Language Models on a Diet: Cost-Efficient Development of Encoders for Closely-Related Languages via Additional Pretraining

Nikola Ljubešić[1,2], Vít Suchomel[3], Peter Rupnik[1], Taja Kuzman[1], Rik van Noord[4]
[1]Jožef Stefan Institute, [2]University of Ljubljana, [3]Masaryk University, [4]University of Groningen
nikola.ljubesic@ijs.si, vit.suchomel@sketchengine.eu,
peter.rupnik@ijs.si, taja.kuzman@ijs.si, r.i.k.van.noord@rug.nl

Abstract

The world of language models is going through turbulent times, better and ever larger models are coming out at an unprecedented speed. However, we argue that, especially for the scientific community, encoder models of up to 1 billion parameters are still very much needed, their primary usage being in enriching large collections of data with metadata necessary for downstream research. We investigate the best way to ensure the existence of such encoder models on the set of very closely related languages – Croatian, Serbian, Bosnian and Montenegrin, by setting up a diverse benchmark for these languages, and comparing the trained-from-scratch models with the new models constructed via additional pretraining of existing multilingual models. We show that comparable performance to dedicated from-scratch models can be obtained by additionally pretraining available multilingual models even with a limited amount of computation. We also show that neighboring languages, in our case Slovenian, can be included in the additional pretraining with little to no loss in the performance of the final model.

Keywords: additional pretraining, named entity recognition, sentiment analysis, causal commonsense reasoning, Croatian, Serbian

1. Introduction

The field of natural language processing is in the middle of a paradigm shift due to the emergence of large language models (LLMs) that showcase impressive capabilities across a diverse range of natural language understanding tasks. While the current front-runners mainly cover English and some other 'large' languages (Zhang et al., 2022; OpenAI, 2023; Touvron et al., 2023), it is just a matter of time for those models to start performing on a similar (or even higher) level for less-resourced languages. One example is the COPA benchmark for South Slavic languages. This task was just partially solvable by smaller non-English language models (Ljubešić and Lauc, 2021), to which GPT-3.5 Turbo has been catching up significantly even for very under-resourced languages such as Macedonian. What is more, GPT-4 was shown to bring the performance for all South Slavic languages to the level of its performance on the English version of the same benchmark.[1]

With these developments, we are placed today in front of a big dilemma. Should we simply wait for large language models to become more parameter- and data-efficient, thereby encompassing our languages of interest with good-enough performance? Alternatively, is there still room for the up-to-1-billion-parameters models that we are able to pretrain with the limited computing capacity available in most of academia? Our claim is that, besides the pure academic endeavor of researching language modelling techniques, which are very needed activities by themselves, on the application side there is still a need for encoder models of the up-to-1-billion-parameters size, primarily for the enrichment of our research data, mostly large corpora, for downstream research. Examples of such enrichment are genre annotation of tens of millions of documents inside the CLASSLA web corpora of South Slavic languages with the X-GENRE Transformer-based classifier (Kuzman et al., 2023), or annotation of billions of tokens of the ParlaMint corpus of parliamentary proceedings with the latest Transformer-based sentiment models (Mochtak et al., 2023).

In addition to concerns that large language models might simply require too much computation (or even more problematic, API calls) to enrich millions of documents, there are additional issues with using large language models for data enrichment for scientific purposes. These considerations are twofold. Firstly, the decoder models do not generate limited classification or regression outputs, but free text, which is often hard to map to the pre-defined set of classes intended for downstream data analysis. And secondly, they perform overall great in zero-shot, in-context learning scenarios, but as the length of the instruction, provided in a prompt, is very limited, it is not pos-

[1] `https://github.com/clarinsi/benchich/tree/main/copa`

SIGUL2024 Workshop, pages 189–203
21-22 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

sible to provide detailed directions on how to separate between less clear cases, as can be achieved via manual annotation of thousands of instances, on which fine-tuned encoder models are based (Kuzman et al., 2023).

Languages in focus In this paper, we search for the best path towards creating well-performing encoder language models with less than a billion parameters for medium-sized languages. We perform our search on the example of the South Slavic pluricentric Serbo-Croatian macro-language (code `hbs` by ISO 639-3, called HBS onward). The HBS macro-language encompasses the following official languages: Bosnian (code `bs` by ISO 639-1), Croatian (`hr` by ISO 639-1), Montenegrin (`cnr` by ISO 639-3) and Serbian (`sr` by ISO 639-1). We investigate the following options: (1) pretraining the models from scratch, as is the case with the BERTić model (Ljubešić and Lauc, 2021), pretrained on more than 8 billion words of Croatian, Bosnian, Montenegrin and Serbian texts, or the cseBERT model (Ulčar and Robnik-Šikonja, 2020), pretrained on Slovenian, English and Croatian texts, and (2) additionally pretraining existing multilingual models, specializing them for the languages of interest.

Research questions To explore the second option, we additionally pretrain base-sized and large-sized XLM-RoBERTa (XLM-R) models (Conneau et al., 2020) with a comparable amount of computation. Furthermore, we compare the model additionally pretrained on HBS data only, as well as a model additionally pretrained on both HBS and Slovenian, a closely-related, but not mutually intelligible South Slavic languages. The main questions that we want to obtain an answer for are the following: (1) Is it possible to achieve performance of dedicated models that were trained-from-scratch (BERTić or cseBERT) by additionally pretraining a multilingual model (XLM-R) for a limited number of steps? (2) How do base and large XLM-R models compare in this approach? (3) Is it beneficial not to additionally pretrain for a single language, but include closely related languages into the additional pretraining as well?

Contributions The contributions of this paper are the following: (1) we expand an existing benchmark (Rupnik et al., 2023)[2] with three additional tasks, one for named entity recognition on four separate datasets, another for sentiment identification on political texts, and a final one on causal commonsense reasoning on two datasets, (2) we build the largest collection of raw HBS text up to this point, measuring 11.5 billion words,[3] (3) we obtain insights into how base and large multilingual models behave as they get additionally pretrained, comparing the pretraining on a single language group (HBS) and the language group extended with a closely related language (Slovenian), and, finally, (4) we release new models for the HBS languages[4] as well as for Slovenian and the HBS languages[5] which achieve comparable or improved performance on the four tasks.

2. Related Work

Given the significant impact of BERT (Devlin et al., 2019), there has been a large push towards similarly effective models for all other languages, especially given the often inferior performance of the multilingual BERT variant for low-resource languages (Wu and Dredze, 2020). Following these findings, researchers started exploring how to cater to low-resource languages. We can see three major approaches: 1) development of monolingual models, 2) development of moderately multilingual models, 3) adapting massively multilingual models to improve their performance on the target language.

Monolingual models Monolingual models are pretrained from scratch on texts in one language. Given the relative simplicity of this approach and the initial effectiveness in terms of downstream performance, many successful monolingual language models (LMs) were developed (de Vries et al., 2019; Martin et al., 2020; Le et al., 2020; Tanvir et al., 2021; Snæbjarnarson et al., 2022). While monolingual models often provided the best performance (Ulčar et al., 2021), in the case of less-resourced languages, the main limitation of this approach is that there might not be enough available data for pretraining.

Moderately multilingual models To mitigate this challenge, development of moderately multilingual models was suggested (Ulčar and Robnik-Šikonja, 2020). In this case, the model is pretrained from scratch as well, but on data from multiple closely-related languages. This approach was used in Ulčar and Robnik-Šikonja (2020), who developed the CroSloEngual BERT (cseBERT) model which was pretrained on three languages: Croatian and Slovenian, which are closely related, and English. Similarly, the BERTić model

[2] `https://github.com/clarinsi/benchich/`

[3] `https://huggingface.co/datasets/classla/xlm-r-bertic-data`

[4] `https://huggingface.co/classla/xlm-r-bertic`

[5] `https://huggingface.co/classla/xlm-r-bertic`

(Ljubešić and Lauc, 2021) was pretrained on four languages that are very closely related and mutually intelligible: Bosnian, Croatian, Serbian and Montenegrin. This model outperformed cseBERT on downstream tasks in Croatian (except on named entity recognition), as was shown in Ulčar et al. (2021), likely because it was trained on significantly more data. Singh et al. (2023) experimented with bilingual models and showed that they outperform the massively multilingual models even if the two languages that are combined for training are very distant, e.g., Slovenian and Basque. Additionally, as these models are multilingual, they can be used in cross-language learning scenarios between the included languages (Ulčar and Robnik-Šikonja, 2020). Furthermore, this is a more cost-efficient approach, as it accommodates multiple low-resource languages with the cost of pretraining a single model.

Adaptation However, both these approaches demand pretraining models from scratch, which is very computationally expensive. To mitigate this, one can benefit from existing massively multilingual pretrained models and simply adapt them to the target low-resource language. There are two main approaches for adaptation of massively multilingual models to specific languages: 1) language-adaptive pretraining and 2) adapters (Pfeiffer et al., 2020). In the case of language-adaptive pretraining the massively multilingual model is additionally pretrained with the masked language modelling (MLM) objective on data in the target language. This method was repeatedly shown to provide better results than the base massively multilingual model on monolingual tasks (Wang et al., 2020; Chau et al., 2020; Snæbjarnarson et al., 2022). An alternative method is adapting massively multilingual models to specific languages by learning modular language-specific representations via adapters (Pfeiffer et al., 2020, 2021). Ebrahimi and Kann (2021) compared the methods of extending XLM-RoBERTa to low-resource languages on multiple NLP tasks in a cross-language zero-shot scenario. They showed that additional pretraining provides the best results, while considering it also to be the simplest method to apply. Moreover, additionally pretraining requires much less pretraining than pretraining a model from scratch, and is thus more cost-efficient. Consequently, we have decided to employ this method in the development of language models for the HBS macro-language and Slovenian language. An additional motivation for this choice is the fact that this particular approach has not yet been explored in the context of South Slavic languages.

3. Additional Pretraining

3.1. Data

In this section, we describe the data used for additional pretraining of the XLM-RoBERTa models. We separately describe the HBS and the Slovenian data collection. These two collections jointly consist of more than 19 billion words of running text. All the data inside each language group are heavily near-deduplicated by using Onion[6] (Pomikálek, 2011) with 5-tuples of words, a 90% duplicate threshold and smoothing disabled. The tool operates on the paragraph level, provided that the paragraphs are available (originally separated either as HTML block elements or empty lines), otherwise on the document level.

HBS For the HBS collection of languages, we compiled, to the best of our knowledge, the largest collection of HBS texts up to this date, consisting of 11.5 billion words of running text. The collection consists, in order of near-deduplication[7], of the recent MaCoCu crawl of the Croatian (Bañón et al., 2023b), Bosnian (Bañón et al., 2023a), Montenegrin (Bañón et al., 2023c) and Serbian web (Bañón et al., 2023d); the text collection on which the BERTić model (Ljubešić and Lauc, 2021) was pretrained – including the hrWaC, slWaC, srWaC, and bsWaC web corpora (Ljubešić and Erjavec, 2011; Ljubešić and Klubička, 2014), the CC100 collection (Conneau et al., 2020), and the Riznica corpus (Brozović Rončević et al., 2018) –; a collection of on-line newspapers donated for the purpose of training the presented models; and the mC4 collection (Xue et al., 2021). The size of each part of the HBS pretraining data is given in Table 1. One should note that while the BERTić data collection was originally 8.39 billion words large, its size has shrunk to 3.82 billion words due to the harsh near-deduplication especially with the recent MaCoCu crawls, which certainly contain older web data as well. A similar phenomenon can be observed for the mC4 dataset, which was originally 1.74 billion words in size, shrinking down to 800 million words only.

Slovenian For Slovenian we primarily, again in the order of near-deduplication, relied on the recent MaCoCu crawl of the Slovenian web (Bañón et al., 2023e), but also included the very large

[6] https://corpus.tools/wiki/Onion

[7] The order of near-deduplication is important because it works on the ”first-come-only-retained” principle, only the first paragraph of mutually similar text being retained, all later occurring paragraphs being removed from the collection.

Dataset	Number of words
MaCoCu HBS	5,490,335,790
BERTić data	3,815,720,806
Online newspaper	1,433,110,363
mC4	799,773,550
Total	11,538,940,509

Table 1: Overview of the pretraining data for the HBS language group.

MetaFida corpora collection (Erjavec, 2023) (including, but not limited to the reference GigaFida corpus (Krek et al., 2020) and the KAS corpus of academic writing (Erjavec et al., 2021)), as well as the mC4 dataset (Xue et al., 2021) and the CC100 dataset (Conneau et al., 2020). An overview is shown in Table 2.

Dataset	Number of words
MaCoCu Slovenian	1,907,662,185
MetaFida	3,257,795,640
mC4	2,263,513,217
CC100	195,989,576
Total	7,624,960,618

Table 2: Overview of the pretraining data for the Slovenian language.

3.2. Methodology

We perform additional pretraining of the massively multilingual XLM-RoBERTa (XLM-R) (Conneau et al., 2020) model in base size (XLM-R-base) and large size (XLM-R-large). The base-sized model we only additionally pretrain on the HBS data collection. Henceforth, this model is referred to as XLM-R-base-BERTić, or XB-BERTić for brevity. The large model, which is pretrained on the HBS data collection, is denoted as XLM-R-large-BERTić, or XL-BERTić. Additionally, the model pretrained on the merged HBS and Slovenian data collection is named XLM-R-large-SloBERTić, or XL-SloBERTić. We perform additional pretraining on the Google Cloud infrastructure, using a single TPUv3 for each pretraining with a batch size of 1,024. We run each pretraining process with a comparable amount of computation. For the base model, we perform 96k steps overall, while for large models we perform 48k steps. We organize each pretraining into 8 rounds and report the results at the end of each round. A description of models with additional pretraining hyperparameters is shown in Table 3.

Name	Data	Steps	Warmup	LR
XB-BERTić	HBS	96k	5k	1e-04
XL-BERTić	HBS	48k	2.5k	1e-04
XL-SloBERTić	HBS + SL	48k	2.5k	1e-04

Table 3: Information on the pretraining hyperparameters and data for the newly introduced models. XB-BERTić is the XLM-R-base model additionally pretrained on HBS data only. XL-BERTić is the XLM-R-large model additionally pretrained on HBS data only. XL-SloBERTić is XML-R-large model additionally pretrained on HBS and Slovenian data.

Dataset	Number of tokens
hr500k	499,635
ReLDI-NormTagNER-hr	89,855
ReLDI-NormTagNER-sr	97,673
SETimes.SR	92,271

Table 4: Sizes of datasets (in tokens), used in the named entity recognition experiments.

4. Evaluation

We evaluate the models on three diverse tasks. We use named entity recognition as a token classification task over two Croatian and two Serbian datasets. Next, we evaluate the models on a sequence regression task in form of a parliamentary sentiment prediction task. Lastly, we evaluate on a sequence pair classification task via the choice of plausible alternatives (COPA) dataset translations into Croatian and Serbian. We describe the three tasks in detail below.

4.1. Datasets

Named Entity Recognition We evaluate the performance of the models on the task of named entity recognition on two languages – Croatian and Serbian. Our benchmark consists of two datasets per language: one for the standard language, another for the non-standard language. Specifically, the following datasets are used:

- Croatian linguistic training corpus hr500k 2.0 (Ljubešić and Samardžić, 2023)
- Croatian Twitter training corpus ReLDI-NormTagNER-hr 3.0 (Ljubešić et al., 2023a)
- Serbian linguistic training corpus SETimes.SR 2.0 (Batanović et al., 2023)
- Serbian Twitter training corpus ReLDI-NormTagNER-sr 3.0 (Ljubešić et al., 2023b)

We use the train, development and test set splits as they are split in the original datasets.

Sentiment Identification For experiments on sentiment, we use the ParlaSent dataset (Mochtak et al., 2023), a dataset of sentences from parliamentary proceedings, manually annotated for sentiment. Specifically, we use the HBS train and test subsets, each of them containing 2,600 sentences annotated with an ordinal 0 (negative) to 5 (positive) schema.

Commonsense Reasoning The Choice of Plausible Alternatives (COPA, Roemmele et al., 2011) is a task in which a model has to choose between two plausible continuations of text, given a premise sentence, and return the more plausible one. This task is part of the SuperGLUE English benchmark (Wang et al., 2019) and has human translations available for Croatian (Ljubešić, 2021) and Serbian (Ljubešić et al., 2022). We use the standard split of 400 training, 100 development and 500 test instances.

4.2. Evaluation Methodology

Baseline Models We compare our newly introduced models to four baseline models: two moderately multilingual models, BERTić (Ljubešić and Lauc, 2021) and cseBERT (Ulčar and Robnik-Šikonja, 2020), and the massively multilingual XLM-RoBERTa (XLM-R) (Conneau et al., 2020) model in base and large size. The BERTić model was pretrained on 8.4 billion words in mostly Croatian, but also very closely related, mutually intelligible languages of Bosnian, Serbian and Montenegrin (Ljubešić and Lauc, 2021). The cseBERT model was pretrained on 5.9 billion tokens, of which 31% were in Croatian, 23% in Slovenian, and the rest in English. The massively multilingual XLM-R model was pretrained on the CommonCrawl multilingual data (Conneau et al., 2020), which consists of 167 billion tokens in 100 languages. In terms of the size, the BERTić and cseBERT models are comparable to the base-sized XLM-R with 12 hidden layers and 768 hidden states, whereas the large-sized XLM-R is approximately three times larger in terms of the number of parameters, and consists of 24 hidden layers and 1,024 hidden states.

Hyperparameter Search For all tasks, we perform hyperparameter searches for the BERTić model, the cseBERT model, the base-sized XLM-R model and the large-sized XLM-R model. For the newly introduced models, the best settings of XLM-R-base are used for XB-BERTić, while the settings of XLM-R-large were used for XL-BERTić and XL-SloBERTić. In both named entity recognition and sentiment identification, we optimize only the learning rate and the number of epochs. The hyperparameter search is performed by evaluating on the development data. For named entity recognition, optimal hyperparameters depend on the NER dataset. We perform a separate hyperparameter search for the Croatian standard dataset and for the Serbian standard dataset because of the difference in size, while we perform a joint hyperparameter search for the two non-standard datasets due to their very similar size and diversity. For sentiment identification, we perform a hyperparameter search on a subset of the training dataset which is marked as validation data, as defined in the ParlaSent dataset (Mochtak et al., 2023). For COPA, we perform a hyperparameter search over learning rate and batch size. During fine-tuning, we always train for 15 epochs. Detailed hyperparameter settings are shown in Section A.1 in the Appendix.

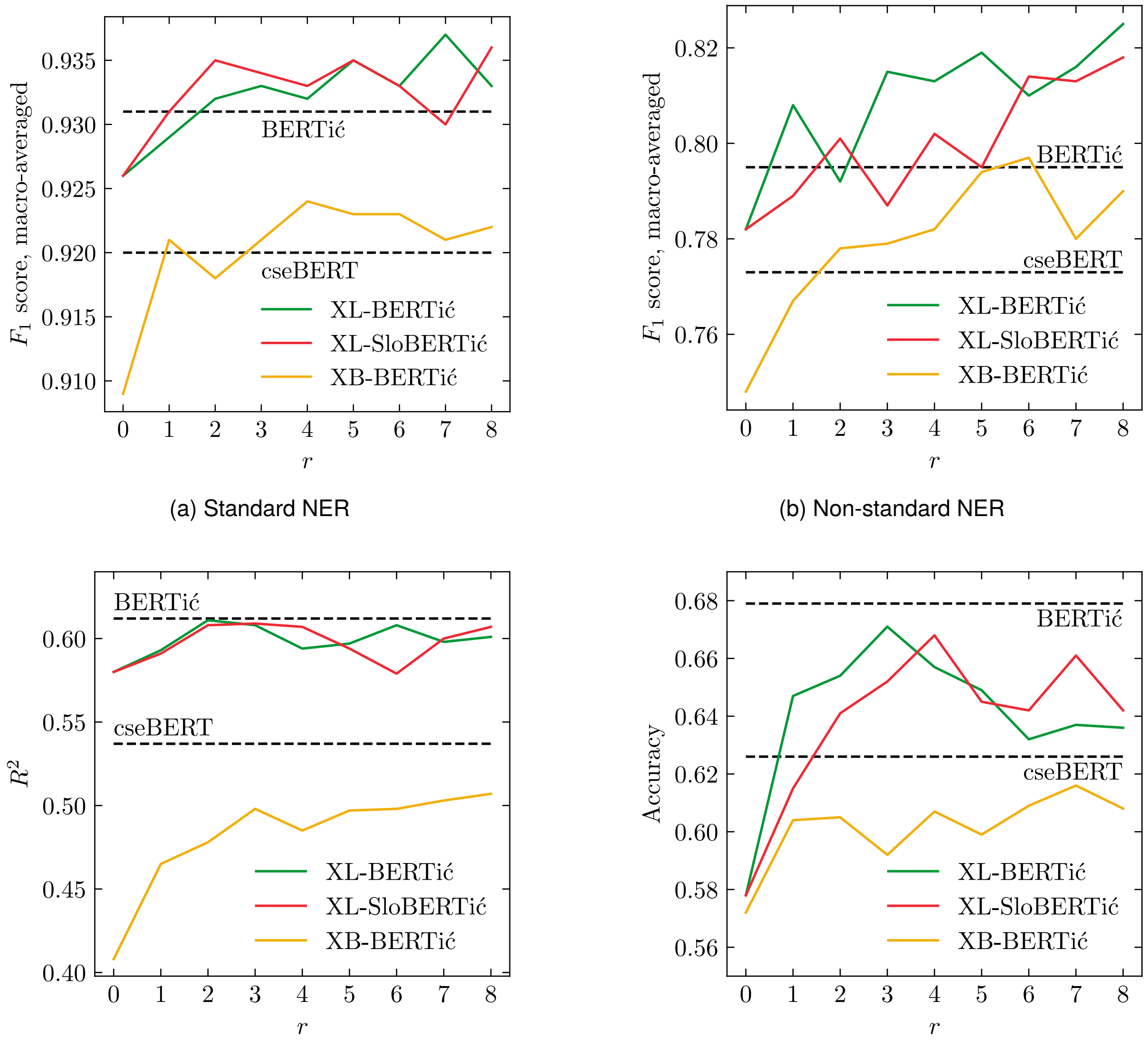
Evaluation Setup For named entity recognition, we train and test each model three times and report aggregated results in the macro F1 score. For sentiment, we perform five runs, and report average R^2 scores. For COPA, we average over 10 runs and report the accuracy score.

5. Results

In this section, we present the results of the evaluation of the newly trained models, compared to the existing models that were trained from scratch, namely BERTić, cseBERT, XLM-R-base and XLM-R-large. We consider these four models as the baseline models. Additionally, to provide insights into the efficiency of pretraining, we do not only evaluate the final pretrained model – we evaluate models, created in 8 rounds of additional pretraining, where base models are updated for 12k steps per round and large models 6k steps, each round corresponding to an identical amount of computation regardless of model size. We evaluate the models on three tasks: the token classification task of named entity recognition, the sequence regression task of sentiment analysis, and the sequence pair classification task in form of the commonsense reasoning benchmark COPA.

5.1. Named Entity Recognition

Given that the named entity recognition task consists of four datasets, here we present a summarized version of the results in form of average results on the two standard and the two non-standard datasets. The full results are available in Section A.2.1 in the Appendix.



(a) Standard NER

(b) Non-standard NER

(c) Sentiment regression

(d) Causal commonsense reasoning

Figure 1: Performance of models on different tasks in relation to the round of additional pretraining. $r = 0$ is referring to round 0, before any additional pretraining, and thus represents the performance of the XLM-RoBERTa-base and XLM-RoBERTa-large models. Subsequent 8 datapoints represent stages of additional pretraining. One round equals 12k steps for the base model (XB-BERTić), and 6k steps for large models (XL-BERTić and XL-SloBERTić), in this way identical amount of computation per round was assured regardless of model size. The performance of cseBERT and BERTić is depicted with a black dashed line.

Standard datasets Figure 1a presents the performance of all the compared models on the two standard named entity recognition datasets. From the baseline models, BERTić performs the best, with a minor difference to cseBERT. XLM-R-large performs between the two models, while the XLM-R-base model underperforms. Once the XLM-R models are additionally pretrained, their performance significantly improves, with the biggest improvements being achieved in the first few rounds of additional pretraining. When we compare the BERTić and the SloBERTić versions of the updated XLM-R-large models to these baselines, we do not see any difference in performance. Full results are published in Section A.2.1 in the Appendix.

Non-standard datasets When the models are evaluated on the two non-standard datasets, results of which are presented in Figure 1b, the picture is somewhat similar to the results on the standard datasets. Among the baseline models, BERTić performs best, with XLM-R-large positioned between BERTić and cseBERT. XLM-R-base again shows significantly lower results. Updating the XLM-R models shows that the models' performances improve most in the first rounds of additional pretraining, with the difference to the standard data that the models' improvement does not completely flatten out, but raises slightly through all of the 8 rounds of additional pretraining. An early hypothesis for this behavior is that non-standard named entity recognition is a harder task and observing more data during additional pretraining has a slight positive effect, one that cannot be observed when performing named entity recognition over standard data. Full results are published in Section A.2.1 in the Appendix.

Overall NER results Overall, on both the standard and the non-standard dataset collections, the additionally updated XLM-R-large improves slightly over the best-performing out of all the baseline models, which is the BERTić model. This improvement is more pronounced on the non-standard datasets.

5.2. Sentiment Identification

Baselines Secondly, we evaluate the models on sentiment identification on parliamentary proceedings. In Figure 1c, we present our results in a comparable manner to the named entity recognition results. The results of the baseline models are comparable to the NER results. That is, BERTić achieves the best results, XLM-R-large falls somewhere between BERTić and cseBERT, while the base-sized XLM-R performs the worst.

Additional pretraining Additional pretraining shows a very similar behavior to the NER results on the standard language datasets. Namely, XLM-R-large models improve their results mostly during the first few rounds of additional pretraining, the improvements being leveled out further. However, a clear difference is that the XLM-R-base model this time achieves improvements throughout all the 8 rounds of additional pretraining. Regarding the difference in performance between the XL-BERTić and the XL-SloBERTić model, the results are comparable to those in the named entity recognition task, with almost no negative impact if significant part of the pretraining was performed on a closely related language. For the overall best results, the updated XLM-R-large model never surpasses, but arrives close to the result of the best-performing BERTić model. Full results are published in Section A.2.2 in the Appendix.

5.3. Commonsense Reasoning

Baselines In this subsection, we present the results over our two commonsense reasoning datasets, COPA-HR and COPA-SR in Figure 1d. If we compare the baseline models, we can see that while BERTić still performs the best, cseBERT now positions itself as the second-best system, in contrast to the results in the two previous tasks. Here, XLM-R-large shows significantly lower performance than BERTić and cseBERT. This is in agreement with previous results, showing that multilingual models for smaller languages, such as Croatian and Serbian, do not perform well on the COPA task (Ljubešić and Lauc, 2021). Interestingly enough, there is not a big difference in performance of the large-sized and the base-sized XLM-R model.

Additional pretraining Once the XLM-R models undergo additional pretraining, their performance exhibits a significant improvement during the initial rounds of updates. However, an unexpected phenomenon occurs thereafter, as the models begin to exhibit a decline in performance compared to the early rounds of updates. Although the performance does not regress to the level observed prior to the additional pretraining, the decrease in performance cannot be disregarded. In the subsequent subsection, we discuss this phenomenon further, together with a concise summary of the results obtained across all three tasks. Full results are published in Section A.2.3 in the Appendix.

5.4. Discussion

Baselines Summarizing the performance of baseline models, we have a clear overall winner – the BERTić model, which obtains the best result on all tasks and datasets. This follows the previous results of Ljubešić and Lauc (2021), but not those of Ulčar et al. (2021), the latter potentially not having invested enough in hyperparameter search for ELECTRA models. cseBERT does come second in one task – commonsense reasoning, while in the two remaining tasks XLM-R-large shows to be more potent. The base-sized XLM-R is regularly the worst performing model.

Performance over time A very interesting trend can be observed when summarizing the results of the additionally pretrained models. What is common to all results, regardless of the task, is that the big improvement in performance comes after just a few rounds of updates. Once pretraining is continued, the behavior of the models is different depending on the task. When the models are fine-tuned for named entity recognition, which can be regarded as the shallowest task, there is visible improvement throughout the whole additional pretraining process. On the sentiment identification task, such continuous improvements cannot be observed and the performance curve flattens out after a few rounds of additional pretraining.

Most interestingly, on the causal commonsense reasoning, which is the most complex task of the three, prolonged training starts to negatively impact the models' performance. Our early hypothesis for this very interesting phenomenon is the following: additional pretraining of XLM-R models just with a single language (group), if performed for long enough, starts to break the multilingual fabric of the model. Considering that the majority of the collective knowledge has been acquired from the "large" languages, which are most prominent in the pretraining data of the XLM-R models, deviating from this shared representation, by pretraining on less prominent languages, results in the loss of crucial profound knowledge required for tasks like commonsense reasoning. The adverse impact is not observable in less complex tasks such as named entity recognition, where the use of shared multilingual knowledge is relatively low, and the additional pretraining compensates for the loss incurred by diverging from the multilingual representations.

Adding related languages Furthermore, in the evaluation, we also compare the performance of the models that were additionally pretrained on the HBS language group with the models where we included also Slovenian in the pretraining data. While Slovenian is closely related to HBS, it is not mutually intelligible with the languages in the HBS language group. The results show that there is no negative impact to the model's performance if closely related languages are also included in the training data, and thus indicate that the cost-efficiency of developing encoder models for less resourced languages can be yet further improved by additionally pretraining on multiple related languages and providing for them all at once.

6. Conclusion

Summary This paper investigates how dedicated monolingual or moderately multilingual encoder models that were pretrained from scratch compare to additionally pretraining massively multilingual encoder models of size up to 1 billion parameters on the example of the HBS language group, comprising the Bosnian, Croatian, Montenegrin and Serbian official languages. The existing and newly introduced models for HBS are evaluated on a benchmark that comprises a token classification task (named entity recognition), a sequence regression task (sentiment analysis) and a sequence pair classification task (causal commonsense reasoning). The benchmark is available at https://github.com/clarinsi/benchich/ and we invite the research community to add additional models to this benchmark.

Our results show that by additionally pretraining the XLM-R-large model performance on the languages of interest increases significantly on all tasks. However, beyond a certain threshold of additional pretraining, the performance gains begin to level off. In fact, for the task of commonsense reasoning, the performance even decreases. Our hypothesis is that the loss in performance through additional pretraining can be attributed to the potential disruption of the multilingual aspect of the original model, where the majority of the language understanding capacity is encoded.

Research questions For our research questions stated in the introduction, we propose the following answers: (1) it is possible to achieve a comparable or even better performance to the language-specific models trained from scratch if one additionally pretrains large multilingual models on the language of interest, (2) large multilingual models regularly perform better than the base-sized models, and (3) no drop in performance can be observed if a significant part of the additional pretraining data consists of a closely related language.

Model and data releases We have decided to publish the two new, additionally pretrained models via HuggingFace – the XL-BERTić model https://huggingface.co/classla/xlm-r-bertic and the XL-SloBERTić model https://huggingface.co/classla/xlm-r-slobertic, both after 48 thousand steps of additional pretraining where most stable results are obtained on all three benchmarking tasks. The reasons for publishing these models are the following: (1) these models perform slightly worse on two, but improve on one task (on both subtasks) to the overall winner of our experiments, the BERTić model, (2) while the BERTić model still performs slightly better on two tasks, we expect for the XL-SloBERTić model to cover both HBS and Slovenian similarly well, including also cross-lingual learning, both of which still have to be confirmed in upcoming experiments, but are sensible expectations, (3) the new XL-BERTić and XL-SloBERTić models were pretrained on newer data, spanning into 2023, while the BERTić model was pretrained on data spanning until 2019, and (4) the XL-BERTić and XL-SloBERTić models are three times the size of the BERTić model, a feature that might be useful in learning some tasks. We also release the 11.5 billion words of HBS data the models were additionally pre-trained on as a HuggingFace Dataset: https://huggingface.co/datasets/classla/xlm-r-bertic-data.

Main takeaway Given the observed results during our experiments, our recommendation for future activities in terms of developing encoder models of up to 1 billion parameters for less-resourced languages is for researchers to take advantage of the existing massively multilingual models and specialize them for the language of interest via additional pretraining. During additional pretraining, it is important that the performance of the model is continuously analysed via evaluation on relevant tasks. This is important as our findings suggest that after a specific amount of additional pretraining, performance could start to deteriorate due to the loss of deeper language understanding that is provided by the multilingual aspect of the model. This "drifting away" phenomenon might be countered by adding some data of large languages to the dataset used for additional pretraining, but this assumption has to be assessed in future research.

On the notion of under-resourcedness We have to note that, while the languages in question are less resourced than most of the European and large world languages, they are still not close to under-resourced on the global scale. All the languages in question have been present during pretraining of the XLM-R models, and we performed experiments with additionally pretraining them on multiple billions of words, most of the world languages cannot come close to. However, we are of the position that there is a significant number of languages that can be helped with the insights provided in this paper.

7. Acknowledgments

This work has received funding from the European Union's Connecting Europe Facility 2014-2020 - CEF Telecom, under Grant Agreement No. INEA/CEF/ICT/A2020/2278341. This communication reflects only the author's view. The Agency is not responsible for any use that may be made of the information it contains.

This work was also funded by the Slovenian Research Agency within the basic research project MEZZANINE "Basic Research for the Development of Spoken Language Resources and Speech Technologies for the Slovenian Language" (J7-4642) and the research programme "Language resources and technologies for Slovene" (P6-0411).

This research was supported with Cloud TPUs from Google's TPU Research Cloud (TRC).

8. Bibliographical References

References

Ethan C Chau, Lucy H Lin, and Noah A Smith. 2020. Parsing with Multilingual BERT, a Small Corpus, and a Small Treebank. *: Findings of the Association for Computational Linguistics: EMNLP 2020*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Abteen Ebrahimi and Katharina Kann. 2021. How to adapt your pretrained multilingual model to 1600 languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4555–4567.

Tomaž Erjavec, Darja Fišer, and Nikola Ljubešić. 2021. The KAS corpus of Slovenian academic writing. *Language Resources and Evaluation*, 55:551–583.

Simon Krek, Špela Arhar Holdt, Tomaž Erjavec, Jaka Čibej, Andraz Repar, Polona Gantar, Nikola Ljubešić, Iztok Kosem, and Kaja Dobrovoljc. 2020. Gigafida 2.0: The reference corpus of written standard Slovene. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3340–3345, Marseille, France. European Language Resources Association.

Taja Kuzman, Igor Mozetič, and Nikola Ljubešić. 2023. Automatic genre identification for robust enrichment of massive text collections: Investigation of classification methods in the era of large language models. *Machine Learning and Knowledge Extraction*, 5(3):1149–1175.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Nikola Ljubešić and Tomaž Erjavec. 2011. hrWaC and slWaC: Compiling web corpora for Croatian and Slovene. In *Text, Speech and Dialogue: 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings 14*, pages 395–402. Springer.

Nikola Ljubešić and Filip Klubička. 2014. bs,hr,srWaC - web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35, Gothenburg, Sweden. Association for Computational Linguistics.

Nikola Ljubešić and Davor Lauc. 2021. BERTić - the transformer language model for Bosnian, Croatian, Montenegrin and Serbian. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 37–42, Kiyv, Ukraine. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Michal Mochtak, Peter Rupnik, and Nikola Ljubešič. 2023. The ParlaSent multilingual training dataset for sentiment identification in parliamentary proceedings. *arXiv preprint arXiv:2309.09783*.

OpenAI. 2023. Gpt-4 technical report. Technical report, OpenAI.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. Adapterfusion: Non-destructive task composition for transfer learning. In *16th Conference of the European Chapter of the Associationfor Computational Linguistics, EACL 2021*, pages 487–503. Association for Computational Linguistics (ACL).

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673.

Jan Pomikálek. 2011. Removing boilerplate and duplicate content from web corpora.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pages 90–95.

Peter Rupnik, Taja Kuzman, and Nikola Ljubešić. 2023. BENCHić-lang: A benchmark for discriminating between Bosnian, Croatian, Montenegrin and Serbian. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 113–120, Dubrovnik, Croatia. Association for Computational Linguistics.

Pranaydeep Singh, Aaron Maladry, and Els Lefever. 2023. Too many cooks spoil the model: Are bilingual models for slovene better than a large multilingual model? In *17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 32–39. Association for Computational Linguistics.

Vésteinn Snæbjarnarson, Haukur Barri Símonarson, Pétur Orri Ragnarsson, Svanhvít Lilja Ingólfsdóttir, Haukur Jónsson, Vilhjalmur Thorsteinsson, and Hafsteinn Einarsson. 2022. A warm start and a clean crawled corpus - a recipe for good language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4356–4366, Marseille, France. European Language Resources Association.

Hasan Tanvir, Claudia Kittask, Sandra Eiche, and Kairit Sirts. 2021. EstBERT: A pretrained language-specific BERT for Estonian. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 11–19, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume

Lample. 2023. LLAMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Matej Ulčar, Aleš Žagar, Carlos S Armendariz, Andraž Repar, Senja Pollak, Matthew Purver, and Marko Robnik-Šikonja. 2021. Evaluation of contextual embeddings on less-resourced languages. *arXiv preprint arXiv:2107.10614*.

M. Ulčar and M. Robnik-Šikonja. 2020. FinEst BERT and CroSloEngual BERT: less is more in multilingual models. In *Text, Speech, and Dialogue TSD 2020*, volume 12284 of *Lecture Notes in Computer Science*. Springer.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Zihan Wang, K Karthikeyan, Stephen Mayhew, and Dan Roth. 2020. Extending Multilingual BERT to Low-Resource Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2649–2656.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *5th Workshop on Representation Learning for NLP, RepL4NLP 2020 at the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 120–130. Association for Computational Linguistics (ACL).

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

9. Language Resource References

Bañón, Marta and Chichirau, Malina and Esplà-Gomis, Miquel and Forcada, Mikel L. and Galiano-Jiménez, Aarón and García-Romero, Cristian and Kuzman, Taja and Ljubešić, Nikola and van Noord, Rik and Pla Sempere, Leopoldo and Ramírez-Sánchez, Gema and Runić, Marija and Rupnik, Peter and Suchomel, Vít and Toral, Antonio and Zaragoza-Bernabeu, Jaume. 2023a. *Bosnian web corpus MaCoCu-bs 1.0*. [link].

Bañón, Marta and Chichirau, Malina and Esplà-Gomis, Miquel and Forcada, Mikel L. and Galiano-Jiménez, Aarón and García-Romero, Cristian and Kuzman, Taja and Ljubešić, Nikola and van Noord, Rik and Pla Sempere, Leopoldo and Ramírez-Sánchez, Gema and Rupnik, Peter and Suchomel, Vít and Toral, Antonio and Zaragoza-Bernabeu, Jaume. 2023b. *Croatian web corpus MaCoCu-hr 2.0*. [link].

Bañón, Marta and Chichirau, Malina and Esplà-Gomis, Miquel and Forcada, Mikel L. and Galiano-Jiménez, Aarón and García-Romero, Cristian and Kuzman, Taja and Ljubešić, Nikola and van Noord, Rik and Pla Sempere, Leopoldo and Ramírez-Sánchez, Gema and Rupnik, Peter and Suchomel, Vít and Toral, Antonio and Zaragoza-Bernabeu, Jaume. 2023c. *Montenegrin web corpus MaCoCu-cnr 1.0*. [link].

Bañón, Marta and Chichirau, Malina and Esplà-Gomis, Miquel and Forcada, Mikel L. and Galiano-Jiménez, Aarón and García-Romero, Cristian and Kuzman, Taja and Ljubešić, Nikola and van Noord, Rik and Pla Sempere, Leopoldo and Ramírez-Sánchez, Gema and Rupnik, Peter and Suchomel, Vít and Toral, Antonio and Zaragoza-Bernabeu, Jaume. 2023d. *Serbian web corpus MaCoCu-sr 1.0*. [link].

Bañón, Marta and Chichirau, Malina and Esplà-Gomis, Miquel and Forcada, Mikel L. and Galiano-Jiménez, Aarón and García-Romero, Cristian and Kuzman, Taja and Ljubešić, Nikola and van Noord, Rik and Pla Sempere, Leopoldo and Ramírez-Sánchez, Gema and Rupnik, Peter and Suchomel, Vít and Toral, Antonio and Zaragoza-Bernabeu, Jaume. 2023e. *Slovene web corpus MaCoCu-sl 2.0*. [link].

Batanović, Vuk and Ljubešić, Nikola and Samardžić, Tanja and Erjavec, Tomaž. 2023. *Serbian linguistic training corpus SETimes.SR 2.0*. [link].

Brozović Rončević, Dunja and Ćavar, Damir and Ćavar, Małgorzata and Stojanov, Tomislav and Štrkalj Despot, Kristina and Ljubešić, Nikola and Erjavec, Tomaž. 2018. *Croatian language corpus Riznica 0.1*. [link].

Erjavec, Tomaž. 2023. *Corpus of combined Slovenian corpora metaFida 1.0*. [link].

Ljubešić, Nikola. 2021. *Choice of plausible alternatives dataset in Croatian COPA-HR*. [link].

Ljubešić, Nikola and Erjavec, Tomaž and Batanović, Vuk and Miličević, Maja and Samardžić, Tanja. 2023a. *Croatian Twitter training corpus ReLDI-NormTagNER-hr 3.0*. [link].

Ljubešić, Nikola and Erjavec, Tomaž and Batanović, Vuk and Miličević, Maja and Samardžić, Tanja. 2023b. *Serbian Twitter training corpus ReLDI-NormTagNER-sr 3.0*. [link].

Ljubešić, Nikola and Samardžić, Tanja. 2023. *Croatian linguistic training corpus hr500k 2.0*. [link].

Ljubešić, Nikola and Starović, Mirjana and Kuzman, Taja and Samardžić, Tanja. 2022. *Choice of plausible alternatives dataset in Serbian COPA-SR*. [link].

Mochtak, Michal and Rupnik, Peter and Meden, Katja and Ljubešić, Nikola. 2023. *The multilingual sentiment dataset of parliamentary debates ParlaSent 1.0*. [link].

A. Appendix

A.1. Hyperparameters

We use the following hyperparameters for fine-tuning the models for the evaluation tasks:

- **Named Entity Recognition**: we use the learning rate of `4e-05`, the train batch size of `32` and the maximum sequence length of `256`. The hyperparameter search showed that optimum number of epochs depends on the size and difficulty level of the named entity dataset. Thus, different numbers of epochs are used depending on the dataset, as shown in Table 5.

- **Sentiment Identification**: the hyperparameter search showed that the optimal epoch number for all models is `15`. We use the train batch size of `32` and the maximum sequence length of `256`. In contrast to the named entity recognition task, the optimum learning rate was shown to depend on the model. Namely, we use `4e-05` for cseBERT and BERTić, and `8e-06` for the base- and large-sized XLM-RoBERTa models and all additionally pretrained models.

- **Commonsense Reasoning**: we performed a hyperparameter search over batch size and learning rate over the baseline models per language. We actually found uniform results. The best settings were a batch size of `8` and learning rate of `1e-05` for a training time of `15` epochs across all models. Note that when averaging over 10 runs, we ignore failed runs, i.e. runs for which the training loss never decreases. We noticed that this occurred more frequently for the models that were trained for longer.

Model	HR-s	Non-s	SR-s
XLM-R-base	5	8	6
XLM-R-large	7	11	13
BERTić	9	10	10
CSEbert	4	7	9

Table 5: Epoch number used for fine-tuning the models on different named entity recognition datasets: standard Croatian (HR-s), standard Serbian (SR-s), and non-standard Croatian and Serbian datasets (Non-s). All XB-BERTić models use the same epoch number as the base-size XLM-RoBERTa model (XLM-R-base), and the other pretrained models (XL-BERTić and XL-SloBERTić) use the same epoch number as the large-sized XLM-RoBERTa model (XLM-R-large).

A.2. Full Results

In the following subsections, we provide more details on the results for all the three tasks, that is, named entity recognition, sentiment identification and commonsense reasoning.

A.2.1. Named Entity Recognition

In this section, we show the results of the evaluation of the models on the named entity recognition task on each of the four evaluated datasets. More precisely, Table 6 shows the results on the standard Croatian dataset, Table 7 on non-standard Croatian dataset, Table 8 on standard Serbian dataset, and Table 9 on non-standard Serbian dataset. We train and test each model three times and report aggregated results, using the macro F1 score.

A.2.2. Sentiment Identification

Table 10 shows the results of evaluation of the models on the task of sentiment identification on

base	large	cseBERT	BERTić	XB-BERTić	XL-BERTić	XL-SloBERTić
0	0	0.918±0.002	0.925±0.003	0.903±0.001	0.919±0.005	0.919±0.005
12	6			0.915±0.001	0.917±0.005	0.920±0.007
24	12			0.911±0.004	0.923±0.004	0.926±0.001
36	18			0.912±0.004	0.918±0.005	0.922±0.005
48	24			0.916±0.007	0.921±0.002	0.926±0.001
60	30			0.916±0.001	0.929±0.005	0.925±0.004
72	36			0.916±0.001	0.929±0.002	0.925±0.004
84	42			0.918±0.004	0.926±0.003	0.927±0.003
96	48			0.917±0.002	0.927±0.001	0.923±0.006

Table 6: Comparison of the models on the NER task on the standard Croatian dataset (hr500k) in terms of macro F1 score, averaged over 3 runs. ‘base’ and ‘large’ correspond to the number of steps performed to additionally pretrain base- or large-sized models, each row therefore requiring equal amount of time on a TPU. Step 0 in the columns for XB-BERTić, XL-BERTić and XL-SloBERTić represents the performance of the models prior to pretraining, i.e., the performance of the XLM-RoBERTa-base and XLM-RoBERTa-large.

base	large	cseBERT	BERTić	XB-BERTić	XL-BERTić	XL-SloBERTić
0	0	0.794±0.006	0.792±0.016	0.763±0.016	0.791±0.014	0.791±0.014
12	6			0.768±0.010	0.810±0.021	0.789±0.034
24	12			0.770±0.018	0.810±0.003	0.805±0.034
36	18			0.790±0.024	0.818±0.015	0.802±0.021
48	24			0.791±0.015	0.810±0.027	0.779±0.024
60	30			0.786±0.015	0.803±0.013	0.802±0.017
72	36			0.806±0.005	0.814±0.005	0.820±0.003
84	42			0.782±0.016	0.797±0.015	0.810±0.008
96	48			0.792±0.018	0.809±0.032	0.812±0.012

Table 7: Comparison of the models on the NER task on the non-standard Croatian dataset (ReLDI-hr) in terms of macro F1 score, averaged over 3 runs. ‘base’ and ‘large’ correspond to the number of steps performed to additionally pretrain base- or large-sized models, each row therefore requiring equal amount of time on a TPU. Step 0 in the columns for XB-BERTić, XL-BERTić and XL-SloBERTić represents the performance of the models prior to pretraining, i.e., the performance of the XLM-RoBERTa-base and XLM-RoBERTa-large.

base	large	cseBERT	BERTić	XB-BERTić	XL-BERTić	XL-SloBERTić
0	0	0.922±0.002	0.936±0.004	0.914±0.004	0.933±0.005	0.933±0.005
12	6			0.926±0.005	0.942±0.003	0.941±0.010
24	12			0.925±0.006	0.941±0.004	0.944±0.003
36	18			0.930±0.001	0.947±0.005	0.946±0.005
48	24			0.932±0.001	0.944±0.001	0.941±0.005
60	30			0.930±0.003	0.942±0.004	0.945±0.006
72	36			0.929±0.006	0.938±0.003	0.941±0.010
84	42			0.924±0.004	0.948±0.008	0.932±0.008
96	48			0.927±0.004	0.940±0.003	0.949±0.003

Table 8: Comparison of the models on the NER task on the standard Serbian dataset (SETimes.SR) in terms of macro F1 score, averaged over 3 runs. ‘base’ and ‘large’ correspond to the number of steps performed to additionally pretrain base- or large-sized models, each row therefore requiring equal amount of time on a TPU. Step 0 in the columns for XB-BERTić, XL-BERTić and XL-SloBERTić represents the performance of the models prior to pretraining, i.e., the performance of the XLM-RoBERTa-base and XLM-RoBERTa-large.

parliamentary data. We train and test each model five times and report average R^2 scores.

A.2.3. Commonsense Reasoning

Tables 11 and 12 show the results of evaluation of the models on the task of commonsense reas-

base	large	cseBERT	BERTić	XB-BERTić	XL-BERTić	XL-SloBERTić
0	0	0.751±0.012	0.798±0.033	0.734±0.024	0.774±0.013	0.774±0.013
12	6			0.765±0.005	0.806±0.006	0.790±0.031
24	12			0.786±0.007	0.775±0.024	0.797±0.014
36	18			0.768±0.024	0.812±0.010	0.772±0.021
48	24			0.772±0.006	0.816±0.026	0.825±0.016
60	30			0.802±0.002	0.834±0.026	0.788±0.021
72	36			0.787±0.018	0.805±0.064	0.809±0.010
84	42			0.779±0.005	0.834±0.018	0.816±0.030
96	48			0.788±0.009	0.841±0.013	0.824±0.006

Table 9: Comparison of the models on the NER task on the non-standard Serbian dataset (ReLDI-sr) in terms of macro F1 score, averaged over 3 runs. 'base' and 'large' correspond to the number of steps performed to additionally pretrain base- or large-sized models, each row therefore requiring equal amount of time on a TPU. Step 0 in the columns for XB-BERTić, XL-BERTić and XL-SloBERTić represents the performance of the models prior to pretraining, i.e., the performance of the XLM-RoBERTa-base and XLM-RoBERTa-large.

base	large	cseBERT	BERTić	XB-BERTić	XL-BERTić	XL-SloBERTić
0	0	0.537±0.006	0.612±0.005	0.408±0.007	0.580±0.014	0.580±0.014
12	6			0.465±0.009	0.593±0.009	0.591±0.010
24	12			0.478±0.006	0.611±0.004	0.608±0.006
36	18			0.498±0.011	0.608±0.009	0.609±0.007
48	24			0.485±0.010	0.594±0.006	0.607±0.008
60	30			0.497±0.003	0.597±0.009	0.594±0.005
72	36			0.498±0.009	0.608±0.012	0.579±0.055
84	42			0.503±0.003	0.598±0.008	0.600±0.006
96	48			0.507±0.008	0.601±0.007	0.607±0.007

Table 10: Comparison of models on the sentiment identification in terms of R^2 scores, averaged over 5 runs. 'base' and 'large' correspond to the number of steps performed to additionally pretrain base- or large-sized models, each row therefore requiring equal amount of time on a TPU. Step 0 in the columns for XB-BERTić, XL-BERTić and XL-SloBERTić represents the performance of the models prior to pretraining, i.e., the performance of the XLM-RoBERTa-base and XLM-RoBERTa-large.

oning on Croatian and Serbian COPA dataset respectively. We train and test each model ten times and report average accuracy scores.

base	large	cseBERT	BERTić	XB-BERTić	XL-BERTić	XL-SloBERTić
0	0	0.645±0.024	0.669±0.016	0.585±0.018	0.571±0.029	0.571±0.029
12	6			0.602±0.021	0.651±0.025	0.616±0.018
24	12			0.607±0.015	0.640±0.036	0.643±0.030
36	18			0.585±0.019	0.656±0.026	0.654±0.027
48	24			0.593±0.015	0.655±0.032	0.668±0.023
60	30			0.589±0.023	0.658±0.033	0.641±0.020
72	36			0.599±0.016	0.635±0.038	0.651±0.027
84	42			0.604±0.024	0.644±0.034	0.656±0.033
96	48			0.599±0.022	0.635±0.031	0.628±0.035

Table 11: Comparison of models on the commonsense reasoning on the Croatian COPA dataset in terms of accuracy scores, averaged over 10 runs. 'base' and 'large' correspond to the number of steps performed to additionally pretrain base- or large-sized models, each row therefore requiring equal amount of time on a TPU. Step 0 in the columns for XB-BERTić, XL-BERTić and XL-SloBERTić represents the performance of the models prior to pretraining, i.e., the performance of the XLM-RoBERTa-base and XLM-RoBERTa-large.

base	large	cseBERT	BERTić	XB-BERTić	XL-BERTić	XL-SloBERTić
0	0	0.607±0.027	0.689±0.024	0.573±0.016	0.570±0.032	0.570±0.032
12	6			0.605±0.016	0.642±0.022	0.613±0.021
24	12			0.603±0.018	0.668±0.033	0.639±0.017
36	18			0.598±0.030	0.685±0.034	0.650±0.022
48	24			0.621±0.015	0.659±0.035	0.667±0.023
60	30			0.609±0.032	0.640±0.030	0.649±0.030
72	36			0.618±0.024	0.629±0.035	0.632±0.028
84	42			0.628±0.024	0.630±0.036	0.666±0.031
96	48			0.617±0.025	0.637±0.021	0.655±0.026

Table 12: Comparison of models on the commonsense reasoning on the Serbian COPA dataset in terms of accuracy scores, averaged over 10 runs. 'base' and 'large' correspond to the number of steps performed to additionally pretrain base- or large-sized models, each row therefore requiring equal amount of time on a TPU. Step 0 in the columns for XB-BERTić, XL-BERTić and XL-SloBERTić represents the performance of the models prior to pretraining, i.e., the performance of the XLM-RoBERTa-base and XLM-RoBERTa-large.