LGBTIQ+-specific topics (Martinc et al., 2021).

Two teams looked at different challenges related to comment analysis:

- One team automated news comment moderation. They compiled and labeled a dataset of English news and social posts, and experimented with cross-lingual transfer of comment labels from English and subsequent supervised machine learning on Croatian and Estonian news comments (Korenčić et al., 2021).
- Another team looked at the diversity of news comment recommendations, motivated by democratic debate. They implemented a novel metric based on theories of democracy and used it to compare recommendation strategies of New York Times comments in English (Reuver and Mattis, 2021).

Finally, one team worked on a generation task:

- The team experimented with several methods for generating headlines, given the contents of a news story. They found that headlines formulated as questions about the story's content tend to be both informative and enticing.

[37] https://github.com/EMBEDDIA/embeddia-nlg-output-corpus

6 Conclusions

This paper presents the contributions of the EMBEDDIA project, including a large variety of tools, new datasets of news articles and comments from the media partners, as well as challenges that were proposed to the participants of the EACL 2021 Hackathon on News Media Content Analysis and Automated Report Generation. The hackathon had six participating teams who addressed different challenges, either from the list of proposed challenges or their own news-industry-related tasks. In the future, the tools and resources described can be used for a large variety of new experiments, and we hope that the proposed challenges will be addressed by the wider NLP research community.

Acknowledgements

This work has been supported by the European Union's Horizon 2020 research and innovation program under grant 825153 (EMBEDDIA).

We would like to thank EventRegistry for providing free access to their data for hackathon participants.

References

Firoj Alam, Ferda Ofli, Muhammad Imran, and Michael Aupetit. 2018. A Twitter tale of three hurricanes: Harvey, Irma, and Maria. *Proc. of ISCRAM, Rochester, USA*.

Khalid Alnajjar and Hannu Toivonen. 2020. Computational generation of slogans. *Natural Language Engineering*, First View:1–33.

David M Blei and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120.

Emanuela Boros, Ahmed Hamdi, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Jose G. Moreno, Nicolas Sidere, and Antoine Doucet. 2020. Alleviating digitization errors in named entity recognition for historical documents. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 431–441, Online. Association for Computational Linguistics.

Joze Bučar, Martin Žnidarsic, and Janez Povh. 2018. Annotated news corpora and a lexicon for sentiment analysis in Slovene. *Language Resources and Evaluation*, 52:895–919.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Boshko Koloski, Senja Pollak, Blaž Škrlj, and Matej Martinc. 2021a. Extending neural keyword extraction with TF-IDF tagset matching. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*. Association for Computational Linguistics.

Boshko Koloski, Elaine Zosa, Timen Stepišnik-Perdih, Blaž Škrlj, Tarmo Paju, and Senja Pollak. 2021b. Interesting cross-border news discovery using cross-lingual article linking and document similarity. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*. Association for Computational Linguistics.

Damir Korenčić, Ipek Baris, Eugenia Fernandez, Katarina Leuschel, and Eva Sánchez Salido. 2021. To block or not to block: Experiments with machine learning for news comment moderation. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*. Association for Computational Linguistics.

Janez Kranjc, Vid Podpečan, and Nada Lavrač. 2012. ClowdFlows: A cloud based scientific workflow platform. In Peter A. Flach, Tijl Bie, and Nello Cristianini, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 7524 of *Lecture Notes in Computer Science*, pages 816–819. Springer Berlin Heidelberg.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. *arXiv cs.CL*. Preprint: 1905.07213.

Gregor Leban, Blaz Fortuna, Janez Brank, and Marko Grobelnik. 2014. Event registry: learning about world events from news. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 107–110.

Leo Leppänen, Myriam Munezero, Mark Granroth-Wilding, and Hannu Toivonen. 2017. Data-driven news generation for automated journalism. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 188–197.

Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. 2019. The FRENK Datasets of Socially Unacceptable Discourse in Slovene and English. In *International Conference on Text, Speech, and Dialogue*, pages 103–114. Springer.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with Words at the National Library of Sweden – Making a Swedish BERT. *arXiv cs.CL*. Preprint: 2007.01658.

Matej Martinc, Petra Kralj Novak, and Senja Pollak. 2019a. Leveraging contextual embeddings for detecting diachronic semantic shift. *arXiv preprint arXiv:1912.01072*.

Matej Martinc, Nina Perger, Andraž Pelicon, Matej Ulčar, Andreja Vezovnik, and Senja Pollak. 2021. EMBEDDIA hackathon report: Automatic sentiment and viewpoint analysis of Slovenian news corpus on the topic of LGBTIQ+. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*. Association for Computational Linguistics.

Matej Martinc, Blaz Skrlj, and Senja Pollak. 2019b. Fake or not: Distinguishing between bots, males and females. In *CLEF (Working Notes)*.

Matej Martinc, Blaž Škrlj, and Senja Pollak. 2020. Tntkid: Transformer-based neural tagger for keyword identification. *arXiv preprint arXiv:2003.09166*.

David Mimno, Hanna Wallach, Jason Naradowsky, David A Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of the 2009 conference on Empirical Methods in Natural Language Processing*, pages 880–889.

Igor Mozetič, Miha Grčar, and Jasmina Smailović. 2016. Multilingual twitter sentiment classification: The role of human annotators. *PLOS ONE*, 11(5):1–26.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual Name Tagging and Linking for 282 Languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Andraž Pelicon, Marko Pranjić, Dragana Miljković, Blaž Škrlj, and Senja Pollak. 2020. Zero-shot learning for cross-lingual news sentiment classification. *Applied Sciences*, 10(17):5993.

Andraž Pelicon, Ravi Shekhar, Matej Martinc, Blaž Škrlj, Matthew Purver, and Senja Pollak. 2021a. Zero-shot cross-lingual content filtering: Offensive language and hate speech detection. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*.

Andraž Pelicon, Ravi Shekhar, Blaž Škrlj, Matthew Purver, and Senja Pollak. 2021b. Investigating cross-lingual training for offensive language detection. *Submitted, to appear*.

Iryna Pentina and M. Tarafdar. 2014. From "information" to "knowing": Exploring the role of social media in contemporary news consumption. *Comput. Hum. Behav.*, 35:211–223.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively Multilingual Transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Francisco Rangel and Paolo Rosso. 2019. Overview of the 7th author profiling task at pan 2019: bots and gender profiling in Twitter. In *Working Notes Papers of the CLEF 2019 Evaluation Labs Volume 2380 of CEUR Workshop*.

Myrthe Reuver and Nicolas Mattis. 2021. Implementing evaluation metrics based on theories of democracy in news comment recommendation (Hackathon report). In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*. Association for Computational Linguistics.

Frankie Robertson, Jarkko Lagus, and Kaisla Kajava. 2021. A COVID-19 news coverage mood map of Europe. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*. Association for Computational Linguistics.

Shane Sheehan, Pierre Albert, Masood Masoodian, and Saturnino Luz. 2019. TeMoCo: A visualization tool for temporal analysis of multi-party dialogues in clinical settings. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 690–695. IEEE.

Shane Sheehan, Saturnino Luz, Pierre Albert, and Masood Masoodian. 2020. TeMoCo-Doc: A visualization for supporting temporal and contextual analysis of dialogues and associated documents. In *Proceedings of the International Conference on Advanced Visual Interfaces*, AVI '20, New York, NY, USA. Association for Computing Machinery.

Ravi Shekhar, Marko Pranjić, Senja Pollak, Andraž Pelicon, and Matthew Purver. 2020. Automating News Comment Moderation with Limited Resources: Benchmarking in Croatian and Estonian. *Journal for Language Technology and Computational Linguistics (JLCL)*, 34(1).

Blaž Škrlj, Andraž Repar, and Senja Pollak. 2019. Rakun: Rank-based keyword extraction via unsupervised learning and meta vertex aggregation. In *Statistical Language and Speech Processing*, pages 311–323, Cham. Springer International Publishing.

STT. 2019. Finnish news agency archive 1992-2018, source (http://urn.fi/urn:nbn:fi:lb-2019041501).

STT, Helsingin yliopisto, and Khalid Alnajjar. 2020. Finnish News Agency Archive 1992-2018, CoNLL-U, source (http://urn.fi/urn:nbn:fi:lb-2020031201).

Matej Ulčar and Marko Robnik-Šikonja. 2020. FinEst BERT and CroSloEngual BERT. In *International*

Conference on Text, Speech, and Dialogue, pages 104–111. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 shared task on the identification of offensive language. In *Proceedings of the GermEval 2018 Workshop (GermEval)*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of NAACL*, pages 1415–1420.

Elaine Zosa and Mark Granroth-Wilding. 2019. Multilingual dynamic topic model. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1388–1396, Varna, Bulgaria. INCOMA Ltd.

Elaine Zosa, Mark Granroth-Wilding, and Lidia Pivovarova. 2020. A comparison of unsupervised methods for ad hoc cross-lingual document retrieval. In *Proceedings of the LREC 2020 Workshop on Cross-Language Search and Summarization of Text and Speech*. European Language Resources Association (ELRA).

A Entities of Interest for Diachronic News Article Analysis Challenge

For the challenge described in Section 4.1.3, each of the media partners provided some people/parties/concepts of their interest. These include the following.

Political parties:

- Estonian (Eskpress meedia): Reformierakond, EKRE, Keskerakond
- Finnish (STT)[38]: Suomen Sosialidemokraattinen Puolue, demarit, SDP, (sd.); Kokoomus, (kok.); Keskusta, (kesk.); Perussuomalaiset, (ps.); Kristillisdemokraatit, KD, (kd.)
- Croatian: Hrvatska demokratska zajednica (HDZ), Socijaldemokratska partija Hrvatske (SDP), Hrvatska narodna stranka (HNS), Most nezavisnih lista (MOST)

Popular people:

- Estonian: Jüri Ratas, Kersti Kaljulaid, Kaja Kallas, Martin Helme
- Croatian: Andrej Plenković (the prime minister), Zoran Milanović (the president), Kolinda Grabar-Kitarović (previous president), Milan Bandić (mayor of Zagreb)

Interesting topics were selected for all three languages to allow also cross-lingual comparisons:

- **corona crisis, pandemics**: Estonian: Koroonakriis, pandeemia; Finnish: korona, koronakriisi, pandemia, koronapandemia; Croatian: korona, koronavirus, korona kriza, pandemija, korona pandemija
- **same sex rights, registered partnership act, marriage referendum**: Estonian:samasooliste õigused, kooseluseadus, abielureferendum; Finnish: tasa-arvoinen avioliitto, rekisteröity parisuhde; Croatian: referendum o braku, životno partnerstvo, civilno partnerstvo
- **financial knowledge, savings, investing, pension:** Estonian: rahatarkus, säästmine, investeerimine, pension; Finnish: sijoittaminen, piensijoittaja, säästäminen, eläke, eläkkeet; Croatian: ulaganje, investiranje, mali ulagači, dionice, ušteđevina, mirovina, penzija
- **doping:** same word in Estonian/Finnish/Croatian.

[38]The names without brackets are names the parties use and the abbreviation inside brackets is the way to mark a mp's / other person's political party within a text. For example Jussi Halla-aho (ps.) said that-