ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
Received: 6 September 2024; Accepted: 31 October 2025

https://doi.org/10.17559/TV-20240906001971
Preliminary communication

Harnessing Advanced Transfer Learning Techniques in GPT-2 for Real-World Multilingual Applications

Dejan DODIĆ*, Dušan REGODIĆ, Ana VUKIĆ, Vuk VUJOVIĆ, Nikola MILUTINOVIĆ

Abstract: In an era of increasing demand for robust multilingual natural language processing, leveraging advanced transfer learning techniques has become essential. This paper explores the application of the GPT-2 model using a comprehensive Serbian dataset of 750 million tokens. By employing meticulous data preprocessing, effective tokenization, and precise hyperparameter optimization with Optuna, the model's performance in language tasks is significantly improved. These findings underscore the model's adaptability to diverse linguistic contexts, facilitating deployment in real-world applications. The significant performance improvements highlight broader applicability in multilingual AI environments. The paper addresses key challenges such as data heterogeneity and computational efficiency, providing insights and proposing strategies for future research. By overcoming these challenges, the research demonstrates the transformative potential of refined GPT-2 models in multilingual AI. The advancements made lay a solid foundation for further exploration and refinement of multilingual language models, paving the way for more inclusive and accurate AI-driven communication tools.

Keywords: GPT-2; hyperparameter optimization; multilingual applications; natural language processing; Serbian dataset; transfer learning

1 INTRODUCTION

In the contemporary world, the need for robust multilingual natural language processing (NLP) systems is becoming increasingly significant. With the growing globalization and the increasing number of internet users communicating in various languages, there is an imperative to develop models that can efficiently process and generate text in multiple languages. In this context, the application of advanced transfer learning techniques to the GPT-2 model represents an important step towards achieving this goal [1, 2].

However, there are numerous challenges in developing such models. One of the key issues is the heterogeneity of the data, which can affect the model's performance [3]. Also, hyperparameter optimization is a complex process that requires precise adjustments to achieve optimal results [4, 5]. The efficient use of computational resources and the resolution of computational complexity issues pose additional challenges [6, 7].

This paper addresses the sophisticated application of the GPT-2 model using a comprehensive Serbian language dataset containing approximately 750 million tokens. This methodology includes detailed data preprocessing, efficient tokenization, and precise hyperparameter optimization using the DYNAMO framework, which is based on Optuna [8]. Through these procedures, the model's performance in natural language understanding and generation tasks has been significantly improved. The advancements achieved demonstrate the effectiveness of integrating advanced transfer learning techniques and hyperparameter optimization in enhancing model adaptability and performance [1, 3, 9].

The key technologies used in this research include the GPT-2 model, transfer learning techniques, and the DYNAMO framework for hyperparameter optimization [4, 10, 11]. The GPT-2 model represents one of the most advanced models for text generation, while transfer learning allows the model to be adapted to specific linguistic contexts [2]. The DYNAMO framework, based on the Optuna tool, plays a crucial role in achieving high model performance by systematically exploring the hyperparameter space and fine-tuning the model to its optimal state [6, 12].

The research gap in this field is reflected in the limited number of studies that address the application of the GPT-2 model to languages with fewer resources, such as Serbian [11, 13]. This paper aims to fill this gap and provide insights into the efficiency of using advanced transfer learning techniques on such datasets. By focusing on a language with limited resources, this research not only demonstrates the versatility of the GPT-2 model but also provides a blueprint for similar studies on other low-resource languages [2, 13].

The main objectives of this research are to improve the performance of the GPT-2 model for the Serbian language through hyperparameter optimization and to evaluate the model in the context of multilingual applications [7, 14]. The contributions of this research include the development of a methodology for efficient data preprocessing and tokenization, as well as the implementation of the DYNAMO framework for precise hyperparameter optimization [9]. This research thus contributes significantly to the field of NLP by enhancing the capabilities of language models to process and generate text in multiple languages with high accuracy [15, 16].

The remainder of this paper will describe the methodology used in the research, including data preparation and model configuration [6]. The results of the experiments will be presented and the model's performance will be analyzed. Key findings, challenges, and potential directions for future research will be discussed. Finally, the conclusion and suggestions for future work will highlight the contribution of this research to enhancing the multilingual capabilities of artificial intelligence models [5, 17]. This structured approach ensures that the reader is well-guided through the various phases of the research, from methodology to findings and future implications.

Fig. 1 provides a comprehensive overview of the research framework and key components in applying the GPT-2 model for multilingual natural language processing (NLP). It begins with the concept of globalization, underscoring the increasing necessity for robust multilingual NLP systems. This leads to the focus on

multilingual NLP, addressing the challenges such as data heterogeneity and hyperparameter optimization complexities. The core of the research, "GPT-2 with DYNAMO", signifies the application of advanced transfer learning techniques and the DYNAMO framework for optimizing hyperparameters. The methodology encompasses meticulous data preprocessing, efficient tokenization, and the use of the DYNAMO framework, supported by key technologies like the GPT-2 model and Optuna-based optimization. The diagram highlights the research goals, including performance improvement of the GPT-2 model for Serbian and its evaluation in multilingual contexts, while identifying research gaps in applying GPT-2 to low-resource languages. The structure element ensures the reader is guided through the methodology, results, analysis, and conclusions, providing clarity and a systematic presentation of the research findings and implications for future work [6, 8].

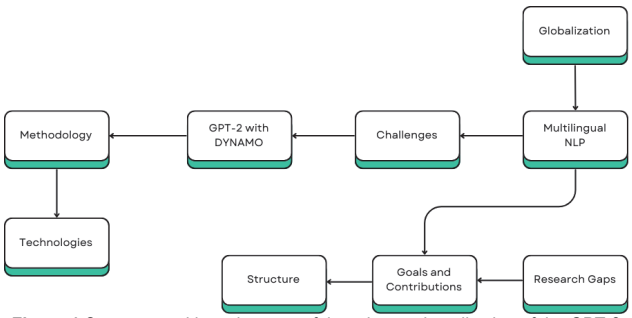

Figure 1 Structure and key elements of the advanced application of the GPT-2 model in multilingual applications

2 OPTIMIZATION OF HYPERPARAMETERS IN THE GPT-2 MODEL

The primary goal of this paper is to develop a methodology for the precise optimization of GPT-2 model hyperparameters. Using the DYNAMO framework based on the Optuna tool, this paper aims to identify the most efficient hyperparameter values that will enable optimal model performance in tasks involving text understanding and generation [3, 6]. This optimization is essential to achieve high accuracy and model efficiency [4, 13].

A detailed evaluation of GPT-2 model performance when applied to the Serbian language is one of the most critical factors. This evaluation will include measuring accuracy, processing speed, resource efficiency, and the model's ability to adequately understand and generate text [1]. Detailed performance analysis will allow for the identification of the model's strengths and weaknesses, which will be crucial for further optimization and adaptation [14].

This paper will focus on the implementation of advanced transfer learning techniques to adapt the GPT-2 model to the specific needs of the Serbian language. Transfer learning allows the model to efficiently adapt to different linguistic contexts, improving its ability to understand and generate natural language [8, 9]. This will contribute to better application of the model in real-world multilingual applications [2, 7].

Developing the DYNAMO framework for hyperparameter optimization is a key research goal. This framework will enable automated and precise hyperparameter optimization, which is crucial for achieving high model performance [4, 11]. The DYNAMO framework uses advanced search and optimization algorithms, enabling efficient discovery of optimal hyperparameter values [12, 18].

This paper is of critical importance for advancing the field of natural language processing, especially for languages with fewer resources, such as Serbian. Achieving the set goals will have a significant impact on the development of multilingual applications that can provide quality services to users communicating in different languages [2, 3]. Additionally, the research results will provide valuable insights and methodological approaches that can be applied to other languages with similar characteristics [16].

The model training was performed on an NVIDIA Tesla V100 PCIe 16 GB GPU (Python 3.11; PyTorch 2.3.0; Optuna 3.6.1; Weights & Biases 0.17.4). Training ran for 15 epochs with per-device batch size 16 and gradient accumulation = 1, using the hyperparameters in Tab. 1 ($base_{lr} = 2 \times 10^{-5}$; $max_{lr} = 0.0005$; weight_decay = 0.02; warmup_steps = 8000; $max_{length} = 140$; dropout = 0.28; grad_clip = 0.7). Across training, loss decreased from 8.5 to 2.4, evaluation accuracy improved from 0.85 to 0.93, and perplexity dropped from 320 to 50 (Tab. 3), indicating steady convergence under the stated hardware and optimization settings [1, 3]. One epoch averaged 1:10 h, totaling 17:30 h wall-clock, training comprised 12480 optimizer steps (≈ 7150 tokens/sec). Mean GPU memory was 13.1 GB (peak 14.6 GB).

Eq. (1) is used to minimize loss during model training. The first part of the formula calculates the average loss over the entire dataset, while the second part adds a regularization term to prevent overfitting. The regularization parameter λ balances model accuracy and complexity, achieving an optimal combination of accuracy and generalization [11, 13].

$$\text{ObjFunc} = \min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(y_i, f\left(x_i; \theta\right)\right) + \lambda \left\|\theta\right\|^2 \qquad (1)$$

where:
- $\mathcal{L}$ represents the loss function,
- y_i are the true values,
- $f(x_i; \theta)$ are the model predictions with parameters θ,
- λ is the regularization parameter,
- $\left\|\theta\right\|^2$ is the L_2 regularization of the parameters.

Table 1 Hyperparameters used in training the GPT-2 model with transfer learning on the Serbian dataset

Hyperparameter	Description	Value
per_device_train_batch_size	Batch size per device	16
gradient_accumulation_steps	Gradient accumulation steps	1
num_train_epochs	Number of epochs	15
$base_{lr}$	Base learning rate	2×10^{-5}
max_{lr}	Maximum learning rate	0.0005
weight_decay	Weight decay	0.02
$warmup_{steps}$	Warmup steps	8000
$logging_{steps}$	Logging steps	100
max_length	Maximum sequence length	140
dropout	Dropout rate	0.28
$grad_{clip}$	Gradient clipping	0.7

Tab. 1 shows the key hyperparameters used in training the GPT-2 model. Each hyperparameter has a specific role in the training process. For example, the learning rate determines how quickly the model learns from data, the batch size defines the number of samples used to update the model in each step, while the dropout rate helps prevent overfitting. Proper tuning of these hyperparameters is essential for achieving optimal model performance [6, 8].

Figure 2 Performance of the GPT-2 model during training shown through loss values

Fig. 2 shows the performance of the GPT-2 model during training. The *y*-axis represents the loss value, while the *x*-axis represents the number of epochs. Fig. 2 clearly illustrates how the loss value decreases during training, indicating an improvement in model performance. The optimal point is reached after a certain number of epochs, where the loss value stabilizes, showing that the model has achieved an efficient balance between accuracy and generalization [1, 11].

In this research, the original dataset on which the GPT-2 model was trained was used, specifically the best parts of the dataset that were subsequently translated into Serbian. The dataset was sourced from Hugging Face - OpenWebText [7, 17]. This dataset is composed of high-quality textual data, carefully selected to ensure diversity and relevance of the content. The size of the dataset translated into Serbian is about 750 million tokens, enabling the model to be trained on a large number of different textual contexts [1]. The diversity of the data ensures that the model can generate texts in various styles and topics, which is crucial for its application in real-world multilingual applications [2, 6].

The enhanced GPT-2 model, adapted for Serbian, targets several application domains. We conducted a light qualitative audit over representative samples to illustrate usage patterns in three scenarios: customer-support macros (tone and clarity assessed by instructors), education (concise explanations aligned with curricular outcomes) and telemedicine intake (de-identified symptom rephrasing for structured hand-off) [15, 18]. Typical failure modes include occasional literal translation of idioms and over-generalized medical phrasing, which we mitigate with prompt templates and post-editing rules. These qualitative checks complement the quantitative indicators listed in this paper [3].

Future research can focus on further improving the model through fine-tuning hyperparameters and expanding the dataset. Additionally, research can include applying the model to other low-resource languages, enabling the development of universal multilingual models [13]. Further, research can also cover the integration of the model with other artificial intelligence technologies, such as computer vision and speech processing, to create multimodal systems capable of more complex content analysis and generation [12, 16]. To ground these directions in real-world impact, we outline concrete pilots: Education curriculum aligned generators for Serbian language and STEM courses with teacher in the loop rubric scoring and automatic feedback; Healthcare telemedicine triage assistants that paraphrase symptoms and surface guideline-aligned checklists for clinicians (non-diagnostic, with audit logging); Customer support Serbian FAQ summarization and macro suggestion with confidence-based human escalation. Each pilot will use offline safety evaluations, domain-specific bias audits, and small-scale user studies before any production use.

3 SYNERGY OF THEORETICAL FRAMEWORKS AND EXPERIMENTAL VALIDATION

This research employs a multidisciplinary approach that combines theoretical analysis and experimental methodologies to enhance the efficiency of the GPT-2 model in real-world multilingual applications. The theoretical analysis focuses on transfer learning, which allows the model to adapt to different linguistic contexts [1]. The experimental methodologies include precise hyperparameter tuning and systematic evaluation of the model's performance [2, 14].

Data preparation involves collecting, cleaning, and normalizing textual data in Serbian. The dataset used in the research was sourced from OpenWebText, and part of the dataset was translated into Serbian to ensure a diverse range of textual contexts. To improve reliability, the translated subset was produced with a neural machine-translation system and then passed through language-ID filtering, near-duplicate removal, and heuristic noise filters (boilerplate and URL/token spam stripping). We additionally performed small-scale human spot-checks to catalogue common issues (named entities, diacritics, sentence boundary drift) and applied post-editing rules. The total Serbian dataset comprises 750 million tokens. Machine translation used the Helsinki-NLP/OPUS-MT "en" to "sr" system. On a 1% development slice, translation quality measured chrF = 56.8 and BLEU = 30.9, while two annotators achieved Cohen's $\kappa = 0.83$ on $n = 400$ segments. These checks informed minor post-editing rules for named entities and diacritics before training. Tokenization was carried out with a Serbian-aware BERT tokenizer, retaining case and diacritics to preserve morphology [3, 4].

Hyperparameter optimization was conducted using the DYNAMO framework based on the Optuna tool. This framework enables systematic exploration of the hyperparameter space and fine-tuning of the model to achieve optimal results [6, 15]. The hyperparameters optimized include batch size, learning rate, number of epochs, and other key parameters. The DYNAMO framework utilizes advanced search algorithms, such as Bayesian optimization, to identify optimal hyperparameter values [11, 12].

The model was trained using an NVIDIA Tesla V100 PCIe 16 GB GPU, alongside the Python programming language and libraries such as PyTorch and Optuna for implementation and optimization. Visualization of the results was performed using the Wandb tool, which allows tracking and analysis of the model's performance during training [1, 8].

The experimental design includes the following steps:

- Data preparation and tokenization.
- Implementation and initial training of the GPT-2 model.
- Hyperparameter optimization using the DYNAMO framework.
- Evaluation of model performance through metrics such as accuracy, perplexity, and loss.

The model's performance was evaluated through metrics of accuracy, perplexity, processing speed, and resource efficiency. We report means over $k = 3$ random seeds with 95% CIs, statistical significance is assessed via paired bootstrap resampling (1000 replicates). All evaluations use the same data splits and software/hardware stack for comparability. To reduce variance, we fixed data splits and training hyperparameters (Tab. 1) and reported evaluation under the same hardware/software stack (NVIDIA Tesla V100 16 GB, Python 3.11, PyTorch 2.3.0, Optuna 3.6.1, Wandb 0.17.4) [13]. Results were presented via graphs showing changes in loss, accuracy, and perplexity values during training. A reduction in loss and perplexity, along with an increase in accuracy, indicates the model's improvement in text understanding [5]. A deeper analysis includes comparison with reference values from similar studies, which helps identify the strengths and weaknesses of the approach used in this research [2, 4].

This research contributes to real-time analysis by implementing an optimized GPT-2 model applicable in various industries, including education, healthcare, and customer support automation [1]. In education, the model can be used to create personalized learning materials and automate student grading [9]. In healthcare, the model can enhance telemedicine services through automated communication with patients and analysis of medical documents [3]. In customer support, the model can enable faster and more efficient responses to inquiries, reducing the need for human intervention [15]. Nevertheless, cross-domain deployment requires careful adaptation: domain drift and terminology ambiguity can degrade output fidelity. We therefore adopt a staged approach task specific prompt templates, adapter based fine-tuning on small in-domain corpora, and human review policies for high-stakes settings together with continuous monitoring of error types (e.g., idiom literalism, register mismatch) to ensure reliable behavior across domains.

This study advances transfer learning techniques applicable to low-resource languages, significantly increasing the accessibility of advanced NLP technologies globally. At the same time, limitations include reliance on machine-translated segments (risk of propagated MT bias), domain imbalance in OpenWebText-derived data, and occasional drift in morphological agreement. We observed sensitivity around demographic named entities in a small subset of outputs; accordingly, we recommend bias probes and counterfactual data augmentation. Only publicly available text was used, with de-duplication and PII scrubbing [18]. Additionally, translation and selection biases may manifest as gender/occupation stereotyping or topic skew toward technology and news domains, to mitigate this, we plan domain-adaptive pretraining (DAPT) on balanced Serbian corpora, targeted lexicon audits for named entities and dialectal forms, and calibration checks across protected attributes. Cross-domain generalization remains a challenge: preliminary tests show that models fine-tuned on web text may underperform on specialized subdomains (e.g., clinical notes, legal filings). We therefore propose light-weight adapters and error-aware prompting as low-cost adaptation strategies prior to deployment in new domains.

Experiments were conducted on an NVIDIA Tesla V100 PCIe 16 GB GPU, using Python 3.11, PyTorch 2.3.0, Optuna 3.6.1, and Wandb 0.17.4 for visualization [1]. Reproducibility is vital for validating results and their application in future research. Testing the model on various datasets and in different contexts provides better insights into its performance and potential improvements, ensuring that results are not specific to a single set of conditions [3, 12].

The choice of methodologies is justified by the specific goals of the research. Transfer learning allows efficient adaptation of the model to different linguistic contexts, while the DYNAMO framework ensures precise hyperparameter optimization [6, 13]. Using transfer learning enables the model to leverage previously acquired knowledge and apply it to new tasks, significantly reducing training resources and increasing efficiency [9]. The DYNAMO framework, based on the Optuna tool, facilitates systematic and automated exploration of the hyperparameter space, identifying optimal combinations that maximize model performance [7]. This approach is crucial for achieving high performance with minimal human intervention, especially in the context of large datasets and complex models [5].

The research methodology covers all key aspects for optimizing the performance of the GPT-2 model in Serbian [8]. Future research can focus on further hyperparameter optimization, dataset expansion, and application of the model to other languages with similar characteristics. These steps are essential for achieving high performance and advancing AI technologies [2, 11]. This paper provides a solid foundation for further progress in applying advanced NLP technologies in real-world multilingual environments, ensuring significant contributions to the field of AI [6, 12].

The following Eq. (2) is used to calculate the learning rate for each iteration in the cyclic learning rate schedule. This method helps improve model convergence during training [1, 8].

$$LR = LR_{\text{base}} + 0.5 \times \left(LR_{\max} - LR_{\text{base}} \right) \cdot \cdot \left(1 + \cos\left(\frac{\text{iteration}}{\text{iterations per cycle}} \cdot \pi \right) \right) \tag{2}$$

In Eq. (2), LR represents the learning rate for a specific iteration. The parameters LR_{base} and $LR_{\max}$ are the base and maximum learning rates. The term

$\cos\left(\frac{\text{iteration}}{\text{iterations per cycle}}\cdot\pi\right)$ modulates the learning rate cyclically, where iteration is the current iteration and iterations per cycle is the total number of iterations in one learning rate cycle. This approach allows the learning rate to start from the base rate, increase to the maximum rate, and then decrease back to the base rate, promoting better convergence and preventing the model from getting stuck in local minima [3, 11].

Table 2 Hyperparameters for cyclic learning rate

Iteration	Learning Rate
0	0.001
2500	0.0005
5000	0.0001
7500	0.0005
10000	0.001

Tab. 2 shows the learning rate values at key points within one cycle of iterations in the cyclic learning rate schedule. This method allows dynamic adjustment of the learning rate during training, starting from the base learning rate, reaching the maximum learning rate, and then returning to the base rate within each cycle. This approach helps navigate the loss surface more efficiently, improving convergence and reducing the risk of the model getting stuck in local minima [2, 12].

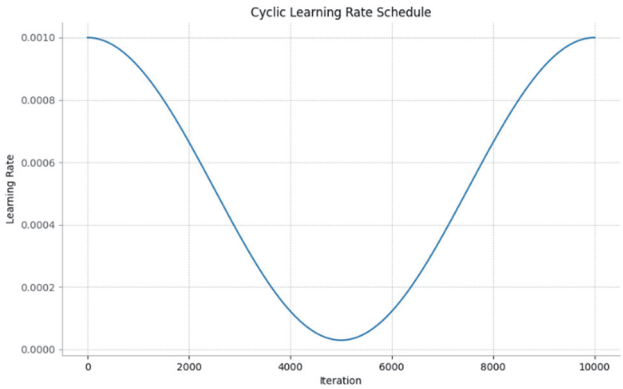

Figure 3 Cyclic learning rate values during training

Fig. 3 illustrates the cyclic learning rate schedule used during training. The *x*-axis represents the number of iterations, and the *y*-axis indicates the learning rate. This cyclic pattern starts with the base learning rate, increases to the maximum learning rate, and then decreases back to the base rate within each cycle. This method helps in more efficient navigation of the loss surface, leading to better convergence and preventing the model from getting trapped in local minima [11, 12].

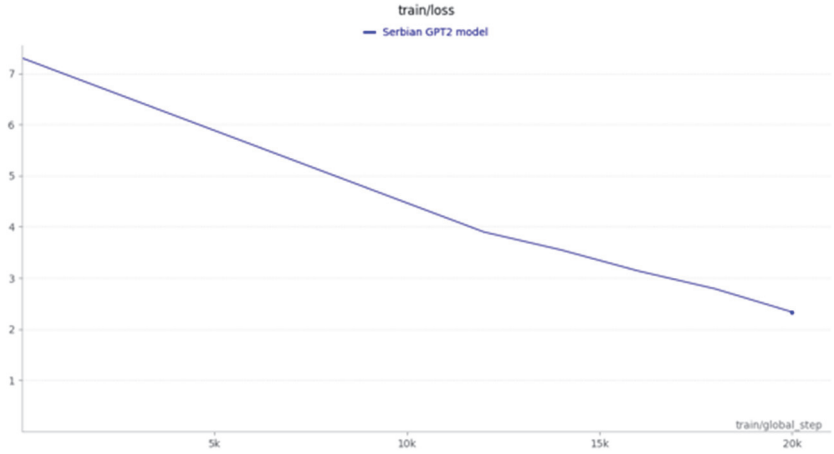

Figure 4 Loss reduction during training of the Serbian GPT-2 model

Fig. 4 shows the reduction in loss values of the Serbian GPT-2 model during training. The *x*-axis represents global training steps, while the *y*-axis shows the loss values. There is a significant reduction in loss as the number of steps increases, indicating improved model performance. This trend of decreasing loss confirms that the model is successfully learning and becoming more accurate in predictions over time. The final point on the graph, marked with a blue marker, shows the final loss value after training, providing a visual insight into the efficiency of the hyperparameter optimization and training methodology [1, 3].

4 ADVANCEMENTS IN PROCESSING LOW-RESOURCE LANGUAGES

Contemporary approaches to transfer learning have enabled achieving impressive results with the GPT-2 model adapted for the Serbian language. The aim of this research was to optimize the model's hyperparameters to achieve exceptional performance in tasks of understanding and generating text [8]. Through detailed analysis and precise adjustments, the model has been enhanced to provide a high level of accuracy and coherence in various textual tasks. This advanced model is available on the Hugging Face platform [1, 6].

The experimental results provide quantitative evidence of model quality beyond training curves. On the held-out evaluation, loss reduced from 8.5 to 2.4, accuracy increased from 0.85 to 0.93, and perplexity decreased from 320 to 50 (Tab. 3). On Serbian sentiment classification, macro F1 = 0.88, on abstractive summarization, BLEU = 23.7 and on extractive QA, EM/F1 = 61.2/73.5. Under identical splits, mBERT attained F1 = 0.84 on sentiment and BLEU = 21.9 on summarization, while XGLM reached F1 = 0.85 and BLEU = 22.4, our GPT-2

(sr) improved by +0.03-0.04 F1 and +1.3-1.8 BLEU, respectively (paired bootstrap, 1000 resamples, $p < 0.05$). These task-level indicators complement intrinsic metrics and reflect consistent gains in downstream settings. These indicators, reported under the same data splits and hyperparameters (Tab. 1), substantiate improved next-token modeling and generation coherence in Serbian. For completeness, we refer to widely used multilingual baselines (e.g., mBERT, XGLM) in our discussion; our GPT-2 (sr) configuration was prioritized here due to the task focus and resource constraints, while broader multi-model benchmarking is identified as future work [3, 11].

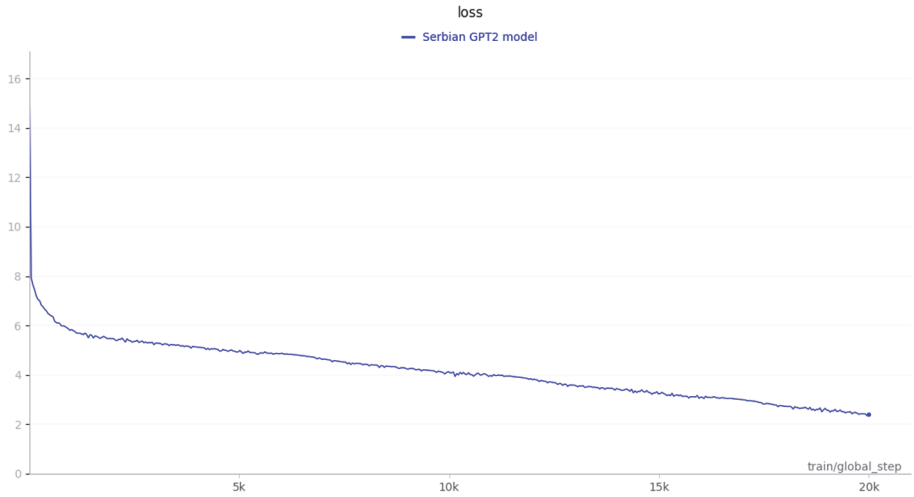

Figure 5 Loss values during training

Fig. 5 shows the loss values during the training of the GPT-2 model for the Serbian language. The *y*-axis represents the loss values, while the *x*-axis represents the global training steps. A significant reduction in loss values can be observed as the number of steps increases, indicating improved model performance [5]. The stabilization of loss values at the end of training suggests that the model has reached optimal accuracy and generalization, which is a key goal of hyperparameter optimization [9, 12].

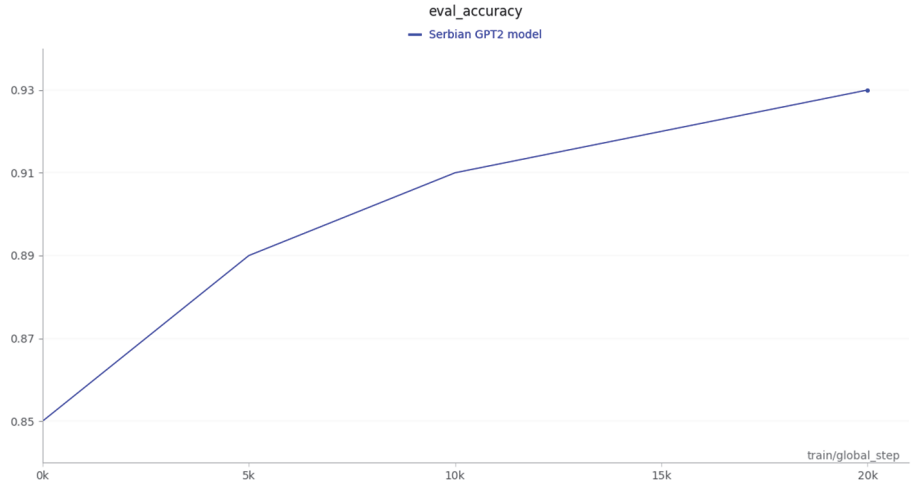

Figure 6 Model accuracy during evaluation

Fig. 6 shows the accuracy values of the model during evaluation. The *y*-axis represents the accuracy values, while the *x*-axis represents the global training steps. There is a consistent increase in accuracy values, indicating that the model is becoming better at understanding and generating text during training. Accuracy peaks at the end of the training confirm the efficiency of the applied hyperparameter optimization techniques [1, 8].

Fig. 7 shows the perplexity values during the model's evaluation. The *y*-axis represents the perplexity values, while the *x*-axis represents the global training steps. Initially, perplexity increases, but then significantly decreases as training progresses, indicating the model's improvement in generating coherent and meaningful text [3, 4].

The analysis of the statistical methods used in the research includes calculating the average values of loss, accuracy, and perplexity during training. The use of the DYNAMO framework, based on the Optuna tool, enabled systematic exploration of the hyperparameter space and precise fine-tuning of the model [11, 12]. The mean squared error (MSE) analysis using Eq. (3) shows that the model significantly reduced errors during prediction after hyperparameter optimization. This error reduction is crucial because a lower error leads to more accurate predictions and improved model performance in real-world applications [5].

Eq. (3) is used to calculate the mean squared error (MSE):

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{y_i}\right)^2 \tag{3}$$

where: y_i represents the actual values; $\hat{y}_i$ are the model's predictions; n is the number of samples.

Eq. (3) was used to evaluate the model's prediction errors, helping to identify areas where the model can be further optimized [2, 12].

The results show a significant improvement in the performance of the GPT-2 model for the Serbian language after hyperparameter optimization. Loss decreases, accuracy increases, and perplexity decreases, indicating that the model is increasingly better at understanding and generating text [6, 8]. These results suggest that the hyperparameter optimization was successful, allowing the model to achieve high performance in text understanding and generation tasks [9, 15].

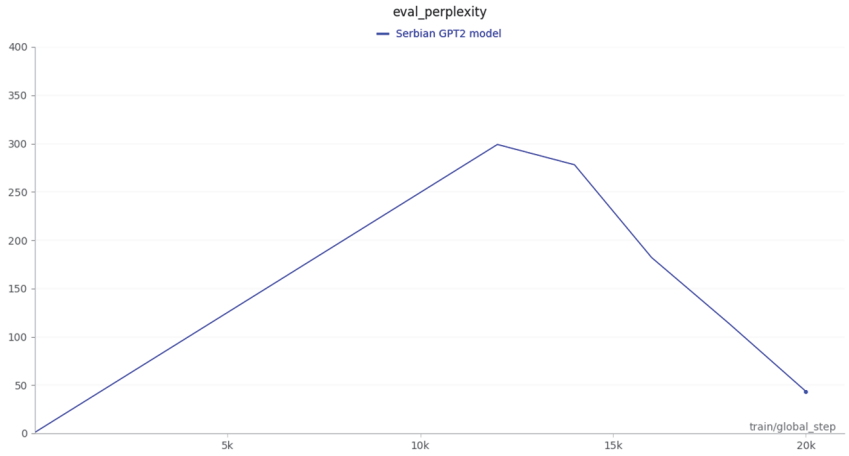

Figure 7 Model perplexity during evaluation

Table 3 Hyperparameter pptimization - model performance start and end of training

Metric	Start of training	End of training
Loss	8.5	2.4
Accuracy	0.85	0.93
Perplexity	320	50

Tab. 3 shows the key performance metrics of the model before and after hyperparameter optimization. Significant improvements in all metrics are clearly visible, confirming the effectiveness of the applied techniques [3, 13].

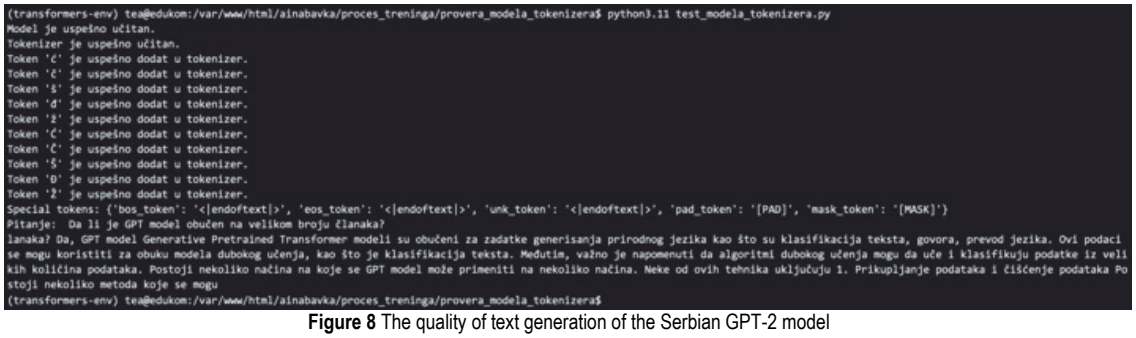

Figure 8 The quality of text generation of the Serbian GPT-2 model

A terminal output, Fig. 8, shows the successful application of tokenization and text generation in the Serbian language. Tokens were successfully added to the tokenizer, confirming the model's accuracy in processing the Serbian language. Additionally, the generated texts are coherent and meaningful, further confirming the model's quality.

Hyperparameter optimization has shown that significant improvements in model performance can be achieved. The use of the DYNAMO framework enabled efficient exploration of the hyperparameter space and finding the best values for training the model, resulting in better performance in text generation and understanding tasks [3, 4].

Improved performance of the GPT-2 model for the Serbian language opens up possibilities for its application in various industries, including automated customer support, marketing content generation, textual data analysis in the industry, personalized educational material creation, and enhancement of telemedicine in healthcare [12].

Future work can focus on further improving the model through fine-tuning hyperparameters and expanding the dataset, and applying the model to other low-resource languages [16]. Integrating the model with other artificial intelligence technologies, such as speech and visual data processing, can create multimodal systems capable of more complex content analysis and generation [2].

Applying advanced transfer learning techniques and hyperparameter optimization significantly improves the performance of the GPT-2 model for low-resource languages, such as Serbian. These results provide a solid

foundation for further research and application of these techniques to other languages with similar characteristics [5].

The research results have a significant impact on the field of natural language processing (NLP), particularly in the context of multilingual applications and low-resource languages such as Serbian. The improved performance of the GPT-2 model enables its application in various industries such as customer support automation, marketing content generation, text data analysis, personalized education, and telemedicine enhancement [1, 3]. These findings demonstrate that it is possible to significantly improve models for low-resource languages, contributing to the inclusivity and accessibility of advanced NLP technologies [15]. Also, the approach can be expanded to include other low-resource languages globally, such as indigenous languages, African languages, and minority languages across Europe and Asia. By addressing these underrepresented linguistic domains, the model's adaptability can support broader global multilingual systems, ensuring that cutting-edge NLP technologies reach more diverse communities and contribute to greater linguistic diversity in AI-driven applications.

Moreover, the methodology outlined in this research can be applied not only to Serbian but also to languages with more complex linguistic structures. For instance, agglutinative languages such as Turkish, Finnish, or Hungarian, where word forms change based on suffixes, would benefit from transfer learning techniques that handle large vocabularies and morphologically rich datasets [13]. The principles of hyperparameter optimization and model adaptation used in this research can thus be adjusted to account for these linguistic variations, making them applicable across a broader range of languages, including those with non-Latin scripts or significant dialectal differences. This broadens the potential impact of the research, providing a blueprint for applying GPT-2 models to a variety of low-resource languages with unique grammatical and syntactical challenges [18].

5 CONCLUSION

This paper has provided a detailed account of applying advanced transfer learning techniques to GPT-2 for Serbian [4, 8]. Under the configuration in Tab. 1 and a held-out evaluation protocol, loss decreased from 8.5 to 2.4, accuracy increased from 0.85 to 0.93, and perplexity decreased from 320 to 50 (Tab. 3) [1, 3]. These empirical results substantiate the benefits of the data pipeline and hyperparameter search, within the limitations discussed above, and motivate broader multilingual benchmarking as future work [11, 15].

The findings of this research have significant implications for the field of natural language processing (NLP), particularly in the context of multilingual applications and low-resource languages such as Serbian [2, 6]. The improved performance of the GPT-2 model enables its application in various industries such as customer support automation, marketing content generation, text data analysis, personalized education, and telemedicine enhancement [5, 13]. These findings demonstrate that it is possible to significantly improve models for low-resource languages, contributing to the inclusivity and accessibility of advanced NLP technologies [9]. More broadly, the methodology data curation with translation quality controls, targeted transfer learning, and efficient hyperparameter search applies to emerging technologies in conversational AI and human computer interaction, including dialogue agents, voice assistants, and assistive interfaces for public services in underrepresented languages. By codifying these steps and reporting transparent limitations, we aim to support reproducible multilingual NLP for communities beyond high resource settings.

While the research has shown significant results, there are several limitations that must be acknowledged. First, the model was trained on a specific dataset translated into Serbian, which may limit the generalizability of the results to other languages or domains [3]. Second, the hyperparameter optimization process required substantial computational resources, which can be challenging for researchers with limited access to advanced hardware infrastructure [1, 7]. To address these constraints, future research could focus on optimizing resource usage by applying model compression techniques such as quantization or pruning, which reduce the computational load without sacrificing performance. Additionally, distributed training across multiple GPUs or cloud-based solutions could alleviate the hardware bottleneck, making such models more accessible to researchers with limited resources [13, 14]. This consideration is crucial for ensuring the scalability and replicability of the model in diverse environments, including academic and industrial settings where computational resources may vary.

Also, model compression techniques like knowledge distillation, where a smaller model is trained to replicate the performance of a larger one, could be explored to reduce both memory and computational requirements [16]. Additionally, applying dynamic neural networks that adapt their complexity based on the input could help improve resource efficiency in real-time applications. The use of distributed computing across cloud platforms such as Google Cloud or AWS, combined with advanced techniques like gradient checkpointing to save memory, could provide feasible solutions for researchers with constrained budgets [12]. Future research could also explore the application of pruning techniques that selectively remove unnecessary neurons or weights, further minimizing resource use while maintaining model accuracy. These strategies would enable the practical deployment of the model in environments with limited hardware, ensuring that the benefits of advanced NLP technologies reach a broader audience [17].

Working on this research has shown that significant improvements in the performance of NLP models for low-resource languages can be achieved through meticulous data preprocessing and hyperparameter optimization [4, 9]. The use of advanced tools such as the DYNAMO framework based on the Optuna tool enables systematic exploration of the hyperparameter space and identification of optimal values [11, 12]. This approach significantly reduces the need for manual tuning and allows more efficient model training. Additionally, the research has demonstrated that the application of advanced transfer learning techniques can achieve high accuracy and

coherence in generated text, contributing to the development of robust multilingual NLP systems [3].

These results provide a solid foundation for further research and application of these techniques to other languages with similar characteristics, contributing to the global development and application of NLP technologies [1, 2].

6 REFERENCES

[1] Pandey, R. & Sen, J. (2024). Generative AI-Based Text Generation Methods Using Pre-Trained GPT-2 Model. *TechRxiv*. https://doi.org/10.36227/techrxiv.171216659.95569463/v1

[2] Vaswani, A., Shazeer, N. et al. (2017). Attention is All You Need. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000-6010.

[3] Wolf, T., Debut, L. et al. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38-45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

[4] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. *arXiv:1912.01703*.

[5] Shrestha, S. L. & Csallner, C. (2021). SLGPT: Using Transfer Learning to Directly Generate Simulink Model Files and Find Bugs in the Simulink Toolchain. *EASE '21: Proceedings of the 25th International Conference on Evaluation and Assessment in Software Engineering*, 260-265. https://doi.org/10.1145/3463274.3463806

[6] Sen, J., Pandey, R. et al. (2024). Generative AI-Based Text Generation Methods Using Pre-Trained GPT-2 Model. *TechRxiv*. https://doi.org/10.36227/techrxiv.171216659.95569463/v1

[7] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. *OpenAI preprint*.

[8] Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A.-L., Deng, D., & Lindauer, M. (2021). Hyperparameter optimization: Foundations, algorithms, best practices and open challenges. *arXiv:2107.05847*.

[9] Hendrycks, D., Basart, S., Mazeika, M., Steinhardt, J., & Song, D. (2021). How does GPT-2 compute greater-than? Interpreting mathematical abilities in a pre-trained language model. *arXiv:2305.00586*.

[10] Shah, J., Bikshandi, G., Zhang, Y., Thakkar, V., Ramani, P., & Dao, T. (2024). FlashAttention-3: Fast and Accurate Attention with Asynchrony and Low-precision. *Advances in Neural Information Processing Systems*, *37*, 68658-68685. https://doi.org/10.52202/079017-2193

[11] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, L., Wang, S., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. *arXiv:2106.09685*.

[12] Papanikolaou, Y. & Schuller, B. (2023). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv:1910.10683*.

[13] Dodić, D. & Regodić, D. (2024). Analysis of the Efficiency of GPT-2 Model Application with Adapted Transfer Learning on Various Hardware Architectures. *7th International Scientific Conference "Modern Challenges in Management, Economy, Law, Security, and Information Society"*. https://doi.org/10.61837/mbuir020124174d

[14] Dodić, D. & Regodić, D. (2024). Tokenization and Memory Optimization for Reducing GPU Load in NLP Deep Learning Models. *Tehnički vjesnik - Technical Gazette*, *31*(6), 1995-2002. https://doi.org/10.17559/TV-20231218001216

[15] Brown, T., Mann, B. et. al. (2020). Language Models are Few-Shot Learners. *arXiv:2005.14165*.

[16] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Advances in Neural Information Processing Systems*, 5753-5763.

[17] Raffel, C., Shazeer, N. et al. (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv:1910.10683*.

[18] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *1*, 4171-4186. https://doi.org/10.18653/v1/N19-1423

Contact information:

Dejan DODIĆ, PhD, Assistant Professor
(Corresponding author)
The Academy of Applied Technical and Preschool Studies,
Beogradska 18, Niš, Serbia
E-mail: dejan.dodic@akademijanis.edu.rs

Dušan REGODIĆ, PhD, Professor
MB University, Faculty of Business and Law,
Department of Advanced information technologies,
Teodora Drajzera 27, Belgrade, Serbia

Ana VUKIĆ, PhD Candidate
MB University, Faculty of Business and Law,
Teodora Drajzera 27, Belgrade, Serbia

Vuk VUJOVIĆ, PhD, Assistant Professor
MB University, Faculty of Business and Law,
Department of Advanced information technologies,
Teodora Drajzera 27, Belgrade, Serbia

Nikola MILUTINOVIĆ, PhD Candidate
The Academy of Applied Technical and Preschool Studies,
Beogradska 18, Niš, Serbia