New Language Models for Serbian

UDC 811.163.41'322.2
DOI 10.18485/infotheca.2024.24.1.1

Mihailo Škorić
mihailo.skoric@rgf.bg.ac.rs
ORCID: 0000-0003-4811-8692
University of Belgrade
Faculty of Mining and Geology
Belgrade, Serbia

ABSTRACT: The paper will briefly present the development history of transformer-based language models for the Serbian language. Several new models for text generation and vectorization, trained on the resources of the Society for Language Resources and Technologies, will also be presented. Ten selected vectorization models for Serbian, including two new ones, will be compared on four natural language processing tasks. The paper will analyze which models are the best for each selected task, how their size and the size of their training sets affects the performance on those tasks, and what is the optimal setting to train the best language models for the Serbian language.

KEYWORDS: language models, Serbian language, vectorization, natural language processing.

PAPER SUBMITTED: 27 January 2024
PAPER ACCEPTED: 21 February 2024

1 Introduction

The beginning of the twenty-first century, brought a sharp increase in the amount of available textual data, followed by a sharp increase in computing power, triggering a wave of research based on the idea of *deep learning* (LeCun, Bengio, and Hinton 2015). In the case of natural language processing, the research culminated in the appearance of the transformer architecture (Vaswani et al. 2017), based on the use of encoders, responsible for text analysis, and decoders, responsible for text synthesis. The first extremely popular model of this type was *BERT*[1] (Devlin et al. 2018), based exclusively on the transformer encoder (encoder-only model). This model had made a major breakthrough on multiple natural language processing tasks,

[1] Bidirectional Encoder Representations from Transformers

primarily the ones based on vectorization of text (word embedding). Its variations, *RoBERTa*[2] (Liu et al. 2019) and *DeBERTa*[3] (He et al. 2020) to this day achieve state-of-the-art results for word embedding, word annotation (e.g. word type marking and named entity recognition) and classification of sentences and documents. On the other hand, the appearance of *GPT* (generative pretrained transformer) (Radford et al. 2018) and *GPT-2* (Radford et al. 2019) popularized language models based on the transformer decoder (decoder-only models), which is a group of models currently developing the fastest. Models that combine the use of encoders and decoders, such as *BART* (Lewis et al. 2020) and *T5* (Raffel et al. 2020), remain underrepresented despite their outstanding results on text transformation tasks e.g. machine translation, document summarization and style transfer.

1.1 Overview of published models for the Serbian language

Transformer-based language models made way into the Serbian language through multilingual models, firstly through *MBERT*[4] (Devlin et al. 2018), and then and through *XLM-RoBERTa*[5] (Conneau et al. 2019), for the training of which about 4 billion tokens from texts written in Serbian or another closely related language (Croatian, Bosnian) were used. The latter model was released in December 2019 in two variants, *base* (279 million parameters) and *large* (561 million parameters). Even today, as one of the largest encoder models, *XLM-RoBERTa* is being used for the processing of Serbian texts and achieves good results, especially after additional, specific training.

In early 2021, a model called *BERTić* (*classla/bcms-bertic*) (Ljubešić and Lauc 2021) was published on the platform *Huggingface*.[6] The model is based on the *ELECTRA* architecture (Clark et al. 2020) (110 million parameters), and it was trained on a corpus of over 8 billion tokens, Bosnian (800 million), Croatian (5.5 billion), Montenegrin (80 million) and Serbian (2 billion).

Later that same year, the first models were trained specifically for Serbian and published as part of a wider linguistic research for the Macedonian language (Dobreva et al. 2022). More precisely, the Serbian version of the *RoBERTa-base* model, *macedonizer/sr-roberta-base* (120 million parameters) and the Serbian version of the *GPT2-small* model, *macedonizer/sr-gpt2* (130

[2] Robustly Optimized *BERT*

[3] Decoding-Enhanced *BERT*

[4] Multilingual *BERT*

[5] Cross-lingual Language Model

[6] Huggingface, the largest web hub for publishing language models.

million parameters) were published. Both of these models were trained on the Serbian Wikipedia corpus and support only the Cyrillic alphabet.

Not long after, a similar undertaking begun, during which five *RoBERTa-base* models were trained for the Serbian language (Cvejić 2022). The initial model, *Andrija/SRoBERTa*, had 120 million parameters and was trained on a small corpus of 18 million tokens known as *Leipzig* (Biemann et al. 2007), while the remaining four models each had 80 million parameters and were trained on an increasingly larger corpus. For the *Andrija/SRoBERTa-base* model, the corpus *OSCAR* (Suárez, Sagot, and Romary 2019) (220 million tokens) was added, for the *Andrija/SRoBERTa-L* model, the *srWAc* (Ljubešić and Klubička 2014) (490 million tokens) was added, for the *Andrija/SRoBERTa-XL* model, a part of the *cc100-hr* (21 billion tokens) and *cc100-sr* (5.5 billion tokens) (Wenzek et al. 2020) corpora were added, while for the model *Andrija/SRoBERTa-F* all of the mentioned corpora were used in their entirety.

By the end of 2022, three experimental generative models for Serbian were published (Škorić 2023). The control model, *procesaur/gpt2-srlat*, was again based on the *GPT2-small* architecture, had 138 million parameters and was trained on a subset of the Society for Language Resources and Technologies corpora (260 million tokens) (Krstev and Stanković 2023). The other two models, *procesaur/gpt2-srlat-sem* and *procesaur/gpt2-srlat-synt*, were created by retraining the control model using two specially prepared corpora with the aim of separately modeling the semantics and the syntax of the text. The three models were then used for the experiment of combining language models on the sentence classification task (Škorić, Utvić, and Stanković 2023).

Early next year, researchers from the University of Niš published the *JelenaTosic/SRBerta* model (75 million parameters) based on the *RoBERTa-base* architecture, trained using the *OSCAR* corpus (Suárez, Sagot, and Romary 2019). What is interesting about this model and its second version (*nemanjaPetrovic/SRBerta*, 120 million parameters), is that they were retrained before publication using texts from the law domain (Bogdanović, Kocić, and Stoimenov 2024).

Between the publication of these two models, the first question-answering model for Serbian, *aleksahet/xlm-r-squad-sr-lat* (Cvetanović and Tadić 2023), was created by adapting the RoBERTa model using the *SQuAD* (Rajpurkar, Jia, and Liang 2018) dataset translated into Serbian.

In mid-2023, two more generative models for Serbian based on the GPT architecture were released. Both were trained on the same dataset: the corpora of the Society for Language Resources and Technologies (Krstev and Stanković 2023), doctoral dissertations downloaded from the NARDUS platform,[7] corpus of public discourse of the Serbian language by the Institute of Serbian Language SANU dubbed *PDRS* (Wasserscheidt 2023), and some additional publicly available corpora from the web, such as the already mentioned *srWAc* (Ljubešić and Klubička 2014) and *cc100-sr* (Wenzek et al. 2020). The total number of tokens in this dataset was about 4 billion. The larger model, *jerteh/gpt2-orao*,[8] has 800 million parameters, it is based on the *GPT2-large* architecture and is currently the largest available model pretrained for the Serbian language. The smaller model, *jerteh/gpt2-vrabac*,[9] has 136 million parameters and it is based on the *GPT2-small* architecture. Both models were trained using the computing resources of the National Platform for Artificial Intelligence of Serbia. In addition to the training corpus, the two models also share a dictionary and a tokenizer, specially equipped to pair Cyrillic and Latin characters, enabling equal support for both alphabets.

After the publication of the 800 million parameter generative model (*jerteh/gpt2-orao*), the focus slowly shifted to retraining available large models for English using Serbian language texts. Hence, two models based on the *Alpaca* (Taori et al. 2023) architecture, *datatab/alpaca-serbian-3b-base* (3 billion parameters) and *datatab/alpaca-serbian-7b-base* (7 billion parameters) were published, while the publication of another 7 billion parameter model based on the *Mistral-7b* (Jiang et al. 2023) architecture, trained on Croatian, Bosnian and Serbian texts numbering 11.5 billion tokens, was announced. The same 11.5 billion tokens corpus was used to retrain the *XLM-RoBERTa-large*. This model was published under the name *classla/xlm-r-bertic* (Ljubešić et al. 2024) and has 561 million parameters, the same number as the original *XLM* model.

Finally, the dataset which was used to train *jerteh/gpt2-orao* and *jerteh/gpt2-vrabac*, was also used to train more encoding models from scratch. The larger model, *jerteh/Jerteh-355*,[10] is based on the *RoBERTa-large* architecture and has 355 million parameters, while the smaller model, *jerteh/Jerteh-81*,[11] is based on the *RoBERTa-base* architecture and has 81 million parameters. As with the *jerteh/gpt2-orao* model, the goal was to

[7] NARDUS – National Repository of Doctoral Dissertations from all Universities in Serbia.

[8] jerteh/gpt2-orao

[9] jerteh/gpt2-vrabac

[10] jerteh/Jerteh-355

[11] jerteh/Jerteh-81

train the models on the highest quality corpora. This paper will present an analysis of the performance of these two models individually, as well as in comparison with the performance of other selected models, in order to establish their place in the hierarchy of Serbian language models for text vectorization.

1.2 The Experiment

In the previous section, it was pointed out that there is a large number of multilingual models that support the processing of the Serbian language to a different extent, and that there are about twenty models that have been prepared specifically for the processing of Serbian. Published models differ from each other by several features: the family (architecture) of the model, the number of parameters, the dictionary or tokenizer on which the model is based, the corpora used for its training, the task on which the model was trained, and the training length. It should be noted that some of the information on models is missing, but also that some available information (primarily the properties of the training set) is not verifiable.

In the following sections, the paper will focus on ten selected encoding models (general type). Basic information about those models will be presented in Section 2, an experiment comparing their performance on four prepared tasks will be presented in Section 3, and the results of the experiments will be presented and discussed in Section 4. Finally, in Section 5, the process of training new models for Serbian will be proposed. This paper will not focus on generative models due to the lack of a reliable (automatic) mechanisms for measuring their performance. Encoder-decoder models specially developed for the Serbian language are yet to be published.

2 Selected encoder-based models

For the purposes of this paper, ten of the previously mentioned models (Section 1) were selected, and will be analyzed in more detail. They include four *SRoBERTa* models, which, due to the fact that they differ only in the training data set, are very suitable for this experiment. Furthermore, the oldest model, *classla/bcms-bertic* and the newest model, *classla/xlm-r-bertic*, published by the Center for South Slavic Languages *CLASSLA*, will be examined, as well as the two most popular multilingual models *xlm-roberta-base* and *xlm-roberta-large.* The last two models to be analyzed, *jerteh-81* and *jerteh-355*, trained on the resources of the Society for Language Resources and Technologies, are presented for the first time in this paper. The basic features of the ten models are shown in Table 1.

No.	1	2	3	4	5	6	7	8	9	10
ID	Andrija/SRoBERTa-base	Andrija/SRoBERTa-L	Andrija/SRoBERTa-XL	Andrija/SRoBERTa-F	classla/bcms-bertic	classla/xlm-r-bertic	xlm-roberta-base	xlm-roberta-large	jerteh/jerteh-81	jerteh/jerteh-355
Tokenizer	SRoBERTa				bertic	XLM-R			jerteh	
Architecture	RoBERTa				ELE.	XLM-R			RoBERTa	
Model Size	80				110	561	279	561	81	355
Dataset Size	500	1000	3750	5700	8400	11500	4000*		4000	
Serbian	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Croatian			✓	✓	✓	✓	✓	✓		
Bosnian					✓	✓	✓	✓		
Montenegrin					✓		✓	✓		

Table 1. Ten selected encoder-based models for Serbian and their features: tokenizer, architecture, model size in millions of parameters and training set size in millions of tokens. The data was taken from the *HuggingFace* platform. *The size of the training set for models 7 and 8 (*xlm-roberta-base*, *xlm-roberta-large*) refers to the part of the set in Serbian, Croatian or other related language. The lower part of the table shows in which of these languages the models were trained.

From the table, as well as from the description of the models in the previous section, it is apparent that the most popular architecture is *RoBERTa* (6 out of 10 selected models), with the additional three models being based on a close-related, *XLM-RoBERTa* architecture. The one remaining model, *bcms-bertic* based on the ELECTRA architecture, is the only selected model not pre-trained on the masked language modeling task (prediction of parts of text masked behind a special label).

The size of the selected models varies from 80 (for four *SRoBERTa* models) to over 560 million parameters (for models based on *XLM-RoBERTa-large*). The size of the training set varies from 500 million for model 1 (*SRoBERTa-base*) to as much as 11.5 billion tokens for model 6 (*classla/xlm-r-bertic*), where it should be noted that this model was not trained from scratch, but rather a *xlm-roberta-large* model addapted on Croatian, Bosnian and Serbian texts. Only four out of ten models were trained exclusively on Serbian texts, namely models 1, 2, 9 and 10, i.e. the first two *SRoBERTa* models, *jerteh/jerteh-81* and *jerteh/jerteh-355.*

It is also important to note that the ten presented models use only four different dictionaries/tokenizers:

X_1 *SRoBERTa* tokenizer – the first 4 models;
X_2 *bertic* tokenizer – model 5;
X_3 *XLM-R* tokenizer – models 6 to 8;
X_4 *jerteh* tokenizer – the last 2 models (9 and 10).

3 Performance evaluation setting

Ten selected models were evaluated on four separate tasks to compare their performance:

T_1 Masked language modeling (guessing missing tokens);
T_2 Calculation of (semantic) sentence similarity;
T_3 Part-of-speech annotation;
T_4 Named entity recognition.

The first two tasks belong to the group of so-called *upstream* tasks, which use models in their basic state, while the other two tasks belong to the group of *downstream* tasks because they require the models to be fine-tuned and evaluated on a specially prepared, task-specific datasets.

3.1 Model evaluation - upstream tasks

As already mentioned, upstream tasks do not require model adaptation, so only preparation of test sets is necessary.

In order to evaluate the models on the masked language modeling task (T_1), a special data set was prepared using texts in which one random token is masked in each sentence behind a mask `<MASK>`. Four sources were used for the textual material:

Y_1 *Dečko*, Serbian translation of the novel *The Adolescent* (*Подросток*) by Dostoyevsky;
Y_2 *Mladić*, alternative translation of *Dečko*;
Y_3 Serbian translation of Jules Verne's novel *Around the World in 80 Days*;
Y_4 Croatian translation of Jules Verne's novel *Around the World in 80 Days*.

The first two sources were not used to train any of the models, while the other two have been available on the web for a long time (Vitas et al. 2008) and were therefore probably used to train most, if not all, of the models listed.

In order that no model has a particular advantage, the texts were tokenized using all four tokenizers (X_1 to X_4) and then masked. Each of the ten models had the task of unmasking each of the sixteen prepared texts (four sources tokenized and masked in four different ways). One token was masked in each sentence, and the models offered three candidates in its place. Every instance where the masked (i.e., requested) token appeared in the set of candidates provided by the model for the given sentence was counted as a successful hit, and the accuracy on this task was used for assessment of the test results.

To evaluate models on the second task, namely, calculating the similarity between sentences (T_2), triplets based on extraction of the same sentence from parallelized novels were used (Y_1 and Y_2, i.e. Y_3 and Y_4). Since the novels were parallelized at the sentence level, it was easy to create pairs of sentences with the same meaning. Each triplet was formed by adding a similar length drawn a different point of the counterpart novel, that sharing as many tokens as possible with the first sentence. Example of a triplet:

To evaluate models on the second task, namely, calculating the similarity between sentences (T_2), triplets based on extraction of the same sentence from parallelized novels were used (Y_1 and Y_2, i.e. Y_3 and Y_4). Since the novels were parallelized at the sentence level, it was easy to create pairs of sentences with the same meaning. Each triplet was formed by adding a

sentence sharing as many tokens as possible with the first sentence, and of a similar length, but extracted from a different point of the counterpart novel. Example of a triplet:

1. "Zaista, ko ne bi obišao svet i za manju cenu?" (control sentence, Y_3: *Around the World in 80 Days*, Serbian)
2. "Doista, nije li i za manje od toga vrijedno izvršiti put oko svijeta?" (pair, Y_4: *Around the World in 80 Days*, Croatian)
3. "He! he! pa konačno zašto ne bi uspio?" (false pair, Y_4: *Around the World in 80 Days*, Croatian)

The models were tasked with recognizing two of the sentences in the assigned triplets that actually match (the similarity between the first and the second sentence has to be greater than between the first and the third), and the accuracy on that task was used to evaluate performance. The similarity between the sentences is calculated as the difference of the number 1 and the cosine distance of the calculated sentence vectors. To compute sentences vectors, the model first assigns vector values to each token in the sentence, and then the value of those vectors is averaged to obtain a vector representation of the sentence.

3.2 Model evaluation - downstream tasks

For the purpose of evaluating the performance of models on the remaining two planned tasks, the models were fine-tuned and tested on specially-prepared datasets. The publicly available set, *SrpKor4Tagging* (Stanković et al. 2020) (three hundred and fifty thousand tagged tokens), was used for the part-of-speech tagging task (T_3), while another publicly available set, *SrpELTeC-gold* (Šandrih Todorović et al. 2021), was used for the named entity recognition task (T_4). For both tasks, the models were fine-tuned on 90% of labeled sentences from each set and tested on the remaining 10%. As both tasks are multi-class classification problems, the F_1-score obtained during the classification of the sentences from the test set were used to access the model performance.

4 Evaluation results

The results of the first test (T_1) i.e. the average accuracy of the selected models on the task of guessing the missing tokens in sixteen prepared texts, are shown in Table 2.

No.	1	2	3	4	5	6	7	8	9	10
ID	Andrija/SRoBERTa-base	Andrija/SRoBERTa-L	Andrija/SRoBERTa-XL	Andrija/SRoBERTa-F	classla/bcms-bertic	classla/xlm-r-bertic	xlm-roberta-base	xlm-roberta-large	jerteh/jerteh-81	jerteh/jerteh-355
X_1-Y_1	0.43	0.63	0.66	0.70	/	0.43	0.46	0.51	0.70	**0.75**
X_1-Y_2	0.43	0.62	0.64	0.69	/	0.42	0.46	0.50	0.69	**0.73**
X_1-Y_3	0.37	0.56	0.59	0.63	/	0.34	0.38	0.43	0.66	**0.72**
X_1-Y_4	0.36	0.55	0.64	**0.68**	/	0.34	0.38	0.42	0.58	0.63
X_2-Y_1	0.36	0.47	0.51	0.54	/	0.47	0.50	0.55	0.57	**0.60**
X_2-Y_2	0.37	0.48	0.51	0.54	/	0.46	0.50	0.54	0.56	**0.59**
X_2-Y_3	0.31	0.41	0.45	0.48	/	0.42	0.45	0.50	0.50	**0.54**
X_2-Y_4	0.31	0.42	0.47	**0.51**	/	0.42	0.46	**0.50**	0.47	**0.51**
X_3-Y_1	0.37	0.49	0.52	0.54	/	0.48	0.50	0.55	0.57	**0.60**
X_3-Y_2	0.37	0.48	0.51	0.54	/	0.46	0.50	0.54	0.57	**0.59**
X_3-Y_3	0.30	0.41	0.44	0.47	/	0.41	0.45	0.49	0.50	**0.54**
X_3-Y_4	0.31	0.42	0.47	**0.51**	/	0.41	0.46	**0.50**	0.47	**0.50**
X_4-Y_1	0.42	0.60	0.63	0.67	/	0.43	0.47	0.51	0.73	**0.78**
X_4-Y_2	0.41	0.58	0.61	0.65	/	0.41	0.45	0.49	0.71	**0.75**
X_4-Y_3	0.35	0.53	0.55	0.60	/	0.33	0.38	0.42	0.69	**0.76**
X_4-Y_4	0.34	0.50	0.58	0.62	/	0.33	0.37	0.41	0.62	**0.66**
average	0.36	0.51	0.55	0.59	/	0.41	0.45	0.49	0.60	**0.64**

Table 2. Model accuracy on the task of guessing masked tokens (three candidates) for each of the sixteen prepared masked texts and on the average. Each masked text is marked (at the beginning of each line) with a unique label representing a combination of a tokenizer (X) and a source (Y). The best result in each row (±1%) is marked in bold.

The results show a clear superiority of the new model, *jerteh-355*, which achieved the best result in thirteen out of sixteen cases, and shared the best result (±1%) in two more cases. Moreover, in nine out of twelve cases, the *jerteh-355* model outperformed models on texts masked by tokenizers of those same models. The only model that managed to surpass it in two cases is *SRoBERTa-F*, which performed best in processing the source (Y_4), written in Croatian (the language included in a large percentage in its training set). However, its average accuracy is lower than the accuracy of *jerteh-81*, the other new model. Model 5 (*classla/bcms-bertic*) was not included in the evaluation on this task because that would put in a disadvantage, as unlike the others, it was not trained on the masked language modeling task.

The results of the second test, calculating sentence similarity (T_2) are shown in Table 3. The values show the model accuracy in recognizing sentences with the same/similar meaning in triplets extracted from two Serbian translations of the same novel, Y_1 and Y_2 (first row of values), from the Serbian and Croatian translations of the same novel, Y_3 and Y_4 (second row), and on the average (third row).

No.	1	2	3	4	5	6	7	8	9	10
ID	Andrija/SRoBERTa-base	Andrija/SRoBERTa-L	Andrija/SRoBERTa-XL	Andrija/SRoBERTa-F	classla/bcms-bertic	classla/xlm-r-bertic	xlm-roberta-base	xlm-roberta-large	jerteh/jerteh-81	jerteh/jerteh-355
Y_1-Y_2	0.93	**0.95**	**0.96**	**0.96**	0.92	0.76	0.90	0.87	**0.95**	**0.95**
Y_3-Y_4	0.83	0.89	**0.92**	**0.91**	0.79	0.66	0.78	0.71	0.89	0.83
average	0.88	0.92	**0.94**	**0.93**	0.85	0.71	0.84	0.79	0.92	0.89

Table 3. Performance of the selected models (accuracy) on the task of recognizing sentences with the same or similar meaning in the triplets extracted from the translations of Dostoyevsky (Y_1-Y_2), Verne (Y_3-Y_4), and their average. In each row, the best result (±1%) is marked in bold.

The results for the first set of triplets are very good for several models, where *SRoBERTa-L*, *SRoBERTa-XL*, *SRoBERTa-F*, *jerteh-81* and *jerteh-355* achieve a similar accuracy of around 95%. The *SRoBERTa-XL* model achieves the best results by a very small margin, but also the best results for the second set of triplets containing sentences in the Croatian language (92% accuracy). Therefore, it also has the best overall performance on this task. The only other model that achieves an accuracy of over 90% for the second set of triplets is *SRoBERTa-F*, which was to be expected, because it was also trained on Croatian texts.

The results achieved by the models (F_1-score) on downstream tasks T_3 (part-of-speech-tagging) and T_4 (named entity recognition) are shown in Table 4.

No.	1	2	3	4	5	6	7	8	9	10
ID	Andrija/SRoBERTa-base	Andrija/SRoBERTa-L	Andrija/SRoBERTa-XL	Andrija/SRoBERTa-F	classla/bcms-bertic	classla/xlm-r-bertic	xlm-roberta-base	xlm-roberta-large	jerteh/jerteh-81	jerteh/jerteh-355
T_3	0.974	0.980	0.982	0.982	**0.986**	**0.987**	0.984	**0.986**	0.985	**0.986**
T_4	0.908	0.922	0.929	0.935	**0.942**	**0.942**	0.933	0.935	0.928	0.928

Table 4. F_1-score achieved by the models on tasks T_3 (part-of-speech-tagging) and T_4 (named entity recognition). In each row, the best result (±0.1%) is marked in bold.

From the results shown, it is apparent that on the task T_3 (part-of-speech-tagging) nine out of ten models perform quite well (F_1-score of over 98%), where the results achieved by the four top-performing models (*classla/bcms-bertic*, *classla/xlm-r-bertic*, *xlm-roberta-large* and *jerteh/jerteh-355*) differ by less than 0.02%, indicating that the models are slowly approaching the upper limits of performance for this task.

When it comes to the results on the last task, T_4 (named entity recognition), the highest performance was achieved by models *classla/bcms-bertic* and *classla/xlm-r-bertic*, but the results across the models are not similar to those for task T_3 ($\sim 4\%$ for task T_4 compared to $\sim 1\%$ for task T_3). However, the performance gap between the best and worst performing model is still significantly smaller than the one on the upstream tasks ($\sim 28\%$ gap for task T_1).

In the following section, the achieved results will be discussed, together with the apparent reasons that led to those results, with the aim of determining the most favorable conditions for training Serbian language models in the future.

5 Discussion

The previously presented evaluation results (tables 2–4) show that there is not a single model (or group of models) that performs best in general, but rather that different models (and model groups) are better (or worse) at different tasks. In this sections results for each task will be accessed individually, with an emphasis on the relationship between the achieved performance and the size of the model, the size of its training set and the quality of that set.

5.1 Masked Language Modeling

The results for the masked language modeling task (Table 2) show a substantial advantage of the *jerteh/jerteh-355* model, with *jerteh/jerteh-81* also achieving good results as the second best-performing model on the average. The most probable cause is that these models used the same training dataset. Correlations of the average accuracy of models and the number of their parameters and the models and the size of their training sets are shown in Figure 1.

At first glance, some prominent exceptions are noticeable, primarily the models based on the *XLM-R* architecture, which achieve some of the worst results on this task. If their results are removed, new trends appear (see Figure 2). Thus, when looking only at the *RoBERTa* models, it seems that the larger the model (albeit not very convincingly) and the larger the set used to train it (very convincingly) the better the model performance.

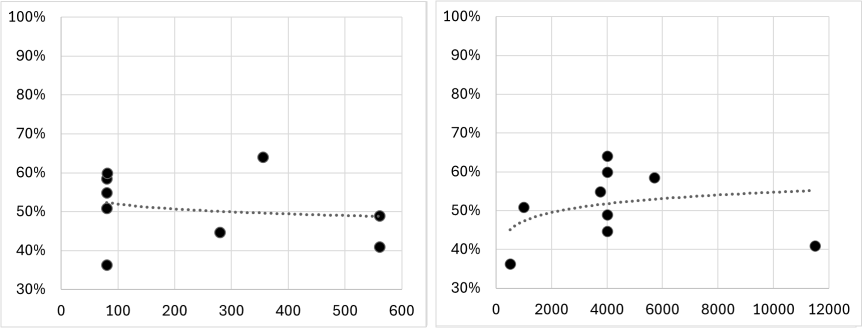

Figure 1. Correlation between each model's accuracy on the masked language modeling task with its size (left), and with the size of its training sets (right). The displayed trend curve corresponds to a logarithmic function.

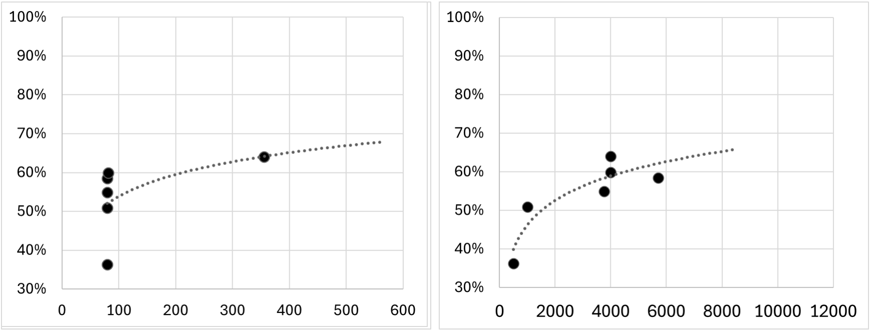

Figure 2. Correlation between each RoBERTa-based model's accuracy on the masked language modeling task with their size (left), and with the size of their training sets (right). The displayed trend curve corresponds to a logarithmic function.

5.2 Calculation of (semantic) sentence similarity

During the evaluation of the T_2 task, it was established that the best results were achieved by the models *SRoBERTa-L*, *SRoBERTa-XL*, *SRoBERTa-F*, *jerteh-81* and *jerteh- 355* when it comes to recognizing sentence pairs in Serbian, and *SRoBERTa-XL* and *SRoBERTa-F* when it comes to recognizing bilingual sentences pairs (Serbian and Croatian, Table 2). This indicates that the key for good sentence embedding is to pre-train the model for the languages being processed. When embedding Serbian sentences only, the best models are those that were previously pre-trained on a sufficiently large set of sentences in Serbian, but when embedding both Serbian and Croatian sentences, models pre-trained on both Serbian and Croatian have an advantage. On the other hand, models based on the *XLM-R* architecture and pre-trained on one hundred world languages are under-performing on this task, probably due to the large noise that the diverse training set produces.

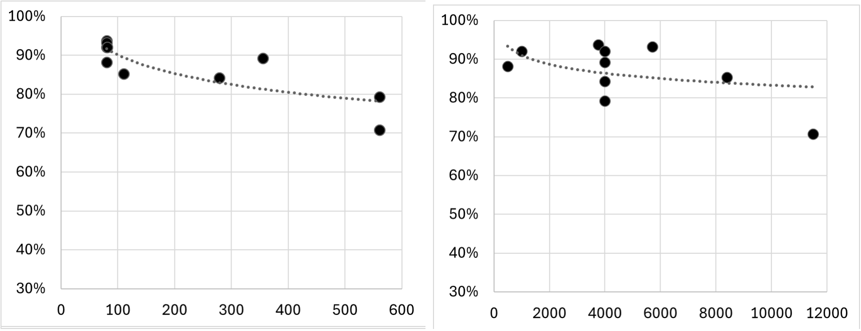

Figure 3. Correlation between each model's accuracy on the sentence embedding task with its size (left), and with the size of its training sets (right). The displayed trend curve corresponds to a logarithmic function.

Figure 3 shows the effect of the model size and the training set size on the performance of the model on this task and, interestingly, the trend lines indicate that the performance decreases with the increase of either parameter. The impact of the dataset size can be attributed (to some extent) to the previously described phenomenon affecting the *XLM-R* architecture. On the other hand, when it comes to the effect of model size, there are additional indicators that smaller models are better for this task, especially for bilingual embeddings where *jerteh/jerteh-81* outperforms *jerteh/jerteh-355*. The reason could be that the smaller model, due to its size, is less adapted to the Serbian language (underfit), but has an aadvantage in generalizing ability.

5.3 Downstream Tasks

Unlike the evaluation on the upstream tasks, the results achieved by the models on the downstream tasks are much more even. Nearly all models achieve great results for part-of-speech tagging (T_3, Table 4), including those based on the *XLM-R* architecture. Moreover, *classla/xlm-r-bertic* and *xlm-roberta-large*, which are *XLM-R*-based are two of the four models that achieve the best results on this task (the other two being *classla/bcms-bertic* and *jerteh/jerteh-355*).

What these four models have in common is that they are either the largest models or models trained on the largest datasets. A positive correlation between the performance and the size of the model, as well as between the performance and the size of the training sets can also be observed in Figure 4.

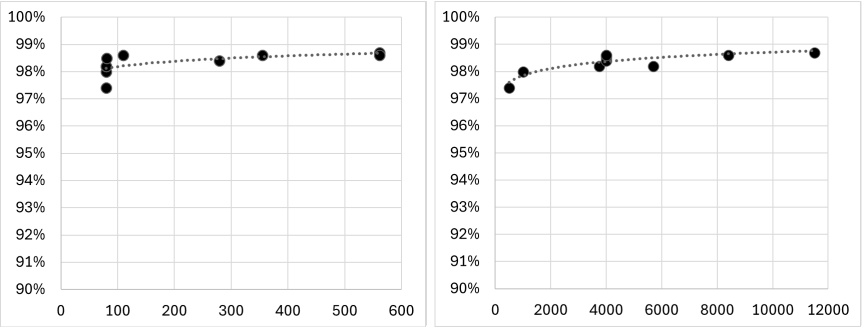

Figure 4. Correlation between each model's F_1-score on part-of-speech tagging with their size (left), and with the size of their training sets (right). The displayed trend curve corresponds to a logarithmic function.

The correlation between the size of the training set is even more obvious in the case of named entity recognition task (Figure 5). The best results on this task (T_4) were achieved by the two models with the largest training sets, with outstanding performance of *XLM-R*-based models. Model size is also shows a (slight) positive correlation with performance on this task.

5.4 Conclusion

When it comes to masked language modeling, it seems that the development of new models for Serbian is going in the right direction. The *jerteh/jerteh-355* model achieves by all means the best results, at least when it comes to working with high-quality texts, even when they are tokenized by unknown

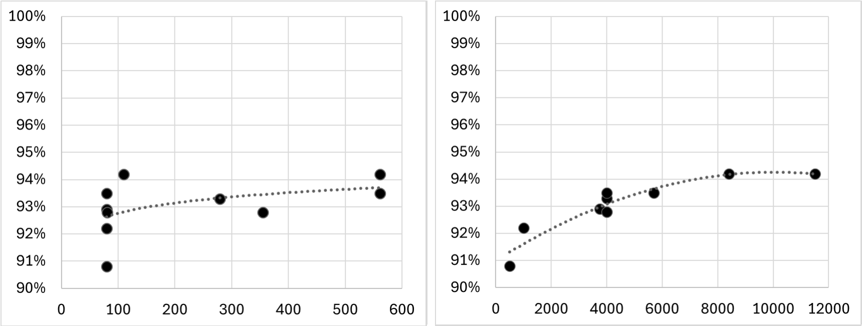

Figure 5. Correlation between each model's F_1-score on named entity recognition with their size (left), and with the size of their training sets (right). The displayed trend curve corresponds to a logarithmic function.

tokenizers (Table 1). Although the training data set size has a positive correlation with the performance of the model (fig.2), the quality of the set should not be neglected, since *jerteh/jerteh-355* and *jerteh/jerteh-81* outperform the *Andrija/SRoBERTa-F* and *Andrija/SRoBERTa-XL* models trained on larger training sets, indicating that web corpora may not always be sufficient to train quality models for this task. This is consistent with the conclusion of another recent research (Li et al. 2023). However, new research should include non-literary sources in the evaluation set, in order to obtain a more comprehensive outlook of the situation.

On the task of calculating the similarity between sentences (sentence embedding) models *Andrija/SRoBERTa-F* and *Andrija/SRoBERTa-XL* stood out, followed by *Andrija/SRoBERTa-L*, *jerteh/jerteh-81* and *jerteh/jerteh-355*, at least when it comes to embedding Serbian sentences (Table 3). What sets these models apart is that they are smaller compared to other models and trained on a larger set, so generalization seems to be the key for this task, that is, larger data sets in combination with smaller models. Also, when it comes to processing sentences of a wider linguistic spectrum (e.g. South Slavic languages), it would be necessary to include sentences from the complete spectrum in the training set or, even better, to adapt the dictionary to map a wider range of tokens, and therefore allow for the correct vectorization of these sentences. New research on this topic should also explore a more recent method of sentence vectorization, for example, using the *sentence transformer* architecture (Reimers and Gurevych 2019).

In the case of both assessed upstream tasks, the performance achieved by the models based on the *XLM-R* architecture is significantly lower than the one of the models based on the *RoBERTa* architecture. In the case of the first task (T_1), this can be explained by their significantly larger token dictionary (which makes the selection of the appropriate token more difficult). However, such an explanation would not be adequate for the second task (T_2). On the other hand, the models based on the *XLM-R* architecture proved to be the best (by a small margin) on downstream tasks, primarily for named entity recognition (T_4). It seems that in order to successfully solve this task, it is best that the model encounters a wide variety of tokens during pre-training, while additional training on Serbian texts brings additional improvements. It seems that in order to improve the performance, it would be optimal to retrain the *XLM-RoBERTa-large* model using the largest and highest quality set of texts in the Serbian language. When it comes to part-of-speech tagging, it seems that any of the new models would be adequate to handle this task.

Acknowledgment

The most important datasets for the training of *GPT2-orao*, *GPT2-vrabac*, *jerteh-81* and *jerteh-355* models were provided by the Society for Language Resources and Technologies.[12]

Computer resources for the training of *GPT2-orao* and *GPT2-vrabac* models were provided by the National Platform for Artificial Intelligence of Serbia.

Computer resources for the training of *jerteh-81* and *jerteh-355* models were provided by the Faculty of Mining and Geology, University of Belgrade.

This research was supported by the Science Fund of the Republic of Serbia, #7276, *Text Embeddings – Serbian Language Applications – TESLA*.

Thank you!

References

Biemann, Chris, Gerhard Heyer, Uwe Quasthoff, and Matthias Richter. 2007. "The Leipzig corpora collection-monolingual corpora of standard size." *Proceedings of corpus linguistic* 2007.

[12] JeRTeh

Bogdanović, Miloš, Jelena Kocić, and Leonid Stoimenov. 2024. "SRBerta-A Transformer Language Model for Serbian Cyrillic Legal Texts." *Information* 15 (2). ISSN: 2078-2489. https://doi.org/10.3390/info15020074.

Clark, Kevin, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. "Electra: Pre-training text encoders as discriminators rather than generators." *arXiv preprint arXiv:2003.10555.*

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. "Unsupervised cross-lingual representation learning at scale." *arXiv preprint arXiv:1911.02116.*

Cvejić, Andrija. 2022. "Prepoznavanje imenovanih entiteta u srpskom jeziku pomoću transformer arhitekture." *Zbornik radova Fakulteta tehničkih nauka u Novom Sadu* 37 (02): 310–315.

Cvetanović, Aleksa, and Predrag Tadić. 2023. "Synthetic Dataset Creation and Fine-Tuning of Transformer Models for Question Answering in Serbian." In *2023 31st Telecommunications Forum (TELFOR),* 1–4. IEEE.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805.*

Dobreva, Jovana, Tashko Pavlov, Kostadin Mishev, Monika Simjanoska, Stojancho Tudzarski, Dimitar Trajanov, and Ljupcho Kocarev. 2022. "MACEDONIZER-The Macedonian Transformer Language Model." In *International Conference on ICT Innovations,* 51–62. Springer.

He, Pengcheng, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. "DEBERTA: Decoding-Enhanced BERT with Disentangled Attention." In *International Conference on Learning Representations.*

Jiang, Albert Q, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. "Mistral 7B." *arXiv preprint arXiv:2310.06825.*

Krstev, Cvetana, and Ranka Stanković. 2023. "Language Report Serbian." In *European Language Equality: A Strategic Agenda for Digital Language Equality,* edited by Georg Rehm and Andy Way, 203–206. Cham: Springer International Publishing. ISBN: 978-3-031-28819-7. https://doi.org/10.1007/978-3-031-28819-7_32.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. “Deep learning.” *nature* 521 (7553): 436–444. https://doi.org/10.1038/nature14539.

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* 7871–7880.

Li, Yuanzhi, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. “Textbooks are all you need ii: phi-1.5 technical report.” *arXiv preprint arXiv:2309.05463.*

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. “Roberta: A robustly optimized BERT pretraining approach.” *arXiv preprint arXiv:1907.11692.*

Ljubešić, Nikola, and Filip Klubička. 2014. “{bs, hr, sr} wac-web corpora of Bosnian, Croatian and Serbian.” In *Proceedings of the 9th web as corpus workshop (WaC-9),* 29–35.

Ljubešić, Nikola, and Davor Lauc. 2021. “BERTić–The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian.” *arXiv preprint arXiv:2104.09243.*

Ljubešić, Nikola, Vit Suchomel, Peter Rupnik, Taja Kuzman, and Rik van Noord. 2024. *Language Models on a Diet: Cost-Efficient Development of Encoders for Closely-Related Languages via Additional Pretraining.* Submitted for review.

Radford, Alec, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. “Improving language understanding by generative pre-training.”

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. “Language models are unsupervised multitask learners.”

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. “Exploring the limits of transfer learning with a unified text-to-text transformer.” *The Journal of Machine Learning Research* 21 (1): 5485–5551.

Rajpurkar, Pranav, Robin Jia, and Percy Liang. 2018. "Know what you don't know: Unanswerable questions for SQuAD." *arXiv preprint arXiv:1806.03822.*

Reimers, Nils, and Iryna Gurevych. 2019. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),* 3982–3992.

Šandrih Todorović, Branislava, Cvetana Krstev, Ranka Stanković, and Milica Ikonić Nešić. 2021. "Serbian ner&beyond: The archaic and the modern intertwinned." In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021),* 1252–1260.

Škorić, Mihailo. 2023. "Композитне псеудограматике засноване на паралелним језичким моделима српског језика." Докторска дисертација. PhD diss., Универзитет у Београду.

Škorić, Mihailo, Miloš Utvić, and Ranka Stanković. 2023. "Transformer-Based Composite Language Models for Text Evaluation and Classification." *Mathematics* 11 (22): 4660.

Stanković, Ranka, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Škorić. 2020. "Machine learning and deep neural network-based lemmatization and morphosyntactic tagging for serbian." In *Proceedings of the Twelfth Language Resources and Evaluation Conference,* 3954–3962.

Suárez, Pedro Javier Ortiz, Benoît Sagot, and Laurent Romary. 2019. "Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures." In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7).* Leibniz-Institut für Deutsche Sprache.

Taori, Rohan, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. "Alpaca: A strong, replicable instruction-following model." *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html* 3 (6): 7.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. “Attention is all you need.” *Advances in neural information processing systems* 30.

Vitas, Duško, Svetla Koeva, Cvetana Krstev, and Ivan Obradović. 2008. “Tour du monde through the dictionaries.” In *Actes du 27eme Colloque International sur le Lexique et la Gammaire,* 249–256.

Wasserscheidt, Philipp. 2023. *Serbian Web Corpus PDRS 1.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1752.

Wenzek, Guillaume, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. “CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data” [in English]. In *Proceedings of the 12th Language Resources and Evaluation Conference,* 4003–4012. Marseille, France: European Language Resources Association, May. ISBN: 979-10-95546-34-4. https://www.aclweb.org/anthology/2020.lrec-1.494.