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: 39 months

D3.7: Final evaluation report on cross-lingual user generated content filtering and analysis technology (T3.4)

Executive summary

This deliverable summarises the overall progress and outputs achieved in Work Package WP3 of the EMBEDDIA project. WP3 aims to develop cross-lingual and multilingual tools for automatic processing of user-generated comments in news media. Work developing these tools was divided into three tasks: comment analysis in Task T3.1, comment filtering in Task T3.2 and comment summarization in Task T3.3; with a fourth Task T3.4 supporting these by producing the required resources and performing evaluation. For each task, we briefly summarize the progress and main results already reported in the respective task-specific deliverables; we then summarise our progress since then, including new results using improved techniques and new analysis tasks; evaluation against new benchmarks; and the development and release of new datasets for evaluation and dissemination. We also describe the results of initial successful integration of our filtering tools into a production environment.

Partner in charge: QMUL

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:	**D3.7**
Deliverable full title:	**Final evaluation report on cross-lingual user generated content filtering and analysis technology**
Deliverable short title:	**Final evaluation report on UGC technology**
Document identifier:	**EMBEDDIA-D37-FinalEvaluationReportOnUGCTechnology-T34-submitted**
Lead partner short name:	**QMUL**
Report version:	**submitted**
Report submission date:	**28/02/2022**
Dissemination level:	**PU**
Nature:	**R = Report**
Lead author(s):	**Matthew Purver (QMUL), Ravi Shekhar (QMUL)**
Co-author(s):	**Linda Freienthal (TEXTA), Marit Asula (TEXTA), Andraž Pelicon (JSI), Mladen Karan (QMUL), Senja Pollak (JSI), Marko Robnik-Šikonja (UL)**
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
10/01/2022	v0.1	Matthew Purver (QMUL)	Initial draft.
10/02/2022	v1.0	Matthew Purver (QMUL), Ravi Shekhar (QMUL)	First complete draft.
11/02/2022	v1.1	Matthew Purver (QMUL)	Submitted for internal review.
11/02/2022	v1.2	Matej Martinc (JSI)	Internal review.
15/02/2022	v1.3	Shane Sheehan (UE)	Internal review.
23/02/2022	v1.4	Matthew Purver (QMUL), Ravi Shekhar (QMUL)	Revision based on internal reviews.
26/02/2022	prefinal	Nada Lavrač (JSI)	Report quality checked.
27/02/2022	final	Matthew Purver (QMUL)	Report finalized.
28/02/2022	submitted	Tina Anžič (JSI)	Report submitted.

Table of Contents

List of abbreviations

API	Application Programming Interface
BERT	Bidirectional Encoder Representations from Transformers
CC-BY	Creative Commons - By Attribution
cseBERT	Croatian/Slovenian/English trilingual BERT
DNN	Deep Neural Network
ExM	Ekspress Meedia
mBERT	Multilingual BERT
MT	Machine Translation
NLP	Natural Language Processing
NN	Neural Network
NYT	New York Times
UGC	User-Generated Content
STY	Styria Media Group

1 Introduction

The EMBEDDIA project aims to improve cross-lingual transfer of language resources and trained models using word embeddings and cross-lingual technologies, with a focus on nine languages: Croatian, English, Estonian, Finnish, Latvian, Lithuanian, Russian, Slovene, and Swedish. Work package WP3 aims to apply EMBEDDIA's cross-lingual advances to help news media companies better serve their audience by understanding and analysing their reactions, and assuring the safety, fairness and integrity of their participation in public internet spaces. The focus is on automatic processing of user-generated content (UGC), primarily the comments readers post under news articles (in contrast to WP4, where the focus is on the news articles themselves). The specific objectives of WP3 are as follows:

- O3.1 Advance cross-lingual context and opinion analysis, via Task T3.1;
- O3.2 Develop cross-lingual comment filtering, via Task T3.2;
- O3.3 Develop techniques for report generation from multilingual comments, via Task T3.3.

The remaining task, T3.4, focuses on resource gathering, benchmarking and evaluation; the objective of this task is therefore to develop resources for evaluation, to assess the performance of the tools and techniques developed in Tasks T3.1-3.3, and to benchmark them against other possible methods.

Task T3.4 was therefore crucial in the early stages of the project, to gather datasets and other resources for use in the development and training of models in T3.1-3.2; this initial work was reported in Deliverable D3.1 at M9. It has become our focus again in the final stages of the project, to evaluate our tools, compare performance across methods, and produce the final data resources used. In this report, we describe the main datasets produced, including details of ongoing efforts to build new supporting resources; evaluate some possible improvements in our models; and compare against a competing baseline method using machine translation. The main contributions presented in this report (in the order of appearance) are as follows:

- A set of new datasets for UGC research: large, multi-year news comment archives in less-resourced languages, publicly released and distributed via CLARIN;
- A new dataset for agreement and stance research, to be released publicly, together with an evaluation of the accuracy of an agreement classifier trained thereon;
- A new evaluation of the improvements that can be gained by varying transfer learning methods in comment filtering tools;
- A new comparison of the EMBEDDIA cross-lingual transfer approach to a machine translation baseline, showing that the EMBEDDIA approach outperforms it as soon as even a small amount of training data can be obtained.
- New public tools for comment filtering, publicly released, integrated into the EMBEDDIA Media Assistant, and tested successfully within a real-world production environment.

This report is split into 6 further sections. Section 2 summarises the motivation and objectives of the tasks in this work package. In Section 3, we describe our progress in the primary task of focus in this deliverable (T3.4), producing and releasing the datasets required. Sections 4, 5 and 6 then describe our progress in using these resources to develop tools for Tasks T3.1, T3.2 and T3.3 respectively. Section 7 then summarises the main outputs of this final phase of our work.

2 Background

This section explains the motivation and aims of this work, including the specific objectives of each of the tasks involved. We do not go into detail on the related research background (the state of the art before the project began), as this has been covered in previous deliverables (see Deliverable D3.1 for a

general overview; D3.2 and D3.4 for comment analysis; D3.3 and D3.6 for comment filtering; and D3.5 for comment reporting).

2.1 Comment analysis (Task T3.1)

Work on user needs in WP6 Task T6.1 identified that news media users would like to have tools that can help perform *comment analysis*, to help them identify a range of meaningful phenomena such as opinions, sentiment and author characteristics, and to help identify fake news and misinformation. Analysis could take a wide range of possible directions, and over the course of the EMBEDDIA project we focused on four:

- author analysis, including gender, age and influence detection;
- sentiment analysis, to detect positive and negative content;
- stance and agreement analysis, to detect the opinions and consensus expressed;
- detection of fake news and misinformation spreading.

Our primary objectives were therefore to develop methods and tools for these four tasks. As the work progressed, a number of subsidiary tasks emerged, including the modelling of comment thread context, detection of topics discussed, integration with context beyond the comments, and detection of bot-authored comments. As across all EMBEDDIA work, we also had the general aim to make the methods developed applicable to less-resourced languages, either via cross-lingual transfer or by choice of methods with low data requirements. More detail of the research background is given in Deliverables D3.2 and D3.4.

2.2 Comment filtering (Task T3.2)

Work on user needs in WP6 Task T6.1 identified the primary need for news media users with regards to UGC to be automatic *comment filtering*, to help media partners deal with their need to quickly moderate large volumes of user-generated comments (see WP6, particularly Deliverable D6.5). The primary requirement is summed up by the user story shown here as Figure 1. This describes the problem that must be solved, and the tool which we aimed to develop in order to solve it.

The objective of Task T3.2 was therefore to develop methods for comment filtering - identifying comments which moderators should block to prevent them appearing on a public news site. Note that hate speech/abuse detection and trolling detection are two major categories on which filtering can be based, but many other phenomena must also be taken into account; and that the ability to label outputs with information about which category a to-be-blocked comment belongs to (i.e. the reason for blocking) is important. As across the whole EMBEDDIA project, it was important that these methods be cross-lingual, allowing them to be developed and trained even when tool and data resources may not be available in the target language (e.g. Croatian or Estonian, for our media partners). More detail of the research background is given in Deliverables D3.3 and D3.6.

2.3 Comment reporting (Task T3.3)

The final objective of WP3 was in *comment reporting*: to develop methods for generating human-readable reports from user comment data, to allow news media users to quickly survey and understand the activity going on in their comment sections.

Rather than formulating this problem directly as a text-to-text generation task, it was split into two conceptually distinct steps: comment analysis and report generation. This facilitates the integration of comment analysis components developed in Tasks T3.1 and T3.2 above, and from other parts of the

Branko is a moderator at 24sata, the largest-circulation daily newspaper in Croatia. 24sata reaches about 2 million readers daily, and many of them post comments about online articles: on an average day, about 8000 comments come in, spread over several hundred articles. Unfortunately, many comments (usually between 5% and 10%) need to be blocked to prevent them appearing online: they might be offensive, dangerous, or legally compromising. This is Branko's job.

Until now, the task of comment filtering and moderation had to be performed almost entirely manually. This is time-consuming and skilled work: the newspaper has a complex moderation policy, as comments may need blocking for a variety of reasons. Some are irrelevant spam or advertising, some contain disinformation, some are threatening or hateful, some obscene or illegal, some written in foreign languages . . . so filtering through them all and making consistent decisions is difficult, especially at peak times when over 1,000 per hour may be coming in. Branko uses a system which flags comments that match a list of blacklisted keywords, but this isn't very accurate and is hard to keep up to date as new topics get discussed. With the current COVID-19 crisis, for example, new kinds of spam, fake stories and ethnically-targeted hate speech emerge very fast, and the word lists can't keep up. That means that Branko largely has to rely on fast reading and experience.

The new EMBEDDIA tools for automated comment moderation have made Branko's job much easier. Comments are filtered in real time, automatically detecting those which are most likely to need blocking, ranking them by severity, and labelling them as to which part of the 24sata policy they seem to break. The final decision is left to Branko, but now he can easily prioritise the worst cases first, and make sure they don't appear on the site, without having to read through all the others. He can then check less severe cases, and can leave unproblematic comments where the classifier is very confident for a less busy time. Branko's final decisions are stored and fed back to the system, so that it learns over time to improve, and to adapt to new vocabulary as new topics and stories develop.

Figure 1: User story from Deliverable D6.5: Comment filtering at 24sata, provided by Croatian EMBEDDIA partner Trikoder (Styria Group).

EMBEDDIA project, while allowing them to be integrated with specific comment summarization methods developed on this task. Our specific objectives on this task were therefore to develop methods for summarizing comments, and then generating reports on the basis of those summaries integrated with other analysis data. Again, given the overall aim of the EMBEDDIA project to produce tools for less-resourced languages, we aimed to develop methods which could work with little or no adaptation to a new language. More detail of the research background is given in Deliverable D3.5.

3 Datasets and resources (Task T3.4)

In this section, we describe the work performed under the supporting Task T3.4 (Resource gathering, benchmarking and evaluation). This task has run from the beginning of the project, to support and evaluate the main tasks, continues to near the end of the project (M38), and is responsible for this deliverable.

3.1 Datasets and resources: early progress

This task has had one previous deliverable, D3.1 at M9, in which we surveyed the existing public datasets available to support WP3, and introduced two new large datasets contributed by media partners ExM and STY. These two new datasets provided the main basis for model development and testing for the EMBEDDIA target languages: they are both large, covering c.10 years of comment archives and including c.30M comments each, and are in the less-resourced languages of interest. The Ekspress

Meedia (ExM) dataset is mostly in Estonian language, taken from Ekspress's publications including Eesti Ekspress, Estonia's highest-circulation daily newspaper. The Styria (STY) dataset is mostly in Croatian language, taken from Styria's news outlets 24sata and Večernji List (two of Croatia's highest-circulation daily newspapers). Both datasets were created with metadata to support comment analysis and filtering experiments, including consistent author IDs and the blocking decisions made by the newspaper moderators. At this early stage, these datasets were available for use within EMBEDDIA but not beyond.

3.2 Datasets and resources: final outputs

As part of Task T3.4 we worked to obtain agreements to publicly release the data, and establish suitable usage licenses and policies for anonymisation under which the data could be safely released (reflected in our Data Management Plan, see Deliverables D8.3 and D8.7).

3.2.1 News Comments Archives

Three general news comment datasets have been made publicly available (including the two introduced in Section 3.1 above and a third from Ekspress's DELFI in Latvian language). 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 or the usernames visible on their sites. User IDs for moderated comments have been removed. These datasets include basic metadata available from the media houses' systems (e.g. comment posting timestamp, associated news article link) and moderation data (label applied by moderator), but no further annotation.

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 has been made publicly available in CLARIN under a CC-BY licence.[1]

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 has been made publicly available in CLARIN under a CC-BY licence.[2]

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 has been made publicly available in CLARIN under a CC-BY licence.[3]

Most work in WP3 has progressed using these major new comment datasets. To our knowledge, they are the first publicly-available comment datasets in their respective languages, and the largest comment datasets available. They have already been used as the basis of research by teams outside the project, in the EMBEDDIA EACL 2021 hackathon (see e.g. Korencic et al., 2021).

[1] http://hdl.handle.net/11356/1401
[2] http://hdl.handle.net/11356/1407
[3] http://hdl.handle.net/11356/1399

3.2.2 Recent progress: stance annotation

Most work in WP3 has used either these new comment datasets, and/or publicly available datasets for specific phenomena (sentiment, fake news detection etc.). In some cases, however, new annotation has been required: in particular, no suitable resources existed to evaluate our work in stance, opinion and agreement detection on comments in the languages of interest, and recent work on T3.4 has been dedicated to developing a new dataset to support this.

Dataset Source We use comments from two newspapers, the New York Times (NYT) in English and 24sata in Croatian. The conversational context that emerges in threads can be key to understanding comments. We intend analysis of this context to be an important part of the task. We therefore impose the following conditions on our dataset selection: (1) A thread must have $\geq$5 comments, with $\geq$2 unique users commenting $\geq$2 times each; (2) An article must have $\geq$2 threads which satisfy the previous condition. Condition (1) makes sure than there is some form of conversation; condition (2) ensures that multiple readers engaged with the article/thread, helping ensure a range of opinions. We randomly selected 50 articles with equal distribution across a range of news categories (articles are tagged as e.g. *Politics*, *Sport*, *Finance* etc.). This ensures that our data is as diverse as possible in terms of vocabulary and topics discussed.

Annotation Annotators were recruited directly, in preference to crowdsourcing, to ensure quality in a relatively complex annotation task. We have already recruited native English and Croatian speakers in London and Zagreb, respectively; all annotators are graduate or higher-level students and paid hourly. An hour of training and 2-3 pilot annotations are used to provide feedback to ensure consistency across annotators.

First the whole thread is shown to the annotator; they are asked to read the thread to get the whole context of the discussion. Each individual comment is then focused on in turn, and annotators must label agreement/dis-agreement between the comment and its antecedent on a 5-point scale (strong/weak disagreement, neutral/none, weak/strong agreement). Where antecedent can not explicitly be determined from the existing thread structure (the 24sata data often underspecifies structure). We annotated the antecedent before the agreement/dis-agreement annotations.

Then in the same screen the annotator must provide a free-text explanation of their decision. The explanation should be chosen so as to contain the main point(s) of agreement/dis-agreement, rather than an overall summary of one or both comments. In Table 1, we report the statistics of the dataset.

Table 1: Data distribution for the New York Times (English) and 24sata (Croatian) datasets.

Source	Train		Val		Test	
	#articles	#comment	#articles	#comment	#articles	#comment
NYT (EN)	30	396	10	175	10	162
24Sata (HR)	30	720	10	321	10	279
Total	60	1116	726	496	730	441

This new dataset is now available to support stance and agreement detection work in T3.1 - see Section 4 below. The data is being made public via the EMBEDDIA GitHub page,[4] and will be deposited on CLARIN once final checks are complete.

In general, the purpose behind producing the resources described in this section is to support the main tasks T3.1, T3.2 and T3.3. In the next sections, we turn to summarising those tasks with their main achievements and outputs, and describing recent progress in improving, evaluating and benchmarking their outputs.

[4] https://github.com/EMBEDDIA/contextual-view

4 Comment analysis (Task T3.1)

This section describes the main achievements of Task T3.1 (comment analysis), and recent progress in evaluating its outputs.

4.1 Main results and achievements

The objectives of this task were to develop methods for analysis of comments from a range of perspectives, suitable for less-resourced languages. Our work focused on a number of specific analysis tasks including author analysis, sentiment and opinion analysis, and fake news and misinformation detection. This task formally ended at M30, and thus the majority of this work is described in the dedicated Deliverable D3.4. The main achievements were as follows:

- Classifiers for author profiling (including age and gender detection) with state-of-the-art accuracy based on text and social network context; our tools ranked 2nd out of 8 teams making 78 submissions in the CLIN 2019 cross-genre author profiling shared task (Martinc & Pollak, 2019), 3rd in the PAN 2019 author profiling shared task (Martinc et al., 2019b), and 2nd in the PAN@CLEF 2020 author profiling shared task (Koloski et al., 2020a).
- Classifiers for bot and gender detection with good accuracy, ranked 16th out of 55 teams in the PAN 2019 bot and gender profiling shared task (Martinc et al., 2019a).
- Cross-lingual classifiers for sentiment detection, using WP1's cross-lingual models to transfer between EMBEDDIA languages with no measurable performance drop (Robnik-Šikonja et al., 2021).
- Multi-lingual classifiers for opinion detection in social media (Tabak & Purver, 2020),
- Methods for incorporating contextual knowledge into DNN classifiers (Rohanian et al., 2019), including incorporating multi-lingual topic models to improve the performance and interpretability of comment classifiers (Zosa et al., 2021).
- Classifiers to detect the spreading of fake news and misinformation, ranked 3rd in the PAN@CLEF shared task (Koloski et al., 2020b) and scoring within 1.5% of the top entry in the CONSTRAINT 2021 shared task (Koloski et al., 2021).

4.2 Recent progress: stance analysis

Since D3.4 we have focused on improving our methods for stance and opinion analysis, making them more effective on news comment data (rather than the social media data used in work up to D3.4) and evaluating their effectiveness on our EMBEDDIA datasets.

Stance is a complex phenomenon (see e.g. Zubiaga et al., 2016) and our evaluation so far therefore focuses on two specific components of the overall task. Based on the annotation of our new dataset (Section 3.2.2 above), the following two tasks have been performed:

Task A: (Dis-)Agreement Classification The simplest version of the task is to classify a given comment as agreeing or disagreeing with its antecedent comment. This can be framed as a classification task over pairs of comments: for any pair, predict the correct label from a three-way choice (agree, disagree, none/mixed).

Task B: Explanation Generation The final task is generative: for each classification decision in Task A, systems must produce a short text explanation of their decision. This text is expected to include the key words/phrases in the comments that make the viewpoints and (dis)agreements clear, but may rephrase or reformulate them freely.

These two tasks test a model's understanding, specifically via explanation (Wiegreffe & Marasović, 2021). This task is envisioned to require fine-grained complex analysis to generate explanations, gauging the potential of state-of-the-art NLP models.

To solve these two tasks, first we created models to perform Task A, and Tasks A & B together. Task A is modelled as a classification task using the EMBEDDIA BERT (cseBERT, Ulčar & Robnik-Šikonja, 2020). For Task A & B together, we used multi-task setting. Specifically, we used shared encoding, in which both classification (Task A) and generation (Task B) were performed together. We also tested zero-shot cross-lingual transfer of the method, i.e. trained on the English data and tested on the Croatian data (i.e. NYT (EN) -> 24sata (HR)) and vice-versa. In Table 2, we report the F1 score on the different settings. We found that model trained with the explanation performs comparatively better (0.80 vs 0.78 and 0.75 vs 0.72), while zero-shot cross-lingual transfer performance is much lower (0.72 vs 0.47 and 0.78 and 0.55). This shows that a model trained and tested on the same language achieves better performance and having human readable explanation further improves performance, but suggests that cross-lingual transfer is challenging. However, this may not just be due to linguistic differences, but could also be attributed to the nature of the comments in different newspapers: NYT comments are longer and highly informative, while 24sata comments are shorter and less formal. Code for these experiments and training models is being made public via the EMBEDDIA GitHub page.[5]

Table 2: F1 score for the Agreement/Disagreement task with and without Explanation.

	Without Explanation	WithExplanation
NYT (EN)	0.72	0.75
24Sata (HR)	0.78	0.80
24Sata (HR) -> NYT (EN)	0.47	0.48
NYT (EN) -> 24Sata (HR)	0.55	0.51

5 Comment filtering (Task T3.2)

This section describes the main achievements of Task T3.2 (comment filtering), and recent progress in evaluating, benchmarking and applying its outputs.

5.1 Main results and achievements up to D3.6

The overall objective of this task was to develop tools for automated news comment filtering in less-resourced languages, particularly in Croatian and Estonian, in the face of a lack of annotated data in those target languages. In this we were successful, producing not only accurate filtering classifiers, but developing general methods by which useful classification tools can be built without requiring significant manual annotation and development effort. This task formally ended at M34, and thus the majority of this work is described in the dedicated Deliverable D3.6. The main achievements were as follows:

- Classifiers for specific filtering categories, hate speech and abuse detection, with good accuracy on standard datasets in well-resourced languages, including 4th place in the SemEval 2019 OffensEval task (Pelicon et al., 2019).
- Classifier models for comment filtering, including hate speech and offensive language detection, trained on real news comment data and in EMBEDDIA less-resourced languages (Pelicon, Shekhar, Martinc, et al., 2021; Pelicon, Shekhar, Škrlj, et al., 2021).
- Classifiers that produce predicted blocking decisions together with their reasons (category of policy rules broken), trained and evaluated against real moderator behaviour on EMBEDDIA news

[5] https://github.com/EMBEDDIA/contextual-view

media partner comment data in EMBEDDIA project languages (Croatian, Estonian) (Shekhar et al., 2020).

- Methods for cross-lingual classifier training, based on embeddings models from WP1, that can equal the accuracy of classifiers trained in the standard monolingual way, but with much less target-language training data (Pelicon, Shekhar, Škrlj, et al., 2021).
- Methods for improving the accuracy, confidence and interpretability of comment filtering classifiers by incorporating topic information (Zosa et al., 2021).
- Implemented multi-lingual classifier code and models, available as dockerized components for integration with the Embeddia Assistant in WP6.
- Implemented multi-lingual classifier code and models, publicly available as research software and as services behind APIs for software integration, and with a Dockerized version available for integration with the EMBEDDIA Media Assistant in WP6.

5.2 Recent progress and improvements

Evaluation of our cross-lingual filtering models showed that performance was good (and can equal conventional monolingual training with a fraction of the target-language training data), but that it is limited by several factors of the cross-lingual transfer. One of the most important is the appropriateness of the base vocabulary of the language model: the word and/or sub-word token vocabulary of a model initially trained on a source language may not be optimal for a different target language.

There is growing evidence that domain-specific training of a language model improves performance compared to the generic language model (Koto et al., 2021; Sachidananda et al., 2021; Tai et al., 2020; Wang et al., 2020; Yao et al., 2021; Zhu et al., 2021). There are two major approaches in this direction: firstly training the language model on the target domain data from scratch, and secondly fine-tuning a generic language model on target domain data. To train the language model from scratch requires a large amount of target domain data, which is rarely available in less-resourced languages. In our work described above, we performed domain specific training by fine-tuning the language model using the EMBEDDIA data. The domain specific fine-tuning of the language model improves the overall performance, however, it suffers from a major drawback that vocabulary used is still from the generic language model. To overcome this, we proposed to extend the vocabulary of the generic language model based on the domain data and then finetune it. Koto et al. (2021) show that average is better than random and sum initialization of new words. In our experimentation, we used the following 4 types of initializations:

- Random: randomly initialize a new word
- Avg: take the average of the words from the existing vocabulary
- Sum: take the sum of the words from the existing vocabulary to initialize a new word
- Max: take the elementwise maximum of the words from the existing vocabulary to initialize a new word

Table 3: F1-Score based on Domain Adaptation. Comparison of mBERT by (Devlin et al., 2019) and the trilingual model cseBERT from EMBEDDIA (Ulčar & Robnik-Šikonja, 2020)

	mBERT	cseBERT
No Domain Adaptation	73.7	75.8
Domain Adaptation with small data (1M Sentences)	75.9	76.2
Domain Adaptation with large data (12M Sentences)	76.2	76.5

First, we used the Croatian 24sata newspaper comment data to fine-tune mBERT and cseBERT. To measure the effectiveness of the domain data size, we trained on two datasets: first, containing all data

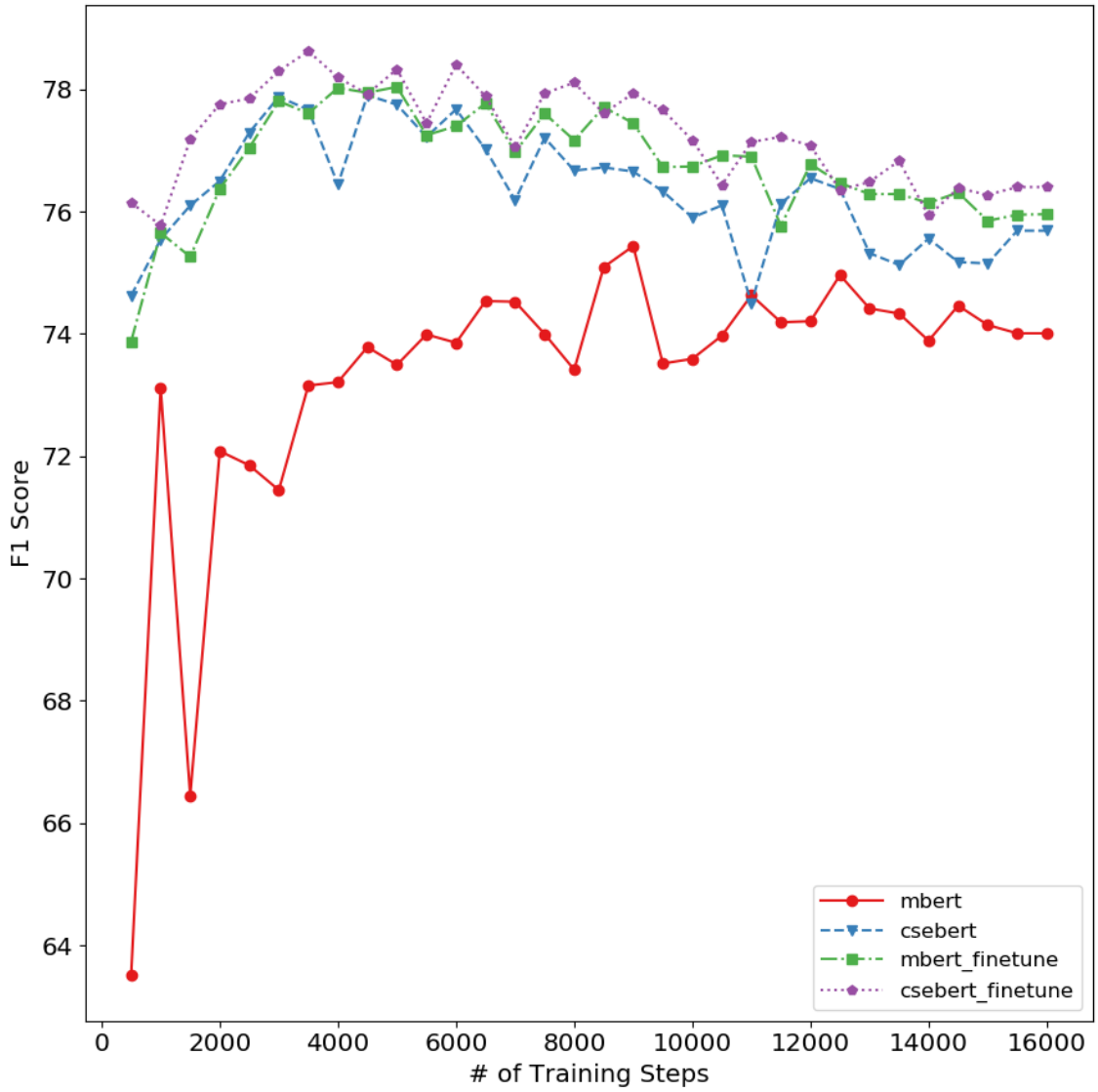

Figure 2: Effect of training steps. Here, we report the change in the F1-score as training progress, where the F1 score is the harmonic mean of the precision and recall. 'mbert/csebert' denotes when it is tuned for the classification task only and 'mbert_finetune/csebert_finetune' when language model is finetuned for the domain adaptation followed by the classification task.

till 2018, and another, containing only 2018 data. For the task training, we used the same split as in our previous experiments (Pelicon, Shekhar, Škrlj, et al., 2021). This allows us to quantify the effect of dataset size on domain-specific language model fine-tuning. The comment moderation task is trained and tested on the same set as used by Pelicon, Shekhar, Škrlj, et al. (2021). From Table 3, we can see that domain fine-tuning improves the performance of both mBERT and cseBERT, while the gain for the mBERT is more significant compared to cseBERT (73.7 vs 76.2 and 75.8 vs 76.5). Since mBERT is trained on 102 languages, domain fine-tuning has a significant effect. Further, having a large dataset (12M Sentences) for domain adaption doesn't improve performance as much. We decided to use small data(1M Sentences) for language model fine-tuning based on these results. Using this small data, we trained our model for 10 epochs and observed the effect of target training steps in Figure 2.

Table 4: F1-Score Comparison based on different data source and vocabulary initialization.

	General Data		24sata Data	
	mBERT	cseBERT	mBERT	cseBERT
Random	68.94	76.30	61.05	76.91
Average	73.53	76.93	73.43	78.03
Sum	72.25	76.27	73.44	76.65
Maximum	73.34	75.05	74.01	75.69

To separate the effect of language from the domain data, we fine-tuned the language model using generic data, following Ulčar & Robnik-Šikonja (2020), and domain-specific 24sata comment data from 2018. From Table 4, we could observe that for the cseBERT, averaging of token embedding performs best, and there is a significant increase by using 24sata when averaging of tokens is used. For all other cases, the improvement is negligible. However the results for mBERT are very interesting: when adding new vocabulary, performance deteriorates for both general and 24sata data. One reason of this could be that due to larger size of mBERT, it requires more number of epochs to learn the representation of the new tokens. This again advocates that having a smaller and focused language model such as the EMBEDDIA BERT (cseBERT, Ulčar & Robnik-Šikonja, 2020) is beneficial.

5.3 Comparison to other baselines

The second direction we have pursued since the formal end of this task has been to compare the cross-lingual performance to an alternative way of creating tools for less-resourced languages: the use of machine translation (MT). In general, our approach on EMBEDDIA has been one of cross-lingual transfer: developing target-language tools by first pre-training multi-lingual language models on unlabelled data (easily available even in most less-resourced languages), then fine-tuning on suitable available labelled datasets in well-resourced languages; the transfer properties of the pre-trained model then allow the tool to perform even in the less-resourced target language.

An alternative approach would be to rely on MT: if an accurate MT system exists for the source and target language pair of interest, one could use it to translate less-resourced target language examples into the well-resourced language (e.g. English), and apply a monolingual English classifier (which could be trained in the standard way using English datasets). This has some notable disadvantages: the use of MT is very computationally demanding; it adds a time overhead, making it harder to use in real-time services; high-quality MT does not exist for all language pairs; and adaptation to language- and culture-specific domain effects would be harder to achieve. However, we investigate it here in order to discover how its performance would compare.

In (Pelicon, Shekhar, Škrlj, et al., 2021), we proposed an intermediate training to improve the overall performance of the comment moderation and compared mBERT and cseBERT models, and showed that cseBERT outperforms mBERT. However, in general, there is a large amount of data available to train in English. In Table 5, we compare different settings on Croatian and Slovenian. First, we train a model on English data and test on MT-translated Croatian and Slovenian test sets, without training on

target-language data. The translation is performed using the Google MT service. We then compared the results with the results from (Pelicon, Shekhar, Škrlj, et al., 2021) in three settings: zero-shot setting (model trained on English data only and tested on original Croatian and Slovenian data), full-data setting (model trained on English data as well as 100% of Slovenian/Croatian data and tested on original Slovenian/Croatian test sets) and 10% data setting (model trained on other languages as well as 10% Slovenian/Croatian data and evaluated on original Slovenian/Croatian data). All results are represented as F1 scores.

Table 5: F1 Score Comparison with Translated Data.

	Slovenian		Croatian	
	mBERT	cseBERT	mBERT	cseBERT
Translated using MT	62.03	64.97	66.00	66.57
Cross-lingual: Zero Shot	59.57	63.98	60.30	67.70
Cross-lingual: Few-shot fine-tuned (10%)	68.22	72.63	66.82	70.91
Cross-lingual: Fully fine-tuned (100%)	72.33	76.78	71.96	76.54

As shown in Table 5, in a pure zero-shot setting (no labelled target-language data used for tuning), the MT approach outperforms the cross-lingual approach for the standard massively multi-lingual mBERT. For the EMBEDDIA cseBERT, the difference is not clear: MT appears to perform better for Slovenian, but worse for Croatian. This suggests that when there is no data for the target language available at all, the MT approach is viable and comparable to the cross-lingual approach with a good base pre-trained language model.

However, in the more realistic setting where some small amount of labelled target-language data can be obtained (we simulate this scenario in the 10% case), cross-lingual training on the target data is effective, giving a significant increase in performance: for Slovenian, c.7.5% absolute F1-score improvement, with c.4.5% improvement for Croatian. Note also that cseBERT always performs better than mBERT. If even small amounts of target-language data can be labelled when implementing a new tool, using the EMBEDDIA cross-lingual approach outperforms the use of MT, as well as reducing the time and computational overhead that would be required in use.

5.4 Final release and real-world exploitation

Finally, we worked on the public application and exploitation of our comment filtering tools, including releasing a range of models publicly for research and commercial exploitation, and working with media partner Styria to integrate a version of our tools into their production systems, and test and evaluate its use by real moderators in their actual work.

5.4.1 Release of new tools

At the time of the last comment filtering deliverable (D3.6), we had publicly released software for the training and development of comment moderation classifiers, together with a pre-built version available to run as a service behind an API, installed via Docker. This pre-built version, although able to process multiple languages, was optimised for English as it was trained on English data using a standard mBERT language model.

Since then, we have added and publicly released pre-built tools optimised for broader multilingual use, and for specific less-resourced EMBEDDIA language combinations. The models now available are as follows:

- Model using general mBERT, trained on English data;
- Model using general mBERT, trained on data in English, German, Estonian, Slovenian, Croatian;

- Model using trilingual CroSloEngualBERT, trained on data in English, Slovenian, Croatian;
- Model using trilingual FinEstBERT, trained on data in English and Estonian.

Each model is available as an API, installable via Docker, and available via the EMBEDDIA Media Assistant at `https://embeddia.texta.ee/`. Direct links to access these models are given in Section 8.

5.4.2 Evaluation in situ

One of the above-mentioned tools for comment moderation was integrated into the production system at 24sata, one of Styria's newspapers and the highest-circulation daily newspaper in Croatia, in December 2021. Its output was integrated into the user interface used by the 24sata newspaper moderators to monitor comments. Until now, moderators blocked comments entirely manually with the aid of a word blacklist: the system's outputs can now be used to highlight comments likely to need blocking, and show the reason in terms of 24sata's moderation policy rules (see details in Shekhar et al., 2020). The moderators use the highlighting to decide which comments to review, speeding up their job of deciding to block a comment if it violates any of the moderation rules. After using the system for more than a month, the moderators reported that the system is effective and saves them time. Moreover, the moderators unanimously stated that they want to keep using it, with some minor modifications to its output and visualisation. We will report on this evaluation process in more detail, together with more details of the feedback from moderators, in Deliverable D6.12 (Final report on EMBEDDIA Assistant platform evaluation).

6 Comment reporting (Task T3.3)

The work on comment reporting finished in M30 and was fully described in Deliverable D3.5. Progress on the natural language generation (NLG) methods used in this work has continued, however, as part of WP5. This work included final evaluation of the NLG methods and their capabilities, including aspects that are shared across all the NLG work in the EMBEDDIA project (including comment reporting), especially the underlying multi-lingual NLG architecture developed. This evaluation is described in Deliverable D5.7, and we therefore do not include it here.

7 Conclusions and further work

The focus of WP3 is on automatic processing of user-generated content (UGC), primarily the comments readers post under news articles, and specifically on methods for analysing, filtering and reporting on those comments. The specific focus of this task T3.4 is on datasets to support development and evaluation of our UGC classifiers, on evaluating the performance of the tools and techniques developed in Tasks T3.1-3.3, and on benchmarking them against other possible methods.

As shown in this report, we have successfully developed a range of suitable datasets, publicly releasing several large collections of news comments in our target less-resourced languages, the first of their kind in Croatian, Estonian and Latvian. Our methods for cross-lingual training have proved effective in developing analysis and filtering tools, that can be applied cross-lingually, by training them initially on datasets available in well-resourced languages like English and fine-tuning them on less-resourced target languages to adapt them quickly with little need for expensive training data. We have also shown that this method outperforms the use of machine translation, giving better accuracy as soon as small amounts of training data in the target language can be obtained.

A comment filtering tool trained using these methods, part of our EMBEDDIA Media Assistant, has now been integrated into a real working industry system at 24sata, used by moderators within their work, with positive results; this will be reported on in full in Deliverable D6.12.

8 Associated outputs

The work described in this deliverable has resulted in the following resources:

Citation	Status	Appendix
Shekhar, R.,& Purver, M. “Exploring Large Scale NYT News Comment Dataset.” (Submitted to LREC)	Under review	(available on request)
Ch Kranti, Shekhar, R.,& Purver, M. “Cry for Help: India Covid Dataset during Delta Wave.” (Submitted to LREC)	Under review	(available on request)
Shekhar, R.,Karan, M., & Purver, M. “Domain Adaptation via Vocabulary Extension for Comment Moderation”	In preparation	(available on request)

Description	URL	Availability
Code and models for comment filtering (Pelicon, Shekhar, Martinc, et al., 2021)	github.com/EMBEDDIA/hackashop2021_comment_filtering	Public (CC0)
Code for cross-lingual training (Pelicon, Shekhar, Škrlj, et al., 2021)	github.com/EMBEDDIA/cross-lingual_training_for_offensive_language_detection	Public (MIT)
Dockerized API for comment filtering (EN training data) (Pelicon, Shekhar, Škrlj, et al., 2021)	github.com/EMBEDDIA/comment-filter	Public (MIT)
Dockerized API for comment filtering (EN/ET training data) (Pelicon, Shekhar, Škrlj, et al., 2021)	github.com/EMBEDDIA/comment-filter-finest-bert-engee	Public (MIT)
Dockerized API for comment filtering (EN/HR/SL training data) (Pelicon, Shekhar, Škrlj, et al., 2021)	github.com/EMBEDDIA/comment-filter-csebert-cse	Public (MIT)
Dockerized API for comment filtering (EN/DE/SL/ET/HR training data) (Pelicon, Shekhar, Škrlj, et al., 2021)	github.com/EMBEDDIA/comment-filter-mbert-multi	Public (MIT)
TEXTA Toolkit including Estonian comment filtering	github.com/EMBEDDIA/texta-rest	Public (GPL)
Code and dataset for stance analysis	github.com/EMBEDDIA/contextual-view	Public (MIT)

Bibliography

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (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)* (pp. 4171–4186).

Koloski, B., Pollak, S., & Škrlj, B. (2020a). Know your neighbors: Efficient author profiling via follower tweets. In *Proceedings of the 11th International Conference of the CLEF Association (CLEF 2020).*

Koloski, B., Pollak, S., & Škrlj, B. (2020b). Multilingual detection of fake news spreaders via sparse matrix factorization. In *Proceedings of the 11th International Conference of the CLEF Association (CLEF 2020).*

Koloski, B., Stepišnik-Perdih, T., Pollak, S., & Škrlj, B. (2021). Identification of COVID-19 related fake news via neural stacking. In T. Chakraborty, K. Shu, H. Bernard, H. Liu, & M. Akhtar (Eds.), *Combating online hostile posts in regional languages during emergency situation. CONSTRAINT 2021* (Vol. 1402). Springer.

Korencic, D., Baris, I., Fernandez, E., Leuschel, K., & Sánchez Salido, E. (2021, April). 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* (pp. 127–133). Online: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2021.hackashop-1.18`

Koto, F., Lau, J. H., & Baldwin, T. (2021, November). IndoBERTweet: A pretrained language model for Indonesian Twitter with effective domain-specific vocabulary initialization. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 10660–10668). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2021.emnlp-main.833` doi: 10.18653/v1/2021.emnlp-main.833

Martinc, M., & Pollak, S. (2019). Pooled LSTM for Dutch cross-genre gender classification. In *Proceedings of the shared task on cross-genre gender prediction in Dutch at CLIN29 (GxG-CLIN29) co-located with the 29th conference on computational linguistics in the Netherlands (clin29).*

Martinc, M., Škrlj, B., & Pollak, S. (2019a). Fake or not: Distinguishing between bots, males and females. In *Proceedings of the 10th International Conference of the CLEF Association (CLEF 2019).*

Martinc, M., Škrlj, B., & Pollak, S. (2019b). Who is hot and who is not? profiling celebs on Twitter. In *Proceedings of the 10th International Conference of the CLEF Association (CLEF 2019).*

Pelicon, A., Martinc, M., & Kralj Novak, P. (2019, June). Embeddia at SemEval-2019 task 6: Detecting hate with neural network and transfer learning approaches. In *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 604–610). Minneapolis, Minnesota, USA: Association for Computational Linguistics.

Pelicon, A., Shekhar, R., Martinc, M., Škrlj, B., Purver, M., & Pollak, S. (2021, April). Zero-shot cross-lingual content filtering: Offensive language and hate speech detection. In H. Toivonen & M. Boggia (Eds.), *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report*

Generation (p. 30-34). Retrieved from http://www.eecs.qmul.ac.uk/~mpurver/papers/pelicon-et-al21eacl.pdf

Pelicon, A., Shekhar, R., Škrlj, B., Purver, M., & Pollak, S. (2021). Investigating cross-lingual training for offensive language detection. *PeerJ Computer Science*, *7*, e559. Retrieved from https://doi.org/10.7717/peerj-cs.559 doi: 10.7717/peerj-cs.559

Robnik-Šikonja, M., Reba, K., & Mozetič, I. (2021). Cross-lingual transfer of sentiment classifiers. *Slovenščina 2.0*, *9*(1), 1–25. doi: https://doi.org/10.4312/slo2.0.2021.1.1-25

Rohanian, M., Hough, J., & Purver, M. (2019). Detecting depression with word-level multimodal fusion. In *Proceedings of INTERSPEECH* (pp. 1443–1447). Graz, Austria: ISCA. (ISSN 1990-9772)

Sachidananda, V., Kessler, J., & Lai, Y.-A. (2021, November). Efficient domain adaptation of language models via adaptive tokenization. In *Proceedings of the second workshop on simple and efficient natural language processing* (pp. 155–165). Virtual: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.sustainlp-1.16 doi: 10.18653/v1/2021.sustainlp-1.16

Shekhar, R., Pranjić, M., Pollak, S., Pelicon, A., & Purver, M. (2020). Automating news comment moderation with limited resources: Benchmarking in Croatian and Estonian. *Journal of Language Technology and Computational Linguistics*, *34*, 49-79. (Special Issue on Offensive Language)

Tabak, T., & Purver, M. (2020). Temporal mental health dynamics on social media. In *Proceedings of the 1st workshop on NLP for COVID-19 (part 2) at EMNLP 2020.* Online: Association for Computational Linguistics. doi: 10.18653/v1/2020.nlpcovid19-2.7

Tai, W., Kung, H., Dong, X. L., Comiter, M., & Kuo, C.-F. (2020). exbert: Extending pre-trained models with domain-specific vocabulary under constrained training resources. In *Findings of the association for computational linguistics: Emnlp 2020* (pp. 1433–1439).

Ulčar, M., & Robnik-Šikonja, M. (2020). FinEst BERT and CroSloEngual BERT: less is more in multilingual models. In *Proceedings of Text, Speech, and Dialogue, TSD 2020.* doi: 10.1007/978-3-030-58323-1_11

Wang, Z., K, K., Mayhew, S., & Roth, D. (2020, November). Extending multilingual BERT to low-resource languages. In *Findings of the association for computational linguistics: Emnlp 2020* (pp. 2649–2656). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.findings-emnlp.240 doi: 10.18653/v1/2020.findings-emnlp.240

Wiegreffe, S., & Marasović, A. (2021). Teach me to explain: A review of datasets for explainable nlp..

Yao, Y., Huang, S., Wang, W., Dong, L., & Wei, F. (2021, August). Adapt-and-distill: Developing small, fast and effective pretrained language models for domains. In *Findings of the association for computational linguistics: Acl-ijcnlp 2021* (pp. 460–470). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.findings-acl.40 doi: 10.18653/v1/2021.findings-acl.40

Zhu, H., Peng, H., Lyu, Z., Hou, L., Li, J.-Z., & Xiao, J. (2021). Travelbert: Pre-training language model incorporating domain-specific heterogeneous knowledge into a unified representation. *ArXiv*, *abs/2109.01048*.

Zosa, E., Shekhar, R., Karan, M., & Purver, M. (2021, September). Not all comments are equal: Insights into comment moderation from a topic-aware model. In G. Angelova, M. Kunilovskaya, R. Mitkov, & I. Nikolova-Koleva (Eds.), *Proceedings of Recent Advances in Natural Language Processing (RANLP)* (pp. 1652–1662). online. Retrieved from https://doi.org/10.26615/978-954-452-072-4_185 doi: 10.26615/978-954-452-072-4_185

Zubiaga, A., Kochkina, E., Liakata, M., Procter, R., & Lukasik, M. (2016). Stance classification in rumours as a sequential task exploiting the tree structure of social media conversations. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 2438–2448). Osaka, Japan: The COLING 2016 Organizing Committee.