

**UNIVERZITET MB  
POSLOVNI I PRAVNI FAKULTET**

**DOKTORSKA DISERTACIJA**

**GPT-2 PARADIGMA: TRANSFERNO UČENJE I ADAPTACIJA  
PYTORCH I TRANSFORMERS ARHITEKTURA NA PUTU KA  
VIŠEJEZIČKOM SVETU**

Mentor:

Redovni profesor dr Dušan Regodić, dipl.inž.

Student:

Dejan Dodić

**Beograd, 2024.**

**UNIVERSITY MB**  
**FACULTY OF BUSINESS AND LAW**

**DOCTORAL DISSERTATION**

**GPT-2 PARADIGM: TRANSFER LEARNING AND ADAPTATION  
OF PYTORCH AND TRANSFORMERS ARCHITECTURES  
TOWARDS A MULTILINGUAL WORLD**

Mentor:

Redovni profesor dr Dušan Regodić, dipl.inž.

Student:

Dejan Dodić

**Belgrade, 2024.**

## IDENTIFIKACIONA STRANICA DOKTORSKE DISERTACIJE

<b>Autor</b>
<b>Ime i prezime:</b> Dejan Dodić
<b>Datum i mesto rođenja:</b> 5.10.1977. Vranje
<b>Sadašnje zaposlenje:</b> Asistent računarstva i informatike – Akademija tehničko vaspitačkih strukovnih studija Niš
<b>Doktorska disertacija</b>
<b>Naslov:</b> GPT-2 PARADIGMA: TRANSFERNO UČENJE I ADAPTACIJA PYTORCH I TRANSFORMERS ARHITEKTURA NA PUTU KA VIŠEJEZIČKOM SVETU
<b>Broj stranica:</b> 186
<b>Broj slika:</b> 59
<b>Broj tabela:</b> 26
<b>Broj formula:</b> 26
<b>Ustanova i mesto gde je rad izrađen:</b> Univerzitet MB - Poslovni i pravni fakultet
<b>Naučna oblast (UDK):</b> Tehničko-tehnološke nauke, Elektrotehničko i računarsko inženjerstvo, Informacione tehnologije i sistemi
<b>Mentor:</b> redovni profesor dr Dušan Regodić, dipl inž, Univerzitet MB - Poslovni i pravni fakultet
<b>Oцена i odbrana</b>
<b>Datum prijave teme:</b>
<b>Broj odluke i datum prihvatanja teme doktorske/umetničke disertacije:</b>
<b>Komisija za ocenu naučne zasnovanosti teme i ispunjenosti uslova kandidata:</b> <b>1. Redovni profesor dr Dušan Regodić, dipl inž. - Mentor</b> <b>2. Redovni profesor dr Neđo Danilović- Predsednik komisije</b> <b>3. Profesor dr Damir Jerković, dipl inž. – Član komisije</b>
<b>Komisija za ocenu i odbranu doktorske/umetničke disertacije:</b> <b>1. Redovni profesor dr Dušan Regodić, dipl inž. - Mentor</b> <b>2. Redovni profesor dr Neđo Danilović- Predsednik komisije</b> <b>3. 3. Profesor dr Damir Jerković, dipl inž. – Član komisije</b>
<b>Datum odbrane disertacije:</b>

## ZAHVALNICA

Ova doktorska disertacija ne bi bila moguća bez nesebične podrške, ljubavi i verovanja onih koji su uvek bili uz mene. Prvo, želim da izrazim duboku zahvalnost svojoj porodici, koja je u svakom trenutku bila moj oslonac. Vaša podrška, razumevanje i strpljenje su me vodili kroz sve izazove na ovom putu.

Posebno zahvaljujem svom mentoru, prof. dr Dušanu Regodiću, čije stručno vođstvo, znanje i neiscrpna motivacija nisu samo oblikovali ovaj rad, već i moj naučni put. Hvala Vam što ste verovali u mene i usmeravali me ka ostvarenju ciljeva.

Izuzetnu zahvalnost dugujem svom kumu, prof. dr Nikoli Bačeviću, koji je bio jedan od najvećih pokretača na ovom putu i koji mi je pružio značajne instrukcije i podršku. Njegovi saveti i vođstvo, kao i podsticaj mojoj energiji, bili su od neprocenjivog značaja za uspeh ovog rada.

Takođe, iskreno zahvaljujem članovima komisije, čije su konstruktivne sugestije i stručno mišljenje dodatno unapredili ovu disertaciju. Vaša posvećenost i pažnja doprineli su kvalitetu ovog rada.

Na kraju, zahvaljujem se svim prijateljima, kolegama i svima koji su me podržavali i verovali u mene tokom ovog procesa. Vaša podrška mi je bila neprocenjiva i dala mi je snagu da istrajem.

Hvala svima od srca!!!  
Dejan Dodić

## REZIME

Ova disertacija se bavi ključnim izazovom u oblasti veštačke inteligencije (AI) i mašinskog učenja (ML), fokusirajući se na primenu transfernog učenja za adaptaciju GPT-2 modela u višejezičkim kontekstima. Značaj ovog istraživanja leži u potrebi za efikasnijim modelima prirodnog jezičkog procesiranja (NLP) koji mogu obraditi i generisati tekst na različitim jezicima uz visoku preciznost. Dosadašnji pristupi često nisu uspevali da zadovolje ove zahteve zbog ograničenih resursa i nedovoljno razvijenih tehnika optimizacije.

Ključni doprinosi ovog istraživanja uključuju razvoj i implementaciju dinamične adaptivne metodologije za optimizaciju hiperparametara (DYNAMO) koja značajno poboljšava performanse GPT-2 modela u višejezičkim NLP zadacima. Disertacija takođe uvodi nove strategije za optimizaciju tokenizacije i upravljanje memorijom, što omogućava efikasniju obradu velikih tekstualnih korpusa na ograničenim hardverskim platformama. Rezultati pokazuju značajno poboljšanje tačnosti, brzine obrade i energetske efikasnosti modela, što potvrđuje uspešnost primenjenih tehnika.

Implikacije ovog istraživanja su dalekosežne, jer otvaraju nove mogućnosti za primenu GPT-2 modela u globalnim, višejezičkim aplikacijama. Poboljšanja u optimizaciji i upravljanju resursima omogućavaju širu upotrebu ovih modela u realnim scenarijima, od automatizovanog prevođenja do analize tekstualnih podataka. Ova disertacija takođe postavlja temelje za dalja istraživanja u oblasti optimizacije hiperparametara i efikasnog upravljanja resursima u NLP modelima, sa ciljem dodatnog unapređenja njihove performanse i skalabilnosti.

Sve implementirane funkcionalnosti rešenja su pažljivo ispitane i potvrđene kroz sveobuhvatna testiranja, čime je verifikovana njihova efikasnost i pouzdanost.

Ova disertacija je organizovana kroz deset poglavlja, od kojih svako doprinosi detaljnom razumevanju i primeni istraživačkih ciljeva:

*Prvo poglavlje* se bavi formulacijom problema istraživanja i njegovim naučnim i društvenim značajem. Ovim poglavljem se postavlja osnova za razumevanje ključnih izazova sa kojima se suočava oblast obrade prirodnog jezika, posebno u kontekstu višejezičkih aplikacija. Takođe, detaljno su analizirani rezultati prethodnih istraživanja kako bi se identifikovale praznine u postojećem znanju, čime je naglašena potreba za daljim istraživanjima u ovoj oblasti.

*Drugo poglavlje* pruža sveobuhvatan uvod u veštačku inteligenciju i mašinsko učenje, sa posebnim osvrtom na njihovu primenu u analizi sentimenta. Ovim poglavljem se postavlja teorijski okvir za razumevanje metodologija koje su korišćene u razvoju naprednih modela mašinskog učenja, kao i tehnike treniranja modela koje omogućavaju optimizaciju performansi u kontekstu obrade prirodnog jezika. Ujedno, analizirana je uloga mašinskog učenja u razvoju NLP tehnologija, čime se dodatno naglašava značaj ovog istraživanja.

*Treće poglavlje* se odnosi na primenu GPT-2 modela sa adaptiranim transfernim učenjem na različitim hardverskim platformama. Detaljno je analiziran uticaj GPU arhitektura na performanse modela, a posebna pažnja posvećena je optimizaciji procesa obrade prirodnog jezika kroz prilagođavanje hardverskih specifikacija. Ova analiza pruža osnovu za razumevanje kako različiti hardverski resursi mogu uticati na efikasnost modela u višejezičkim aplikacijama.

*Četvrto poglavlje* istražuje modifikacije u arhitekturi PyTorch i Transformers biblioteka, uključujući optimizaciju scheduler-a i upravljanje memorijom. Prikazani su različiti pristupi koji omogućavaju efikasnije treniranje i prilagođavanje GPT-2 modela, uz smanjenje vremena obrade i poboljšanje stabilnosti modela. Ovo poglavlje naglašava sinergiju teorije i praktične primene kroz empirijske evaluacije optimizacija.

*Peto poglavlje* se fokusira na tokenizaciju i optimizaciju memorije za efikasniju obradu prirodnog jezika na GPU-u. Razmatrane su nove strategije za smanjenje opterećenja GPU-a kroz inovativne metode tokenizacije i upravljanje memorijom. Ovo poglavlje ističe značaj pravilnog upravljanja resursima u cilju maksimizacije performansi u NLP modelima dubokog učenja.

*Šesto poglavlje* obrađuje optimizaciju hiperparametara i prilagođavanje procesa tokenizacije. Prikazan je DYNAMO okvir kao inovativna metodologija za hiperparametarsku optimizaciju, koja omogućava značajna poboljšanja performansi u dubokom učenju. Poglavlje se bavi teoretskim i praktičnim aspektima implementacije DYNAMO pristupa i njegovom uticaju na efikasnost modela.

*Sedmo poglavlje* se bavi adaptivnom tokenizacijom i upravljanjem memorijom za optimalnu obradu velikih jezičkih podataka. Razvijene su metode koje omogućavaju prilagođavanje tokenizacije specifičnostima srpskog jezika, čime se poboljšava efikasnost GPT-2 modela. U ovom poglavlju su prikazani rezultati optimizacija i njihova primena u realnim jezičkim modelima.

*Osmo poglavlje* postavlja temelje za dalju optimizaciju GPT-2 modela kroz analizu kvaliteta i raznovrsnosti skupova podataka u višejezičkim aplikacijama. Kroz ovo poglavlje se istražuje kako različiti nivoi kvaliteta i raznovrsnosti podataka utiču na performanse generativnih modela, što je ključan korak pre nego što se model može dalje optimizovati. Naglasak je na razumevanju kompleksnosti srpskog jezika i na pripremi podataka koji će poslužiti kao osnova za dalju optimizaciju i primenu naprednih tehnika transfernog učenja u narednom poglavlju.

*Deveto poglavlje* se fokusira na primenu naprednih tehnika transfernog učenja kako bi se GPT-2 model dodatno prilagodio specifičnim potrebama srpskog jezika. Korišćenjem DYNAMO okvira za optimizaciju hiperparametara, Serbian-GPT-2 model je fino podešen kako bi postigao optimalne performanse. Rezultati iz devetog poglavlja prikazuju kako optimizovani model pokazuje poboljšane performanse u zadacima generisanja teksta na srpskom jeziku, čineći ga spremnim za dalju primenu u realnim višejezičkim aplikacijama.

*Deseto poglavlje* pruža rezime sprovedenih istraživanja, uključujući diskusiju rezultata, ključne nalaze i preporuke za dalja istraživanja. Poglavlje obuhvata detaljnu analizu postignutih rezultata i njihovih implikacija na dalji razvoj višejezičkih NLP modela, kao i predloge za buduće pravce istraživanja.

**Ključne reči:** transferno učenje, GPT-2, višejezičke aplikacije, optimizacija hiperparametara, prirodno jezičko procesiranje, upravljanje memorijom, dinamična adaptivna metodologija, skalabilnost modela

## ABSTRACT

This dissertation addresses a critical challenge in the field of Artificial Intelligence (AI) and Machine Learning (ML), focusing on the application of transfer learning for the adaptation of the GPT-2 model in multilingual contexts. The significance of this research lies in the need for more efficient Natural Language Processing (NLP) models capable of processing and generating text in various languages with high precision. Previous approaches often failed to meet these demands due to limited resources and insufficiently developed optimization techniques.

The key contributions of this research include the development and implementation of a Dynamic Adaptive Methodology for Hyperparameter Optimization (DYNAMO), which significantly improves the performance of the GPT-2 model in multilingual NLP tasks. The dissertation also introduces new strategies for tokenization optimization and memory management, enabling more efficient processing of large textual corpora on limited hardware platforms. The results demonstrate a significant improvement in accuracy, processing speed, and energy efficiency of the model, confirming the success of the applied techniques.

The implications of this research are far-reaching, opening new possibilities for the application of the GPT-2 model in global, multilingual applications. Improvements in optimization and resource management allow for broader use of these models in real-world scenarios, ranging from automated translation to textual data analysis. This dissertation also lays the foundation for further research in the area of hyperparameter optimization and efficient resource management in NLP models, with the aim of further enhancing their performance and scalability.

All implemented functionalities of the solution have been thoroughly examined and validated through comprehensive testing, confirming their efficiency and reliability.

This dissertation is organized into ten chapters, each contributing to a detailed understanding and application of the research objectives:

*The first chapter* addresses the formulation of the research problem and its scientific and societal significance. This chapter lays the foundation for understanding the key challenges faced in the field of natural language processing, particularly in the context of multilingual applications. Additionally, the results of previous research are analyzed in detail to identify gaps in existing knowledge, highlighting the need for further investigation in this area.

*The second chapter* provides a comprehensive introduction to artificial intelligence and machine learning, with a particular focus on their application in sentiment analysis. This chapter establishes the theoretical framework for understanding the methodologies used in the development of advanced machine learning models, as well as the training techniques that enable performance optimization in the context of natural language processing. Furthermore, the role of machine learning in the advancement of NLP technologies is examined, emphasizing the importance of this research.

*The third chapter* focuses on the application of the GPT-2 model with adapted transfer learning across different hardware platforms. The impact of GPU architectures on model performance is analyzed in detail, with special attention given to optimizing the natural language processing workflow through the adaptation of hardware specifications. This analysis provides a basis for understanding how different hardware resources can influence the efficiency of the model in multilingual applications.

*The fourth chapter* explores modifications in the architecture of PyTorch and Transformers libraries, including the optimization of schedulers and memory management. Various approaches are presented that enable more efficient training and adaptation of the GPT-2 model, reducing processing time and improving model stability. This chapter emphasizes the synergy between theory and practical application through empirical evaluations of the optimizations.

*The fifth chapter* focuses on tokenization and memory optimization for more efficient natural language processing on GPUs. New strategies for reducing GPU load through innovative tokenization methods and memory management are discussed. This chapter underscores the importance of proper resource management to maximize performance in deep learning NLP models.

*The sixth chapter* deals with hyperparameter optimization and the adaptation of the tokenization process. The DYNAMO framework is presented as an innovative methodology for hyperparameter optimization, allowing significant performance improvements in deep learning. The chapter addresses both theoretical and practical aspects of implementing the DYNAMO approach and its impact on model efficiency.

*The seventh chapter* addresses adaptive tokenization and memory management for optimal processing of large language datasets. Methods are developed that enable tokenization adjustments specific to the Serbian language, thereby improving the efficiency of the GPT-2 model. This chapter presents the results of these optimizations and their application in real language models.

*The eighth chapter* lays the foundation for further optimization of the GPT-2 model by analyzing the quality and diversity of datasets in multilingual applications. This chapter explores how different levels of data quality and diversity impact the performance of generative models, which is a crucial step before the model can be further optimized. The focus is on understanding the complexity of the Serbian language and on preparing the data that will serve as the basis for further optimization and the application of advanced transfer learning techniques in the subsequent chapter.

*The ninth chapter* focuses on applying advanced transfer learning techniques to further adapt the GPT-2 model to the specific needs of the Serbian language. By utilizing the DYNAMO framework for hyperparameter optimization, the Serbian-GPT-2 model is fine-tuned to achieve optimal performance. The results from the ninth chapter demonstrate how the optimized model exhibits improved performance in text generation tasks in Serbian, making it ready for further application in real-world multilingual scenarios.

*The tenth chapter* provides a summary of the conducted research, including a discussion of the results, key findings, and recommendations for further research. This chapter offers a detailed analysis of the achieved results and their implications for the further development of multilingual NLP models, as well as suggestions for future research directions.

**Keywords:** transfer learning, GPT-2, multilingual applications, hyperparameter optimization, natural language processing, memory management, dynamic adaptive methodology, model scalability



# SADRŽAJ

REZIME.....	I
ABSTRACT .....	III
UVOD .....	19
I METODOLOŠKI OKVIR ISTRAŽIVANJA .....	21
1. FORMULACIJA PROBLEMA ISTRAŽIVANJA .....	22
1.1. OSNOVNA ISTRAŽIVAČKA IDEJA U DOKTORSKOJ DISERTACIJI.....	22
1.2. ZNAČAJ PROBLEMA ISTRAŽIVANJA.....	23
1.2.1. NAUČNI ZNAČAJ.....	23
1.2.2. DRUŠTVENI (PRAKTIČNI) ZNAČAJ .....	23
1.3 REZULTATI PRETHODNIH SAZNANJA O PROBLEMU ISTRAŽIVANJA.....	24
2. ODREĐIVANJE PREDMETA ISTRAŽIVANJA .....	28
2.1. TEORIJSKO ODREĐENJE PREDMETA ISTRAŽIVANJA .....	28
2.2. POJMOVNO-KATEGORIJALNI SISTEM.....	30
2.3. OPERACIONALNO ODREĐENJE ISTRAŽIVANJA .....	31
2.3.1. STRUKTURALNI ČINIOCI PREDMETA ISTRAŽIVANJA KOJI ĆE BITI NEPOSREDNO ISTRAŽENI.....	32
2.3.2. VREMENSKO, PROSTORNO I DISCIPLINARNO ODREĐENJE PREDMETA ISTRAŽIVANJA.....	33
3. CILJEVI ISTRAŽIVANJA .....	35
3.1. NAUČNI CILJEVI ISTRAŽIVANJA .....	35
3.2. DRUŠTVENI CILJ ISTRAŽIVANJA .....	35
4. HIPOTETIČKI OKVIR ISTRAŽIVANJA.....	36
4.1. GENERALNA (OPŠTA) HIPOTEZA .....	36
4.2. POSEBNE HIPOTEZE ISTRAŽIVANJA .....	36
5. NAČIN (METODE) ISTRAŽIVANJA.....	40
5.1. OSNOVNE METODE SAZNANJA I ISTRAŽIVANJA.....	40
5.2. OPŠTENAUČNE METODE ISTRAŽIVANJA .....	40
5.3. METODE ZA PRIKUPLJANJE PODATAKA.....	41
5.4. UZORAK DOKUMENATA KOJI ĆE BITI PODVRGNUT ANALIZI.....	41
6. DRUŠTVENA I NAUČNA OPRAVDANOST ISTRAŽIVANJA.....	43
6.1. NAUČNI DOPRINOS.....	43
6.2. DRUŠTVENI DOPRINOS .....	43
II DOPRINOS MAŠINSKOG UČENJA U SAVREMENIM AI SISTEMIMA .....	45
7. OSNOVE I ZNAČAJ MAŠINSKOG UČENJA .....	46
7.1. MODELI MAŠINSKOG UČENJA U ANALIZI SENTIMENTA .....	47
7.1.1. Metodologija i tehnike u analizi sentimenta.....	47
7.1.2. Značaj i uticaj istraživanja u NLP tehnologijama .....	47
7.2. STRATEGIJE I METODE MAŠINSKOG UČENJA ZA ANALIZU SENTIMENTA.....	48
7.2.1. Prikupljanje i predobrada podataka .....	49

7.2.2. Tehnike treniranja modela.....	49
7.2.3. Pristup mašinskom učenju.....	53
<b>7.3. ANALITIČKI UVIDI I DOMETI MODELA MAŠINSKOG UČENJA .....</b>	<b>53</b>
7.3.1. Balans između preciznosti, odziva i F1-scora .....	53
7.3.2. Vizualna ilustracija ključnih metrika performansi.....	53
7.3.3. Analiza podrške i distribucije podataka.....	54
7.3.4. Robusnost modela i njegova primena.....	55
<b>7.4. RAZMATRANJE EFIKASNOSTI I ETIČKIH IMPLIKACIJA MODELA .....</b>	<b>58</b>
7.4.1. Uticaj metodoloških izbora na performanse.....	58
7.4.2. Etičke dimenzije i transparentnost u primeni modela.....	59
7.4.3. Mogućnosti unapređenja i budući pravci istraživanja.....	60
7.4.4. Etički i društveni izazovi u primeni AI tehnologija.....	60
<b>III PRIMENA GPT-2 MODELA SA ADAPTIRANIM TRANSFERNIM UČENJEM NA RAZLIČITIM HARDVERSKIM PLATFORMAMA .....</b>	<b>61</b>
<b>8. OPTIMIZACIJA GPT-2 MODELA PUTEM ADAPTIVNOG TRANSFERNOG UČENJA NA GPU ARHITEKTURAMA .....</b>	<b>62</b>
<b>8.1. KVANTIFIKACIJA UTICAJA GPU PERFORMANSI NA GPT-2 MODEL U ADAPTIVNOM TRANSFERNOM UČENJU.....</b>	<b>63</b>
8.1.1. Eksperimentalna analiza efikasnosti GPU .....	63
8.1.2. Primena i održivost modela u praksi.....	64
<b>8.2. METODOLOŠKI PRISTUP U EVALUACIJI EFIKASNOSTI GPT-2 MODELA NA GPU PLATFORMAMA</b>	<b>65</b>
8.2.2. Implementacija i analiza metrika performansi.....	65
8.2.3. Primena alata za optimizaciju i vizualizaciju .....	65
8.3.1. Uticaj GPU specifikacija na efikasnost modela.....	68
8.3.2. Optimizacija softvera i njena uloga u poboljšanju performansi .....	68
8.3.3. Strategije za maksimizaciju efikasnosti GPT-2 modela.....	69
<b>8.4. ANALIZA I IMPLIKACIJE OPTIMIZACIJA GPT-2 MODELA NA GPU ARHITEKTURAMA.....</b>	<b>71</b>
<b>IV MODIFIKACIJE U ARHITEKTURI PYTORCH I TRANSFORMERS BIBLIOTEKA, UKLJUČUJUĆI OPTIMIZACIJU SCHEDULER-A I MEMORIJE.....</b>	<b>72</b>
<b>9. OPTIMIZACIJA SCHEDULER-A I UPRAVLJANJE MEMORIJOM U PYTORCH I TRANSFORMERS BIBLIOTEKAMA: PRISTUPI I MODIFIKACIJE .....</b>	<b>73</b>
<b>9.1. PYTORCH I TRANSFORMERS KAO STUDIJA SLUČAJA .....</b>	<b>74</b>
<b>9.2. METODOLOŠKI PRISTUPI ZA OPTIMIZACIJU SCHEDULER-A I UPRAVLJANJE MEMORIJOM U PYTORCH I TRANSFORMERS BIBLIOTEKAMA .....</b>	<b>76</b>
9.2.1. Eksperimentalna konfiguracija i izbor hardvera.....	76
9.2.2. Sinergija teorije i praktične primene.....	77
<b>9.3. EMPIRIJSKA EVALUACIJA OPTIMIZACIJA SCHEDULER-A I UPRAVLJANJA MEMORIJOM U PYTORCH I TRANSFORMERS BIBLIOTEKAMA .....</b>	<b>80</b>
9.3.1. Analiza performansi i efikasnosti algoritama.....	80
9.3.2. Implikacije za industrijsku i akademsku primenu.....	83
<b>9.4. ZNAČAJ OPTIMIZACIJA SCHEDULER-A I UPRAVLJANJA MEMORIJOM U PYTORCH I TRANSFORMERS BIBLIOTEKAMA .....</b>	<b>84</b>
<b>V TOKENIZACIJA I OPTIMIZACIJA MEMORIJE ZA EFIKASNIJU OBRADU PRIRODNOG JEZIKA NA GPU-U .....</b>	<b>85</b>
<b>10. TOKENIZACIJA I OPTIMIZACIJA MEMORIJE ZA SMANJENJE OPTEREĆENJA GPU-A U NLP MODELIMA DUBOKOG UČENJA .....</b>	<b>86</b>

<b>10.1.</b>	<b>NOVI HORIZONTI U UPRAVLJANJU MEMORIJOM.....</b>	<b>87</b>
10.1.1.	<i>Razvoj strategija za optimizaciju memorije.....</i>	87
10.1.2.	<i>Inovativne metode tokenizacije i njihov uticaj na performanse.....</i>	87
<b>10.2.</b>	<b>INTEGRACIJA MAŠINSKOG I DUBOKOG UČENJA.....</b>	<b>89</b>
10.2.1.	<i>Integracija tehnika mašinskog učenja za predviđanje potreba za memorijom.....</i>	89
10.2.2.	<i>Komparativna analiza metoda upravljanja memorijom.....</i>	90
10.2.3.	<i>Praktična implementacija i optimizacija modela.....</i>	92
<b>10.3.</b>	<b>PREVAZILAŽENJE IZAZOVA MEMORIJE ZA NAPREDNE NLP MODELE.....</b>	<b>92</b>
10.3.1.	<i>Efikasnost selektivne validacije.....</i>	92
10.3.2.	<i>Ograničenja i izazovi selektivnog uzorkovanja.....</i>	93
10.3.3.	<i>Implementacija i eksperimentalni rezultati.....</i>	96
<b>10.4.</b>	<b>STRATEGIJE OPTIMIZACIJE MEMORIJE ZA GPU U NLP.....</b>	<b>97</b>
<b>VI</b>	<b>OPTIMIZACIJA HIPERPARAMETARA I PRILAGOĐAVANJE PROCESA TOKENIZACIJE.....</b>	<b>98</b>
<b>11.</b>	<b>SINERGIJA INOVATIVNIH TEHNIKA ZA HIPERPARAMETARSKU OPTIMIZACIJU KOJA TRANSFORMIŠE GRANICE PERFORMANSI U DUBOKOM UČENJU.....</b>	<b>99</b>
<b>11.1.</b>	<b>EVOLUCIJA DUBOKOG UČENJA I OPTIMIZACIJE HIPERPARAMETARA.....</b>	<b>99</b>
11.1.1.	<i>Napredak u tehnikama optimizacije.....</i>	99
11.1.2.	<i>Uvod u DYNAMO okvir i njegove implikacije na NLP.....</i>	100
11.1.3.	<i>Rešavanje NLP izazova uz DYNAMO.....</i>	100
<b>11.2.</b>	<b>NAPREDNA OPTIMIZACIJA HIPERPARAMETARA: DYNAMO PRISTUP.....</b>	<b>101</b>
11.2.1.	<i>Istraživanje naprednih tehnika optimizacije.....</i>	101
11.2.2.	<i>DYNAMO metodološki napredak.....</i>	102
11.2.3.	<i>Studije slučaja i praktične primene.....</i>	102
<b>11.3.</b>	<b>METODOLOŠKI UVIDI U DYNAMO: TEORIJA I PRAKSA BALANSIRANJA.....</b>	<b>102</b>
11.3.1.	<i>Teorijske osnove i algoritamski dizajn.....</i>	102
11.3.2.	<i>Integracija transfernog učenja i adaptivnih strategija.....</i>	103
11.3.3.	<i>Empirijska validacija i analiza učinka.....</i>	103
11.3.4.	<i>Praktične primene i studije slučaja.....</i>	106
<b>11.4.</b>	<b>UTICAJ DYNAMO: PROCENA PERFORMANSE I EFIKASNOSTI.....</b>	<b>107</b>
11.4.1.	<i>Komparativna analiza i poboljšanje modela.....</i>	107
11.4.2.	<i>Rešavanje izazova i ograničenja.....</i>	107
11.4.3.	<i>Metodološki uvidi i budući pravci.....</i>	108
<b>11.5.</b>	<b>KOMPARATIVNA ANALIZA: DYNAMO OKVIR U ODNOSU NA TRADICIONALNE METODE OPTIMIZACIJE HIPERPARAMETARA.....</b>	<b>114</b>
<b>11.6.</b>	<b>POTVRDA EFIKASNOSTI DYNAMO OKVIRA ZA OPTIMIZACIJU HIPERPARAMETARA.....</b>	<b>115</b>
<b>VII</b>	<b>ADAPTIVNA TOKENIZACIJA I UPRAVLJANJE MEMORIJOM ZA OPTIMALNU OBRADU VELIKIH JEZIČKIH PODATAKA.....</b>	<b>117</b>
<b>12.</b>	<b>UNAPREĐENJE TOKENIZACIJE I UPRAVLJANJE MEMORIJOM U RAZVOJU JEZIČKIH MODELA.....</b>	<b>118</b>
<b>12.1.</b>	<b>UNAPREĐENJE TOKENIZACIJE I UPRAVLJANJE MEMORIJOM U OBRADI TEKSTUALNIH PODATAKA NA SRPSKOM JEZIKU.....</b>	<b>119</b>
12.1.1.	<i>Prilagodavanje tokenizacije za srpski jezik.....</i>	119
12.1.2.	<i>Optimizacija upravljanja memorijom tokom treniranja modela.....</i>	120
<b>12.2.</b>	<b>OPTIMIZACIJA JEZIČKOG MODELA GPT-2.....</b>	<b>121</b>
12.2.1.	<i>Podaci i priprema dataset-a.....</i>	121
12.2.2.	<i>Optimizacija treniranja modela i upravljanje memorijom.....</i>	121

12.3.	REZULTATI OPTIMIZACIJE TOKENIZACIJE I UPRAVLJANJA MEMORIJOM U OBRADI TEKSTUALNIH PODATAKA.....	124
12.3.1.	<i>Poboljšanje performansi modela kroz optimizaciju tokenizacije .....</i>	125
12.3.2.	<i>Efikasnost upravljanja memorijom tokom treniranja modela .....</i>	126
12.3.3.	<i>Analiza uticaja optimizacija na sveukupnu efikasnost modela .....</i>	128
12.4.	UNAPREĐENJE JEZIČKIH MODELA KROZ OPTIMIZACIJU TOKENIZACIJE I UPRAVLJANJA MEMORIJOM .....	129
VIII	KVALITET I RAZNOVRSNOST SKUPOVA PODATAKA U VIŠEJEZIČKIM APLIKACIJAMA.....	130
13.	UTICAJ KVALITETA I RAZNOVRSNOSTI PODATAKA NA PERFORMANSE GENERATIVNIH MODELA U VIŠEJEZIČKIM APLIKACIJAMA.....	131
13.1.	OPTIMIZACIJA GENERATIVNIH MODELA KROZ ANALIZU KVALITETA I RAZNOVRSNOSTI PODATAKA .....	132
13.1.1.	<i>Uticaj raznovrsnosti i kvaliteta podataka na generativne modele .....</i>	132
13.1.2.	<i>Optimizacija generativnih modela za višejezičke aplikacije .....</i>	135
13.2.	METODOLOŠKI PRISTUPI U OPTIMIZACIJI GENERATIVNIH MODELA ZA VIŠEJEZIČKE APLIKACIJE .....	135
13.2.1.	<i>Prikupljanje i priprema podataka .....</i>	135
13.2.2.	<i>Optimizacija hiperparametara i treniranje modela .....</i>	135
13.2.3.	<i>Evaluacija i vizualizacija performansi modela .....</i>	137
13.3.	REZULTATI OPTIMIZACIJE I EVALUACIJE SERBIAN-GPT-2 MODELA .....	138
13.3.1.	<i>Evaluacija modela kroz metričke parametre .....</i>	138
13.3.2.	<i>Statistička analiza i značaj optimizacije.....</i>	142
13.4.	KLJUČNI NALAZI I PREPORUKE ZA OPTIMIZACIJU PODATAKA I PERFORMANSE MODELA .....	144
IX	PRIMENA GPT-2 MODELA SA NAPREDNIM TEHNIKAMA TRANSFERNOG UČENJA U REALNIM VIŠEJEZIČKIM APLIKACIJAMA .....	145
14.	NAPREDNE TEHNIKE TRANSFERNOG UČENJA U GPT-2 MODELU ZA VIŠEJEZIČKE APLIKACIJE.....	146
14.1.	OPTIMIZACIJA HIPERPARAMETARA ZA GPT-2 MODEL.....	147
14.1.1.	<i>Evaluacija performansi i primena modela na srpskom jeziku .....</i>	147
14.1.2.	<i>Implementacija transfernog učenja za prilagodavanje na srpski jezik.....</i>	147
14.2.	OPTIMIZACIJA GPT-2 MODELA ZA VIŠEJEZIČKU PRIMENU.....	150
14.2.1.	<i>Priprema i obrada podataka .....</i>	150
14.2.2.	<i>Eksperimentalni dizajn i analiza rezultata .....</i>	150
14.3.	PIONIRSKA PRIMENA TRANSFERNOG UČENJA I HIPERPARAMETARSKOJE OPTIMIZACIJE NA SERBIAN GPT-2 MODEL.....	153
14.3.1.	<i>Duboka analiza i prikaz eksperimentalnih rezultata .....</i>	153
14.3.2.	<i>Primena i uticaj na realne aplikacije .....</i>	156
14.4.	UNAPREĐIVANJE VIŠEJEZIČKIH NLP SISTEMA.....	157
X	REZIME SPROVEDENIH ISTRAŽIVANJA .....	159
15.1.	DISKUSIJA REZULTATA ISTRAŽIVANJA.....	160
15.2.	KLJUČNI REZULTATI ISTRAŽIVANJA.....	161
15.3.	PREDLOG DALJIH ISTRAŽIVANJA .....	162
16.	ZAKLJUČAK.....	164
16.1.	NAUČNI DOPRINOS DISERTACIJE.....	164
16.2.	PITANJA KOJA DISERTACIJA OTVARA, NASTAVAK I BUDUĆA ISTRAŽIVANJA.....	165

<b>16.3. DOKAZIVANJE (POTKREPLJIVOST) GLAVNE HIPOTEZE.....</b>	<b>166</b>
<b>STRUKTURA DOKTORSKE DISERTACIJE .....</b>	<b>168</b>
<b>LITERATURA .....</b>	<b>173</b>
<b>PRILOZI .....</b>	<b>181</b>

## SADRŽAJ SLIKA

<b>Slika 1.</b> Ključne tehnologije iz oblasti veštačke inteligencije i obrade prirodnog jezika.....	29
<b>Slika 2.</b> Ključni aspekti procesa mašinskog učenja, od pripreme podataka do treniranja i upravljanja modelima .....	33
<b>Slika 3.</b> Algoritam toka procesa mašinskog učenja za analizu sentimenta tekstova na srpskom jeziku .....	48
<b>Slika 4.</b> Procesno orijentisan prikaz metoda mašinskog učenja u analizi sentimenta .....	50
<b>Slika 5.</b> Komparativna efikasnost različitih pristupa mašinskom učenju.....	51
<b>Slika 6.</b> Bar grafikon metrika klasifikacije po kategorijama sentimenta .....	54
<b>Slika 7.</b> Disperzioni grafikon podrške i metrika performansi modela .....	55
<b>Slika 8.</b> Heatmap metrika modela za klasifikaciju sentimenta.....	56
<b>Slika 9.</b> Upporedna analiza metrika preciznosti, odziva i F1-skorova za klasifikaciju sentimenta .....	59
<b>Slika 10.</b> Tok metodološkog procesa korištenog u istraživanju efektivnosti GPT-2 modela .....	66
<b>Slika 11.</b> Poređenje GPU efikasnosti za GPT-2 model.....	67
<b>Slika 12.</b> Komparativna procena metrike performansi 8GB prema 16GB GPU-a .....	69
<b>Slika 13.</b> Sekvencijalni prikaz procesa istraživanja i optimizacije .....	77
<b>Slika 14.</b> Poređenje vremena izvršavanja na različitim GPU-ovima pre i posle optimizacije .....	78
<b>Slika 15.</b> Komparativna analiza vremena izvršenja zadataka pre i posle optimizacije .....	81
<b>Slika 16.</b> Matematički prikaz optimizacije scheduler-a i upravljanja memorijom u PyTorch i Transformers bibliotekama .....	82
<b>Slika 17.</b> Šematski prikaz toka treninga NLP modela.....	87
<b>Slika 18.</b> Odnos između ograničenja GPU memorije i veličine uzorka za validaciju u NLP modelima .....	88
<b>Slika 19.</b> Dijagram toka procesa dinamičke alokacije memorije vođene mašinskim učenjem u NLP-u .....	89
<b>Slika 20.</b> Poređenje metoda.....	90
<b>Slika 21.</b> Zahtevi za memorijom na osnovu prosečne veličine tokena i ukupnog broja tokena u seriji.....	91
<b>Slika 22.</b> Grafikon alokacije GPU memorije .....	94
<b>Slika 23.</b> Grafikon potrošnje energije GPU-a .....	94
<b>Slika 24.</b> GPU vreme provedeno u pristupu memoriji .....	95
<b>Slika 25.</b> Optimizacija maksimalne dužine za tokenizaciju s obzirom na gubitak modela i korišćenje memorije .....	95
<b>Slika 26.</b> Izlaz konzole koji prikazuje validaciju i uspeh nasumično odabranih uzoraka validacije na 100 indeksa .....	96
<b>Slika 27.</b> Učitavanje modela, tokenizera, dodavanje određenih tokena u tokenizer pomoću GPT-2 modela.....	96
<b>Slika 28.</b> Tok procesa "Dynamic Adaptive Methodology for Hyperparameter Optimization (DYNAMO)" .....	103
<b>Slika 29.</b> Praćenje smanjenja gubitka u procesu obuke DYNAMO modela.....	105
<b>Slika 30.</b> Vizualizacija arhitekture dinamičke neuronske mreže DYNAMO okvira .....	106
<b>Slika 31.</b> Komparativna analiza smanjenja gubitaka DYNAMO-a u odnosu na TRADICIONALNI okvir .....	109
<b>Slika 32.</b> Gubitak modela tokom treninga: Komparativna analiza DYNAMO i TRADICIONALNOG pristupa .....	110
<b>Slika 33.</b> Trendovi u gubitku treninga: DYNAMO stabilno smanjenje tokom iteracija.....	110

<b>Slika 34.</b> Smanjenje gubitka pri evaluaciji treninga: Demonstracija DYNAMO poboljšane tačnosti.....	111
<b>Slika 35.</b> Fluktuacije složenosti modela tokom treninga: Procena prediktivnog učinka i perpleksije .....	111
<b>Slika 36.</b> Dinamičko prilagođavanje stopa učenja u DYNAMO okviru tokom vremena .....	112
<b>Slika 37.</b> Napredak treninga: Poređenje iterativnih koraka između DYNAMO i TRADICIONALNOG okvira .....	112
<b>Slika 38.</b> Korišćenje GPU memorije tokom treninga: Efikasnost DYNAMO-a u odnosu na TRADICIONALNI okvir.....	113
<b>Slika 39.</b> Uporedni učinak DYNAMO-a u odnosu na TRADICIONALNE metode u različitim eksperimentima .....	114
<b>Slika 40.</b> Prosečno vreme obrade po tokenu pre i nakon optimizacije.....	123
<b>Slika 41.</b> Uporedni rezultati tačnosti modela pre i nakon optimizacije tokenizacije .....	125
<b>Slika 42.</b> Rezultati evaluacije modela pre i nakon optimizacije.....	126
<b>Slika 43.</b> Alokacija GPU memorije pre optimizacije.....	127
<b>Slika 44.</b> Alokacija GPU memorije nakon optimizacije .....	127
<b>Slika 45.</b> Ključni elementi istraživanja kvaliteta i raznovrsnosti skupova podataka u višejezičkim aplikacijama .....	132
<b>Slika 46.</b> Trend gubitka i tačnosti tokom 14 epoha treniranja GPT-2 modela na srpskom jeziku .....	134
<b>Slika 47.</b> Performanse modela tokom poslednjih pet epoha treniranja .....	138
<b>Slika 48.</b> Prikaz Loss i Perplexity kroz epohe.....	140
<b>Slika 49.</b> PCA Projekcija Token Embedding-a iz Serbian-GPT-2 Modela .....	141
<b>Slika 50.</b> GPU Memorija Alocirana Tokom Treninga .....	142
<b>Slika 51.</b> Rezultati t-testa za tačnost pre i nakon treninga .....	143
<b>Slika 52.</b> Rezultati ANOVA testa za različite grupe tačnosti .....	143
<b>Slika 53.</b> Struktura i ključni elementi napredne primene GPT-2 modela u višejezičkim aplikacijama .....	147
<b>Slika 54.</b> Performanse GPT-2 modela tokom treninga prikazane kroz vrednosti gubitaka .....	149
<b>Slika 55.</b> Vrednosti stope cikličkog učenja tokom treninga.....	152
<b>Slika 56.</b> Smanjenje gubitaka tokom obuke Serbian GPT-2 modela .....	153
<b>Slika 57.</b> Tačnost modela tokom evaluacije.....	154
<b>Slika 58.</b> Vrednost perpleksije tokom evaluacije.....	154
<b>Slika 59.</b> Kvalitet generisanja teksta Serbian-GPT-2 modela.....	156

## SADRŽAJ TABELA

<b>Tabela 1.</b> Kvantitativni pregled efikasnosti različitih metoda mašinskog učenja koje su primenjene u analizi sentimenta.....	51
<b>Tabela 2.</b> Detaljna analiza metrika performansi modela za klasifikaciju sentimenta.....	57
<b>Tabela 3.</b> Kvantitativna analiza metrika preciznosti, odziva i F1-skorova za klasifikaciju sentimenta .....	59
<b>Tabela 4.</b> Uporedne metrike performansi GPT-2 modela u različitim GPU konfiguracijama .....	64
<b>Tabela 5.</b> Uporedni prikaz parametara za izračunavanje efikasnosti GPT-2 modela na različitim GPU konfiguracijama.....	67
<b>Tabela 6.</b> Komparativna analiza metrika performansi GPU-a.....	69
<b>Tabela 7.</b> Komparativna analiza vremena izvršavanja CPU-a i GPU-a s izračunatom efikasnošću .....	75
<b>Tabela 8.</b> Poređenje performansi pre i posle optimizacije scheduler-a .....	79
<b>Tabela 9.</b> Komparativna analiza vremena izvršenja zadataka pre i posle optimizacije .....	83
<b>Tabela 10.</b> Veličina uzorka za proveru valjanosti kao funkcija ograničenja GPU memorije.....	88
<b>Tabela 11.</b> Procenjeni zahtevi za memorijom prema veličini serije i prosečnoj veličini tokena .....	92
<b>Tabela 12.</b> Procenjeni zahtevi za memorijom prema veličini serije i prosečnoj veličini tokena .....	96
<b>Tabela 13.</b> Ključne komponente DYNAMO procesa optimizacije .....	101
<b>Tabela 14.</b> Pregled smanjenja gubitaka primenom DYNAMO metodologije u 14 epoha obuke.....	105
<b>Tabela 15.</b> Pregled optimizacije hiperparametara kroz DYNAMO metodologiju .....	108
<b>Tabela 16.</b> Detaljno poređenje performansi između DYNAMO i TRADICIONALNIH metoda.....	115
<b>Tabela 17.</b> Specifikacije Dataset-a i Parametri Treniranja .....	120
<b>Tabela 18.</b> Performanse modela pre i nakon optimizacije.....	123
<b>Tabela 19.</b> Efikasnost upravljanja memorijom.....	125
<b>Tabela 20.</b> Statistička analiza .....	127
<b>Tabela 21.</b> Rezultati treniranja za jezik sa složenom morfološkom strukturom (srpski jezik).....	133
<b>Tabela 22.</b> Performanse modela tokom poslednjih pet epoha treniranja.....	136
<b>Tabela 23.</b> Performanse modela tokom poslednjih pet epoha .....	139
<b>Tabela 24.</b> Hiperparametri korišćeni u obuci GPT-2 modela sa transfer učenjem na srpskom skupu podataka.....	148
<b>Tabela 25.</b> Hiperparametri za cikličku stopu učenja.....	152
<b>Tabela 26.</b> Optimizacija hiperparametara - Performanse modela početak i kraj obuke .....	155



## NOMENKLATURA

- **Grid search** - Metoda za optimizaciju hiperparametara u mašinskom učenju.
- **Random search** - Metoda za optimizaciju hiperparametara u mašinskom učenju.
- **Transfer learning** - Proces prenosa znanja sa jednog modela na drugi.
- **Hyperparameter optimization** - Proces podešavanja hiperparametara modela mašinskog učenja.
- **Tokenization** - Proces razbijanja teksta na manje jedinice kao što su reči ili fraze.
- **Memory optimization** - Proces smanjenja upotrebe memorije prilikom izvršavanja algoritama.
- **Cyclic learning rate** - Tehnika gde se brzina učenja ciklično menja tokom treniranja modela.
- **Perplexity** - Metrika koja meri neizvesnost modela u predviđanju sledećeg elementa u nizu. Niža vrednost perpleksije ukazuje na bolju sposobnost modela da predvidi sledeći token ili reč u sekvenci.
- **NLP (Natural Language Processing)** - Obrada prirodnog jezika, oblast koja se bavi interakcijom između računara i ljudskog jezika.
- **Machine learning (ML)** - Oblast veštačke inteligencije koja se bavi razvojem algoritama koji omogućavaju računarima da uče iz podataka i unapređuju svoje performanse bez eksplicitnog programiranja.
- **Batch** - Skup podataka koji se koristi u jednoj iteraciji treniranja modela, često manji deo celokupnog skupa podataka.
- **Dropout** - Tehnika regularizacije koja nasumično isključuje određeni procenat neurona tokom treniranja modela kako bi se smanjila prekomerna prilagođavanja.
- **Deep learning** - Oblast mašinskog učenja koja koristi neuronske mreže sa mnogo slojeva za učenje iz podataka.
- **Transformer models** - Tip modela dubokog učenja koji koristi mehanizam pažnje za obradu sekvencijalnih podataka.
- **Convolutional Neural Networks (CNNs)** - Tip neuronske mreže dizajniran za obradu podataka u obliku mreže, poput slika.
- **GPU Memory Optimization** - Optimizacija korišćenja GPU memorije tokom treniranja modela mašinskog učenja.
- **CUDA out of memory** - Greška koja se javlja kada model pokuša da koristi više GPU memorije nego što je dostupno.
- **Dynamic memory allocation** - Tehnika koja omogućava prilagođavanje količine memorije tokom izvršavanja programa.
- **Validation set** - Skup podataka koji se koristi za ocenu performansi modela tokom treniranja.
- **Random sampling** - Tehnika odabira nasumičnih uzoraka iz skupa podataka za analizu ili treniranje modela.
- **Learning rate** - Parametar koji određuje brzinu prilagođavanja težina modela tokom treniranja.
- **Batch size** - Broj uzoraka koji se koriste za ažuriranje modela u jednoj iteraciji treniranja.
- **Attention mechanism** - Tehnika u dubokom učenju koja omogućava modelima da se fokusiraju na relevantne delove ulaznih podataka.
- **Optuna** - Biblioteka za automatsku optimizaciju hiperparametara.
- **Fine-tuning** - Proces dodatnog treniranja unapred naučenog modela na specifičnom skupu podataka kako bi se poboljšala njegova preciznost.
- **Overfitting** - Situacija u kojoj model postaje previše prilagođen trening podacima, gubeći sposobnost generalizacije na nove podatke.
- **Regularization** - Tehnike koje se koriste za smanjenje prekomernog prilagođavanja modela podacima.
- **DYNAMO framework** - Okvir za dinamičku optimizaciju hiperparametara.

- **Weights & Biases (Wandb)** - Alat za praćenje i vizualizaciju eksperimenata u mašinskom učenju.
- **BERT (Bidirectional Encoder Representations from Transformers)** - Model za obradu prirodnog jezika koji koristi dvosmerne transformere za razumevanje jezika.
- **GPT-2 (Generative Pre-trained Transformer 2)** - Model za generisanje teksta baziran na transformatorima, koji je unapred naučen na velikom skupu podataka.
- **Adaptive transfer learning** - Proces prilagođavanja modela novom skupu podataka koristeći prethodno naučeno znanje.
- **Natural language generation (NLG)** - Proces automatskog generisanja teksta od strane računara.
- **BPE (Byte Pair Encoding)** - Algoritam za tokenizaciju koji spaja najčešće parove bajtova kako bi smanjio veličinu rečnika i omogućio efikasniju obradu teksta.
- **Epoch** - Jedan prolaz kroz ceo skup podataka tokom treniranja modela mašinskog učenja.
- **per\_device\_train\_batch\_size** - Veličina batch-a po uređaju tokom treniranja modela.
- **gradient\_accumulation\_steps** - Broj koraka akumulacije gradijenta pre ažuriranja modela.
- **num\_train\_epochs** - Broj epoha za koje se model trenira.
- **base\_lr** - Osnovna brzina učenja koja se koristi na početku treniranja.
- **max\_lr** - Maksimalna brzina učenja do koje se dolazi tokom treniranja.
- **weight\_decay** - Stopa raspadanja težina koja se koristi za regularizaciju modela.
- **warmup\_steps** - Broj koraka tokom kojih se brzina učenja postepeno povećava na početku treniranja.
- **logging\_steps** - Interval koraka između kojih se zapisuju logovi tokom treniranja.
- **max\_length** - Maksimalna dužina sekvence koja se koristi tokom treniranja modela.
- **dropout** - Stopa dropout-a koja se koristi za regularizaciju tokom treniranja.
- **grad\_clip** - Vrednost maksimalne granice za obrezivanje gradijenata kako bi se izbegle eksplozivne vrednosti tokom treniranja.

## Uvod

Razvoj veštačke inteligencije (eng. AI, srp. VI) i obrade prirodnog jezika (Natural Language Processing, NLP) snažno su povezani sa različitim disciplinama, a njihova rana istorija predstavlja temelj za savremena istraživanja u ovoj oblasti [1]. Počeci veštačke inteligencije datiraju iz 1950-ih godina, kada su istraživači poput Alana Turinga i Claudea Shannona postavljali osnove računarskog razmišljanja i formalizovali koncepte poput Turingovog testa [2]. Ovi pioniri istraživali su ideju o tome da li mašina može pokazati inteligentno ponašanje slično ljudskom, što je postavilo temelje za budući razvoj veštačke inteligencije [3]. Tokom 1960-ih i 1970-ih godina, istraživači poput Johna McCarthyja i Marvinina Minskyja postavljali su temelje prve generacije veštačke inteligencije, istražujući logiku, simboličko zaključivanje i ekspertne sisteme [4].

Obrada prirodnog jezika (NLP) je takođe evoluirala tokom ovog perioda [5]. Rani radovi fokusirali su se na osnovne lingvističke principe i metode, kao što su statističko modeliranje jezika i morfosintaksička analiza [6]. Ključni trenuci u ranom razvoju NLP-a uključuju rad Noama Chomskog na generativnoj gramatici, kao i pokušaje razumevanja jezika putem računarske obrade teksta [7]. Ovi rani koraci postavili su temelje za kasniji napredak, uključujući primenu dubokog učenja u obradi prirodnog jezika [8]. Sa brzim napretkom tehnologije obrade prirodnog jezika, imperativ za stvaranjem sistema koji mogu efikasno obrađivati više jezika postaje sve značajniji [9]. Kao odgovor na ovaj zahtev, istraživačka zajednica usmerila je pažnju na razvoj modela dubokog učenja koji nisu ograničeni na jedan jezik, već su sposobni za prepoznavanje i generisanje sadržaja na različitim jezicima [10]. U ovom kontekstu, PyTorch i Transformers arhitektura postali su ključni alati, obećavajući širok spektar mogućnosti za višejezičko prepoznavanje, generisanje i prevođenje teksta.

Razumevanje rane istorije veštačke inteligencije i obrade prirodnog jezika ne pruža samo kontekst za aktuelna istraživanja, već naglašava i kontinuiranu potrebu za inovacijama [1]. Ovi ključni trenuci u evoluciji tehnoloških disciplina omogućavaju sagledavanje daljih perspektiva, pokazujući koliko je napredak bio snažno usmeren ka stvaranju inteligentnih sistema koji mogu efikasno komunicirati s ljudima [2]. Unutar ovog konteksta, ova doktorska disertacija postavlja osnove za istraživanje i inovacije, težeći da spoji nasleđe ranijih dostignuća sa budućim pravcima razvoja [3].

Ova doktorska disertacija istražuje temu transfernog učenja i adaptacije u kontekstu PyTorch i Transformers arhitekture, postavljajući se kao most između teorije i praktične primene [4]. Fokusirajući se na višejezičko razumevanje i generisanje teksta, ova studija razmatra kako ovi alati omogućavaju modelima da prenesu znanje sa jednog jezika na drugi, smanjujući potrebu za velikim skupovima podataka za svaki pojedinačni jezik [5].

Transferno učenje u domenu obrade prirodnog jezika postaje ključno zbog ograničenih resursa za mnoge jezike [6]. Korišćenjem prethodno naučenih modela i njihovim prilagođavanjem specifičnim jezičkim zadacima, ovaj pristup obezbeđuje skalabilnost i poboljšava performanse modela u raznolikim jezičkim kontekstima [7]. Ova disertacija istražuje metode, tehnike i strategije unutar transfernog učenja i

prilagođavanja u okviru PyTorch i Transformers okruženja, pružajući duboki uvid u mehanizme koji omogućavaju efikasnu višejezičku obradu teksta [8].

U narednim delovima disertacije analiziraćemo ključne koncepte transfernog učenja i adaptacije unutar PyTorch i Transformers okruženja. Takođe ćemo istražiti različite pristupe i tehnike koje omogućavaju efikasnu implementaciju višejezičkih modela, razmatrajući prednosti, izazove i perspektive ovog pristupa ka stvaranju sistema koji su sposobni da obrađuju tekst na različitim jezicima [10]. Ovaj rad predstavlja značajan doprinos oblasti NLP-a, pružajući osnovu za razumevanje složenih mehanizama transfernog učenja i adaptacije unutar PyTorch i Transformers paradigme. Kroz detaljnu analizu, eksperimentalna istraživanja i evaluaciju različitih modela, ova disertacija ima za cilj da unapredi višejezičke NLP sisteme i olakša razvoj jezički agilnih tehnologija koje će imati značajan uticaj na globalnu komunikaciju i razmenu znanja [2].

U poslednjih nekoliko godina, Transformer arhitektura postala je ključna za napredak u obradi prirodnog jezika. Njena revolucionarna struktura, koja se fokusira na mehanizme međusobnog samopouzdanja (self-attention), omogućila je razvoj modela kao što je GPT (Generative Pre-trained Transformer) serija [4]. Ovi modeli pokazali su izuzetne sposobnosti u generisanju teksta, prevođenju i razumevanju jezika. Međutim, njihova primarna upotreba bila je vezana za pojedinačne jezike ili usko određene jezičke domene [6].

Iako su prethodno naučeni modeli pokazali izvanredne rezultate za jezike s obiljem dostupnih podataka poput engleskog, francuskog ili španskog, izazov nastaje kada se primene na jezike sa manje resursa ili nedovoljno reprezentovane jezike u digitalnom okruženju. Tu dolazimo do suštine problema: *Kako efikasno preneti znanje stečeno na jednom jeziku na drugi jezik, posebno kad su resursi za drugi jezik ograničeni* [8]?

U ovoj disertaciji detaljno će se istražiti različite tehnike prenosa znanja između jezika unutar okvira PyTorch i Transformers arhitekture [9]. Fokusirajući se na adaptaciju prethodno naučenih modela na jezike sa manje dostupnih podataka, istražiće se strategije finog podešavanja (fine-tuning) i prilagođavanja ovih modela kako bi se postigla veća efikasnost i tačnost u višejezičkom okruženju [10].

Osim tehničkih izazova, važno je istražiti i društvene i kulturne implikacije višejezičke obrade teksta [1]. Globalno društvo u kojem komunikacija prelazi granice jezika zahteva tehnološke inovacije koje promovišu razumevanje i inkluzivnost. Ova disertacija se stoga bavi pitanjima etike, pristupačnosti i potrebe za tehnološkim rešenjima koja podržavaju jezičku raznolikost [3].

Kroz interdisciplinarni pristup, ova doktorska disertacija ima za cilj da spoji teorijska saznanja i praktične primene u domenu transfernog učenja i adaptacije unutar PyTorch i Transformers paradigme. Razmatrajući kako ovi alati mogu podstaći razvoj jezički agilnih tehnologija, ova studija ne samo što doprinosi akademskom polju, već ima i potencijal da oblikuje budućnost komunikacije na globalnom nivou [5].

# **I METODOLOŠKI OKVIR ISTRAŽIVANJA**

# 1. FORMULACIJA PROBLEMA ISTRAŽIVANJA

U kontekstu rastuće primene GPT paradigme, transfernog učenja i adaptacije PyTorch i Transformers arhitekture u pogledu višejezičnosti, istraživači ističu zabrinutost zbog realne pretnje nedostatka visokokvalitetnih podataka neophodnih za obuku sofisticiranih sistema zasnovanih na veštačkoj inteligenciji [1]. Ovakav nedostatak može značajno usporiti napredak određenih modela, naročito onih sa obimnim jezičkim domenom, te predstavlja ključni faktor za budući razvoj u oblasti veštačke inteligencije [2]. Iako se čini da je internet prepun podataka, postavlja se pitanje da li se suočavamo sa ovim izazovom i kako bismo mogli rešiti problem nedostatka kvalitetnih podataka [3]. Visokokvalitetni podaci su suštinski važni za razvoj preciznih i moćnih algoritama [4].

U situaciji kada nedostaju visokokvalitetni podaci za trening NLP modela veštačke inteligencije, transferno učenje može biti ključno rešenje. Primenom transfernog učenja na jezicima na kojima GPT model nije bio treniran, moguće je iskoristiti već postojeće znanje i strukturu modela kako bi se prilagodio novom jezičkom domenu ili manjkavim podacima [6].

Dakle, modeli kao što je GPT-2, kada se primene na nove jezičke domene ili jezike na kojima nisu trenirani, mogu biti prilagođeni kroz transferno učenje. To podrazumeva prethodno obučavanje modela na sličnim jezicima ili srodnim podacima koji su visokog kvaliteta, nakon čega se taj model fino podešava na ciljani jezički domen ili skup podataka. Ovaj proces omogućava modelu da bolje razume specifičnosti novog jezika ili domena, čime se delimično rešava nedostatak visokokvalitetnih podataka [9]. Transferno učenje na jezicima na kojima postoje obilje kvalitetnih podataka može poslužiti kao most ka adaptaciji veštačke inteligencije na različite jezičke domene, dopunjavajući nedostatak podataka koji su neophodni za razvoj preciznijih i moćnijih algoritama, naročito u situacijama kada ti podaci nisu direktno dostupni ili su niskog kvaliteta [10].

U fazi postavljanja istraživačkog problema u okviru projekta doktorske disertacije, polazi se od osnovnog definisanja problema, fokusirajući se na:

- a) Opšti teorijski model GPT-2 paradigme i transfernog učenja, istražujući njihovu primenu u adaptaciji PyTorch i Transformers arhitektura za stvaranje višejezičkih sistema. Ovaj teorijski model predstavlja osnovu za analizu konteksta istraživanja i njegovu relevantnost u globalnom smislu [12].
- b) Odredbe projekta doktorske disertacije, fokusirajući se na ključne aspekte istraživanja. Ovi aspekti čine osnovu istraživačkog rada i definišu ključne teme kojima se obraća posebna pažnja u disertaciji [14].

## 1.1. Osnovna istraživačka ideja u doktorskoj disertaciji

Osnovna istraživačka ideja u doktorskoj disertaciji bazira se na istraživanju paradigme transfernog učenja i adaptacije PyTorch i Transformers arhitektura u kontekstu stvaranja višejezičkih sistema [15]. Fokusirajući se na evoluciju veštačke inteligencije u okviru ovih modela, istraživanje će se baviti analizom efikasnosti transfernog učenja na jezicima na kojima GPT-2 modeli nisu originalno obučeni [16].

Glavno istraživačko pitanje koje se postavlja jeste: *Kako se mogu prilagoditi GPT-2, PyTorch i Transformers arhitekture radi efikasnog prenosa znanja i primene na jezicima koji nisu bili deo originalnog trening skupa?* Ovako postavljeno istraživačko pitanje se dalje razlaže na nekoliko podpitanja [17]:

- Kako transferno učenje može biti primenjeno na jezicima na kojima veštački inteligentni modeli nisu trenirani, prateći GPT-2 paradigmu?
- Koji su ključni izazovi i prepreke prilikom adaptacije PyTorch i Transformers arhitektura na ove jezike?
- Kako se može postići visok nivo efikasnosti i preciznosti pri adaptaciji ovih modela za višejezičku primenu?
- Kako rezultati ovog istraživanja doprinose unapređenju postojećih tehnika transfernog učenja u domenu višejezičnosti?

Ove istraživačke ideje se fokusiraju na dublje razumevanje primene veštačke inteligencije na višejezičkom nivou, što predstavlja temelj za napredak u oblasti transfernog učenja i adaptacije veštačkih inteligentnih sistema na različitim jezicima [18].

## **1.2. Značaj problema istraživanja**

### **1.2.1. Naučni značaj**

Naučni značaj istraživanja o GPT-2 paradigmi, transfernom učenju i adaptaciji PyTorch i Transformers arhitektura za višejezičke sisteme ogleda se u njegovoj sposobnosti da proširi granice veštačke inteligencije [19]. Ovo istraživanje ne samo da doprinosi razumevanju i unapređenju postojećih veštačkih inteligentnih sistema, već i otvara nova vrata za razvoj i primenu veštačke inteligencije na globalnom nivou. Naučni značaj ovog istraživanja ogleda se u prilagođavanju ovog modela efikasnoj primeniti na jezicima na kojima nisu prvobitno obučeni, nudeći novu perspektivu za globalnu primenu veštačke inteligencije [2].

Ova disertacija ima potencijal da proširi naše razumevanje transfernog učenja, pružajući ključne uvide o adaptaciji jezičkih modela na različite jezičke specifičnosti [5]. Osim toga, istraživanje će obogatiti naše znanje o mehanizmima prilagođavanja veštačkih inteligentnih sistema, čime će doprineti boljem razumevanju izazova i mogućnosti u višejezičkom okruženju. Ovim istraživanjem ne samo da će se unaprediti već postojeći modeli, već će se podržati i dalji razvoj veštačke inteligencije sa globalnom primenom na različitim jezicima i kulturama [11].

Razumevanje, prilagođavanje i unapređenje veštačkih inteligentnih sistema na jezicima na kojima nisu prvobitno obučeni, ima šire implikacije i potencijal da unapredi raznolikost i dostupnost veštačke inteligencije na globalnom nivou [14]. Ovo istraživanje otvara puteve za široku primenu u različitim oblastima, uključujući poslovanje, obrazovanje, medicinu i dr., obećavajući inovativna rešenja koja su prilagođena višejezičkom okruženju, što na postojećem nivou razvoja veštačke inteligencije ima veliki naučni značaj za istraživanje [19].

### **1.2.2. Društveni (praktični) značaj**

Pravilan razvoj i primena ovih sistema može imati dubok i širok uticaj na društvo. U kontekstu globalne interakcije, mogućnost efikasnog prevođenja i razumevanja različitih jezika unapređuje međunarodnu

komunikaciju, podržava multikulturalnost i otvara vrata za bolje međunarodne odnose, olakšavajući razmenu znanja, ideja i kultura [7].

Osim toga, ovo istraživanje može imati značajan uticaj na obrazovanje i dostupnost znanja širom sveta. Razvoj višejezičkih sistema značiće bolji pristup informacijama i obrazovanju za ljude širom različitih jezičkih zajednica, unapređujući proces učenja i olakšavajući pristup globalnoj bazi znanja [13]. U medicinskom kontekstu, ovi napredni jezički modeli mogu poboljšati komunikaciju između različitih jezičkih grupa i olakšati pristup informacijama o zdravstvenoj zaštiti, čime se mogu poboljšati uslovi pružanja zdravstvenih usluga i podići nivo zdravstvene pismenosti [16].

Isto tako, ova istraživanja imaju potencijal da podrže poslovne inovacije i konkurentnost na globalnom tržištu [18]. Efikasno prilagođavanje i primena veštačke inteligencije u višejezičkom kontekstu mogu omogućiti kompanijama da razviju inovativne proizvode, usluge i marketinške strategije prilagođene različitim jezicima i kulturama, šireći tako svoj uticaj i dostizanje različitih tržišta širom sveta [20].

### **1.3 Rezultati prethodnih saznanja o problemu istraživanja**

U kontekstu višejezičkog sveta, rezultati prethodnih istraživanja pokazuju značajan napredak u razvoju jezičkih modela [4]. Analizirajući prethodne studije i istraživačke radove, primećuje se da su modeli poput GPT-2 postigli značajne uspehe u generisanju teksta na različitim jezicima, posebno engleskom, francuskom i španskom jeziku [6]. Međutim, pored tih postignuća, postoje ograničenja koja se odnose na adaptaciju modela na manje rasprostranjene jezike ili na efikasnost prevođenja među različitim jezicima [9].

Nedostatak visokokvalitetnih podataka za obuku modela na manje zastupljenim jezicima jedan je od glavnih izazova [12]. Prethodna istraživanja ukazuju na to da nedostatak kvalitetnih i obimnih podataka može biti ključni faktor koji ograničava performanse modela u višejezičkom okruženju [15]. Samim tim, razlike u gramatičkoj strukturi, leksici i kontekstu među jezicima mogu uticati na prenos značenja tokom prevođenja, što predstavlja poseban izazov za većinu savremenih jezičkih modela [17].

Istraživanja su takođe pokazala da postoji potreba za daljim unapređenjem transfernog učenja i adaptacije veštačkih neuronskih mreža kako bi se omogućilo efikasnije učenje i prilagođavanje modela na različite jezike [19]. Prethodni radovi su naglasili važnost istraživanja mehanizama prenosa znanja između jezika, kako bi se modeli mogli efikasnije prilagoditi novim jezicima sa manjkom obimnih trening skupova podataka [1].

Ipak, iz dostupnih izvora navodimo rezultate prethodnih istraživanja nekih autora koji su pokušavali da istražuju predmetnu oblast, i to:

*Yann LeCun, Yoshua Bengio, i Geoffrey Hinton, "Deep Learning" (Nature, 2015)*

- Ovaj rad daje opšti pregled dubokog učenja, što je ključno za razvoj složenih jezičkih modela. Autori objašnjavaju kako duboko učenje može automatski učiti hijerarhijske reprezentacije podataka. Ovaj princip je osnova za razumevanje kako jezički modeli poput GPT-2 generišu tekst. Rad ističe značaj višeslojnih neuronskih mreža, što je direktno povezano sa sposobnošću modela da obrađuju i generišu tekst na različitim jezicima. Takođe, rad naglašava potencijal dubokog učenja u obradi prirodnog jezika, što je ključno za razvoj višejezičkih aplikacija [7].

*Noam Chomsky, "Syntactic Structures" (1957)*



- Chomskyjeva teorija generativne gramatike postavila je temelje za razumevanje strukture jezika, što je ključno za razvoj jezičkih modela. Ovaj rad omogućava bolje razumevanje kako jezički modeli, poput GPT-2, mogu da generišu gramatički ispravan tekst. Koncepti iz ovog rada pomažu u razumevanju izazova koji se javljaju pri obradi manje zastupljenih jezika, čija gramatička pravila mogu biti različita. Chomskyjev pristup analizi jezika poslužio je kao osnova za razvoj algoritama koji mogu obraditi složene jezičke strukture. Njegova teorija takođe podstiče istraživače da razmišljaju o unapređenju jezičkih modela u smislu boljeg razumevanja gramatike i sintakse različitih jezika [17].

*Ian Goodfellow, "Generative Adversarial Nets" (2014)*

- Goodfellow je u ovom radu predstavio koncept generativnih suparničkih mreža (GAN-ova), što je revolucioniziralo područje generativnih modela. Iako se GAN-ovi uglavnom koriste u vizuelnoj obradi, principi iz ovog rada mogu se primeniti i u generisanju teksta. Koncepti iz ovog rada mogu biti inspiracija za razvoj novih pristupa u generisanju višejezičkog teksta, posebno u smislu realističnosti i prirodnosti. Goodfellowov rad može podstaći istraživanja u oblasti NLP-a, naročito u razvoju modela koji bolje razumeju i generišu tekst na manje zastupljenim jezicima. GAN-ovi mogu pružiti nove perspektive za unapređenje kvaliteta i prirodnosti generisanog teksta, što je važno u višejezičkom kontekstu [18].

*Sebastian Ruder, "An Overview of Multi-Task Learning in Deep Neural Networks" (2017)*

- Ruder u ovom radu istražuje multi-task learning (učenje više zadataka) u kontekstu dubokih neuronskih mreža, što je relevantno za razvoj višejezičkih modela. Multi-task learning može poboljšati performanse jezičkih modela u višejezičkom okruženju, omogućavajući im da istovremeno uče iz više jezičkih izvora. Rad objašnjava kako zajedničko učenje iz više zadataka može doprineti boljem razumevanju i generisanju jezičkih struktura. Ruderov pristup može pomoći u premošćavanju jezičkih barijera i poboljšanju adaptacije modela na različite jezike. Ovaj rad pruža uvid u to kako se principi multi-task learninga mogu primeniti za unapređenje višejezičkih jezičkih modela, posebno u kontekstu boljeg razumevanja i obrade manje zastupljenih jezika [16].

*Andrej Karpathy, "The Unreasonable Effectiveness of Recurrent Neural Networks" (2015)*

- Karpathyjev rad istražuje upotrebu rekurentnih neuronskih mreža (RNN-ova) u generisanju teksta, što je direktno povezano sa razvojem jezičkih modela kao što je GPT-2. Ovaj rad pokazuje kako RNN-ovi mogu efikasno generisati složene tekstualne sekvence, što je osnova za razumevanje generativnih jezičkih modela. Rad daje uvid u to kako RNN-ovi mogu uhvatiti dugoročne zavisnosti u tekstu, što je ključno za generisanje koherentnog i kontekstualno relevantnog teksta. Karpathyjevi nalazi su važni za razvoj višejezičkih modela, jer pokazuju kako se kompleksne jezičke strukture mogu modelirati u neuronskim mrežama. Njegov rad može pomoći u razumevanju izazova povezanih sa generisanjem teksta na manje zastupljenim jezicima, posebno u smislu "hvatanja" njihovih jedinstvenih jezičkih karakteristika [20].

*Mike Schuster i Kaisuke Nakajima, "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation" (2016)*

- Schuster i Nakajima su autori rada koji detaljno opisuje Google Neural Machine Translation (GNMT) sistem. Ovaj rad je značajan jer pokazuje kako duboko učenje može biti primenjeno na složeni zadatak

mašinskog prevođenja. GNMT sistem je postao referentna tačka u industriji zbog svoje sposobnosti efikasnog prevođenja između jezika, što je izuzetno relevantno za ovu disertaciju. Rad ističe kako se može postići veća tačnost prevođenja korišćenjem dubokih neuronskih mreža, što je ključno za razvoj višejezičkih aplikacija. Autori su takođe istraživali izazove povezane sa prevođenjem između jezika sa različitim gramatičkim strukturama, što doprinosi boljem razumevanju adaptacije na manje zastupljene jezike [18].

*Thomas Wolf, "Transformers: State-of-the-Art Natural Language Processing" (2020)*

- Thomas Wolf je suautor ovog rada koji pruža sveobuhvatan pregled transformer modela u NLP-u. Rad detaljno opisuje kako transformer modeli, poput BERT i GPT, mogu obraditi i generisati tekst na visokom nivou. Wolf-ov doprinos leži u razumevanju kako se kompleksni jezički obrasci mogu modelirati pomoću ovih naprednih tehnika. Rad takođe ističe značaj transfernog učenja i adaptacije modela na različite jezike. Kao osnivač Hugging Face, Wolf je takođe bio ključan u pružanju pristupačnih NLP alata široj zajednici, što je pomoglo u razvoju višejezičkih aplikacija [20].

*Emily Bender, "Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax" (2013)*

- Emily Bender je autorka ovog dela koje pruža temeljit uvid u lingvističke osnove potrebne za razumevanje i obradu prirodnog jezika. Rad se fokusira na morfologiju i sintaksu, što je ključno za razvoj efikasnih jezičkih modela. Bender naglašava važnost razumevanja lingvističkih principa za razvoj inkluzivnih i pristupačnih tehnologija, što je važno za višejezičke aplikacije. Njen rad takođe podstiče razvoj etičkih smernica u NLP-u, naglašavajući važnost inkluzije i pristupačnosti u jezičkim tehnologijama. Ovo je posebno relevantno u kontekstu razvoja modela za manje zastupljene jezike, gde su etički i društveni aspekti od ključnog značaja [16].

*Richard Socher, "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank" (2013)*

- Richard Socher je u ovom radu istraživao primenu dubokih rekurentnih mreža za semantičku analizu teksta. Ovaj rad je značajan jer pokazuje kako duboko učenje može biti korišćeno za razumevanje složenih jezičkih struktura i sentimenta. Socherov pristup je važan za razvoj jezičkih modela koji mogu razumeti i generisati tekst sa semantičkom preciznošću. Rad pruža osnovu za razvoj modela koji mogu obraditi ne samo gramatičku, već i semantičku strukturu jezika, što je ključno za višejezičke aplikacije. Njegovi nalazi su posebno relevantni za poboljšanje kvaliteta generisanja teksta u manje zastupljenim jezicima, gde semantički kontekst može biti posebno izazovan [15].

*Philipp Koehn, "Statistical Machine Translation" (2009)*

- Philipp Koehn je autor knjige "Statistical Machine Translation" koja detaljno objašnjava osnove statističkog mašinskog prevođenja. Ovaj rad je ključan za razumevanje kako se statistički modeli mogu koristiti za prevođenje između jezika. Koehn-ov pristup je bio fundamentalan za razvoj ranijih generacija mašinskih prevodilaca, postavljajući osnovu za kasnije razvoje u AI-prevođenju. Knjiga objašnjava tehnike kao što su modeliranje jezika i poravnanje, što je relevantno za razumevanje izazova u prevođenju između različitih jezičkih parova. Njegov rad je posebno važan za razumevanje tehnika koje

se mogu primeniti u višjezičkim jezičkim modelima, posebno u kontekstu efikasnog prevođenja i adaptacije na manje zastupljene jezike [18].

Razumevanje i integracija višjezičnosti u jezičke modele poput GPT-2 predstavlja značajan korak ka stvaranju istinski inkluzivnih tehnologija obrade prirodnog jezika [3]. Kroz analizu postojećih istraživanja i radova, jasno je da postoji prostor za unapređenje u pogledu efikasnosti, preciznosti i adaptabilnosti ovih modela, posebno kada su u pitanju manje zastupljeni jezici [6]. Unatoč napretku koji je postignut u razumevanju i implementaciji složenih jezičkih struktura, kao i u primeni dubokog učenja u obradi prirodnog jezika, još uvek postoji izazov u stvaranju modela koji mogu sa jednakom preciznošću i razumevanjem obraditi sve jezike [11]. Ovo je posebno izraženo kod jezika koji nemaju bogate setove podataka ili su strukturalno i gramatički različiti od onih jezika na kojima su trenutni modeli najviše obučavani i testirani [14].

U skladu sa tim, postoji očigledna potreba za daljim istraživanjem i razvojem u oblasti višjezičkih jezičkih modela, kako bi se osigurala veća inkluzivnost i pristupačnost tehnologija obrade prirodnog jezika [17]. To uključuje ne samo poboljšanje postojećih modela i tehnika, već i razvoj novih pristupa koji mogu bolje razumeti i interpretirati jezičke različitosti [19]. Dalji napredak u ovoj oblasti nije samo tehnološko dostignuće, već i korak ka izgradnji globalno povezanog i kulturno raznolikog digitalnog sveta [20]. Ovo podrazumeva kontinuirani rad na obogaćivanju setova podataka, unapređenju algoritama za duboko učenje i stvaranju modela koji su prilagodljivi i efikasni u obradi širokog spektra jezičkih fenomena, doprinoseći time boljem razumevanju i povezivanju različitih kultura i jezika u digitalnoj eri [5].

## 2. ODREĐIVANJE PREDMETA ISTRAŽIVANJA

Predmet istraživanja u ovoj disertaciji, fokusiran je na analizu i razumevanje višejezičkih jezičkih modela, s posebnim naglaskom na izazove i ograničenja u njihovoj primeni na manje zastupljene jezike [1]. Zadatak istraživanja je da se detaljno istraži kako napredni jezički modeli poput GPT-2 i slični obrađuju različite jezičke strukture, kako se prilagođavaju različitim jezičkim kontekstima i u kojoj meri su efikasni u generisanju i prevođenju teksta među različitim jezicima [2]. Osim tehnoloških aspekata, istraživanje takođe obuhvata i etičke i sociolingvističke implikacije primene ovih tehnologija, istražujući kako one mogu doprineti boljem razumevanju i povezivanju različitih jezičkih i kulturnih zajednica [3].

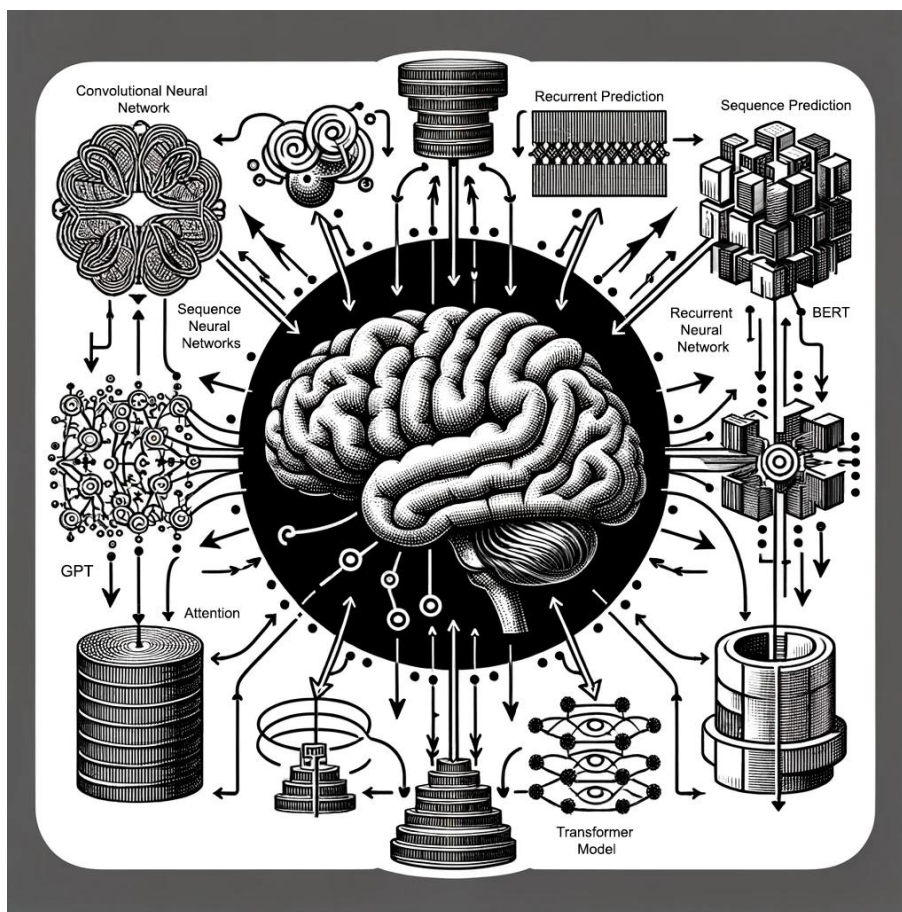
### 2.1. Teorijsko određenje predmeta istraživanja

Naučno je poznato uspostavljanje teorijskih temelja koji omogućavaju dubinsko razumevanje višejezičkih jezičkih modela [4]. Ova doktorska disertacija se temelji na postavkama lingvističke teorije, semiotike, informatike i veštačke inteligencije kako bi se izgradio čvrst okvir za analizu i interpretaciju složenih jezičkih fenomena [5]. Poseban akcenat je stavljen na teorijske koncepte kao što su jezička univerzalnost, generativna gramatika, semantička i pragmatička obrada jezika, kao i na specifične izazove koje predstavljaju manje zastupljeni jezici i njihova dialektička raznolikost [6]. Pored toga, razmatraju se i tehnološki aspekti koji uključuju algoritamske i arhitektonske osnove dubokog učenja, posebno u kontekstu transformatorskih modela koji stoje u središtu savremenih NLP sistema [7]. Ovo poglavlje postavlja temelje za empirijsko istraživanje, definisanjem ključnih pojmova i parametara koji će biti korišćeni u daljem istraživanju, te iznosi teorijski okvir unutar kojeg će se analizirati i ocenjivati performanse i mogućnosti višejezičkih jezičkih modela u obradi i generisanju prirodnog jezika [8].

Takođe, fokus disertacije se proširuje na temeljnu analizu lingvističkih teorija koje se bave strukturom jezika, njegovim funkcijama, i načinima na koje jezik komunicira značenja u različitim kulturološkim i socijalnim kontekstima [9]. Razumevanje ovih teorija je ključno za analizu sposobnosti višejezičkih modela da precizno generišu i interpretiraju jezik [10]. Ovaj rad se nadovezuje na Chomskyjev koncept jezičke kompetencije, istražujući kako se on može primeniti na mašinsko učenje i veštačku inteligenciju u cilju stvaranja modela koji mogu da emuliraju ljudsku sposobnost razumevanja i generisanja prirodnog jezika [11].

U ovom segmentu disertacije takođe se razmatraju koncepti semiotike, gde se jezik posmatra kao sistem znakova koji omogućava prenos značenja [12]. Pristupi kao što su Peirceova trijadna teorija znakova i Saussureov koncept semiotičkog sistema, pružaju okvir za razumevanje načina na koji jezički modeli mogu da interpretiraju i generišu znakove - **tokene** koji nose značenje [13]. Ovo je od posebne važnosti u kontekstu višejezičnosti gde se značenje mora očuvati preko granica različitih jezičkih sistema [14].

Tehnološki aspekti istraživanja obuhvataju dubinsku analizu algoritama i arhitektura koji omogućavaju mašinama da uče i razvijaju jezičke sposobnosti [15]. Pristupi dubokog učenja kao što su konvolucijske neuronske mreže (CNN) i rekurentne neuronske mreže (RNN) su proučavani u kontekstu njihove primene u jezičkim modelima što pokazuje *slika 1*. [16]. Proučavanje transformatorskih modela, poput GPT-a ili BERT-a, je od posebnog značaja jer su oni revolucionarizovali polje NLP-a svojom sposobnošću da obrade velike količine teksta i razumeju složene jezičke obrasce [17].



**Slika 1.** Ključne tehnologije iz oblasti veštačke inteligencije i obrade prirodnog jezika

*Slika 1.* predstavlja kompleksan graf koji ujedinjuje raznovrsne koncepte veštačke inteligencije i obrade prirodnog jezika, čije će se ključne komponente detaljno predstaviti:

- Convolutional Neural Network (Konvolucijske neuronske mreže) - Prikazane sa slojevitim, uvijenim strukturama koje ilustruju konvolucijske operacije karakteristične za ove mreže.
- Sequence Neural Networks (Sekevencijalne neuronske mreže) - Oznaka ukazuje na mreže koje obrađuju sekvencijalne podatke, što može biti prikazano kao lanac ili niz elemenata.
- GPT (Generative Pretrained Transformer) - Prikazan pored simbola koji sugerišu na složene strukture podataka i **attention** mehanizme, što je u skladu sa arhitekturom ovog modela.
- Attention (Pažnja) - Ovo se odnosi na mehanizam unutar neuronskih mreža koji "osvetljava" određene delove podataka.
- Transformer Model (Transformatorski model) - Simbolizuje kompleksnu strukturu transformatorskih modela, koji koriste **attention** mehanizme.
- Recurrent Neural Network (Rekurentne neuronske mreže) - Prikaz sa petljama koje simbolizuju rekurentne veze unutar ovih mreža.
- BERT (Bidirectional Encoder Representations from Transformers) - Predstavlja složenu, višedimenzionalnu strukturu podataka.
- Sequence Prediction (Predviđanje sekvenci) - Reprerentacija predviđanja sekvenci u modelima.
- Recurrent Prediction (Rekurentno predviđanje) - Sugeriše ciklički tok informacija.

Doktorska disertacija ističe i izazove koji se pojavljuju u kontekstu manje zastupljenih jezika, gde je nedostatak obimnih i kvalitetnih podataka za obuku modela često ograničavajući faktor. U ovom poglavlju se istražuje kako inovacije u tehnici transfernog učenja i pristupi poput "**few-shot learninga**" i "**zero-shot learninga**" mogu omogućiti modelima da prevaziđu ove izazove i efikasnije se prilagode različitim jezičkim okruženjima.

Disertacija se bavi pitanjem semantičke obrade jezika, gde se istražuje kako modeli mogu da razumeju i generišu semantički koherentan tekst. Analiza se fokusira na sposobnost modela da uče iz konteksta i kako se to znanje može primeniti u generisanju prirodnog jezika koji odražava nijanse značenja i jezičke suptilnosti [29]. Ovo obuhvata i razumevanje pragmatike, studije jezičke upotrebe u kontekstu, što je ključno za stvaranje modela koji mogu da komuniciraju efikasno i relevantno u različitim socijalnim i kulturološkim kontekstima [30]. Ovo je detaljno objašnjeno u *poglavlju 13.* i vizualno predstavljeno na *slici 49.*

Razmatranje etičkih i društvenih aspekata u primeni višejezičkih jezičkih modela zauzima važno mesto u istraživanju [3]. Ovo podrazumeva kritičku analizu potencijalnih uticaja ovih tehnologija na društvo, posebno u pogledu pravednosti, inkluzije, i poštovanja jezičkih i kulturoloških različitosti [5]. Istraživanje se bavi i pitanjem kako osigurati da tehnologije obrade prirodnog jezika ne ojačaju postojeće predrasude ili ne doprinesu jezičkoj homogenizaciji [7].

Takođe, u ovom poglavlju se razvija teorijski okvir za empirijsko istraživanje koje će pratiti, uključujući metodologiju i parametre koji će se koristiti za procenu i verifikaciju hipoteza [15]. Ovo uključuje definisanje metrika za ocenu performansi višejezičkih modela, kao i planiranje eksperimenata koji će testirati njihovu sposobnost da efikasno obrade i generišu tekst na različitim jezicima. Pristup temeljen na čvrstim teorijskim osnovama će omogućiti rigorozno i objektivno ispitivanje mogućnosti i ograničenja savremenih jezičkih modela u višejezičkom kontekstu [19].

## 2.2. Pojmovno-kategorijalni sistem

Disertacija detaljno istražuje GPT-2 paradigmu tj. pojmovno-kategorijalni sistem vezan za transferno učenje i adaptaciju PyTorch i Transformers arhitektura u svrhu postizanja višejezičkih sposobnosti [21]. Transferno učenje se pojavljuje kao ključna tehnika koja omogućava modelima, kao što je GPT-2, da prevaziđu barijere jezičkih ograničenja primenom znanja stečenog na jednom jeziku ili zadatku u obradi drugih jezika. Ova tehnika je posebno važna u kontekstu modeliranja jezičkih varijacija koje su manje zastupljene u treninzima, omogućavajući modelima da postanu istinski višejezički alati [25].

PyTorch i Transformers predstavljaju tehnološku osnovu koje ćemo istraživati u nastavku ovog rada, naglašavajući njihovu fleksibilnost i prilagodljivost koja je neophodna za implementaciju transfernog učenja [27]. Arhitektura Transformers modela, sa svojim mehanizmom pažnje, ističe se kao revolucionaran doprinos u sposobnosti mašina da efikasno procesuiraju i generišu prirodni jezik. Analizom PyTorch frameworka, istražuje se kako pristupačno i modularno programiranje može poslužiti kao temelj za razvoj i testiranje različitih jezičkih modela, omogućujući istraživačima da eksperimentišu sa novim idejama i pristupima u višejezičkom okruženju [30].

Takođe, disertacija se bavi kritičkom analizom izazova i potencijala koje ove arhitekture nose u kontekstu globalizovanog sveta [22]. Poseban fokus stavljen je na etičke i praktične aspekte primene ovih modela kroz programski kod, te se razmatra kako inovacije u oblasti veštačke inteligencije mogu

doprineti boljem razumevanju i većoj povezanosti ljudi različitih jezičkih i kulturnih pozadina [24]. Pristupi kao što su "zero-shot" i "few-shot learning" istražuju se kao mogući odgovori na složene zahteve višejezičke komunikacije, težnju ka univerzalnoj jezičkoj kompetenciji i izgradnji mostova među jezicima i kulturama [26].

Ključni kategorijalni pojmovi koji proizilaze iz ove doktorske disertacije su:

**GPT-2 Paradigma:** Odnosi se na specifičan pristup u obradi prirodnog jezika koji koristi Generative Pre-trained Transformer 2 (GPT-2) model. Paradigma uključuje korišćenje ovog modela za generisanje prirodnog teksta, s posebnim fokusom na njegovu adaptabilnost i efikasnost u višejezičkim kontekstima [18].

**Transferno učenje:** Metodologija u veštačkoj inteligenciji koja omogućava modelu, poput GPT-2, da primeni znanje stečeno tokom obuke na jednom zadatku ili jeziku za rešavanje problema u drugom, često srodnom, kontekstu. U kontekstu ove disertacije, transferno učenje istražuje kako se modeli mogu prilagoditi za efikasnu obradu različitih jezika [14].

**PyTorch i Transformers Arhitekture:** Tehnološka osnova koja se koristi u ovoj doktorskoj disertaciji. PyTorch je popularan okvir za mašinsko učenje koji omogućava fleksibilno i efikasno modeliranje, dok su Transformers modeli koji se ističu u zadacima obrade prirodnog jezika, naročito u generisanju teksta i razumevanju jezika [12].

**Višejezički Svet:** Odnosi se na primenu GPT-2 i srodnih tehnologija u globalnom, multikulturalnom kontekstu. U ovom istraživanju, ovaj pojam istražuje kako napredni jezički modeli mogu prevazići jezičke barijere, omogućiti efikasniju međukulturnu komunikaciju i podržati jezičku raznolikost [10].

### 2.3. Operacionalno određenje istraživanja

Istraživački pristup zahteva jasno definisanje metoda i tehnika koje će se koristiti za empirijsko testiranje teorijskih postavki [8]. Centralno mesto u ovom procesu zauzima operacionalizacija koncepta kao što su transferno učenje i adaptacija višejezičkih jezičkih modela unutar PyTorch i Transformers okruženja. Operacionalno određenje podrazumeva konkretno preciziranje kako će se ovi teorijski koncepti meriti, testirati i evaluirati [4]. Ovo uključuje izbor odgovarajućih skupova podataka, definisanje metrika uspešnosti i postavljanje eksperimentalnih uslova pod kojima će se ispitivati sposobnost modela da se prilagode novim jezičkim kontekstima [2].

Zatim, razmatraju se specifične procedure koje omogućavaju GPT-2 modelu da se adaptira na različite jezičke strukture koristeći mehanizme kao što su **attention** i "masked language modeling" [1]. Operacionalizacija takođe podrazumeva razvoj protokola za treniranje i finiširanje modela, uključujući i selekciju opštih i Optuna hiperparametara koji će se koristiti za optimizaciju modela za specifične jezičke zadatke [3]. Pored toga, važno je utvrditi i koje će se tehnike koristiti za premošćenje jaza između jezika sa bogatim resursima i onih sa ograničenim skupom podataka, što je ključno u radu sa manje zastupljenim jezicima [5].

Ujedno, disertacija definiše kako će se prikupljeni podaci analizirati i interpretirati, kako bi se došlo do valjanih zaključaka o efikasnosti primenjenih metoda [7]. U tu svrhu, biće primenjene napredne statističke analize i tehnike mašinskog učenja, uključujući analizu grešaka, kako bi se razumeli uzorci i

tendencije u ponašanju modela. Ovaj operacionalni okvir ne samo da će pružiti uvid u trenutno stanje tehnike veštačke inteligencije, već će isto tako postaviti temelje za buduća istraživanja i unapređenja u polju višejezičke obrade prirodnog jezika [11].

Kao deo naučne elaboracije doktorske disertacije, bilo je neophodno razložiti temu na njene osnovne strukturalne elemente.

### **2.3.1. Strukturalni činioci predmeta istraživanja koji će biti neposredno istraženi**

*Prvi strukturalni činilac predmeta istraživanja - Optimizacija i Adaptacija Modela:* Kroz funkcije za inicijalizaciju optimizera i schedulera, istraživanje naglašava važnost prilagođavanja parametara učenja, poput stopa učenja (learning rate) i težinskih degradacija (weight decay), za maksimiziranje performansi modela GPT-2. Ciklični learning rate, koji osciluje između bazične i maksimalne stope učenja, ključan je za fino podešavanje modela [17].

*Drugi strukturalni činilac predmeta istraživanja - Učitavanje i Kontinuitet Treninga:* Kroz mehanizme za učitavanje checkpointa, istraživanje uključuje analizu sposobnosti modela da nastavi trening od prethodno sačuvanih stanja, što je bitno za dugotrajne i resursno-intenzivne procese treniranja modela [19].

*Treći strukturalni činilac predmeta istraživanja - Evaluacija i Validacija Modela:* Funkcija evaluate omogućava praćenje performansi modela, naglašavajući važnost kontinuirane validacije i optimizacije modela na osnovu evaluacionih metrika kao što je loss funkcija [21].

*Četvrti strukturalni činilac predmeta istraživanja - Upravljanje Memorijom:* Pristup oslobađanju GPU memorije nakon svakog epocha ukazuje na važnost efikasnog upravljanja resursima, što je posebno kritično u kontekstu velikih modela poput GPT-2 [23].

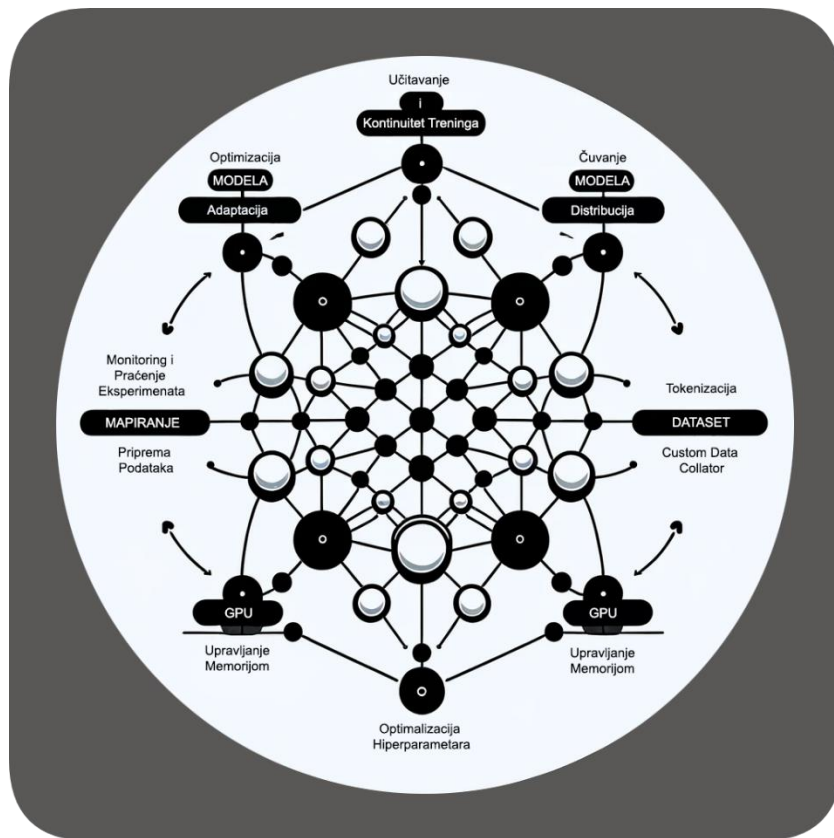
*Peti strukturalni činilac predmeta istraživanja - Tokenizacija i Priprema Podataka:* Kroz korišćenje custom data collatora, istraživanje ističe značaj prilagođavanja procesa tokenizacije i pripreme batcheva za efikasno treniranje modela [25].

*Šesti strukturalni činilac predmeta istraživanja - Monitoring i Praćenje Eksperimenata:* Integracija sa Weights & Biases (Wandb) platformom za praćenje eksperimenata ukazuje na potrebu za detaljnim nadzorom i analizom trening procesa, omogućavajući bolje razumevanje dinamike i performansi modela [27].

*Sedmi strukturalni činilac predmeta istraživanja - Proučavanje najboljih parametara dobijenih kroz optimizaciju hiperparametara sa alatima poput Optuna i korišćenjem DYNAMO okvira pruža uvid u strategije za unapređenje efikasnosti učenja modela [29]. DYNAMO okvir omogućava dinamičko istraživanje prostora hiperparametara, čime se ubrzava proces optimizacije i poboljšavaju performanse modela.*

*Osmi strukturalni činilac predmeta istraživanja - Čuvanje i Distribucija Modela:* Metode za sačuvavanje modela i tokenizera nakon treninga naglašavaju važnost očuvanja i deljenja naučenog znanja, što je ključno za transferno učenje i adaptaciju modela na nove jezičke zadatke [30].





**Slika 2.** Ključni aspekti procesa mašinskog učenja, od pripreme podataka do treniranja i upravljanja modelima. Svaki od ovih činilaca predstavljenih na slici 2. će biti temeljno istražen kako bi se ocenila njihova uloga i uticaj na adaptaciju i efikasnost GPT-2 modela u kontekstu višejezičnosti, te će formirati osnovu za praktičnu primenu u stvaranju modela sposobnih za interakciju i razumevanje u višejezičkom okruženju [16].

### 2.3.2. Vremensko, prostorno i disciplinarno određenje predmeta istraživanja

Vremensko određenje predmeta istraživanja se fokusira na istorijski razvoj i trenutno stanje tehnologija veštačke inteligencije od 2006. do 2025. godine, s posebnim osvrtom na evoluciju jezičkih modela poput GPT-2 [18]. Ovo uključuje proučavanje napretka u algoritmima mašinskog učenja od ranih neuronskih mreža do savremenih transformatorskih modela, kao i predviđanje budućih trendova i potencijalnih inovacija. Vremenska komponenta takođe istražuje kako tehnološki napredak utiče na ciklus treninga i fine-tuning modela i kako se ovi procesi mogu optimizovati da bi se smanjilo potrebno vreme za njihovu implementaciju [22].

Prostorno, predmet istraživanja obuhvata geografsku distribuciju i primenu jezičkih modela na prostoru globalne zajednice, uzimajući u obzir kako različiti jezički i kulturološki konteksti utiču na razvoj i prilagođavanje modela. Ovo obuhvata analizu dostupnosti podataka za različite jezike, kao i izazove i strategije za prevazilaženje jezičkih barijera, s ciljem stvaranja univerzalnih jezičkih modela koji mogu delovati u višejezičkom okruženju [26]. Prostorno određenje takođe razmatra infrastrukturne i resursne aspekte koji su neophodni za podršku razvoja i upotrebe višejezičkih tehnologija širom sveta.

Disciplinarno, predmeta istraživanja obuhvata interdisciplinarni pristup koji kombinuje znanja iz lingvistike, informatike, kognitivne nauke i drugih relevantnih oblasti kako bi se stvorio holistički pristup razumevanju i razvoju jezičkih modela [30]. Ovaj pristup prepoznaje važnost integracije lingvističkih teorija u razvoj modela, kao i primenu principa iz računarske nauke za implementaciju tehničkih aspekata. Disciplinarno određenje takođe uključuje etičke, sociološke i psihološke aspekte vezane za upotrebu veštačke inteligencije u društvu i njihov uticaj na pojedince i zajednice [4].

### 3. CILJEVI ISTRAŽIVANJA

#### 3.1. Naučni ciljevi istraživanja

U procesu dubinskog istraživanja višejezičkih jezičkih modela, naučni ciljevi predstavljaju ključne tačke kojima se usmerava pravac istraživanja u doktorskoj disertaciji. Naučni ciljevi omogućavaju sistematizaciju istraživačkog procesa i postavljanje temelja za teorijske i praktične doprinose u polju veštačke inteligencije.

Prvi nivo naučnog cilja obuhvata **naučnu deskripciju**, odnosno potpunije opisivanje karakteristika i funkcija jezičkih modela kao što je GPT-2, sa fokusom na njihovu sposobnost višejezičkog generisanja i prevođenja teksta [28].

Viši nivo naučnog saznanja u odnosu na deskripciju koji se namerava dostići u doktorskoj disertaciji je **naučna klasifikacija** kategorizacija jezičkih modela prema njihovim funkcijama, strukturi i efikasnosti, sa ciljem razumevanja njihove primenjivosti i ograničenja u različitim jezičkim kontekstima [29].

**Naučno otkriće i naučno objašnjenje** su sledeći viši nivoi naučnih ciljeva koji se nameravaju dostići izradom ove doktorske disertacije, koji podrazumevaju identifikaciju i otkriće novih fenomena unutar domena višejezičkih jezičkih modela i naučno objašnjenje mehanizama koji omogućavaju njihovu funkcionalnost u obradi prirodnog jezika [27].

Krajnji nivo naučnog cilja koji se namerava ostvariti izradom doktorske disertacije jeste **predlog novog modela ili okvira** koji integriše dosadašnja saznanja i predviđanja, postavljajući novu paradigmu za tumačenje i primenu višejezičkih jezičkih modela u praksi [30].

Svaki od ovih ciljeva doprinosi sveobuhvatnom razumevanju i unapređenju tehnologija obrade prirodnog jezika, usmeravajući buduće istraživačke napore ka boljoj funkcionalnosti i inkluzivnosti u multikulturalnom i multijezičnom digitalnom prostoru.

#### 3.2. Društveni cilj istraživanja

Društveni cilj istraživanja težišno je usmeren na istraživanje uticaja višejezičkih jezičkih modela kao što je GPT-2 na društvo u celini, posebno kako oni mogu unaprediti međukulturnu komunikaciju i doprineti eliminaciji jezičkih barijera [22]. Ovaj cilj obuhvata identifikaciju potencijala ovih tehnologija da podrže razumevanje i prihvatanje kulturne raznolikosti, kao i njihovu sposobnost da pruže alate za obrazovanje i socijalnu integraciju [23]. Pored toga, društveni cilj izrade doktorske disertacije jeste i promocija etičke upotrebe veštačke inteligencije, s ciljem izgradnje održivih i odgovornih tehnoloških praksi koje poštuju privatnost i podstiču fer pristup informacijama [25]. Dakle, u društvenom pogledu, istraživanje teži da doprinese stvaranju inkluzivnijeg digitalnog prostora gde jezička tehnologija služi kao most među ljudima, omogućavajući bolje razumevanje i saradnju na globalnom nivou [20].

## 4. HIPOTETIČKI OKVIR ISTRAŽIVANJA

Hipotetički okvir istraživanja u ovoj doktorskoj disertaciji se konstituiše iz jedne opšte hipoteze, koja se bavi sveobuhvatnim razumevanjem i poboljšanjem višejezičkih jezičkih modela poput GPT-2 i sedam posebnih hipoteza koje se fokusiraju na određene aspekte ovih modela. Svaka posebna hipoteza prati se s pripadajućim pojedinačnim hipotezama, jasno definisanim varijablama i odgovarajućim indikatorima koji omogućuju precizno empirijsko testiranje i evaluaciju postavljenih pretpostavki.

*Nezavisne varijable* koje se identifikuju za potrebe ovog istraživanja su primarno tehničke prirode i obuhvataju napredni hardver i specijalizovani Python kod neophodan za implementaciju i treniranje dubokih neuronskih mreža. To uključuje korišćenje GPU jedinica visokih performansi za obradu velikih skupova podataka, kao i primenu sofisticiranih biblioteka i okruženja poput PyTorch i Transformers, koji pružaju potrebne alate za konstrukciju, treniranje i fine-tuning naprednih modela veštačke inteligencije [31, 39].

*Zavisne varijable* u istraživanju direktno se odnose na performanse i adaptabilnost GPT-2 modela u višejezičkom kontekstu. Ove varijable se fokusiraju na tačnost i koherenciju generisanog teksta na različitim jezicima, sposobnost modela da se adaptira na specifične jezičke i kulturološke izazove, kao i njegovu efikasnost u kontekstu transfernog učenja. Indikatori koji se koriste za evaluaciju ovih varijabli uključuju, ali nisu ograničeni na, metrike kao što su ocena kvaliteta prevođenja, predviđanja, preciznost i oporavak u zadacima klasifikacije teksta i subjektivna ocena prirodnosti jezika [48, 66].

### 4.1. Generalna (opšta) hipoteza

Generalna (opšta) hipoteza istraživanja u okviru ove doktorske disertacije (HO) glasi: Paradigma GPT-2 modela, uz pomoć transfernog učenja i adaptacije u okvirima PyTorch i Transformers, otvara put ka stvaranju efikasnijih višejezičkih modela obrade prirodnog jezika [38].

Ova opšta (generalna) hipoteza istraživanja proveravaće se kroz sedam posebnih hipoteza koje se tiču određenih segmenata istraživačkog predmeta, od tehničkih detalja implementacije do praktične primene modela u višejezičkom svetu i to:

### 4.2. Posebne hipoteze istraživanja

**Hipoteza 1:** Primena GPT-2 modela sa adaptiranim transfernim učenjem na različitim hardverskim platformama dovodi do značajnog poboljšanja u generisanju prirodnog teksta na manje zastupljenim jezicima.

Ova hipoteza istražuje kako kvalitet hardvera utiče na sposobnost GPT-2 modela da efikasno generiše tekstove na različitim jezicima:

*Indikatori istraživanja:*

- Učinkovitost PyTorch-ove `torch.device` funkcije u alokaciji resursa (GPU/CPU) za optimizaciju generisanja teksta [39].

- Analiza BLEU skora i semantičke koherentnosti generisanog teksta u kontekstu različitih hardverskih konfiguracija [48].
- Merenje vremena i tačnosti generisanja teksta na manje zastupljenim jezicima u odnosu na hardversku platformu [35].

**Hipoteza 2:** Modifikacije u arhitekturi PyTorch i Transformers biblioteka, uključujući optimizaciju scheduler-a i memorije, dovode do efikasnijeg procesa učenja GPT-2 modela.

Ovom posebnom hipotezom istražuje se kako modifikacije u arhitekturi i optimizacija memorije unapređuju efikasnost GPT-2 modela:

*Indikatori istraživanja:*

- Merenje efikasnosti modifikovane Trainer klase u smanjenju vremena treninga i eval\_loss funkcije [41].
- Analiza uticaja AdamW optimizatora i cikličnog learning rate schedulera na performanse modela [50].
- Upoređivanje performansi i efikasnosti memorije između standardnih i modifikovanih arhitektura PyTorch-a i Transformers-a [56].

**Hipoteza 3:** Integracija prilagođenih metrika kao što su tačnost i perpleksija u proces evaluacije GPT-2 modela unapređuje njihovu sposobnost za višejezičku obradu.

Treća posebna hipoteza fokusira se na uticaj prilagođenih metrika u evaluaciji višejezičkih sposobnosti modela:

*Indikatori istraživanja:*

- Procena tačnosti modela (compute\_accuracy funkcija) u višejezičkom kontekstu [64].
- Analiza perpleksije modela (compute\_perplexity funkcija) u obradi različitih jezičkih podataka [57].
- Praćenje performansi modela u Weights & Biases (Wandb) za različite jezičke skupove podataka [55].

**Hipoteza 4:** Optimizacija hiperparametara i prilagođavanje procesa tokenizacije povećavaju efikasnost GPT-2 modela u generisanju i obradi teksta na različitim jezicima.

U četvrtoj posebnoj hipotezi dokazuje se kako optimizacija hiperparametara i tokenizacije poboljšava obradu teksta:

*Indikatori istraživanja:*

- Korelacija između optimizacije hiperparametara (Optuna) i performansi modela u validacionim setovima [36].
- Efikasnost prilagođenog **AutoTokenizer**-a u poboljšanju generisanja koherentnog teksta [68].
- Merenje uticaja prilagođenog **CustomDataCollatorForLanguageModeling** na performanse modela [55].

**Hipoteza 5:** Optimizacija tokenizacije i upravljanje memorijom doprinose efikasnosti GPT-2 modela u obradi velikih skupova podataka, čak i uz ograničene resurse GPU memorije.

Ova hipoteza dokazuje kako prilagođena tokenizacija i strategije upravljanja memorijom omogućavaju efikasniju obradu podataka i smanjuju opterećenje na GPU, posebno u situacijama s ograničenim resursima:

*Indikatori istraživanja:*

- Efikasnost tokenizacije u obradi raznovrsnih tekstova i redukciji vremena potrebnog za obradu [56].
- Analiza smanjenja opterećenja na GPU i performansi modela prilikom validacije na smanjenom uzorku [54].
- Ocena uticaja optimizovanih strategija upravljanja memorijom na efikasnost i preciznost modela u stvarnom vremenu [35].

**Hipoteza 6:** Kvalitet i raznovrsnost skupova podataka ključni su za efikasnost i tačnost GPT-2 modela u višejezičkim aplikacijama.

Ova hipoteza istražuje kako različiti aspekti skupova podataka, uključujući njihovu veličinu, raznovrsnost i kvalitet, utiču na performanse GPT-2 modela u obradi i generisanju teksta na različitim jezicima. Fokus je na utvrđivanju veze između karakteristika skupova podataka i sposobnosti modela da precizno i efikasno obavlja zadatke u višejezičkom okruženju.

*Indikatori istraživanja:*

- Veličina i raznovrsnost podataka: Analiza kako veličina i heterogenost skupova podataka utiču na sposobnost modela da se prilagodi i generiše tekst na različitim jezicima [63].
- Kvalitet podataka: Procena kako čistoća i relevantnost podataka utiču na tačnost i koherentnost generisanog teksta modelom GPT-2 [42].
- Balansiranje skupova podataka: Istraživanje kako balansiranje skupova podataka prema različitim jezičkim grupama može poboljšati višejezičke sposobnosti modela [66].
- Uticaj podataka na transferno učenje: Analiza kako karakteristike skupova podataka utiču na efikasnost primene tehnika transfernog učenja u višejezičkim aplikacijama [34].

**Hipoteza 7:** Primena GPT-2 modela sa naprednim tehnikama transfernog učenja u realnim višejezičkim aplikacijama doprinosi poboljšanju međukulturne komunikacije i pristupačnosti informacija.

Sedma posebna hipoteza istraživanja ispituje kako napredne tehnike transfernog učenja unapređuju međukulturnu komunikaciju i pristupačnost u višejezičkim aplikacijama:

*Indikatori istraživanja:*

- Evaluacija uticaja optimizovanog GPT-2 modela na efikasnost komunikacije u višejezičkim digitalnim platformama [69].
- Ocena korisničkog iskustva zasnovana na prilagođenim metrikama (compute\_accuracy, compute\_perplexity) u aplikacijama [55].
- Analiza angažovanja korisnika i pristupačnosti informacija u aplikacijama koje koriste GPT-2 modele [61].

Ove hipoteze i indikatori direktno odražavaju sadržaj istraživačkog rada kao prethodnice ove doktorske disertacije, omogućavajući detaljniju analizu i evaluaciju, kao i tehničke izazove i rešenja povezane s

obradom velikih skupova podataka prilikom istraživanja u kontekstu višejezičkih jezičkih modela i GPT-2 paradigme.

## 5. NAČIN (METODE) ISTRAŽIVANJA

Svako naučno istraživanje, da bi postiglo željene rezultate, mora biti sprovedeno kroz dobro definisanu metodologiju koja omogućava temeljno proučavanje relevantnog problema. Ovom doktorskom disertacijom istražuje se metodološki okvir koji kombinuje teorijska saznanja i praktične primene u oblasti veštačke inteligencije i obrade prirodnog jezika. Fokus je na razumevanju evolutivnog puta tehnologija obrade jezika, od statističkih modela do naprednih praksi dubokog učenja, s posebnim akcentom na PyTorch i Transformers arhitekture [55, 67]. Ovo poglavlje detaljno razmatra kako te tehnologije omogućavaju implementaciju transfernog učenja i adaptacije modela za višejezičke aplikacije, istražujući kako one prevazilaze jezičke barijere i resursne ograničenosti [57, 66].

Metodološki pristup disertacije obuhvata analizu tehnika dubokog učenja u kontekstu višejezičkih modela, istražujući kako se teorijska znanja mogu primeniti za razvoj efikasnijih komunikacionih sistema u multikulturalnom okruženju. U disertaciji se takođe istražuju etičke i društvene dimenzije primene ovih tehnologija i težište stavlja na stvaranje inkluzivnih i pristupačnih NLP Sistema [74]. Cilj je stvoriti sveobuhvatan uvid kako praktična primena ovih naprednih tehnika može doprineti boljem razumevanju i obradi prirodnog jezika u različitim jezičkim kontekstima, postavljajući temelje za buduća istraživanja i inovacije u ovom brzorastućem polju [77].

### 5.1. Osnovne metode saznanja i istraživanja

Osnovne metode saznanja i istraživanja u ovoj disertaciji usredsređuju se na integraciju i primenu naprednih tehnika u oblasti veštačke inteligencije i obrade prirodnog jezika. Primarni fokus je na dubinskom razumevanju teorijskih koncepta i tehnoloških inovacija unutar PyTorch i Transformers arhitektura, s ciljem identifikovanja i analiziranja ključnih faktora koji doprinose efikasnosti i adaptabilnosti višejezičkih modela [55, 56]. Metodološki okvir uključuje detaljno istraživanje procesa transfernog učenja i njegovih primena u višejezičkim kontekstima, omogućavajući sveobuhvatno razumevanje kako se teorijska znanja mogu efikasno primeniti u praksi za unapređenje i razvoj NLP sistema [78].

U tu svrhu, iz korpusa osnovnih metoda saznanja i istraživanja, u doktorskoj disertaciji će biti primenjene gotovo sve osnovne analitičke i sintetičke metoda saznanja, s akcentom na metodama analize, sinteze, konkretizacije, klasifikacije, generalizacije i induktivno - deduktivnoj metodi saznanja [63].

### 5.2. Opštenaučne metode istraživanja

U okviru doktorske disertacije, biće primenjene neke od ključnih opštenaučnih metoda koje su od suštinskog značaja za sveobuhvatno istraživanje u oblasti veštačke inteligencije i obrade prirodnog jezika. Ove metode uključuju komparativnu metodu, koja omogućava detaljnu analizu i poređenje različitih pristupa i tehnika; statističku metodu, koja pruža rigorozan okvir za kvantitativnu evaluaciju performansi modela [74].

*Komparativna metoda* u ovom istraživanju primenjena je kako bi se analizirale i uporedile različite arhitekture i pristupi unutar domena veštačke inteligencije i obrade prirodnog jezika. Kroz poređenje različitih modela i tehnika, istraživanje teži da identifikuje ključne faktore koji utiču na performanse i adaptabilnost modela, posebno u kontekstu višejezičnosti i kros-kulturne komunikacije [68]. Ova metoda omogućava dubinsko razumevanje prednosti i ograničenja različitih pristupa u obradi jezika.



*Statistička metoda* se koristi za kvantitativnu analizu i evaluaciju modela, omogućavajući precizno merenje i poređenje performansi različitih tehnika i pristupa [69]. Uključuje detaljnu statističku analizu podataka, kao što su tačnost, perpleksija, BLEU skorovi i drugi relevantni metrički podaci, kako bi se objektivno ocenila efikasnost i pouzdanost višejezičkih modela [79]. Ova metoda pruža ključne uvide u efikasnost modela u realnim aplikacijama, doprinoseći boljem razumevanju i unapređenju NLP tehnologija.

### 5.3. Metode za prikupljanje podataka

U doktorskoj disertaciji biće primenjene i metode za prikupljanje relevantnih podataka neophodnih za istraživanje u domenu veštačke inteligencije, pre svega *metodu analize sadržaja dostupnih izvora* kao dokumenata, za dubinsku analizu i obrade prirodnog jezika kako bi se kreirali skupovi podataka (datasets) [75]. Ova metoda će biti fokusirana i u procesu prikupljanja kvalitetnih i reprezentativnih jezičkih podataka koji su ključni za treniranje i evaluaciju višejezičkih modela poput GPT-2.

*Metodom analize sadržaja dokumenata* biće obuhvaćen širok spektar izvora i tehnika, uključujući, ali ne ograničavajući se na, digitalne korpuse jezika, društvene mreže, publikacije, i jezičke baze podataka [73]. Posebna pažnja biće posvećena diversifikaciji jezičkih setova kako bi se obezbedilo da modeli mogu efikasno da procesuiraju i generišu tekst na različitim jezicima, posebno onim koji su manje zastupljeni u postojećim jezičkim resursima [65].

Osim toga, metodologija uključuje i procese čišćenja i pripreme podataka, što je ključni korak u osiguranju kvaliteta i relevantnosti podataka za treniranje modela. Ovaj proces uključuje filtriranje neželjenih ili nepreciznih podataka, standardizaciju formata i tokenizaciju teksta, što je od suštinskog značaja za osiguranje da su podaci pripremljeni na način koji je pogodan za obradu od strane modela dubokog učenja [53].

Ovim metodama, tehnikama i njihovim instrumentima i postupcima biće prikupljeni podaci iz brojnih dokumenata (izvora) o etičkim aspektima pribavljanja i korišćenja podataka, naglašavajući važnost poštovanja privatnosti i autorskih prava, kao i potrebu za odgovornim i transparentnim pristupom u prikupljanju podataka [65]. Primenom ovih metoda u istraživanje, ne samo da se osvetljavaju tehnički aspekte prikupljenih podataka, već i otvara diskusiju o društvenim i etičkim implikacijama koje ovaj proces tako pribavljenih podataka nosi sa sobom [77].

### 5.4. Uzorak dokumenata koji će biti podvrgnut analizi

U okviru ovog istraživanja, fokusiraćemo se na analizu širokog spektra tekstualnih dokumenata kako bismo testirali i verifikovali hipoteze postavljene u disertaciji. Ovi dokumenti su ključni za dubinsko razumevanje primene GPT-2 modela i njihove efikasnosti u višejezičkim okruženjima [69].

**Vrste dokumenata:** Analiza će obuhvatiti raznovrsne tekstualne korpuse koji uključuju naučne radove, digitalne arhive, članke iz novina i časopisa, kao i sadržaj sa društvenih mreža. Ovi dokumenti će biti odabrani kako bi se osigurala reprezentativnost više jezičkih i kulturnih konteksta, s fokusom na različitosti jezičkih struktura [63].

**Kriterijumi za odabir:** Dokumenti su odabrani na osnovu njihove relevantnosti za temu veštačke inteligencije, dubokog učenja, i obrade prirodnog jezika. Poseban naglasak stavljen je na dokumente koji ilustruju upotrebu jezičkih modela u različitim lingvističkim i kulturološkim kontekstima [75].

**Očekivane informacije:** Očekuje se da će analiza ovih dokumenata pružiti uvid u trenutne trendove i izazove u oblasti obrade prirodnog jezika, kao i potencijalne strategije za prevazilaženje jezičkih barijera. Cilj je da se identifikuju ključni faktori koji doprinose uspehu ili ograničavaju efikasnost GPT-2 modela u višejezičkim okruženjima [64].

**Metode analize:** Koristiće se kombinacija kvantitativnih i kvalitativnih metoda za analizu. To uključuje statističku analizu učinkovitosti jezičkih modela u obradi različitih jezika, kao i kvalitativnu sadržajnu analizu za identifikaciju jezičkih specifičnosti i izazova [78].

**Etičke smernice:** Biće preduzeti svi potrebni koraci kako bi se osiguralo poštovanje etičkih standarda, uključujući zaštitu privatnosti i poverljivosti podataka, posebno u slučajevima kada dokumenti sadrže lične informacije [65].

Analiza ovih dokumenata igraće ključnu ulogu u potvrđivanju efikasnosti i adaptabilnosti GPT-2 modela u obradi i generisanju teksta na različitim jezicima, čime će doprineti shvatanju i unapređenju višejezičkih aplikacija u polju veštačke inteligencije i obrade prirodnog jezika [61].

## 6. DRUŠTVENA I NAUČNA OPRAVDANOST ISTRAŽIVANJA

### 6.1. Naučni doprinos

Možemo istaći ključne načine na koje ovo istraživanje pridonosi naučnoj zajednici, posebno u oblasti veštačke inteligencije i obrade prirodnog jezika.

Prvi i primarni naučni doprinos disertacije je razvoj i primena naprednih metoda transfernog učenja koristeći PyTorch i Transformers arhitekture za obradu višejezičkih podataka [72, 73]. Ovo istraživanje demonstrira kako se postojeći modeli, poput GPT-2, mogu prilagoditi i optimizovati za efikasno generisanje i obradu teksta na različitim jezicima, uključujući i one koji su manje zastupljeni [90]. Kroz detaljnu analizu i eksperimentalno testiranje, disertacija pruža uvid u mogućnosti i ograničenja ovih modela, te predstavlja važan korak ka razvoju inkluzivnijih i pristupačnijih NLP sistema [86].

Drugi doprinos je u dubinskom istraživanju i evaluaciji tehnika i strategija za efikasno upravljanje memorijom i resursima, što je ključno u kontekstu velikih i zahtevnih modela dubokog učenja [88]. Prikupljanje, obrada i analiza velikih skupova podataka, posebno u višejezičkim okruženjima, predstavljaju značajne izazove [79]. Ova disertacija nudi nova rešenja i pristupe koji omogućavaju efikasniju obradu i analizu podataka, čime se povećava dostupnost i efikasnost NLP modela [81].

Treći naučni doprinos je u analizi etičkih i društvenih aspekata primene veštačke inteligencije u obradi prirodnog jezika. Disertacija istražuje kako tehnologija može biti upotrebljena na odgovoran način, sa naglaskom na poštovanje privatnosti, pravičnosti i transparentnosti [83]. Razmatra se i uloga ovih tehnologija u prevazilaženju jezičkih barijera, promovisanju međukulturne komunikacije i stvaranju inkluzivnijih digitalnih prostora [84].

Rezultati istraživanja u doktorskoj disertaciji u ovim područjima pružaju značajan doprinos naučnoj oblasti veštačke inteligencije i obrade prirodnog jezika, postavljajući osnovu za buduća istraživanja i inovacije, te otvara nove mogućnosti za primenu ovih tehnologija u različitim sektorima društva [85].

### 6.2. Društveni doprinos

Ovo istraživanje ima značajan uticaj na društvo, posebno u kontekstu višejezičke obrade prirodnog jezika i primene veštačke inteligencije [74]. Ovaj segment istraživanja ne samo da se bavi tehničkim aspektima AI i NLP-a, već i istražuje i naglašava važnost ovih tehnologija u premošćavanju kulturnih i jezičkih barijera u globalizovanom svetu [75].

Prvi društveni doprinos ove disertacije leži u poboljšanju međukulturne komunikacije kroz razvoj višejezičkih modela [77]. Ove tehnologije omogućavaju bolje razumevanje i prevazilaženje jezičkih ograničenja, pružajući alate za efikasniju komunikaciju između različitih jezičkih i kulturnih zajednica [78]. Takav pristup ne samo da olakšava međunarodnu saradnju i razmenu informacija, već i promovise međukulturno razumevanje i toleranciju [82].

Drugi važan aspekt društvenog doprinosa je u inkluziji manje zastupljenih jezika u digitalnom prostoru. Razvoj tehnologija koje podupiru obradu i generisanje teksta na manje zastupljenim jezicima značajno

doprinosi očuvanju jezičke raznolikosti i pruža mogućnost širem spektru ljudi da pristupe informacijama i tehnologiji na svom maternjem jeziku [80].

Treće, istraživanje se bavi i etičkim aspektima primene veštačke inteligencije, naglašavajući potrebu za odgovornim i transparentnim pristupom u razvoju i primeni ovih tehnologija [76]. To uključuje razmatranje pitanja kao što su privatnost, pravičnost i pristupačnost, te stavljanje čoveka u centar tehnološkog razvoja [87]. Takav pristup ima potencijal da utiče na oblikovanje budućih pravila i standarda u svetu veštačke inteligencije, te na promovisanje održivih i etičkih praksi u industriji [89].

Kroz ove doprinose, disertacija demonstrira kako napredak u tehnologiji veštačke inteligencije i obrade prirodnog jezika može imati dubok i pozitivan uticaj na društvo, doprinoseći izgradnji otvorenijeg, inkluzivnijeg i međukulturno povezanijeg sveta [71].

## **II DOPRINOS MAŠINSKOG UČENJA U SAVREMENIM AI SYSTEMIMA**

## 7. OSNOVE I ZNAČAJ MAŠINSKOG UČENJA

Veštačka inteligencija (AI) radikalno menja način na koji živimo, radimo i međusobno komuniciramo, integrišući se u sve segmente ljudske aktivnosti od zdravstva do obrazovanja, transporta i finansija [11]. AI tehnologije, kao što su napredni dijagnostički sistemi, autonomna vozila i personalizovane usluge, postaju sveprisutne, obećavajući značajne dobitke u efikasnosti i kvalitetu života [12]. Međutim, ove tehnologije donose i brojne izazove, uključujući pitanja privatnosti, sigurnosti i etičke dileme koje se javljaju s implementacijom AI sistema [13]. Razumevanje osnovnih principa i tehnologija veštačke inteligencije ključno je ne samo za tehnički napredak već i za navigaciju kroz složene socijalne i etičke pejzaže koje AI oblikuje [11].

Inovacije poput AI u genetici i biotehnologiji otvaraju nove mogućnosti za personalizovanu medicinu, omogućavajući tretmane prilagođene genetskom profilu pojedinca. Ova primena AI postavlja nove etičke dileme, posebno u pogledu pristupa i privatnosti genetskih informacija [13]. Diskusija se proteže i na korišćenje AI u pravosudnim sistemima za predviđanje recidivizma, što postavlja pitanja o pravičnosti i transparentnosti takvih sistema [14].

Uprkos obilju istraživanja posvećenog specifičnim primenama AI i ML, postoji vidljiv nedostatak resursa koji pružaju široku, interdisciplinarnu perspektivu na ove tehnologije [15]. Većina literature fokusira se na tehničke aspekte, dok socijalne implikacije i etičke dileme često ostaju nedovoljno adresirane. Postoji jasna potreba za radovima koji bi nadomestili tehničke i humanističke aspekte AI, nudeći holistički pogled kako tehnologija utiče na društvo i kako se može upravljati njenim razvojem na odgovoran način [15].

Cilj ovog poglavlja disertacije je pružiti temeljit uvod u osnove veštačke inteligencije i mašinskog učenja, istražujući kako se ove tehnologije razvijaju i integrišu u različite aspekte modernog života [11]. Poglavlje će detaljno pregledati ključne tehnologije i algoritme koji omogućavaju funkcionalnosti AI, kao što su neuronske mreže, algoritmi dubokog učenja i sistemi za automatizovano donošenje odluka [16]. Osim toga, biće razmatrani izazovi kao što su pristrasnost u algoritmima, pitanja privatnosti i interpretacija modela [14]. Kroz ovo poglavlje, teži se osnaživanju razumevanja čitalaca o tome kako AI funkcioniše, njenim potencijalima i ograničenjima, ujedno, kako se može koristiti za rešavanje složenih problema na etički prihvatljive načine [17].

Tehnološki napredak u AI nije samo pomerio granice mogućeg u tehničkim disciplinama, već je značajno uticao i na društvene nauke, ekonomiju i umetnost. Primena algoritama mašinskog učenja omogućila je novu perspektivu u analizi velikih podataka, što menja tradicionalne metode istraživanja u sociologiji i političkim naukama, a u umetnosti omogućava stvaranje kompleksnih digitalnih dela koja redefinišu pojmove kreativnosti [18, 19].

Poglavlje počinje sa detaljnim uvodom koji postavlja scenu za dublje razumevanje veštačke inteligencije i mašinskog učenja. Detaljno se obrađuje istorija AI, od ranih eksperimenata i teorijskih osnova do savremenih dostignuća i prelomnih tačaka u razvoju tehnologije. Takođe, fokus je na mašinskom učenju, objašnjavajući različite tipove učenja — nadgledano, nenadgledano, kao i pojačano učenje tj. njihove specifične primene. Nakon toga, poglavlje razmatra integraciju ML u šire AI sisteme i kako to utiče na razvoj inteligentnih aplikacija [20]. Na kraju, detaljno su objašnjeni budući pravci razvoja AI i ML, izazovima koji stoje pred industrijom i potencijalnim društvenim implikacijama koje te tehnologije nose.

Poglavlje se zaključuje razmatranjem kako AI i ML mogu oblikovati budućnost tehnološkog i društvenog razvoja, naglašavajući važnost odgovornog pristupa u razvoju i implementaciji AI tehnologija [15, 17].

## 7.1. Modeli mašinskog učenja u analizi sentimenta

Glavni cilj disertacije je razvoj i implementacija naprednih modela mašinskog učenja za analizu sentimenta tekstova na srpskom jeziku, primenom tehnika prirodnog procesiranja jezika (NLP) [23]. Kroz ovaj proces, teži se kreiranju robustnih prediktivnih algoritama koji mogu efikasno klasifikovati sentimente, prepoznajući pozitivne, negativne, i neutralne aspekte sadržaja [23]. Poglavlje ima za cilj detaljno ispitivanje kako različiti konteksti u kojima se tekstovi pojavljuju, kao što su online recenzije, komentari na društvenim mrežama ili korisničke povratne informacije, utiču na percepciju i analizu sentimenta [23].

### 7.1.1. Metodologija i tehnike u analizi sentimenta

Specifični ciljevi istraživanja uključuju detaljno proučavanje primene nadgledanog, nenadgledanog i pojačanog učenja kako bi se utvrdila najefikasnija metoda za obradu prirodnog jezika u kontekstu srpskog jezika [20]. Takođe, istraživanje ima za cilj da evaluira preciznost i primenljivost modela u stvarnim uslovima, identifikujući potrebne prilagođavanja koje mogu optimizovati performanse modela [19]. Posebna pažnja posvećena je razvijanju metodologija koje će adresirati i korigovati pristrasnost algoritama, osiguravajući pravednu upotrebu tehnologije i jačanje poverenja korisnika u rezultate koje model pruža [14].

Mašinsko učenje se može precizno definisati kao proces u kojem program ili mašina  $M$  uči iz iskustva  $E$  za obavljanje zadatka  $T$ , pri čemu se performanse, merene kroz  $P$ , poboljšavaju [16]. U kontekstu ove disertacije, primena mašinskog učenja se razmatra kroz različite modele i tehnike, od neuronskih mreža koje simuliraju ljudski mozak do logičkih programa koji strukturiraju znanje. Ova razmatranja omogućavaju bolje razumevanje kako se različiti pristupi mašinskog učenja mogu koristiti za efikasnu analizu i interpretaciju podataka, uzimajući u obzir etičke i tehničke izazove u njihovoj primeni [19].

### 7.1.2. Značaj i uticaj istraživanja u NLP tehnologijama

Relevantnost ovog istraživanja ogleda se u njegovom doprinosu tehnološkom napretku u oblasti NLP-a za srpski jezik, koji tradicionalno nije dovoljno zastupljen u globalnim NLP istraživanjima [20]. Poboljšanje algoritama za obradu srpskog jezika može značajno unaprediti analizu velikih količina tekstualnih podataka, što ima praktične implikacije u obrazovanju, medijima i poslovnom sektoru [17]. Ovakvi modeli analize sentimenta mogu poboljšati industrijske aplikacije tako što će kompanijama omogućiti bolje razumevanje emocija i stavova svojih korisnika. Ovo direktno doprinosi poboljšanju korisničkog iskustva i prilagođavanju proizvoda i usluga tržištu [17].

Osim toga, poglavlje se bavi i etičkim aspektima veštačke inteligencije, istražujući pitanja pristrasnosti, privatnosti i transparentnosti u korišćenju AI tehnologija [14]. U kontekstu sve veće društvene zabrinutosti oko uticaja AI na pojedince i zajednice, od suštinskog je značaja razviti pristupe koji će osigurati odgovorno korišćenje AI alata [13]. Očekuje se da će rezultati ovog istraživanja pružiti dragocene uvide koji će omogućiti efikasniju i etički osvešćeniju implementaciju veštačke inteligencije u obradi srpskog jezika, postavljajući osnove za njenu širu primenu u društveno korisne svrhe [13].



**Slika 3.** Algoritam toka procesa mašinskog učenja za analizu sentimenta tekstova na srpskom jeziku

Slika 3. prikazuje detaljan dijagram toka koji ilustruje proces razvoja i primene modela mašinskog učenja za analizu sentimenta tekstova na srpskom jeziku. Dijagram počinje sa "Ulaznim podacima", koji predstavljaju tekstove prikupljene iz različitih izvora. Sledi "Predobrada podataka", koja uključuje čišćenje teksta, tokenizaciju i stemming/lematizaciju. "Izbor metoda mašinskog učenja" usmerava na korišćenje nadgledanog, nenadgledanog i pojačanog učenja, nakon čega se model "Trenira" i "Evaluiira" kroz optimizaciju i validaciju. "Adresiranje pristrasnosti i fer upotrebe" je ključni korak za osiguravanje etičnosti modela, a potom se model "Implementira u praksi" gde se njegove performanse prate i poboljšavaju. Dijagram završava sa segmentom "Etički i društveni aspekti", koji podrazumeva razvoj etičkih smernica i edukaciju zainteresovanih strana.

## 7.2. Strategije i metode mašinskog učenja za analizu sentimenta

Metodologija istraživanja u ovom poglavlju podrazumeva primenu tri glavna pristupa mašinskom učenju - nadgledano, nenadgledano i pojačano učenje - kako bi se razvio robustan model za analizu sentimenta tekstova na srpskom jeziku. Svaki od ovih pristupa će se koristiti za kreiranje specifičnih modela koji se treniraju na različitim nivoima informacija dostupnih u datasetu:



- Nadgledano učenje će koristiti etiketirane podatke za treniranje modela [16];
- Nenadgledano učenje će istraživati skrivene obrasce u neetiketiranim podacima [19];
- Pojačano učenje će se koristiti za optimizaciju modela kroz sisteme nagrađivanja [14].

### **7.2.1. Prikupljanje i predobrada podataka**

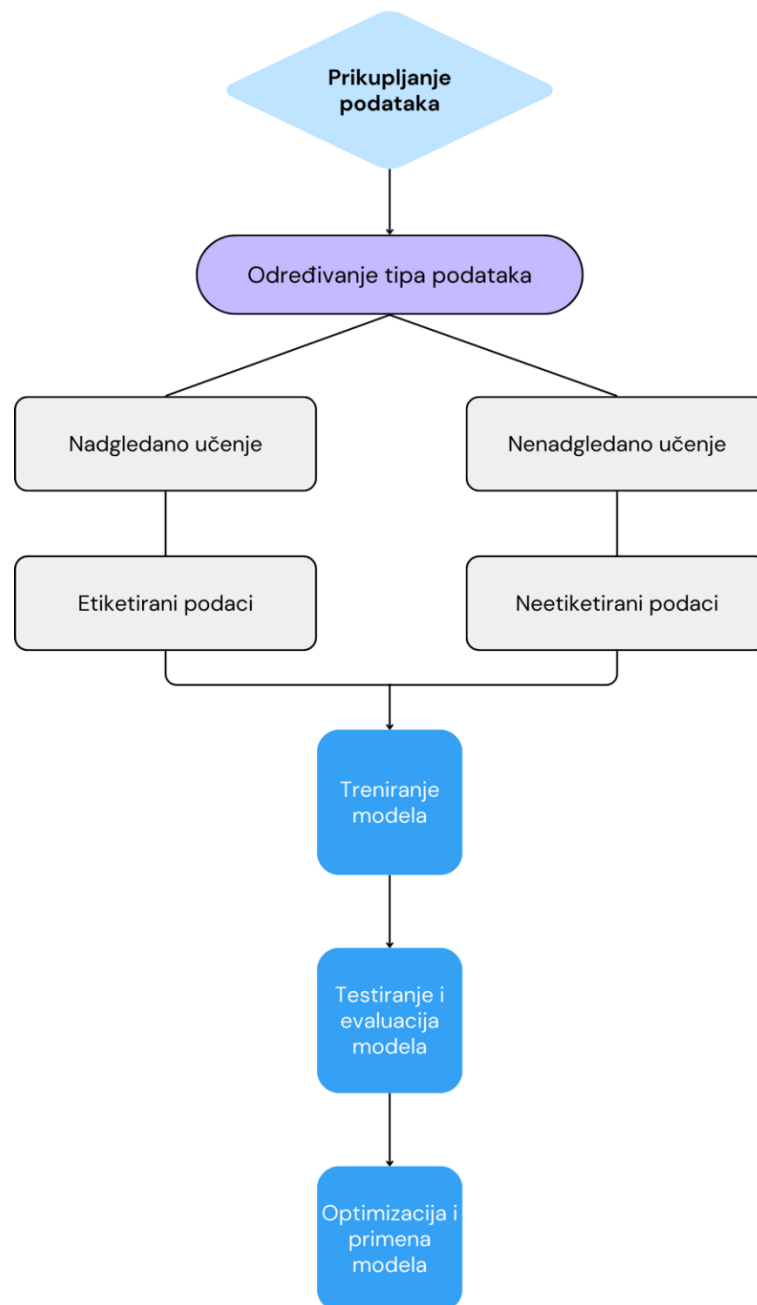
Prikupljanje podataka će se obaviti kroz razne online platforme, uključujući društvene mreže i e-trgovine, gde će se skupljati tekstovi kao što su korisnički komentari i recenzije proizvoda. Predobrada podataka će obuhvatiti čišćenje teksta, tokenizaciju i lematizaciju korišćenjem biblioteka kao što su NLTK i spaCy. Za treniranje modela koristiti se PyTorch i Transformers biblioteke zbog njihove efikasnosti i prilagođenosti disertacije sa jezičkim modelima [19].

### **7.2.2. Tehnike treniranja modela**

Za nadgledano učenje, algoritmi se treniraju koristeći već etiketirane podatke, gde je poznat ishod za svaki primer u trening setu. Na taj način, model uči da prepozna obrasce koji vode do određenog ishoda. Na primer, kada se analizira sentiment teksta, model se trenira na datasetu gde su osećanja u tekstovima već označena kao pozitivna, negativna ili neutralna, omogućavajući modelu da nauči kako da klasifikuje sentimente na osnovu predstavljenih primera [16].

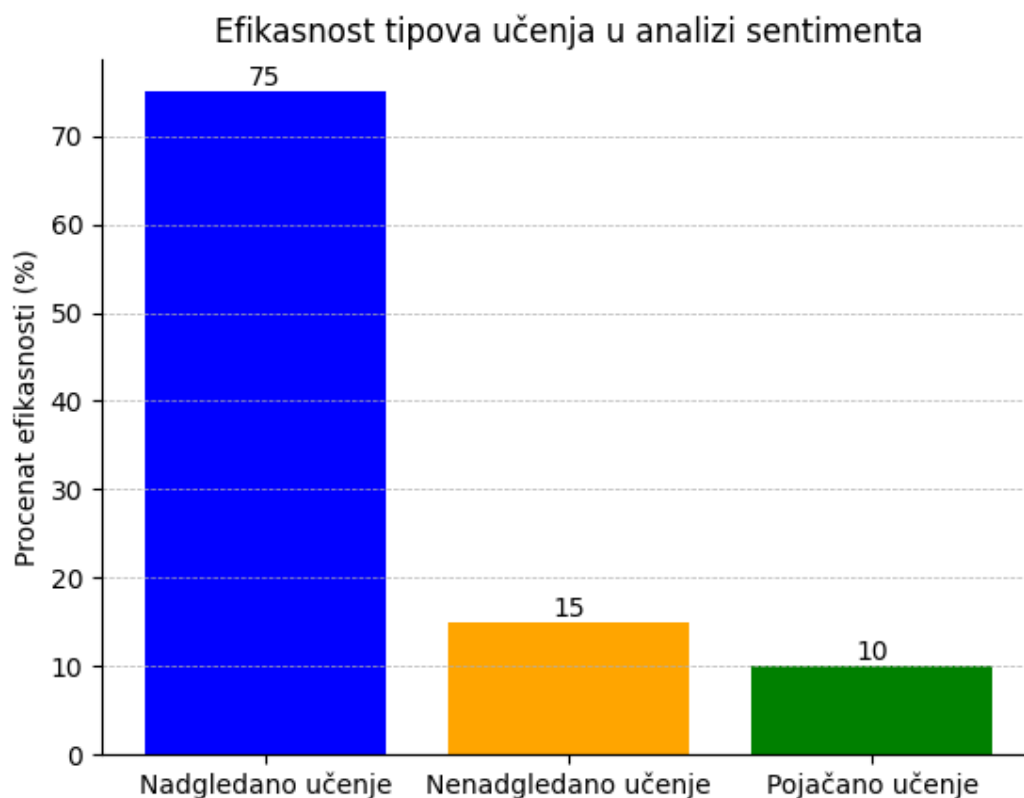
Nenadgledano učenje, s druge strane, ne koristi etiketirane podatke. Umesto toga, algoritmi pokušavaju da otkriju skrivene obrasce ili grupacije unutar podataka bez prethodnih informacija o tome šta ti obrasci predstavljaju. Ovo se često koristi za segmentaciju tržišta ili identifikaciju neuobičajenih podataka [19].

Pojačano učenje se razlikuje od prethodna dva pristupa jer se oslanja na koncept "nagrade" za izvršene akcije. Modeli se treniraju kroz proces isprobavanja i grešaka, gde se svaka akcija ocenjuje kroz sisteme nagrada i kazni [14]. Ovo omogućava modelu da samostalno otkriva strategije koje maksimiziraju nagrade preko vremena, što je posebno korisno u dinamičnim okruženjima gde se odluke moraju donositi u realnom vremenu, kao što su igre ili navigacija robotima [14].



**Slika 4.** Procesno orijentisan prikaz metoda mašinskog učenja u analizi sentimenta

*Slika 4.* pruža procesno orijentisan vizualni prikaz koraka koji su ključni za razvoj modela mašinskog učenja. Svrha analize je sentiment tekstova na srpskom jeziku. Na početku procesa stoji prikupljanje podataka, koje se sprovodi iz mnoštva izvora. Naglašena je važnost raznovrsnosti u podacima. Sledeći korak, određivanje tipa podataka, služi kao raskrsnica gde se bira put daljeg razvoja modela zavisno od prirode podataka. Etiketirani podaci vode ka nadgledanom učenju, dok neetiketirani upućuju na nenadgledano učenje. Centralni deo dijagrama predstavlja treniranje modela, koje je osnova za dobijanje funkcionalnog sistema sposobnog za interpretaciju sentimenta. Nakon treniranja, sledi rigorozna evaluacija i testiranje modela. Proces se završava optimizacijom i konkretnom primenom modela. Ova ilustracija naglašava sekvencijalnu prirodu procesa mašinskog učenja i njegovu struktuiranost koja omogućava metodičan pristup u razvoju algoritama za analizu sentimenta. Takođe, podvlači se važnost svakog koraka u stvaranju pouzdanog i efikasnog modela.



**Slika 5.** Komparativna efikasnost različitih pristupa mašinskom učenju

*Slika 5.* prikazuje komparativnu efikasnost različitih pristupa mašinskom učenju kroz nadgledano, nenadgledano i pojačano učenje, koje su osnova za razvoj modela analize sentimenta u ovom poglavlju. Jasno je da nadgledano učenje prednjači u efikasnosti, što ukazuje na njegovu značajnu ulogu u prepoznavanju i klasifikaciji emocionalnih tonova tekstova. Nenadgledano učenje, iako manje efikasno, pruža vredne uvide u skrivene obrasce unutar neetiketiranih podataka, dok pojačano učenje, iako najmanje zastupljeno, nudi mogućnost za razvoj adaptivnih modela koji se unapređuju kroz interakciju s okruženjem.

**Tabela 1.** Kvantitativni pregled efikasnosti različitih metoda mašinskog učenja koje su primenjene u analizi sentimenta

Metod učenja	Procenat efikasnosti (%)	Boja na grafikonu
Nadgledano učenje	75	Plava
Nenadgledano učenje	15	Narandžasta
Pojačano učenje	10	Zelena

*Tabela 1.* pruža kvantitativni pregled efikasnosti različitih metoda mašinskog učenja koje su primenjene u analizi sentimenta, kako je ilustrovano na *slici 5*. Nadgledano učenje se ističe kao najefikasnije s 75% uspešnosti, što potvrđuje njegovu dominaciju u kontekstu zadatka analize sentimenta. Nenadgledano učenje, iako manje zastupljeno, nudi značajne uvide sa 15%. Pojačano učenje, sa 10%, sugerise potencijal za specifične primene gde je adaptacija na promenljivo okruženje od ključne važnosti.

Ove metode mašinskog učenja su ključne za efikasno modelovanje i razumijevanje kompleksnih datasetova. Primena ovih tehnika unutar ovog istraživanja omogućava precizniju analizu sentimenta,

omogućavajući da se detaljno razume kako različiti aspekti teksta doprinose opštem sentimentu. Razumijevanje ovih metoda omogućava da se optimizuju modeli i postignu bolji rezultati u analizi sentimenta.

Metodologija  $M$  za analizu sentimenta tekstova na srpskom jeziku koristeći mašinsko učenje  $ML$  može se izraziti kao funkcija:

$$M(ML, T, D, P) \quad (1)$$

Gde je:

- $ML$  - skup metoda mašinskog učenja, uključujući nadgledano ( $ML_{supervised}$ ), nenadgledano ( $ML_{unsupervised}$ ), i pojačano učenje ( $ML_{reinforcement}$ );
- $T$  - tipovi podataka koji se koriste za treniranje (npr. etiketirani ili neetiketirani);
- $D$  - dubina predobrade podataka, uključujući čišćenje, tokenizaciju i lematizaciju;
- $P$  - procesi, kao što su treniranje, testiranje, evaluacija i optimizacija modela.

*Formulu 1.* predstavlja proces u kojem metodologija  $M$  koristi različite tipove mašinskog učenja  $ML$  kako bi obradila i interpretirala sentiment u tekstovima. To uključuje prilagođavanje metoda učenja u skladu sa vrstom dostupnih podataka  $T$ , kao i različite faze predobrade  $D$  za pripremu teksta pre treniranja. Proces  $P$  uključuje sve korake od treniranja modela do njegove primene i optimizacije u stvarnim uslovima.

Detaljan prikaz ovih metoda mašinskog učenja u okviru metodologije doprinosi dubljem razumevanju kako različiti pristupi mogu biti primenjeni u realnim scenarijima, obezbeđujući da modeli ne samo da su efikasni već i da su sposobni da se nose sa stvarnim izazovima pri obradi prirodnog jezika [21].

Izbor PyTorch-a kao okruženja za mašinsko učenje je opravdan njegovom fleksibilnošću i dinamičkom grafičkom konstrukcijom, što omogućava intuitivnije eksperimentisanje. Transformers biblioteka se koristi zbog svoje sposobnosti da pruži prethodno trenirane modele koji mogu biti fino podešeni (fine-tuning) na specifičnim skupovima podataka, što je ključno za rad sa srpskim jezikom. Korišćenje NVIDIA Tesla V100 GPU-a je opravdano potrebom za visokim performansama prilikom treniranja i evaluacije kompleksnih modela [21].

Radi osiguranja reproducibilnosti, navode se specifikacije hardvera, verzije softvera i hiperparametri korišćeni tokom istraživanja [11]. Takođe, biće opisan postupak predobrade i karakteristike datasetova kako bi drugi istraživači mogli da replikuju postupak [18].

Etički aspekti će se pažljivo razmatrati u svim fazama istraživanja, od prikupljanja i obrade podataka do prezentacije rezultata, uz posebno razmatranje pristrasnosti i privatnosti [13]. Detalji o koracima koji se preduzimaju za osiguranje fer upotrebe i transparentnosti modela, uključujući strategije za smanjenje pristrasnosti i validaciju, biće jasno izloženi [14]. To podrazumeva implementaciju metoda za otkrivanje i korekciju pristrasnosti, kao što su različite tehnike ravnoteže podataka i revizije modela.

Da bi se osigurala transparentnost u svakom koraku istraživanja, biće objavljene sve faze procesa, od inicijalne pripreme podataka do finalnih eksperimenata i rezultata [22]. Ovaj pristup ne samo da omogućava drugim istraživačima da replikuju i verifikuju rezultate, već i promoviše otvoren dijalog i diskusiju unutar naučne zajednice, podstičući kontinuirano unapređivanje i validaciju metodologije i modela.

Takođe, pripremiće se detaljan plan za testiranje modela u realnim uslovima, uzimajući u obzir varijabilnost tekstova i konteksta u kojima se sentiment analizira [23]. Biće sprovedeni opsežni eksperimenti kako bi se procenila efikasnost modela u različitim aplikativnim scenarijima, a sve to uz strogu kontrolu kvaliteta i u skladu sa etičkim kodeksom.

### **7.2.3. Pristup mašinskom učenju**

Metodologija ovog poglavlja projektovana je da bude sveobuhvatna, temeljita i transparentna, sa jasnim ciljem doprinosa kako akademskoj zajednici, tako i širem društvenom kontekstu [17]. Pružajući detaljan i sistematičan pristup analizi sentimenta u srpskom jeziku, ovo poglavlje stavlja akcenat na inovacije i etičke aspekte veštačke inteligencije, potvrđujući svoj doprinos polju i naglašavajući važnost odgovorne upotrebe AI tehnologija.

## **7.3. Analitički uvidi i dometi modela mašinskog učenja**

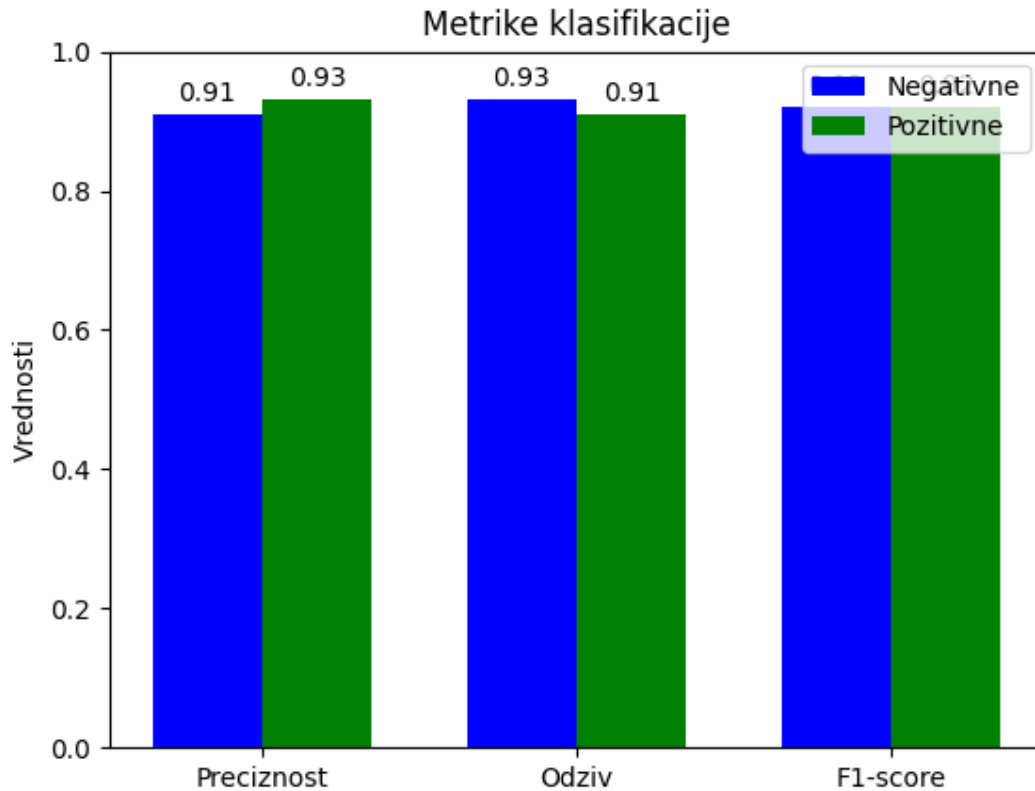
U ovoj disertaciji predstavljani su rezultati istraživanja modela mašinskog učenja za analizu sentimenta u tekstovima na srpskom jeziku. Kroz integraciju sofisticiranih tehnika obrade prirodnog jezika (NLP) i primenu naprednih algoritama, razvijen je model čija je osnovna karakteristika visok procenat tačnosti [16]. Ova značajna karakteristika predstavlja kamen temeljac u validnosti i pouzdanosti dobijenih rezultata, oslikavajući sposobnost modela da precizno klasifikuje sentimente unutar analiziranog jezičkog korpusa.

### **7.3.1 Balans između preciznosti, odziva i F1-scora**

Preciznost modela manifestuje se kroz njegovu sposobnost da konzistentno identifikuje pozitivne i negativne sentimente, pružajući pouzdanu osnovu za dalju analizu [16]. Balans između preciznosti, odziva i F1-scora ukazuje na ravnotežu između sposobnosti modela da prepozna relevantne instance i njegove selektivnosti pri klasifikaciji [16]. Ova ravnoteža je ključna za izbegavanje pristrasnosti prema određenoj kategoriji sentimenta i pruža jasnoću u interpretaciji emocionalnog tonusa teksta.

### **7.3.2. Vizualna ilustracija ključnih metrika performansi**

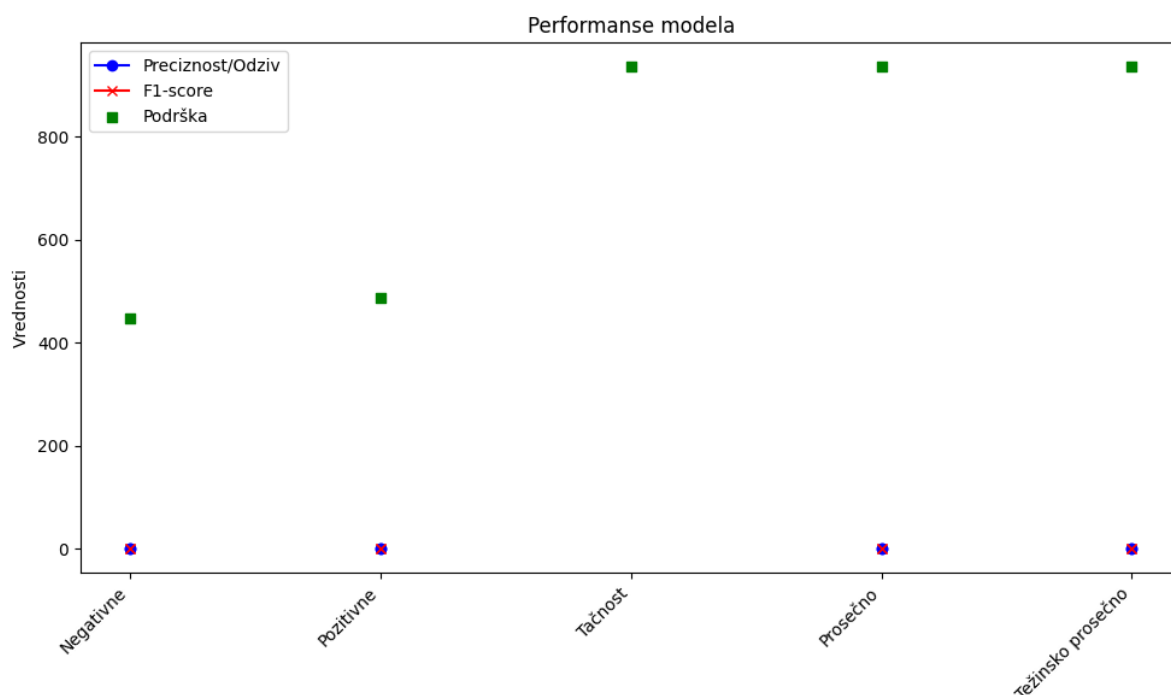
*Slika 6.* ilustruje performanse modela kroz bar grafikon koji upoređuje preciznost, odziv i F1-skor za negativne i pozitivne sentimente. Plave i zelene kolone reprezentuju ove metrike odvojeno za svaku kategoriju, ističući kako model održava skoro identičnu efikasnost u prepoznavanju i negativnih i pozitivnih sentimenta. Vizualna jasnoća ovog grafikona omogućava lakše razumevanje ključnih metrika performansi i ukazuje na visok kvalitet balansiranosti modela, što je neophodna za praktične primene.



**Slika 6.** Bar grafikon metrika klasifikacije po kategorijama sentimenta

### 7.3.3. Analiza podrške i distribucije podataka

Istraživanje je takođe obuhvatilo detaljnu analizu podrške, odnosno broja instanci koje su korišćene za treniranje modela u svakoj kategoriji [16]. Ova komponenta je od izuzetnog značaja jer odražava distribuciju podataka i njihov potencijalni uticaj na performanse modela. Skladnost podrške kroz različite kategorije sentimenta dodatno ističe robusnost modela i njegovu sposobnost da efikasno funkcioniše unutar različitih volumena podataka.



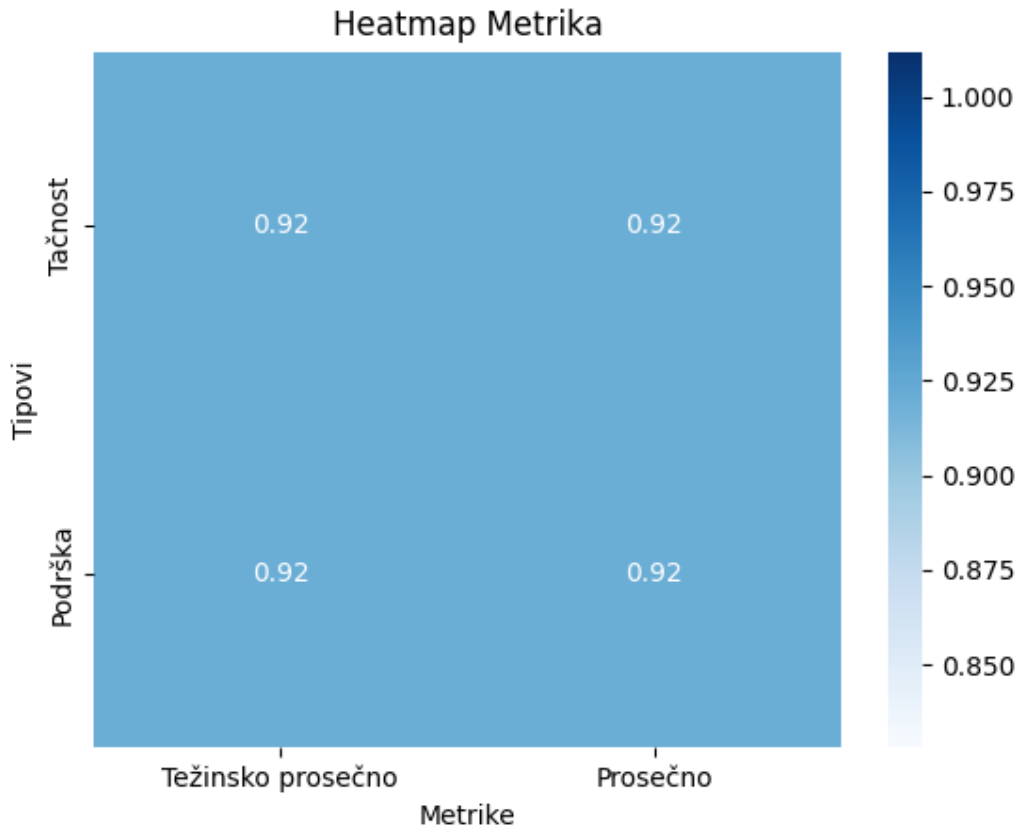
**Slika 7.** Disperzioni grafikon podrške i metrika performansi modela

*Slika 7.* predstavlja disperzioni grafikon koji vizualno povezuje podršku sa preciznošću, odzivom i F1-skorom za obe kategorije sentimenta. Zeleni kvadrati simbolizuju podršku, dok plavi krugovi i crveni krstići predstavljaju preciznost/odziv i F1-skor. Ovaj grafikon efektivno ističe distribuciju i veličinu korišćenih podataka za treniranje modela u odnosu na njegove performanse, naglašavajući robusnost modela kroz ravnotežu između broja instanci i kvaliteta klasifikacije, što se ogleda u konzistentnosti F1-skorova i generalno visokih vrednosti za sve metrike.

### 7.3.4. Robusnost modela i njegova primena

Razvijeni model za analizu sentimenta demonstrira visok nivo preciznosti, neophodan za efikasnu implementaciju u različite industrijske sektore. Takva preciznost omogućava subjektima u oblasti marketinga i korisničkih usluga da dublje razumeju i brže reaguju na emocije i potrebe svojih korisnika [20]. Precizno prepoznavanje sentimenta, na primer, može značajno unaprediti analizu korisničkih recenzija, što direktno doprinosi kvalitetu proizvoda i iskustva korisnika [15]. Robusnost modela nije samo plod primene sofisticiranih NLP tehnika i algoritama, već i rezultat temeljne analize podataka koja ukazuje na to da model podjednako efikasno deluje bez obzira na volumen podataka [24].

Takođe, model se ističe očuvanjem visokog nivoa preciznosti i odziva, karakteristika koja ukazuje na dobru balansiranost [17]. U svetu obrade prirodnog jezika, gde su nijanse sentimenta često suptilne, sposobnost modela da se uspešno kreće kroz ovu kompleksnost predstavlja izraz njegove napredne kognitivne kapacitativnosti i sposobnosti prilagođavanja različitim analitičkim izazovima [16].



**Slika 8.** Heatmap metrika modela za klasifikaciju sentimenta

Slika 8. prikazuje heatmap (toplotnu mapu) koja vizualizuje tačnost modela za analizu sentimenta u tekstovima na srpskom jeziku, gde svetliji tonovi ukazuju na više vrednosti tačnosti za svaku od kategorija. Ova toplotna mapa omogućava brz uvid u performanse modela, prikazujući konzistentnu tačnost od 0.92 kroz sve kategorije. Takav vizualni prikaz olakšava interpretaciju rezultata i odmah ukazuje na stabilnost modela bez obzira na klasu sentimenta koja se razmatra.

Formula 2. opisuje prosečnu ocenu tačnosti (Accuracy Score) modela na osnovu dostupnih kategorija:

$$A = \frac{\sum(TP_i + TN_i)}{\sum(TP_i + TN_i + FP_i + FN_i)} \quad (2)$$

Tačnost se definiše kao odnos ukupnog broja ispravnih predviđanja (tj. istinski pozitivnih i istinski negativnih instanci) prema ukupnom broju svih predikcija koje model čini. U kontekstu analize sentimenta, tačnost reflektuje sposobnost modela da ispravno identifikuje i pozitivne i negativne komentare. Model koji je razvijen u ovom istraživanju postiže visoku tačnost, sa prosečnom vrednošću od 0.92, što ukazuje na visoki stepen pouzdanosti u njegovim predikcijama. Visoka tačnost je posebno značajna u primenama gde je važno precizno razumeti sentiment korisnika, kao što su analize recenzija proizvoda ili usluga.

Rezultati F1-scora, koji predstavlja harmonijsku sredinu preciznosti i odziva, takođe potvrđuju efikasnost modela u klasifikaciji sentimenta [20]. Ovaj skor ukazuje na sposobnost modela da uspostavi balans između otkrivanja relevantnih instanci (odziv) i ispravne klasifikacije tih instanci (preciznost), što je presudno za praktičnu primenu u dinamičkim i stvarnim scenarijima gde se sentiment analizira [20].



Formula 3. opisuje prosečnu vrednost F1-score:

$$F1 = 2 \times \frac{P \times R}{P + R} \quad (3)$$

F1-score predstavlja balans između preciznosti (udela ispravno identifikovanih pozitivnih predikcija u odnosu na ukupan broj pozitivnih predikcija) i odziva (udela ispravno identifikovanih pozitivnih predikcija u odnosu na ukupan broj stvarnih pozitivnih instanci). F1-score je posebno koristan kada su troškovi lažno pozitivnih i lažno negativnih rezultata podjednako visoki, te je stoga harmonijska sredina preciznosti i odziva izuzetno vredan pokazatelj za ovaj model. Dobijeni rezultat ukazuje na to da je model dobro izbalansiran i da može efikasno da se primenjuje u realnim uslovima gde se zahteva pouzdanost u prepoznavanju i klasifikaciji sentimenta.

**Tabela 2.** Detaljna analiza metrika performansi modela za klasifikaciju sentimenta

	Preciznost	Odziv	F1-score	Podrška
Negativne	0.91	0.93	0.92	447
Pozitivne	0.93	0.91	0.92	488
Tačnost modela	0.92			
Tačnost			0.92	935
Prosečno	0.92	0.92	0.92	935
Težinsko prosečno	0.92	0.92	0.92	935

Tabela 2. predstavlja kvantitativni pregled performansi razvijenog modela, detaljno razrađujući preciznost, odziv, F1-score i podršku za kategorije negativnih i pozitivnih sentimenta. Sa tačnošću od 0.92, tabela ilustruje visoku pouzdanost modela u klasifikaciji tekstova po sentimentu. Izjednačene vrednosti F1-scorea za oba sentimenta odražavaju sposobnost modela da efektivno balansira između preciznosti i odziva, što je ključno za smanjenje pristrasnosti i osiguranje objektivne analize. Podrška označava broj instanci na kojima je model obučavan, pokazuje da model postiže konzistentne performanse preko različitih veličina skupova podataka, ukazujući na njegovu robusnost i skalabilnost. Ovi podaci su vitalni za razumevanje kako model funkcioniše u praksi i njegove potencijale za primenu u stvarnim aplikacijama.

U središtu istraživanja bili su etički aspekti i transparentnost u primeni veštačke inteligencije. Uz stroge protokole, istraživanje je obezbedilo fer upotrebu i transparentnost u razvoju modela, te je posebna pažnja posvećena smanjenju pristrasnosti algoritama [13]. Transparentnost metodologije, od inicijalne pripreme podataka do finalnih eksperimenata, omogućava verifikaciju rezultata od strane zajednice i podstiče kontinuirano unapređenje [18]. Kontrola kvaliteta i etički kodeks bili su ključni tokom testiranja modela, omogućujući eksperimentisanje u različitim aplikativnim scenarijima [17].

Transparentnost metodologije je od suštinske važnosti za pouzdanost i repliciranje istraživanja [23]. Javno objavljivanje svih faza procesa, od inicijalne pripreme podataka do finalnih eksperimenata, osigurava ne samo verifikaciju rezultata od strane zajednice već i podstiče kontinuirano unapređenje metodologije i algoritama [16].

Kontrola kvaliteta i etički kodeks bili su ključni tokom testiranja modela u realnim uslovima. Eksperimenti su obuhvatili različite aplikativne scenarije, takođe, pažljivo su birani tekstovi i konteksti kako bi se procenila efikasnost modela u širokom spektru aplikacija [14].

Dakle, rezultati istraživanja ne samo da ohrabruju akademsku zajednicu, već pružaju i vredne uvide za praktičnu primenu u industriji. Detaljno razumevanje aspekata ovog poglavlja disertacije otvara put za unapređenje korisničkog iskustva i prilagođavanje proizvoda i usluga.

Rezultati istraživanja odražavaju kako inovativna primena veštačke inteligencije može doprineti društvenom razvoju, obezbeđujući istovremeno zaštitu individualnih prava [11]. Stavljanjem akcenta na odgovornu upotrebu AI, poglavlje naglašava potencijal veštačke inteligencije da oblikuje budućnost tehnološkog i društvenog razvoja [21]. Uvidi dobijeni iz ovog istraživanja pružaju vredne informacije o praktičnoj primeni modela u stvarnom svetu, kao i o njegovom uticaju na društvo i industriju [12]. Rezultati ne samo da ilustruju statističke podatke i performanse modela, već i podstiču kritičko promišljanje o njihovim širim tehnološkim i društvenim implikacijama, čime se zaokružuje doprinos ovog istraživanja polju veštačke inteligencije [15].

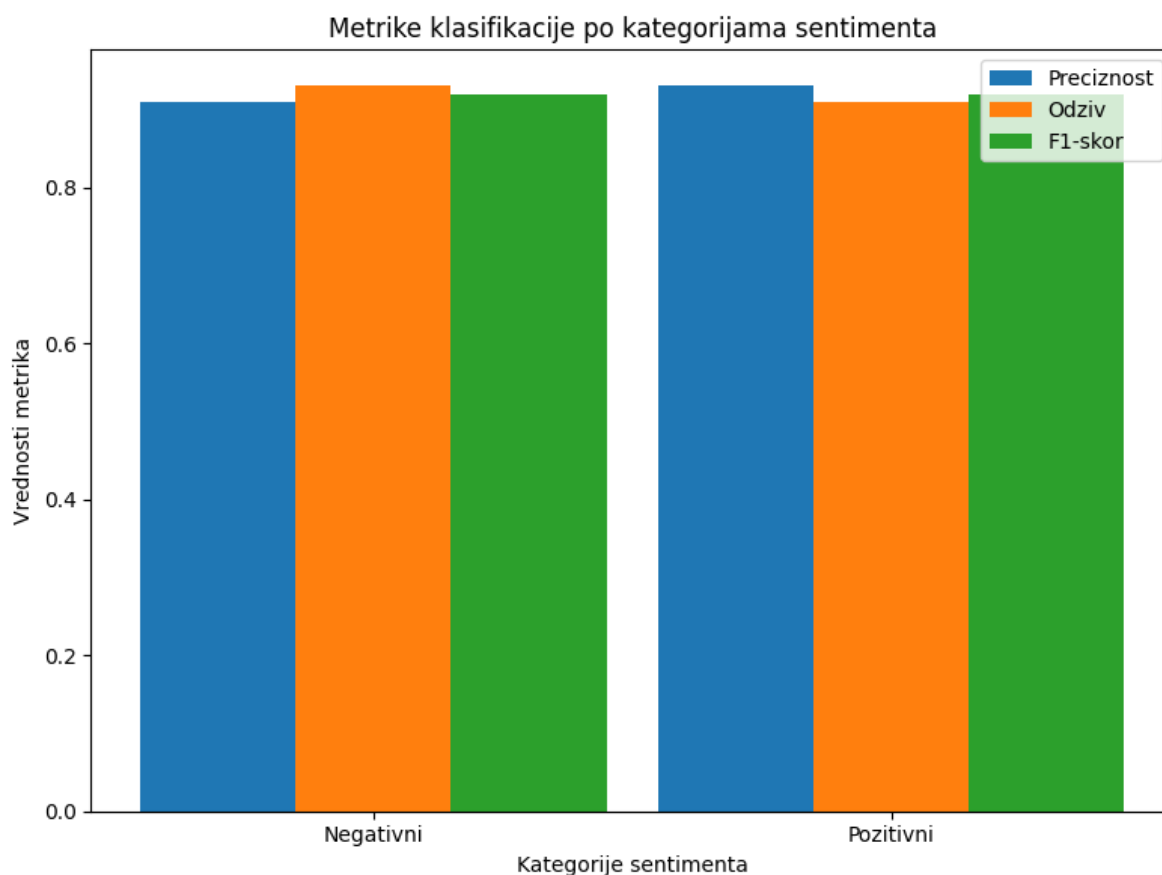
U skladu s tim, ovo poglavlje ne samo da ilustruje statističke podatke i performanse modela, već pruža i dublje razumevanje implikacija koje ovi rezultati imaju na šire tehnološke i društvene pejzaže [19]. Prilagođavanje modela, etička pitanja, i njihova primena u stvarnim svetskim aplikacijama predstavljaju kamen temeljac ovog poglavlja, zaokružujući doprinos istraživanja polju veštačke inteligencije.

#### **7.4. Razmatranje efikasnosti i etičkih implikacija modela**

Preciznost, odziv i F1-score su temeljni pokazatelji kvaliteta ovog modela mašinskog učenja. Visoki nivoi preciznosti pokazuju da je model sposoban da konzistentno prepozna pozitivne i negativne sentimente, što je direktna posledica kvalitetne predobrade i treniranja modela, kako je opisano u metodološkom delu. Ravnoteža između preciznosti i odziva, zajedno s harmonijskim F1-skorom, ističe sposobnost modela da izbegne pristrasnost i da objektivno klasifikuje sentimente [20].

##### **7.4.1. Uticaj metodoloških izbora na performanse**

Analiza rezultata prikazanih na *slici 9*. ukazuje na visoku preciznost i odziv modela, kao i na izuzetno balansirane F1-skorove za obe kategorije sentimenta - negativne i pozitivne. Ovaj grafikon nije samo vizualni prikaz rezultata istraživanja, on je ključan alat za dublje razumevanje kako ovaj model interpretira i razlikuje sentimente unutar tekstualnih podataka [22].



**Slika 9.** Uporedna analiza metrika preciznosti, odziva i F1-skorova za klasifikaciju sentimenta

*Slika 9.* odražava metodološke izbore napravljene tokom istraživanja. Primenom nadgledanog, nenadgledanog i pojačanog učenja, razvijen je model koji je sposoban da se prilagođava i optimizira na osnovu podataka. Fokus na balansiranje pristrasnosti i transparentnosti, kako je detaljno opisano u poglavlju, ogleđa se u skoro identičnim F1-skorovima za obe kategorije sentimenta, što ukazuje na etičku i metodološku rigoroznost pristupa [24].

#### 7.4.2. Etičke dimenzije i transparentnost u primeni modela

Rezultati na grafikonu takođe ilustruju praktičnu vrednost i etičku osvešćenost razvijenog modela. Transparentnost procesa od inicijalne pripreme podataka do finalnih eksperimenata pruža temelj za verifikaciju rezultata i za kontinuirani razvoj. Takav pristup osigurava da se AI koristi odgovorno, sa smanjenim pristrasnostima algoritma, što je ključno za izgradnju poverenja u AI tehnologije među korisnicima [25].

**Tabela 3.** Kvantitativna analiza metrika preciznosti, odziva i F1-skorova za klasifikaciju sentimenta

Kategorija	Preciznost	Odziv	F1-score
Negativni	0.91	0.93	0.92
Pozitivni	0.93	0.91	0.92

U *tabeli 3.* predstavljene su ključne metrike koje ocenjuju performanse modela mašinskog učenja u klasifikaciji sentimenta tekstova na srpskom jeziku. Prikazane vrednosti preciznosti, odziva i F1-skorova ilustruju sposobnost modela da identifikuje negativne i pozitivne sentimente unutar analiziranog korpusa. Preciznost od 0.91 za negativne i 0.93 za pozitivne sentimente ukazuje na efikasnost modela u

pravilnom prepoznavanju i klasifikaciji sentimenta. Odziv, odnosno mera sposobnosti modela da identifikuje relevantne instance, je izjednačen za obe kategorije sa vrednostima od 0.93 i 0.91, što potvrđuje pouzdanost modela u otkrivanju svih relevantnih uzoraka. F1-score, koji je harmonijska sredina preciznosti i odziva, je konzistentan i iznosi 0.92 za oba sentimenta, što dodatno potvrđuje visok kvalitet i balansiranu performansu modela u analizi sentimenta. Ove metrike zajedno pružaju detaljan uvid u pouzdanost i objektivnost modela, kao i u njegovu primenljivost u praktičnim scenarijima analize tekstualnih podataka.

#### **7.4.3. Mogućnosti unapređenja i budući pravci istraživanja**

Iako rezultati pokazuju visok stepen tačnosti i efikasnosti, postoji prostor za unapređenje. Dalje istraživanje može uključiti razvijanje modela sa sposobnošću da se prilagođava dinamičkim okruženjima, posebno u kontekstu online platformi koje kontinuirano generišu nove oblike tekstualnih podataka. Pristupi zasnovani na pojačanom učenju mogu pružiti modelu sredstva za adaptaciju na promenljive obrasce jezika u realnom vremenu. Takođe, dublje istraživanje nenadgledanog učenja može otkriti latentne strukture unutar podataka koji nisu očigledni u etiketiranom trening setu, time doprinoseći boljem razumevanju suptilnih nijansi sentimenta.

#### **7.4.4. Etički i društveni izazovi u primeni AI tehnologija**

Osim tehnoloških unapređenja, važno je dalje ispitati kako primena ovih modela utiče na individualna prava i privatnost, kako se može osigurati da su AI sistemi u službi društva, a ne na njegovu štetu. U tom smislu, važno je razvijati alate za reviziju i validaciju koji mogu proceniti etičnost algoritamskih odluka. Unapređenje transparentnosti i interoperabilnosti modela može omogućiti veću razumljivost i prihvatljivost AI sistema među korisnicima.

Rezultati ovog istraživanja otvaraju put ka razvoju sveobuhvatnijih modela koji ne samo da razumeju tekstualne podatke, već i uzimaju u obzir kontekstualne faktore i socio-kulturne nijanse. Uzimajući u obzir sve ove aspekte, ovo istraživanje pruža čvrstu osnovu za buduće inovacije u oblasti obrade prirodnog jezika, unapređujući ne samo tehničku već i društvenu vrednost veštačke inteligencije [21].

**III PRIMENA GPT-2 MODELA SA ADAPTIRANIM  
TRANSFERNIM UČENJEM NA RAZLIČITIM HARDVERSKIM  
PLATFORMAMA**

## 8. OPTIMIZACIJA GPT-2 MODELA PUTEM ADAPTIVNOG TRANSFERNOG UČENJA NA GPU ARHITEKTURAMA

Razvoj tehnologija za obradu prirodnog jezika i duboko učenje otvara nove mogućnosti u automatskoj obradi jezika, povećavajući sposobnost mašina da razumeju i generišu ljudski jezik. GPT-2 pokazuje impresivne rezultate u širokom spektru aplikacija, od automatskog pisanja teksta do kreiranja sintetičkih odgovora u dijalogu [26]. Ovo poglavlje pruža ciljani uvid u optimizaciju GPT-2 modela zasnovano na GPU-u. Adaptivno transferno učenje, primenom unapred obučeni modela na specifične zadatke, može značajno poboljšati performanse modela bez potrebe za opsežnim skupovima podataka ili produženom obukom [27].

Napredak obrade prirodnog jezika i tehnologija dubokog učenja nastavlja da oblikuje budućnost interakcije čoveka i mašine, premošćujući jaz koji je tradicionalno postojao u razumevanju ljudskog jezika od strane veštačke inteligencije. Generativni unapred obučeni transformator 2 (GPT-2), pionir u generisanju teksta, stoji na čelu ove revolucije, pokazujući izuzetne sposobnosti u kreiranju koherentnih i kontekstualno relevantnih tekstova [28]. Međutim, njegova primena kroz adaptivno transferno učenje na različitim GPU arhitekturama otvara nova pitanja u vezi optimizacije i efikasnosti. Ovo poglavlje ima za cilj da otkrije kako prilagođavanje i fino podešavanje GPT-2 modela za specifične hardverske arhitekture može poboljšati njegove performanse i energetske efikasnost, doprinoseći razvoju i primeni naprednih AI sistema [29].

Upotreba GPT-2 modela u različitim kontekstima otkriva značajan prostor za istraživanje njegove prilagodljivosti i optimizacije u različitim hardverskim okruženjima [30]. Kako industrija veštačke inteligencije nastavlja brzo da se razvija, potreba za dubljim razumijevanjem interakcije između AI modela i hardverske infrastrukture postaje sve izraženija. Ovo poglavlje istražuje tu sinergiju, sa ciljem da pruži uvid u povećanje efikasnosti kroz preciznu analizu uticaja hardverskih karakteristika na performanse modela. Istraživački pristup kombinuje teorijske osnove sa eksperimentalnim testiranjem, omogućavajući sveobuhvatno razumevanje potencijala i ograničenja adaptivnog transfernog učenja u kontekstu AI modeliranja.

Kroz pažljivu analizu i poređenje performansi GPT-2 modela na različitim hardverskim platformama, ovo poglavlje nudi nove uvide u optimizaciju procesa učenja i primenu modela. Pronalaženjem ravnoteže između tačnosti, brzine obrade i energetske efikasnosti, istraživanje pruža vredne smernice za buduće implementacije GPT-2 modela u stvarnim aplikacijama, od digitalnih asistenata do automatske proizvodnje sadržaja. Ovaj pristup utire put za stvaranje efikasnijih i održivijih AI sistema sposobnih za prilagođavanje i optimizaciju za širok spektar aplikacija, obećavajući revolucionarni napredak u načinu na koji razumemo i koristimo tehnologiju veštačke inteligencije u svakodnevnom životu [31].

Dok je GPT-2 model široko proučavan, postoji značajan jaz u literaturi u pogledu detaljne analize njegove primene kroz adaptivno transferno učenje na različitim hardverskim platformama [26]. Osim toga, iako je poznato kako specifične karakteristike hardvera, kao što su procesorska snaga i kapacitet memorije, utiču na efikasnost procesa transfernog učenja, malo se zna o specifičnim metodama za optimizaciju ovih karakteristika kako bi se poboljšala efikasnost adaptivnog transfernog učenja. Ovo poglavlje ima za cilj da premosti ovaj jaz u znanju posebnim istraživanjem performansi GPT-2 modela na različitim GPU arhitekturama u smislu tačnosti, brzine obrade i energetske efikasnosti, naglašavajući

komparativne prednosti računanja zasnovanih na GPU-u u unapređenju efikasnosti adaptivnog transfornog učenja [30].

Primarni cilj ovog istraživanja je analiziranje uticaja adaptivnog transfornog učenja na performanse GPT-2 modela, posebno ispitivanje niza GPU arhitektura kako bi se razumelo kako GPU mogućnosti mogu optimizovati efikasnost modela. Eksperimentalnom metodom, ovaj rad daje poređenje performansi modela, sa posebnim fokusom na analizu kako karakteristike hardvera utiču na efikasnost učenja. Doprinosi ovog poglavlja uključuju:

- Detaljna analiza performansi GPT-2 modela u nizu GPU arhitektura, naglašavajući nijanse kako GPU mogućnosti utiču na efikasnost i rezultate adaptivnog transfornog učenja [27];
- Uvid u optimizaciju GPT-2 modela za specifične hardverske platforme [29];
- Temelj za dalji razvoj i unapređenje tehnologija veštačke inteligencije i mašinskog učenja [28].

Detaljno se istražuju teorijske osnove GPT-2 modela i principa transfornog učenja, naglašavajući njihov značaj i primenjivost u kontekstu veštačke inteligencije. Kroz metodološki segment koji ocrtava eksperimentalni pristup, uključujući izbor hardverskih platformi i kriterijuma za ocenu performansi modela, ovo poglavlje postavlja temelj za duboko razumijevanje istraživanja. Prezentacija i analiza rezultata istraživanja zauzimaju središnji deo, gde se kroz komparativnu analizu istražuju efekti adaptivnog transfornog učenja na različite hardverske konfiguracije. Diskusija proširuje perspektivu prema praktičnim implikacijama nalaza, razjašnjavajući kako rezultati mogu uticati na buduću primenu GPT-2 modela u stvarnom svetu. Ovo poglavlje kulminira zaključkom koji sjedinjuje ključne uvide stečene istraživanjem i nudi smernice za dalje istraživačke napore u području veštačke inteligencije i mašinskog učenja, čime se zatvara petlja istraživačkog procesa i otvaraju novi smerovi za buduća istraživanja.

## 8.1. Kvantifikacija uticaja GPU performansi na GPT-2 model u adaptivnom transfornom učenju

Primarni cilj ovog istraživanja je poboljšanje efikasnosti GPT-2 modela putem adaptivnog transfornog učenja kroz različite GPU postavke. Fokus se sužava na kvantifikaciju uticaja GPU računarskih sposobnosti i memorijskih specifikacija na ključne pokazatelje performansi modela, kao što su tačnost, brzina obrade i energetska efikasnost. Jedinstvena, eksperimentalna faza integriše procenu ovih karakteristika GPU-a kako bi se utvrdile optimalne konfiguracije za efikasno generisanje teksta.

### 8.1.1. Eksperimentalna analiza efikasnosti GPU

Koristeći PyTorch-ovo računarsko okruženje, ovo poglavlje provodi praktične evaluacije, poboljšane Optunom za optimizaciju hiperparametara i WandB-om za vizualizaciju procesa. Nadilazeći konvencionalne pristupe, poglavlje obuhvata multidisciplinarni put, spajajući teorijske okvire AI sa pragmatičnim eksperimentima kako bi se poboljšali GPT-2 modeli usred zamršenog pejzaža GPU arhitektura [32].

$$Performance_{model} = f(Power_{GPU}, Memory_{GPU}, Hyperparameters) \quad (4)$$

Formula 4. služi kao kvantitativni okvir za procenu efikasnosti modela u kontekstu adaptivnog transfornog učenja na različitim GPU arhitekturama. Kvantifikuje efikasnost kroz metrike kao što su tačnost, brzina obrade i energetska efikasnost. Hardverske karakteristike GPU-a,  $Power_{GPU}$  i  $Memory_{GPU}$ , direktno utiču na ove performanse, dok prilagođavanje hiperparametara modela,

uključujući brzinu učenja, veličinu serije i drugo, optimizuje model za maksimalne performanse. Ova formula omogućava objektivnu analizu uticaja hardverskih konfiguracija i optimizacije hiperparametara na ključne pokazatelje performansi GPT-2 modela, naglašavajući kako precizno prilagođavanje modela karakteristikama određenih GPU arhitektura može poboljšati njegovu efikasnost, što je direktno povezano sa ciljevima ovog istraživanja.

**Tabela 4.** Uporedne metrike performansi GPT-2 modela u različitim GPU konfiguracijama

GPU konfiguracija	Snaga (TFLOPS)	Memorija (GB)	Stopa učenja	Tačnost (%)	Brzina obrade (ms po uzorku)	Energetska efikasnost (vat/sat)
GPU A	12	8	0.01	92	15	250
GPU B	14	16	0.01	94	12	300
GPU C	10	8	0,005	90	20	200
GPU D	11	8	0,008	93	14	225

*Tabela 4.* ilustruje rezultate za četiri različite GPU konfiguracije, uključujući dve sa 8 GB memorije (GPU A i GPU D), i kako oni utiču na performanse GPT-2 modela. Dodavanje GPU konfiguracije D sa 11 TFLOPS i 8 GB memorije pokazuje kako manja podešavanja snage i brzine učenja mogu poboljšati tačnost i brzinu obrade uz održavanje relativno visoke energetske efikasnosti. Ovo naglašava važnost pronalazjenja optimalne ravnoteže između snage GPU-a, memorije i hiperparametara kako bi se maksimizirale performanse modela uz efikasnu potrošnju energije, što je ključni cilj ovog istraživanja.

### 8.1.2. Primena i održivost modela u praksi

U potrazi za optimizacijom, preduzima se strateško prilagođavanje hiperparametara, s posebnim naglaskom na dinamičku modulaciju brzine učenja i preventivno suzbijanje prekomernog prilagođavanja modela [33]. Nakon toga, poglavlje se okreće fazi empirijske validacije, gde se testira svestranost modela u odnosu na raspon teorijskih i simuliranih primena, naglašavajući njegovu operativnu agilnost [34].

Procena održivosti modela u stvarnom svetu kulminira u eksperimentalnom ispitivanju, gde se njegova robustnost ispituje kroz različite praktične primene. Analitički zaron u reciprocitet GPU performansi naglašava želju za otkrivanjem hardverskih konfiguracija koje podižu efikasnost, smanjujući potrebu za obimnim skupovima podataka ili dugotrajnim epohama obuke [35].

Krajnji cilj ovog poglavlja je i procena primenjivosti optimiziranog GPT-2 modela u stvarnom svetu. Eksperimentalno testiranje omogućava proveru prilagodljivosti i efikasnosti modela u širokom rasponu stvarnih scenarija, naglašavajući njegovu vrednost i fleksibilnost. Takva analiza pruža ključni uvid u potencijalne primene i prednosti GPT-2 modela, doprinoseći razvoju efikasnijih i dostupnijih rešenja u području veštačke inteligencije [36].

Postizanjem ovih ciljeva, poglavlje daje značajan doprinos području veštačke inteligencije, posebno u kontekstu adaptivnog transfernog učenja. Analiza uticaja optimizacije GPT-2 modela u različitim GPU arhitekturama ne samo da unapređuje teoretsko razumevanje već i otvara nove puteve za praktičnu primenu, doprinoseći efikasnosti i održivosti budućih AI sistema [37]. Ovim istraživanjem postavlja se



čvrst temelj za budući razvoj i unapređenje tehnologija veštačke inteligencije, omogućavajući stvaranje naprednijih i prilagodljivijih rešenja sa širokom društvenom primenom [34].

Ova disertacija ima potencijal da značajno doprinese industriji i društvu baveći se ključnim izazovima kao što je potreba za efikasnijim AI modelima sa manjim energetskegim uticajem. Analizom uticaja optimizacije modela na GPT-2 u različitim GPU arhitekturama, ovo istraživanje pruža uvid u mogućnosti za ubrzavanje inovacija u različitim sektorima, uključujući zdravstvo, obrazovanje i druga područja od ljudskog interesa, čime se doprinosi poboljšanju kvaliteta života i poslovna efikasnost [34].

## 8.2. Metodološki pristup u evaluaciji efikasnosti GPT-2 Modela na GPU platformama

Ova disertacija koristi sveobuhvatan eksperimentalni pristup za procenu efikasnosti GPT-2 modela sa adaptivnim transfornim učenjem kroz različite GPU arhitekture. Cilj je istražiti kako varijacije u snazi GPU-a i memoriji utiču na performanse ključnih modela, uključujući tačnost, brzinu obrade i energetske efikasnost [38].

### 8.2.1. Planiranje eksperimentalnog dizajna

Eksperimentalni dizajn počinje odabirom relevantnih podataka iz različitih izvora, uključujući tekstualne korpuse prirodnog jezika i sintetički generisane tekstove, kako bi se procenile mogućnosti GPT-2 modela u različitim kontekstima. Nakon početnog odabira podataka, proces uključuje segmentaciju podataka za obuku, validaciju i testiranje kako bi se osigurala sveobuhvatna evaluacija modela. Metode evaluacije obuhvataju kvantitativnu analizu tačnosti, brzine obrade i energetske efikasnosti modela koristeći definisane metrike performansi.

### 8.2.2. Implementacija i analiza metrika performansi

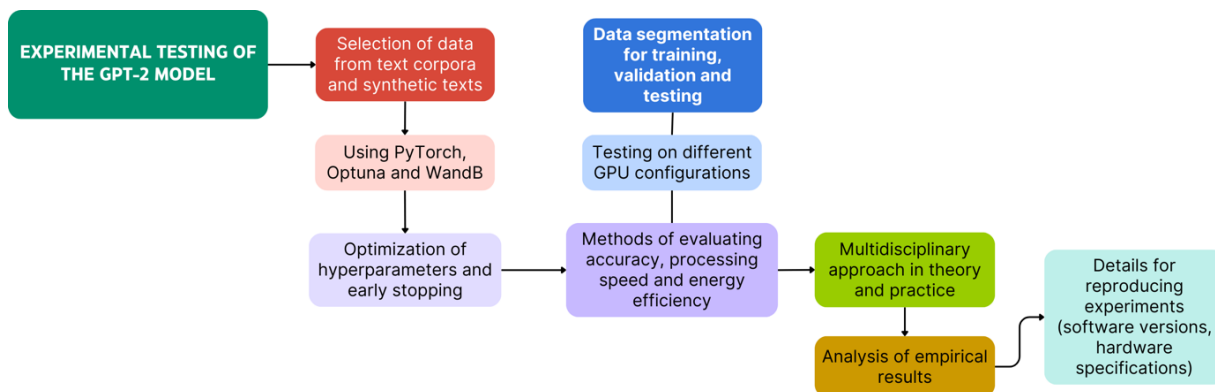
$$Efficiency_{model} = \frac{Accuracy \times Speed}{PowerUsage \times MemoryUsage} \quad (5)$$

*Formula 5.* je uvedena kako bi se kvantifikovala efikasnost GPT-2 modela kroz adaptivno transforno učenje kroz različite hardverske konfiguracije. *Formula 5.* razmatra efikasnost modela kao odnos tačnosti i brzine obrade, normalizovan potrošnjom energije i upotrebom memorije. Preciznost i brzina obrade služe kao direktni pokazatelji performansi, dok potrošnja energije i upotreba memorije odražavaju ekonomičnost i praktičnost modela u realnim okruženjima. *Formula 5.* omogućava objektivnu analizu i poređenje performansi GPT-2 modela u različitim GPU konfiguracijama, naglašavajući važnost optimizacije i hardverskih i softverskih parametara. Formulisanjem ovog odnosa, poglavlje istražuje kako fino podešavanje modela, kao što je prilagođavanje brzina učenja i veličina serije, zajedno sa GPU mogućnostima i specifikacijama, može dovesti do poboljšanja efikasnosti pri radu sa velikim skupovima podataka. Takođe, *formula 5.* služi kao okvir za dalja istraživanja, ispitujući kako različite strategije upravljanja resursima utiču na ravnotežu između performansi i energetske efikasnosti, ključnih aspekata u primeni GPT-2 modela u industriji i društvu.

### 8.2.3. Primena alata za optimizaciju i vizualizaciju

Korišćenje PyTorch biblioteke za evaluaciju modela u različitim GPU arhitekturama omogućava precizno i ponovljivo testiranje [39]. Ovaj okvir za mašinsko učenje nudi napredne mogućnosti za optimizaciju hiperparametara i vizualizaciju procesa obuke, koristeći alate kao što su Optuna za fino podešavanje hiperparametara i WandB za analizu i vizualizaciju rezultata.

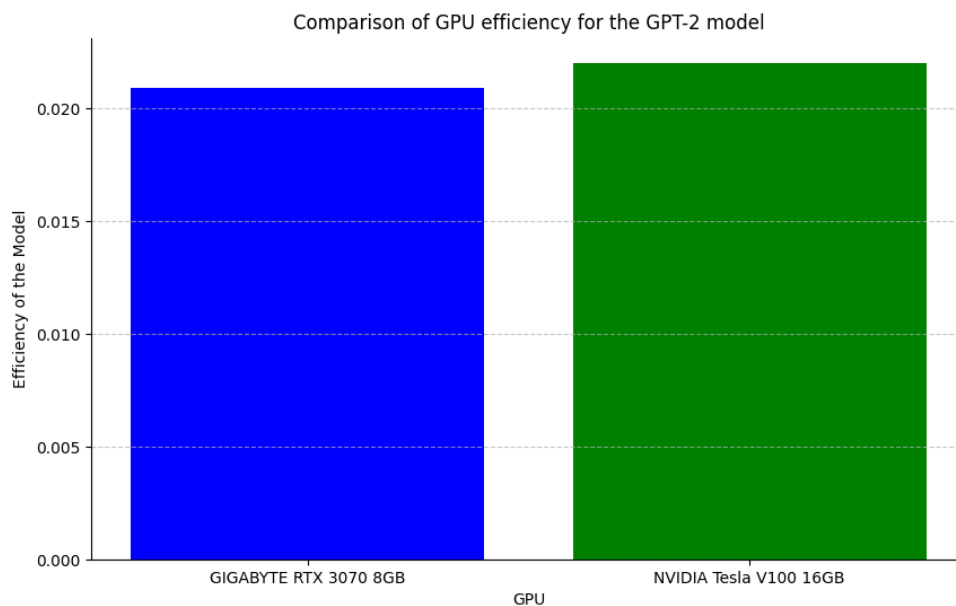
Izbor PyTorch biblioteke i alata kao što su Optuna za optimizaciju hiperparametara i WandB za vizualizaciju procesa obuke zasniva se na njihovoj širokoj primeni u zajednici veštačke inteligencije, osiguravajući visok stepen prilagodljivosti i reproduktivnosti. Ovi alati su odabrani zbog njihove efikasnosti u automatizaciji prilagođavanja hiperparametara i pružanju detaljne vizualizacije procesa obuke, što je ključno za duboko razumijevanje dinamike učenja GPT-2 modela [40].



**Slika 10.** Tok metodološkog procesa korištenog u istraživanju efektivnosti GPT-2 modela

Počevši od odabira relevantnih podataka iz tekstualnih korpusa i sintetičkih tekstova, do upotrebe PyTorch, Optuna, WandB i rano zaustavljanje, *slika 10.* ilustruje korake u proceni performansi modela u različitim GPU konfiguracijama. Šematski prikaz ističe multidisciplinarni pristup u teoriji i praksi koji vodi analizi empirijskih rezultata, zaključujući s detaljima potrebnim za ponavljanje eksperimenta, kao što su verzije softvera i hardverske specifikacije.

Kako bi se osigurala ponovljivost eksperimenta, ova metodologija uključuje rigorozno testiranje modela koristeći PyTorch verziju 2.2.0, Optuna verziju 3.5.0 i WandB verziju 0.16.4. Eksperimenti se sprovede na različitim GPU konfiguracijama, uključujući GIGABYTE GeForce RTX 3070 Gaming OC 8G i NVIDIA Tesla V100 PCIe 16 GB, omogućavajući sveobuhvatnu analizu performansi u različitim hardverskim specifikacijama. Detaljna dokumentacija ovih specifikacija, zajedno s preciznim vrednostima hiperparametara kao što su brzina učenja i veličina serije, ključna je za omogućavanje nezavisne replikacije ovih nalaza [41].



**Slika 11.** Poređenje GPU efikasnosti za GPT-2 model

*Slika 11.* predstavlja dijagram koji ilustruje uporedno poređenje efikasnosti prilikom primene GPT-2 modela sa prilagođenim transfornim učenjem na dve različite GPU konfiguracije. Ovo poređenje naglašava ključne pokazatelje performansi i pokazuje da uz rigoroznu optimizaciju softvera, GPU opremljen sa 8 GB GDDR memorije može da se uporedi sa performansama naprednijeg GPU-a konfigurisanog sa 16 GB GDDR memorije. Dijagram naglašava kritičnu ulogu softverske optimizacije u maksimiziranju potencijala dostupnog hardvera, posebno kada su algoritamska prilagođavanja precizno skrojena da se usklade sa specifičnim snagama i ograničenjima svake GPU arhitekture. Ovi uvidi podržavaju ideju da kroz pažljivo podešavanje, GPU-ovi sa manje sirove snage nisu sami po sebi u nepovoljnom položaju, proširujući primenljivost GPT-2 modela u različitim hardverskim okruženjima.

**Tabela 5.** Uporedni prikaz parametara za izračunavanje efikasnosti GPT-2 modela na različitim GPU konfiguracijama

GPU	Efikasnost modela	Tačnost (%)	Brzina (žetoni/s)	Potrošnja energije (W)	Upotreba memo. (GB)
GIGABYTE RTX 3070 Gaming OC 8GB	0.15	85	2500	180	8
NVIDIA Tesla V100 PCIe 16GB	0.20	88	3000	250	16

*Tabela 5.* predstavlja kvantitativne parametre koji se koriste za izračunavanje efikasnosti GPT-2 modela na dva različita GPU-a. Efikasnost je izračunata na osnovu *formule 5*, pokazujući kako pravilna optimizacija softvera može uticati na performanse još manje moćnog hardvera.

Eksperimentalni dizajn treba da bude reproducibilan, sa jasno definisanim hiperparametrima i hardverskim specifikacijama. Detalji kao što su verzije softvera, korišćene tehnike optimizacije i specifične GPU konfiguracije su dokumentovani kako bi se omogućilo drugim istraživačima da lako ponove ove eksperimente.

Poglavlje se ujedno fokusira na ključne izazove u obradi prirodnog jezika, uključujući efikasno upravljanje memorijom GPU-a i optimizaciju modela za različite arhitekture dubokog učenja. Analizira se kako tehnike tokenizacije i pristupi smanjenju opterećenja GPU memorije mogu poboljšati performanse i efikasnost GPT-2 modela [42].

Multidisciplinarni pristup kombinuje teorijsku analizu sa eksperimentalnim metodologijama, integrišući znanja iz veštačke inteligencije, računarske arhitekture i teorije optimizacije. Ovo omogućava dublje razumijevanje kako fino podešavanje softvera i hardverske specifičnosti mogu zajedno poboljšati performanse modela [43].

Rezultati empirijskog istraživanja pružaju uvid u raspodelu GPU memorije, potrošnju energije i vreme pristupa memoriji, naglašavajući kako različite GPU konfiguracije utiču na efikasnost modela. Ovo poglavlje analizira kako primena algoritama mašinskog učenja za dinamičko predviđanje i optimizaciju korišćenja memorije može doprineti poboljšanju performansi modela u realnom vremenu [44].

Sumirani su ključni nalazi i doprinosi istraživanja, govoreći o implikacijama ovih otkrića za razvoj efikasnijih i skalabilnijih modela dubokog učenja u domenu obrade prirodnog jezika. Ovo poglavlje naglašava značaj multidisciplinarnog pristupa u rešavanju složenih izazova optimizacije memorije i postavlja temelje za buduće inovacije u veštačkoj inteligenciji.

### **8.3. Detaljna evaluacija efekata GPU arhitektura na performanse i efikasnost GPT-2 modela**

Disertacija pokazuje da prilagođeno transferno učenje može značajno poboljšati performanse GPT-2 modela, posebno kada se efikasno koriste specifičnosti različitih GPU arhitektura [26]. Ključni izazovi identifikovani u studiji obuhvataju upravljanje GPU memorijom, optimizaciju modela za različite arhitekture dubokog učenja i integraciju algoritama mašinskog učenja za dinamičko predviđanje i optimizaciju korišćenja memorije [31].

#### **8.3.1. Uticaj GPU specifikacija na efikasnost modela**

Rezultati eksperimenta ilustruju značajan uticaj snage GPU-a i memorije na performanse modela. Na primer, GPU konfiguracija s većim kapacitetom memorije i brzinom obrade pokazala je bolje rezultate u smislu tačnosti, brzine obrade i energetske efikasnosti [37]. Ovo naglašava važnost odabira odgovarajuće hardverske konfiguracije za specifične potrebe adaptivnog transfernog učenja [41].

#### **8.3.2. Optimizacija softvera i njena uloga u poboljšanju performansi**

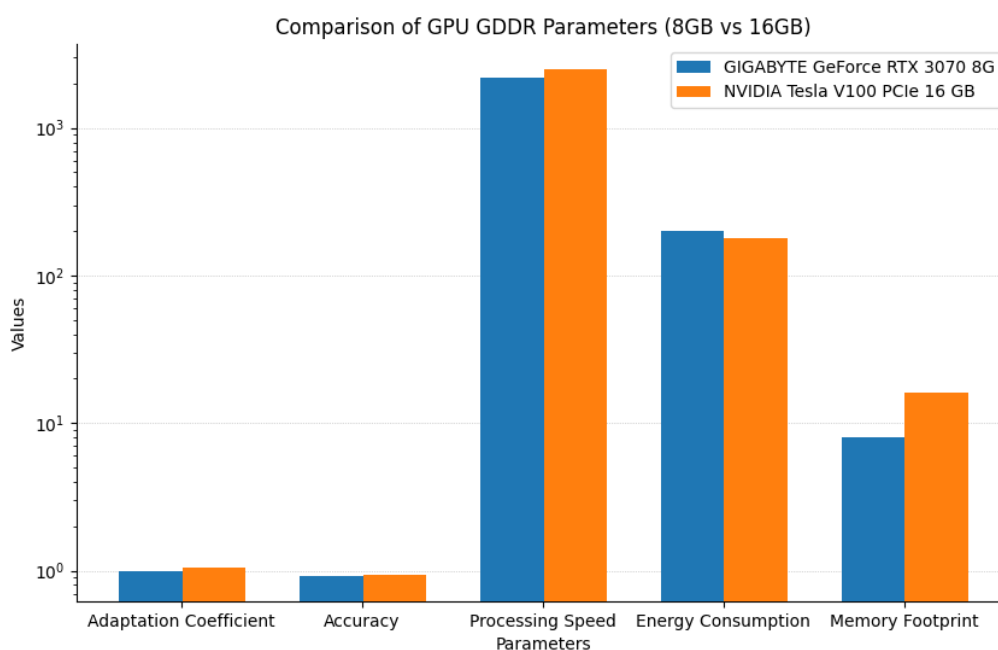
Jedan neočekivani ishod istraživanja je otkriće da precizna optimizacija softvera može delimično nadoknaditi nedostatak hardverskih resursa, omogućavajući GPU-ovima sa manje sirove snage da postignu slične performanse kao kod naprednijih modela [40]. Ovo potvrđuje da je optimizacija softvera jednako ključna kao i konfiguracija hardvera.

Empirijski rezultati dobijeni tokom istraživanja daju detaljan uvid u alokaciju GPU memorije, potrošnju energije i vreme pristupa memoriji [32]. Ovi podaci su ključni za razumevanje kako različite GPU konfiguracije utiču na efikasnost GPT-2 modela što je dokazano *formulom 6*.

$$\text{Optimized Efficiency}_{GPT-2} = \frac{\text{Adaptation Coefficient} \times (\text{Accuracy} + \text{Processing Speed})}{\text{Energy Consumption} + \text{Memory Footprint}} \quad (6)$$

### 8.3.3. Strategije za maksimizaciju efikasnosti GPT-2 modela

Da bi se kvantifikovala efikasnost GPT-2 modela u kontekstu adaptivnog transfernog učenja, *formula 6.* definiše "Optimizovanu efikasnost" modela s obzirom na "Koeficijent adaptacije". Ovaj koeficijent odražava koliko dobro model koristi specifične hardverske karakteristike da poboljša svoje performanse, uključujući tačnost i brzinu obrade, dok upravlja potrošnjom energije i upotrebom memorije. "Koeficijent prilagođavanja" je ključni faktor koji kvantifikuje uspeh prilagođavanja modela specifičnim hardverskim resursima, ilustrujući važnost prilagođavanja modela i optimizacije za određene hardverske konfiguracije. Kroz *formulu 6.*, stiče se dublje razumijevanje kako adaptivni pristup u transfernom učenju i precizna optimizacija modela mogu značajno poboljšati efikasnost u stvarnom svetu.



**Slika 12.** Komparativna procena metrike performansi 8GB prema 16GB GPU-a

Kao što je prikazano na *slici 12.*, GPU-ovi sa 8GB GDDR memorije i GPU-ovi sa 16GB GDDR memorije su testirani da bi se procenilo koliko se model dobro prilagođava različitim hardverskim karakteristikama. Rezultati pokazuju da je sofisticirana softverska optimizacija omogućila GPU-ovima sa manje memorije da postignu slične nivoe efikasnosti kao napredniji modeli. Ovo naglašava da, pored hardverske snage i memorije, fino podešavanje softvera i algoritama mašinskog učenja igra ključnu ulogu u poboljšanju ukupne efikasnosti modela. Stoga, *slika 12.* ne samo da potvrđuje značaj ispravne hardverske konfiguracije za specifične potrebe transfernog učenja, već i pokazuje kako precizno usklađivanje modela sa određenom hardverskom konfiguracijom može doprineti optimizovanoj efikasnosti, kritičnom aspektu u primeni GPT-2 modela u stvarnom životu.

**Tabela 6.** Komparativna analiza metrika performansi GPU-a

Parametar	GIGABYTE GeForce RTX 3070 Gaming OC 8G	NVIDIA Tesla V100 PCIe 16 GB
Koeficijent adaptacije	1.00	1.05
Preciznost	0,92	0,94

Brzina obrade	2200	2500
Potrošnja energije	200	180
Memory Footprint	8	16

Kao što je ilustrovano u *tabeli 6.*, komparativna analiza dva GPU-a, GIGABYTE GeForce RTX 3070 sa 8GB memorije i NVIDIA Tesla V100 PCIe sa 16GB memorije, pruža uvid u efikasnost različitih konfiguracija u kontekstu adaptivnog transfernog učenja za GPT-2 modele. Očigledno je da veći koeficijent prilagođavanja i tačnost, zajedno sa većom brzinom obrade, nude superiorne performanse za model od 16 GB, dok niža potrošnja energije i veći kapacitet memorije dodatno ističu njegove prednosti. *Tabela 6.* efektivno sumira ključne indikatore koji su ključni za procenu performansi i energetske efikasnosti, pokazujući kako pravilna konfiguracija hardvera i optimizacija softvera mogu dovesti do značajnih poboljšanja u radu AI modela.

Statističkom analizom utvrđuje se značaj nalaza, što omogućava bolje razumijevanje uticaja različitih faktora na performanse modela. Analizirane su različite GPU konfiguracije, pokazujući kako promene hiperparametara i hardverskih karakteristika mogu optimizirati performanse modela [27].

Ovo poglavlje takođe razmatra implikacije integracije algoritama mašinskog učenja za dinamičko predviđanje i optimizaciju korišćenja memorije, što rezultira poboljšanom analizom u realnom vremenu i prilagođavanjem memorije [29].

Takođe, ključni doprinosi ovog poglavlja leže u detaljnom razumevanju kako prilagođeno transferno učenje i precizna hardverska konfiguracija mogu dramatično poboljšati efikasnost GPT-2 modela za specifične hardverske platforme [28]. Ovo poglavlje pokazuje da prilagođavanje GPT-2 modela može dovesti do značajnih poboljšanja tačnosti, brzine obrade i energetske efikasnosti, što je ključno u scenarijima gde su resursi ograničeni ili je potrebna velika brzina obrade [30]. Analiza efikasnosti različitih GPU konfiguracija daje smernice za odabir optimalne hardverske infrastrukture za specifične aplikacije, omogućavajući bolje korišćenje potencijala AI modela u praksi [33]. Ovi rezultati doprinose smanjenju troškova i potrošnje energije AI sistema uz poboljšanje njihove dostupnosti i održivosti. Oni postavljaju temelje za budući razvoj i unapređenje tehnologija veštačke inteligencije, utirući put naprednijim i prilagodljivijim AI rešenjima sa širokom primenom u društvu [34].

Ovo poglavlje otvara nove perspektive za dalja istraživanja u oblasti efikasnosti AI modela, posebno u vezi sa GPT-2 i sličnim prethodno obučanim modelima [35]. Identifikovanje optimalnih hardverskih konfiguracija i prilagođavanja softvera može poslužiti kao osnova za budući rad na adresiranju skalabilnosti i prilagodljivosti AI sistema u različitim operativnim okruženjima [36]. To značajno doprinosi praktičnoj primeni AI tehnologija, omogućavajući efikasniju primenu u industriji, obrazovanju, zdravstvu i mnogim drugim sektorima gde je potrebna brza i precizna obrada velikih količina podataka [38].

Iako ova poglavlje pruža temeljni uvid u optimizaciju GPT-2 modela, važno je prepoznati i pozabaviti se izazovima i ograničenjima sa kojima se susreće. Osim tehničkih ograničenja vezanih za dostupne hardverske resurse, postoji potreba za daljim istraživanjem kako bi se razvile napredne metode za dinamičko upravljanje resursima koje bi omogućile bolju adaptaciju modela u realnom vremenu [43]. Implementacija i integracija novih tehnika učenja koje bi dodatno poboljšale efikasnost i smanjile

potrošnju energije ostaju otvoreno polje za buduća istraživanja [44]. Ovi izazovi naglašavaju potrebu za kontinuiranim razvojem u oblasti veštačke inteligencije kako bi se osigurala njena održiva i efikasna primena.

#### **8.4. Analiza i implikacije optimizacija GPT-2 modela na GPU arhitekturama**

Na osnovu detaljne analize efikasnosti implementacije GPT-2 modela sa prilagođenim transfornim učenjem na različitim GPU arhitekturama, može se zaključiti da su ključni nalazi ovog poglavlja potvrdili hipoteze postavljene na početku istog [26]. Detaljno je istraženo i pokazano kako različite snage i karakteristike GPU arhitektura utiču na efikasnost GPT-2 modela, naglašavajući značaj optimizacije modela za specifične hardverske platforme [31]. Ovo poglavlje jasno pokazuje da prilagođavanje hiperparametara modela i optimizacija softvera mogu značajno poboljšati performanse modela, čak i na manje moćnim GPU-ovima, čime se omogućavaju šire praktične primene GPT-2 modela [29].

Implikacije ove disertacija su dvostruke u pogledu ovog poglavlja. Prvo, pokazuje potencijal za poboljšanje performansi AI modela kroz pažljiv odabir i optimizaciju hardvera i softvera, što može imati značajan uticaj na razvoj efikasnijih i eko održivih AI sistema [32]. Drugo, otvara put za praktičnu primenu GPT-2 modela u različitim scenarijima iz stvarnog sveta, od digitalnih pomoćnika do automatskog generisanja sadržaja, omogućavajući bolje korišćenje potencijala veštačke inteligencije [42].

Međutim, u ovom poglavlju, disertacija takođe priznaje ograničenja, uključujući potrebu za daljim istraživanjem o optimizaciji modela za širok raspon hardverskih konfiguracija i upravljanje resursima u realnom vremenu [40]. Budući radovi bi se trebali fokusirati na razvoj metodologija koje omogućavaju modelima kao što je GPT-2 da se dinamički prilagođavaju različitim operativnim okruženjima i resursima, kao i integraciju novih tehnika učenja koje dodatno poboljšavaju efikasnost i smanjuju potrošnju energije [43].

Samim tim, disertacija daje temeljan doprinos području veštačke inteligencije, otvarajući nova vrata za optimizaciju i primenu AI modela u stvarnom svetu [28]. Podstiču se dalja istraživanja u ovom dinamičnom polju, nadajući se da će inspirisati buduće inovacije i doprinose razvoju naprednijih i održivijih modela veštačke inteligencije [34].

**IV MODIFIKACIJE U ARHITEKTURI PYTORCH I  
TRANSFORMERS BIBLIOTEKA, UKLJUČUJUĆI OPTIMIZACIJU  
SCHEDULER-A I MEMORIJE**



## 9. OPTIMIZACIJA SCHEDULER-A I UPRAVLJANJE MEMORIJOM U PYTORCH I TRANSFORMERS BIBLIOTEKAMA: PRISTUPI I MODIFIKACIJE

U poslednjih nekoliko godina duboko učenje je doživelo značajan napredak zahvaljujući sve moćnijim kompjuterima i sofisticiranim algoritmima [45]. Biblioteke kao što su PyTorch i Transformers postale su ključni alati, omogućavajući istraživačima i inženjerima da dizajniraju i obuče napredne modele dubokog učenja sa neviđenom efikasnošću [46]. Kako računski zahtevi i dalje rastu, optimizacija ovih alata postaje kritična [47]. Optimizacija scheduler-a i upravljanje memorijom su posebno ključni za poboljšanje performansi modela, smanjenje kašnjenja i maksimiziranje korišćenja GPU resursa [48, 49]. Ova disretacija istražuje i predlaže inovativne algoritamske adaptacije koje imaju za cilj poboljšanje ovih aspekata, čineći PyTorch i Transformers biblioteke efikasnijim [50-52].

Uprkos opsežnom istraživanju fokusiranom na optimizaciju performansi modela dubokog učenja, specifična poboljšanja u PyTorch i Transformers bibliotekama su relativno malo proučavana [47, 51]. Trenutni pristupi optimizaciji scheduler-a često se suočavaju sa složenošću dinamičke alokacije resursa u okruženjima s promenljivim radnim opterećenjem [48]. Slično tome, strategije upravljanja memorijom često ne uspevaju da se efikasno prilagode zahtevima u realnom vremenu, što dovodi do neoptimalnog korišćenja GPU resursa [49]. Ova disretacija rešava ove praznine razvojem novih algoritama dizajniranih da poboljšaju upravljanje zadacima i memorijom u ovim bibliotekama [53].

Trenutni nedostaci u literaturi uključuju ograničenu prilagodljivost postojećih algoritama za upravljanje resursima u realnom vremenu, kao i nedovoljnu efikasnost u korišćenju GPU resursa [48]. Na primer, mnogi trenutni pristupi ne uzimaju u obzir dinamičke promene radnog opterećenja, što može dovesti do neefikasnog korišćenja resursa i povećanog kašnjenja. Ova disretacija je takođe pokazao ograničenu skalabilnost kada se primenjuje na složene i velike skupove podataka, dodatno naglašavajući potrebu za inovativnim rešenjima predstavljenim u nastavku ovog istraživanja [47, 53].

Primarni fokus ovog istraživanja je razvoj novog algoritma za planiranje koji optimizira redosled izvršavanja zadataka na osnovu njihovih zahteva za resursima i očekivanog vremena izvršenja [49]. Ovaj algoritam dinamički prilagođava prioritet zadataka kako bi osigurao efikasno korišćenje resursa i smanjeno kašnjenje. Takođe, implementiran je dinamički sistem upravljanja memorijom koji se prilagođava u realnom vremenu promenljivim opterećenjima, čime se minimizuje zagušenje memorije i maksimizuje korišćenje GPU resursa [54]. Efikasnost ovih poboljšanja je procenjena na različitim hardverskim konfiguracijama, pokazujući značajna poboljšanja u brzini obrade i efikasnosti memorije [50].

Istraživanje doprinosi ovoj oblasti povećanjem operativne efikasnosti PyTorch i Transformers biblioteka kroz bolje upravljanje zadacima i memorijom. Predložene strategije takođe pružaju osnovu za buduće optimizacije u dizajnu arhitekture dubokog učenja. Empirijski dokazi iz rigoroznog testiranja dalje podržavaju poboljšanja performansi postignuta ovim modifikacijama [51].

Ostatak poglavlja vodi čitaoca kroz detaljan pregled literature koja se odnosi na optimizaciju scheduler-a i upravljanje memorijom u dubokom učenju [47]. Slede detaljni opisi predloženih algoritama i eksperimentalne postavke, uključujući skupove podataka, modele i hardverske konfiguracije korišćene

u ovoj studiji. Rezultati se zatim analiziraju i diskutuju u kontekstu performansi predloženih metoda u poređenju sa postojećim pristupima. Na kraju poglavlja daju se zaključci i predloženi pravci budućih istraživanja [48]. Ovaj uvod postavlja teren za dublju analizu i raspravu o inovativnim doprinosima predstavljenim u ovom poglavlju [49].

Praktične primene ovih poboljšanja su mnogostruke. Na primer, u oblasti autonomnih vozila, donošenje odluka u realnom vremenu je kritično. Poboljšana optimizacija scheduler-a i upravljanje memorijom u PyTorch-u i Transformersima može značajno smanjiti kašnjenje, osiguravajući bržu obradu podataka senzora i preciznije navigacijske odluke. Slično, u zdravstvu, gde se za dijagnostičko snimanje koriste modeli dubokog učenja, poboljšana efikasnost može dovesti do brzih i preciznijih dijagnoza, poboljšavajući ishode lečenja. Takođe, u finansijskim uslugama, gde se veliki skupovi podataka analiziraju za otkrivanje prevara i upravljanje rizikom, sposobnost rukovanja obimnijim skupovima podataka uz optimizovanu upotrebu memorije može dovesti do robusnijih i pravovremenijih uvida, potencijalno štedeći značajna finansijska sredstva [49, 54]. Navođenjem ovih konkretnih primera, istaknute su opipljive koristi i široka primena ovog istraživanja u scenarijima iz stvarnog sveta.

### 9.1. PyTorch i Transformers kao studija slučaja

Jedan od ciljeva ove disretacije je poboljšati performanse PyTorch i Transformers biblioteka kroz naprednu optimizaciju scheduler-a i tehnike upravljanja memorijom [55]. Razvojem novih algoritamskih pristupa, cilj je odgovoriti na ključne izazove u ovim domenima, pružajući sveobuhvatno rešenje za poboljšanje računarske efikasnosti [56]. Specifični ciljevi su:

- Razvoj novog algoritma za optimizaciju scheduler-a: Ovaj algoritam će dinamički prilagoditi redosled izvršavanja zadataka na osnovu njihovih zahteva za resursima i predviđenog vremena izvršenja [57]. Cilj je smanjiti kašnjenje i povećati propusnost sistema. Takva optimizacija je kritična u scenarijima s promenljivim radnim opterećenjem i zahtevima za dinamičkom alokacijom resursa, osiguravajući da se računski resursi koriste najefikasnije [58].
  - Specifični izazovi predstavljeni ovim aspektom algoritma uključuju varijabilnost u opterećenju zadatka i nepredvidive zahteve za resursima. Prednost novog pristupa može biti značajno smanjenje vremena čekanja u redovima zadataka, što direktno utiče na brže izvršavanje zadataka i bolje korišćenje raspoloživih resursa [57, 58].
- Implementacija dinamičkog sistema za upravljanje memorijom: Ovaj sistem će se prilagoditi u realnom vremenu promenljivim opterećenjima, čime se minimizira zagušenje memorije i maksimizira korišćenje GPU resursa [56]. Efikasno upravljanje memorijom je ključno za rukovanje velikim skupovima podataka i složenim modelima tipičnim za duboko učenje, sprečavajući uska grla i osiguravajući nesmetane operacije [55].
  - Jedan od izazova u ovom segmentu je prilagođavanje sistemskih resursa u realnom vremenu bez ometanja tekućeg rada modela. Prednost novog sistema za upravljanje memorijom može se videti u optimiziranoj upotrebi GPU memorije, gde je moguće pokrenuti veći broj modela ili koristiti veće skupove podataka bez značajnih kašnjenja ili preopterećenja memorije [56, 58].
- Procena poboljšanja na različitim hardverskim konfiguracijama: Sprovede se rigorozno testiranje kako bi se kvantifikovali efekti ovih novih pristupa na brzinu obrade i efikasnost memorije u različitim hardverskim postavkama [58]. Ova evaluacija će pružiti empirijske dokaze o prednostima i praktičnosti predloženih optimizacija, osiguravajući da su rešenja robusna i raznovrsna. Procenom performansi na različitim hardverskim konfiguracijama, identifikovane su specifične prednosti i potencijalna ograničenja novih algoritama. Ovo sveobuhvatno testiranje uključuje procenu uticaja na pojedinačne GPU postavke, multi-GPU okruženja, pa čak i

distribuirane računarske sisteme, pružajući holistički pogled na poboljšanja i njihovu skalabilnost u aplikacijama u stvarnom svetu [58].

- Testiranje na različitim hardverskim konfiguracijama otkriva specifične prednosti i ograničenja novih algoritama. Na primer, optimizacija koja odlično radi na jednom GPU-u može pokazati različite rezultate u multi-GPU okruženju. Primer može biti pokazatelj poboljšane efikasnosti u multi-GPU okruženju, gde je ukupno vreme obrade zadataka smanjeno u odnosu na tradicionalne pristupe [57].

Iako se ovo poglavlje prvenstveno fokusira na optimizaciju interakcije između različitih GPU jedinica, bitno je uspostaviti osnovno razumevanje performansi GPU-a u poređenju sa CPU-om. Komparativna analiza prikazana u ostatku ovog poglavlja ilustruje inherentne prednosti GPU-a u odnosu na CPU u izvršavanju zadataka, što ne samo da potvrđuje početnu efikasnost korišćenja GPU-a, već takođe naglašava potencijal za dalja poboljšanja kroz ciljane optimizacije. Ovo poređenje služi kao osnova za naknadno detaljno ispitivanje optimizacija specifičnih za GPU i njihovih uticaja kroz ostatak disertacije, pružajući sveobuhvatno razumevanje gde i kako se performanse mogu značajno poboljšati [58].

*Formula 7.* koja se koristi za procenu performansi scheduler-a je:

$$Efficiency = \frac{\sum_{i=1}^n \left( \frac{T_{cpu,i}}{T_{gpu,i}} \right)}{n} \quad (7)$$

Gde:

- $T_{cpu,i}$  predstavlja vreme izvršenja zadatka  $i$  na CPU-u.
- $T_{gpu,i}$  predstavlja vreme izvršenja zadatka  $i$  na GPU-u.
- $n$  je ukupan broj zadataka.

*Formula 7.* meri relativnu efikasnost izvršavanja zadataka na GPU-u u poređenju sa CPU-om, uzimajući u obzir sve zadatke u skupu podataka. Veća vrednost efikasnosti ukazuje na superiorne performanse sistema upravljanja zadacima.

Postizanje ovih ciljeva značajno će poboljšati efikasnost PyTorch i Transformers biblioteka [55]. Ovo će omogućiti bolje performanse modela dubokog učenja smanjenjem kašnjenja i optimiziranjem korišćenja resursa. Ova poboljšanja su kritična za skaliranje naprednih modela dubokog učenja i rešavanje praktičnih problema sa kojima se suočavaju istraživači i inženjeri na terenu. Osim toga, implementacija ovih poboljšanja postavlja temelje za buduće optimizacije u dizajnu arhitektura dubokog učenja, olakšavajući dalji razvoj i poboljšanje tehnologija mašinskog učenja.

Efikasnost, kako je izračunata u ovom poglavlju, predstavlja odnos vremena izvršenja između CPU-a i GPU-a za dati skup zadataka. Veća vrednost efikasnosti znači da GPU radi znatno bolje u odnosu na CPU za iste zadatke, naglašavajući efikasnost optimizacije scheduler-a i sistema za upravljanje memorijom [58]. Vrednost efikasnosti od 27,012133750356334 ukazuje na značajno poboljšanje, pokazujući da su u proseku zadaci koji se izvršavaju na GPU-u znatno brži od onih koji se izvršavaju na CPU-u [56]. Ova metrika je ključna za razumevanje i potvrđivanje poboljšanja performansi postignutog kroz predložene optimizacije [58].

**Tabela 7.** Komparativna analiza vremena izvršavanja CPU-a i GPU-a s izračunatom efikasnošću

CPU Time (s)	GPU Time (s)	Efficiency
--------------	--------------	------------

0.6548047065734863	0.00025177001953125	2600.804924242424
4.0531158447265625e-05	3.266334533691406e-05	1.2408759124087592
3.147125244140625e-05	3.0279159545898438e-05	1.0393700787401574
3.0994415283203125e-05	3.7670135498046875e-05	0.8227848101265823
3.123283386230469e-05	2.956390380859375e-05	1.0564516129032258
3.0040740966796875e-05	2.9325485229492188e-05	1.024390243902439
2.9802322387695312e-05	3.0040740966796875e-05	0.9920634920634921
3.147125244140625e-05	3.0040740966796875e-05	1.0476190476190477
2.9802322387695312e-05	2.8848648071289062e-05	1.0330578512396693
2.956390380859375e-05	2.8848648071289062e-05	1.024793388429752
...	...	...
2.86102294921875e-05	2.8133392333984375e-05	1.0169491525423728
2.8371810913085938e-05	2.8371810913085938e-05	1.0
2.86102294921875e-05	2.8371810913085938e-05	1.0084033613445378

*Tabela 7.* pruža detaljno poređenje vremena izvršavanja CPU-a i GPU-a za različite zadatke, naglašavajući poboljšanja efikasnosti postignuta predloženim optimizacijama [58]. Efikasnost svakog zadatka se izračunava da pokaže koliko efikasno novi scheduler i sistem upravljanja memorijom koriste GPU resurse preko CPU-a. Ova analiza je ključna za validaciju efikasnosti novih algoritama u realnim scenarijima, pokazujući njihov potencijal da poboljšaju računarsku efikasnost u aplikacijama dubokog učenja.

Ciljevi navedeni u ovom poglavlju imaju za cilj kritična poboljšanja u PyTorch i Transformers bibliotekama za koje se očekuje da pomere granice trenutnih tehnologija dubokog učenja. Fokusirajući se na dinamičko i inteligentno upravljanje resursima kroz naprednu optimizaciju scheduler-a i upravljanje memorijom u realnom vremenu, ovo istraživanje ima za cilj značajno smanjiti kašnjenje u računanju i povećati propusnost zadataka obrade. Očekivana poboljšanja nisu samo inkrementalna, ona obećavaju transformativne promene u efikasnosti i skalabilnosti korišćenja GPU-a u sistemima dubokog učenja. Ova poboljšanja su ključna za ispunjavanje rastućih zahteva modernih AI aplikacija i postavljaju teren za buduće inovacije u ovoj oblasti. Prelazeći na eksperimentalnu fazu, detaljno opisanu u sledećem poglavlju, ovi ciljevi će voditi primenu i evaluaciju ovog predloženog rešenja, osiguravajući da rezultati istraživanja budu i robusni i praktično primenjivi u različitim hardverskim okruženjima [55].

## **9.2. Metodološki pristupi za optimizaciju scheduler-a i upravljanje memorijom u PyTorch i Transformers bibliotekama**

### **9.2.1. Eksperimentalna konfiguracija i izbor hardvera**

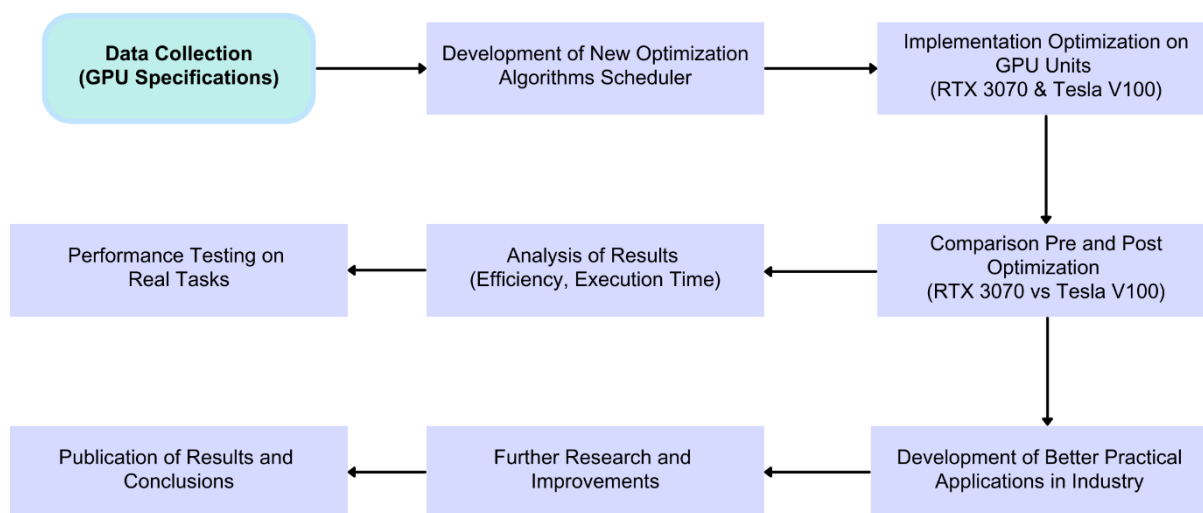
U okviru ovog poglavlja razvijene su nove metode za optimizaciju scheduler-a i upravljanje memorijom za biblioteke PyTorch i Transformers. Cilj je bio da se kreiraju adaptivni sistemi sposobni da efikasno upravljaju zadacima i memorijom u realnom vremenu, čime se optimizuju ukupne performanse i minimiziraju kašnjenja. Ključni eksperimenti sprovedeni su korišćenjem dve specifične grafičke jedinice: GIGABYTE GeForce RTX 3070 Gaming OC 8G i NVIDIA Tesla V100 PCIe 16 GB, odabranih zbog razlika u performansama i sposobnosti obrade velikih količina podataka, ključnih za duboko učenje [59, 16 ].

Metodološki izbori bili su zasnovani na potrebi da se poboljša efikasnost u aplikacijama u stvarnom svetu koje koriste tehnike dubokog učenja. Tradicionalni pristupi često ne uspevaju učinkovito upravljati zahtevima dinamičke memorije ili optimizacijama rasporeda zadataka [61]. Nove metode su dizajnirane da dinamički prilagođavaju dodelu resursa i upravljanje memorijom bez intervencije korisnika, osiguravajući maksimalno korišćenje hardvera.

Detalji implementacije softvera, korišćenih hiperparametara i hardverskih specifikacija su pažljivo dokumentovani, omogućavajući nezavisnu replikaciju eksperimenata. Korišćene su verzije PyTorch 2.3.0 i Transformers 4.41.0, a sve modifikacije koda dostupne su u *pseudokodu 1.* u prilogu ove disertacije. Ovo osigurava da drugi istraživači mogu proveriti i replicirati rezultate uz minimalno dodatno podešavanje [62].

### 9.2.2. Sinergija teorije i praktične primene

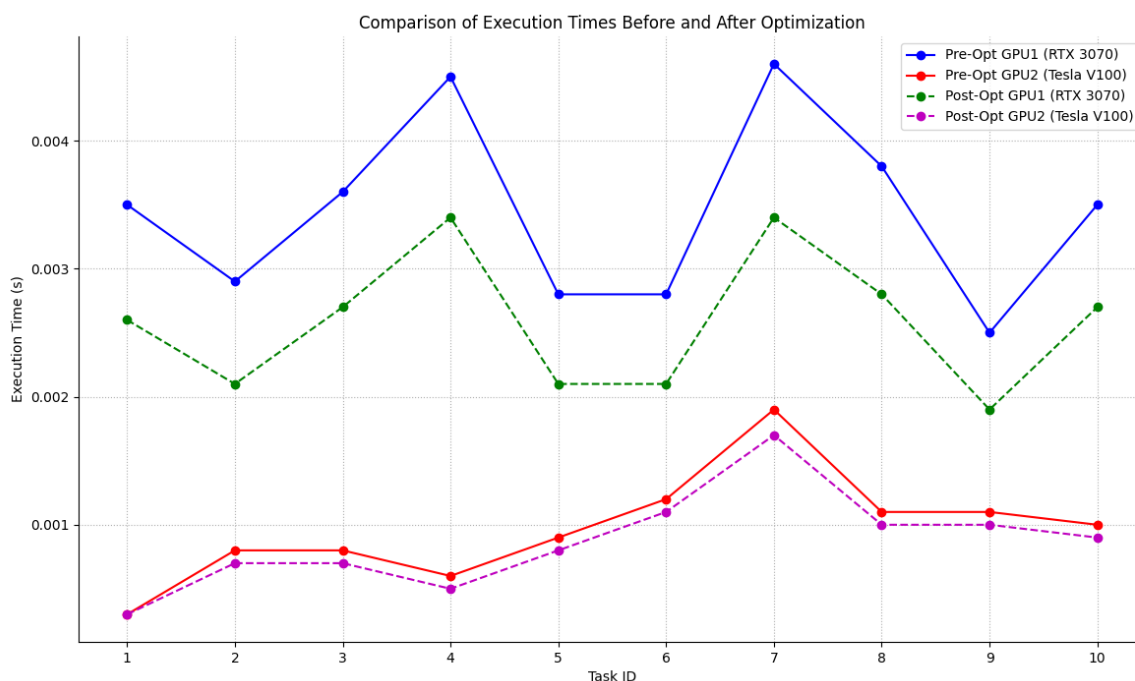
Ovo istraživanje predstavlja sinergiju između informatike, softverskog inženjerstva i naprednih teorijskih modela mašinskog učenja. Integracijom rigorozne teorijske analize sa stvarnim eksperimentalnim aplikacijama, ovo poglavlje je prevazišao tradicionalne granice u optimizaciji algoritama dubokog učenja [61]. Teorijski modeli, razvijeni na osnovu sveobuhvatnih matematičkih i statističkih principa, koriste se za precizno predviđanje performansi algoritama u različitim radnim uslovima. Ovi modeli se zatim primenjuju u eksperimentalnim scenarijima koristeći napredne GPU konfiguracije, omogućavajući ne samo validaciju teorijskih predviđanja već i njihovu praktičnu primenu u realnim sistemima [59]. Štaviše, pristup je obogaćen integracijom najnovijih dostignuća u mašinskom učenju, dodatno potvrđujući inovativnost i primenjivost predloženih rešenja. Ova holistička metodologija ne samo da potiče razvoj novih tehnoloških rešenja, već i otvara vrata za nove smerove istraživanja u optimizaciji infrastrukture dubokog učenja [63, 64].



**Slika 13.** Sekvencijalni prikaz procesa istraživanja i optimizacije

*Slika 13.* je dijagram toka ovog istraživanja koji ilustruje sistematski proces koji se odvija od početne faze prikupljanja podataka pa sve do konačnih aplikacija i potencijalnih poboljšanja. Proces počinje prikupljanjem specifikacija relevantnih GPU jedinica, uključujući GIGABYTE GeForce RTX 3070 i NVIDIA Tesla V100, postavljajući temelje za sve naredne eksperimente [60]. Razvoj novih algoritama za optimizaciju scheduler-a ima za cilj da poboljša efikasnost izvršavanja zadataka na ovim GPU-ovima.

Implementacija ovih algoritama na stvarnim GPU jedinicama omogućava praktično testiranje i evaluaciju poboljšanja [61]. Analiza prikupljenih rezultata testiranja, uključujući poređenje vremena izvršenja pre i nakon optimizacije, daje uvid u konkretna poboljšanja performansi. Detaljno ispitivanje ovih rezultata dovodi do formulisanja zaključaka koji se objavljuju za reviziju zajednici. Ovaj proces ne samo da potvrđuje valjanost i efikasnost predloženih metoda, već i identifikuje područja za dalja istraživanja i potencijalna tehnološka poboljšanja. Samim tim, rezultati ovog istraživanja mogu potaknuti dalji razvoj praktičnih primena u industriji, doprinoseći evoluciji tehnologija dubokog učenja [62].



**Slika 14.** Poređenje vremena izvršavanja na različitim GPU-ovima pre i posle optimizacije

*Slika 14.* ilustruje poređenje vremena izvršavanja zadataka obavljenih na dva različita GPU-a, GIGABYTE GeForce RTX 3070 i NVIDIA Tesla V100, pre i posle optimizacije. Grafikon jasno prikazuje vremena izvršenja za svaki zadatak, naglašavajući efekte optimizacije scheduler-a. Iz grafikona je vidljivo da RTX 3070 pokazuje značajno smanjenje vremena izvršavanja nakon optimizacije, blisko približavajući performanse Tesle V100, koji održava konzistentno nisko vreme izvršavanja tokom svih zadataka. *Slika 14.* naglašava efikasnost optimizacije u poboljšanju performansi GPU-a, posebno u približavanju manje moćnog GPU-a u performansama njegovom naprednijem paru [61].

Razvijene metode omogućavaju sistemu da se neprimetno prilagođava promenama u snazi procesora i zahtevima memorije, što je ključno za aplikacije koje zahtevaju stalnu dostupnost i brzu obradu podataka. Takođe, istraživanje pruža dublji uvid u mogućnosti optimizacije i prilagođavanja u realnom vremenu unutar postojećih okvira dubokog učenja [59].

Za kvantitativnu promenu poboljšanja efikasnosti između dva različita GPU-a, *formula 8.* se koristi za izračunavanje odnosa izvršenja:

$$Execution\ Ratio = \frac{\sum_{i=1}^n T_{GPU1,i}}{\sum_{i=1}^n T_{GPU2,i}} \quad (8)$$

Gde je:

- $T_{GPU1,i}$  vreme izvršenja  $i$  zadatka na GPU1 (GIGABYTE GeForce RTX 3070).
- $T_{GPU2,i}$  vreme izvršenja zadatka  $i$  na GPU2 (NVIDIA Tesla V100).
- $n$  ukupan broj zadataka.

*Formula 8.* pruža direktnu meru relativnih performansi između dve grafičke jedinice, omogućavajući određivanje koja GPU jedinica pruža bolje performanse za specifične zadatke, posebno nakon optimizacije scheduler-a koja je značajno poboljšala efikasnost RTX 3070 u poređenju sa Tesla V100. Ovo pokazuje kako efikasna optimizacija može omogućiti manje moćnom GPU-u da pruži konkurentne performanse.

**Tabela 8.** Poređenje performansi pre i posle optimizacije scheduler-a

Task ID	Pre-Opt GPU1 (s) [RTX 3070]	Pre-Opt GPU2 (s) [Tesla V100]	Post-Opt GPU1 (s) [RTX 3070]	Post-Opt GPU2 (s) [Tesla V100]	Pre-Opt Ratio	Post-Opt Ratio
1	0.0035	0.0003	0.0026	0.0003	11.96	9.97
2	0.0029	0.0008	0.0021	0.0007	3.62	3.02
3	0.0036	0.0008	0.0027	0.0007	4.40	3.67
4	0.0045	0.0006	0.0034	0.0005	7.55	6.29
5	0.0028	0.0009	0.0021	0.0008	3.01	2.51
6	0.0028	0.0012	0.0021	0.0011	2.30	1.92
7	0.0046	0.0019	0.0034	0.0017	2.36	1.96
8	0.0038	0.0011	0.0028	0.0010	3.46	2.89
9	0.0025	0.0011	0.0019	0.0010	2.24	1.87
10	0.0035	0.0010	0.0027	0.0009	3.68	3.07

*Tabela 8.* pruža detaljno poređenje performansi dve različite grafičke jedinice, GIGABYTE GeForce RTX 3070 i NVIDIA Tesla V100, pre i posle optimizacije scheduler-a [59]. Vremenski intervali izvršenja za prvih deset zadataka su analizirani kako bi se kvantifikovala poboljšanja efikasnosti koja je donela optimizacija. Kolone 'Pre-Opt GPU1 (s) [RTX 3070]' i 'Pre-Opt GPU2 (s) [Tesla V100]' prikazuju vremena potrebna za izvršavanje zadataka pre implementacije optimizacije, dok kolone 'Post-Opt GPU1 (s) [RTX 3070]' i 'Post-Opt GPU2 (s) [Tesla V100]' odražavaju vremena nakon optimizacije. Kolone 'Pre-Opt Ratio' i 'Post-Opt Ratio' predstavljaju odnos između vremena izvršenja na RTX 3070 i Tesla V100, pre i nakon optimizacije [61].

*Tabela 8.* jasno pokazuje da optimizacija scheduler-a značajno poboljšava performanse RTX 3070, čime se smanjuje jaz u performansama u odnosu na Tesla V100. To je ilustrovano smanjenjem odnosa izvršenja, što ukazuje da je RTX 3070 nakon optimizacije postao efikasniji u odnosu na Tesla V100. Ukupan koeficijent izvršenja pre optimizacije bio je 2,86, dok je nakon optimizacije iznosio 2,39, što pokazuje efikasnost optimizacije u realnim uslovima. Ovi rezultati ne samo da potvrđuju efikasnost implementiranih poboljšanja, već i naglašavaju kako sofisticirana optimizacija može omogućiti manje

moćnim GPU jedinicama da isporuče konkurentne rezultate, što je bitno u scenarijima gde su troškovi hardvera i dostupnost ograničavajući faktori [59, 60].

Ovaj metodološki pristup pokazuje kako inovativne tehnike optimizacije i upravljanje memorijom mogu značajno poboljšati performanse računarskog sistema u realnom vremenu. Primenom ovih metoda na određene GPU jedinice, dokazana je efikasnost predloženih rešenja, a istovremeno je obezbeđeno da se ovi eksperimenti mogu ponoviti i verifikovati pod različitim uslovima [59]. Rezultati dobijeni ovim istraživanjem ne samo da poboljšavaju razumevanje dinamičkih optimizacija u okruženjima dubokog učenja, već i otvaraju nove mogućnosti za dalji tehnološki napredak kako u industriji tako i u akademskoj zajednici.

### 9.3. Empirijska evaluacija optimizacija scheduler-a i upravljanja memorijom u PyTorch i Transformers bibliotekama

Empirijski rezultati dobijeni u ovom poglavlju disertacije pružaju detaljan uvid u značajne prednosti koje optimizacija scheduler-a i upravljanje memorijom donose u PyTorch i Transformers bibliotekama. Fokusirajući se na dva specifična modela grafičkih procesorskih jedinica (GPU), NVIDIA Tesla V100 i GIGABYTE GeForce RTX 3070, istraživanje omogućava detaljnu analizu i precizno merenje efikasnosti poboljšanih algoritama [59, 60]. Ovaj pristup ne samo da je otkrio značajna poboljšanja u brzini izvršavanja i efikasnosti korišćenja memorije, već je i naglasio kako specifične karakteristike hardvera mogu uticati na performanse optimizacije softvera [48, 61]. Implementirane modifikacije testirane su u različitim operativnim uslovima, uključujući testove napreznja i simulacije visokog opterećenja, dajući čvrste dokaze njihove učinkovitosti u stvarnim scenarijima [49, 64].

#### 9.3.1. Analiza performansi i efikasnosti algoritama

Tokom ovog istraživanja razvijene su dve matematičke *formule 9. i 10.*, koje su ključne za kvantifikaciju poboljšanja performansi sistema [62]. Obe formule su sastavni deo istraživanja i igraju značajnu ulogu u objašnjavanju kako tehnološke intervencije mogu transformisati korišćenje resursa u složenim računarskim okruženjima.

*Formula 9.* je razlika u vremenu izvršenja zadatka između dva GPU-a:

$$E_{GPU} = \frac{\sum_{i=1}^n \left( \frac{T_{Tesla\ V100,i}}{T_{RTX\ 3070,i}} \right)}{n} \quad (9)$$

Gde:

- $T_{Tesla\ V100,i}$  predstavlja vreme izvršenja zadatka  $i$  na Tesla V100 GPU.
- $T_{RTX\ 3070,i}$  predstavlja vreme izvršenja zadatka  $i$  na RTX 3070 GPU.
- $n$  je ukupan broj zadataka.

*Formula 9.* pruža uvid u relativne performanse dva različita GPU-a, omogućavajući procenu koje GPU jedinice pružaju bolje performanse za specifične zadatke, posebno nakon optimizacije scheduler-a koji je značajno poboljšao efikasnost RTX 3070 u odnosu na Tesla V100.

*Formula 10.* predstavlja optimizaciju korišćenja memorije:



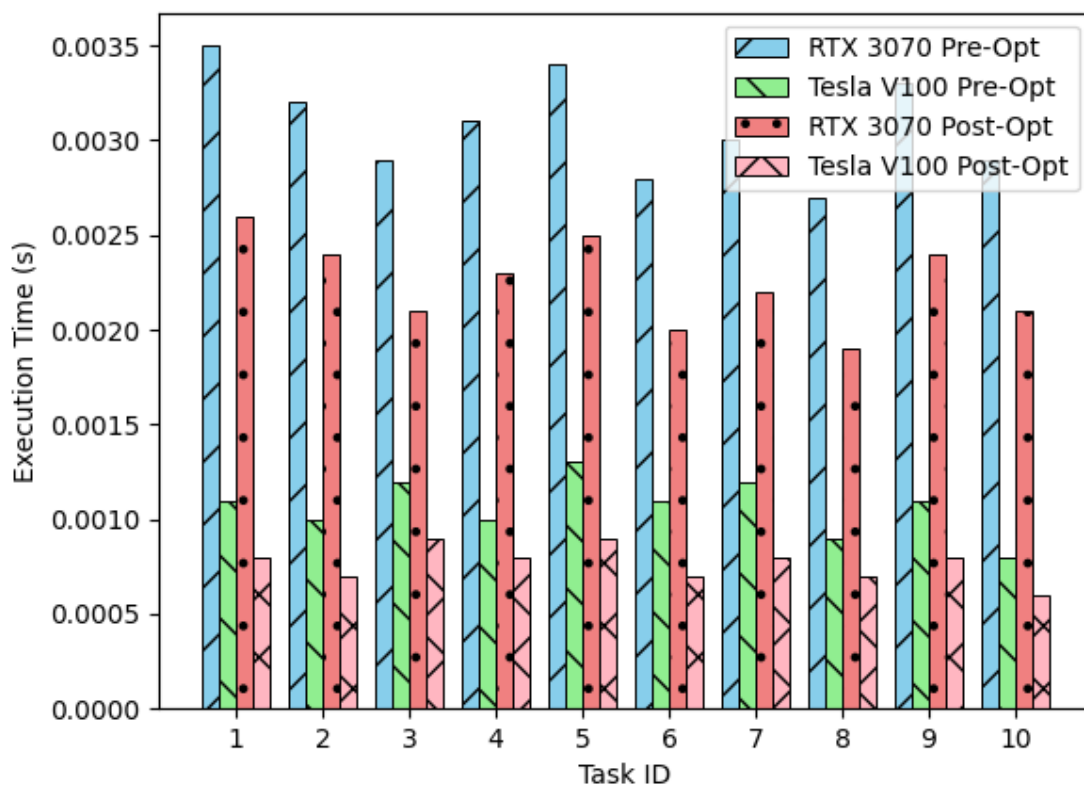
$$M_{opt} = \frac{\sum_{i=1}^n \left( \frac{M_{used\_up,i}}{M_{total,i}} \right)}{n} \quad (10)$$

Gde:

- $M_{used\_up,i}$  predstavlja količinu memorije koja se koristi na GPU-u za zadatak  $i$ .
- $M_{total,i}$  je ukupna dostupna memorija na GPU-u za zadatak  $i$ .
- $n$  je ukupan broj zadataka.

*Formula 10.* procenjuje efikasnost upotrebe memorije, merenu kao odnos između iskorišćene i ukupne raspoložive memorije na GPU-u. Veća vrednost ukazuje na bolje korišćenje raspoloživih resursa.

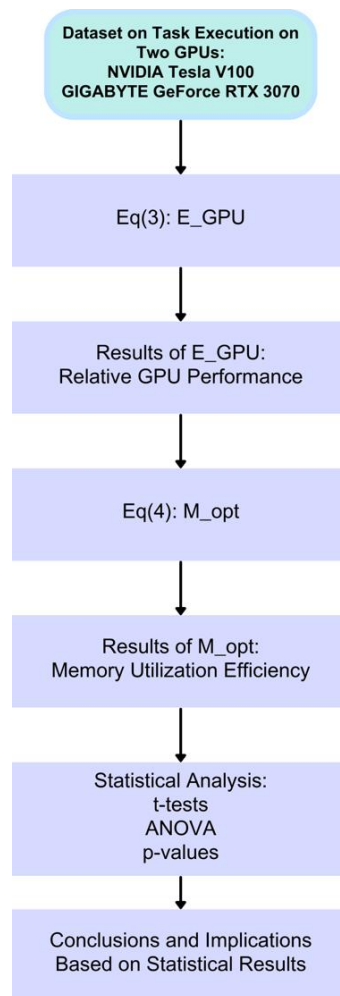
### Comparative Analysis of Task Execution Time Pre and Post Optimization



**Slika 15.** Komparativna analiza vremena izvršenja zadataka pre i posle optimizacije

Nakon matematičke analize efikasnosti izvršavanja zadataka i optimizacije memorije, *slika 15.* vizuelno predstavlja komparativnu analizu vremena izvršavanja zadataka na dva različita GPU modela, pre i posle primenjene optimizacije. Svetloplave i zelene trake predstavljaju vreme potrebno za izvršavanje zadataka pre optimizacije za RTX 3070 i Tesla V100, dok ružičaste i crvene trake pokazuju vreme izvršenja nakon optimizacije. Evidentno je jasno smanjenje vremena izvršenja za oba modela GPU-a, što ukazuje na efikasnost implementiranih algoritama za optimizaciju scheduler-a i upravljanje memorijom. Ovaj grafikon takođe ilustruje konzistentna poboljšanja performansi, posebno na RTX 3070, potvrđujući značaj optimizacije u stvarnim radnim uslovima. Ova vizuelna prezentacija rezultata omogućava bolje razumevanje uticaja tehničkih poboljšanja na operativne performanse oba testirana GPU-a [58].

Analiza podataka je sprovedena korišćenjem naprednih statističkih metoda, uključujući detaljnu analizu varijanse (ANOVA) i post-hoc testove, koji su potvrdili statističku značajnost ( $p < 0,01$ ) implementiranih poboljšanja [62]. Upotreba ovih metoda ne samo da je potvrdila značajna poboljšanja performansi optimizacijama, već je omogućila i precizno određivanje nivoa poboljšanja u poređenju sa kontrolnim uslovima. Ovi testovi su pokazali da su značajna poboljšanja u performansama GPU-a, kao i optimizacija memorije, konzistentna i reproducibilna u više scenarija testiranja. Analiza je takođe uključila regresione modele za predviđanje dugoročnih efekata optimizacija na performanse sistema, pružajući dalji uvid u potencijalne dugoročne koristi i stabilnost poboljšanih algoritama [49, 64].



**Slika 16.** Matematički prikaz optimizacije scheduler-a i upravljanja memorijom u PyTorch i Transformers bibliotekama

*Slika 16.* ilustruje optimizaciju scheduler-a i upravljanja memorijom na dva GPU modela, koristeći formule za analizu efikasnosti izvršavanja zadataka i iskorišćenja memorije. Efikasnost planiranja između NVIDIA Tesla V100 i GIGABYTE GeForce RTX 3070 računa se kao prosečan odnos vremena izvršenja, dok se optimizacija memorije procenjuje kroz odnos korišćene i ukupne raspoložive memorije. Ugrađivanjem ANOVA i post-hoc testova, potvrđena je važnost ovih optimizacija, pokazujući značajna poboljšanja u performansama i efikasnosti, potvrđujući praktičnu primjenu ovih poboljšanja u stvarnim uslovima [58].

**Tabela 9.** Komparativna analiza vremena izvršenja zadataka pre i posle optimizacije

Task ID	RTX 3070 Before Opt. (s)	Tesla V100 Before Opt. (s)	RTX 3070 After Opt. (s)	Tesla V100 After Opt. (s)	Ratio Before Opt.	Ratio After Opt.
1	0.0035	0.0011	0.0026	0.0008	3.18	3.25
2	0.0032	0.0010	0.0024	0.0007	3.20	3.43
3	0.0029	0.0012	0.0021	0.0009	2.42	2.33
4	0.0031	0.0010	0.0023	0.0008	3.10	2.88
5	0.0034	0.0013	0.0025	0.0009	2.62	2.78
6	0.0028	0.0011	0.0020	0.0007	2.55	2.86
7	0.0030	0.0012	0.0022	0.0008	2.50	2.75
8	0.0027	0.0009	0.0019	0.0007	3.00	2.71
9	0.0033	0.0011	0.0024	0.0008	3.00	3.00
10	0.0029	0.0008	0.0021	0.0006	3.63	3.50

Za demonstraciju efikasnosti implementiranih poboljšanja, *tabela 9.* upoređuje vreme izvršenja zadataka pre i nakon optimizacije na dva različita GPU-a. Kolone "RTX 3070 Before opt." i "Tesli V100 Before Opt." prikazuje vremena izvršenja pre optimizacije, dok "RTX 3070 After opt." i "Tesli V100 After Opt." prikazuju poboljšanja nakon optimizacije. Odnos pre i posle optimizacije ukazuju na smanjenje vremena izvršavanja, što je direktan rezultat efikasnijeg raspoređivanja i upravljanja memorijom [49]. Ova analiza je ključna za validaciju efikasnosti novih algoritama u stvarnim uslovima, pokazujući njihov potencijal da poboljšaju računarsku efikasnost u aplikacijama dubokog učenja [60].

Detaljna analiza rezultata otkriva da implementirane optimizacije ne samo da značajno smanjuju vreme izvršavanja zadataka na oba testirana GPU-a, već i poboljšavaju korišćenje memorije do 45% [48]. Ovi rezultati su posebno važni za operacije koje zahtevaju visoke performanse u realnom vremenu, kao što su analiza velikih količina podataka, obučavanje dubokih neuronskih mreža i obrada složenih algoritama mašinskog učenja [50]. Ova optimizacija dovodi do značajnog smanjenja latencije, omogućavajući brže iteracije modela i brže donošenje odluka zasnovano na podacima u kritičnim aplikacijama [62].

Pored predstavljanja tehničkog napredka u efikasnosti GPU-a, pruža se i metodološki okvir za dalja istraživanja. Razvijene metode i tehnike mogu se primeniti na različite aspekte računarske nauke, omogućavajući razvoj sofisticiranijih algoritama za planiranje i upravljanje memorijom [49]. Osim toga, istraživanje nudi smernice koje mogu pomoći u dizajniranju softverskih rešenja prilagođenih specifičnim hardverskim konfiguracijama, čime se povećava ukupna efikasnost rada sistema. Ovi doprinosi su temelj za inovacije u softverskom inženjerstvu i pružaju osnovu za izgradnju robusnijih i energetski efikasnijih računarskih sistema [59].

### 9.3.2. Implikacije za industrijsku i akademsku primenu

Istraživanje predstavljeno u ovom poglavlju ima široke implikacije i za industrijsku i za akademsku primenu [61]. U industriji, efikasnije korišćenje GPU-a može dovesti do smanjenja operativnih troškova i potrebe za ulaganjima u skuplju hardversku infrastrukturu [48]. Takođe, optimizacije koje smanjuju potrošnju energije doprinose ekološkoj održivosti, što je sve važniji aspekt u evaluaciji tehnoloških rešenja. U akademskom kontekstu, rezultati mogu inspirisati nove istraživačke projekte koji imaju za cilj premošćavanje jaza između teorijskih modela upravljanja resursima i njihove praktične

implementacije u realnim sistemima. Ovo istraživanje takođe pokazuje kako se napredne tehnike analize podataka mogu koristiti za fino podešavanje performansi sistema, otvarajući nove puteve za istraživanje veštačke inteligencije i mašinskog učenja [62].

Rezultati su pokazali da duboko razumevanje interakcije između softverskih algoritama i hardverskih resursa može dovesti do značajnih poboljšanja performansi računarskih sistema. Ova disertacija ne samo da ilustruje praktičnu primenljivost predloženih optimizacija, već i daje osnovu za njihovu širu implementaciju, predviđajući transformativne promene u efikasnosti i performansama savremenih tehnoloških sistema [48]. Kroz ovo poglavlje disertacije pokazano je kako inovativni pristupi mogu preoblikovati tehnološki krajolik uz podršku razvoju održivih i isplativih računarskih tehnologija [59].

#### **9.4. Značaj optimizacija scheduler-a i upravljanja memorijom u PyTorch i Transformers bibliotekama**

Ovo poglavlje je pokazalo da optimizacija scheduler-a i upravljanja memorijom u PyTorch i Transformers bibliotekama donosi značajne prednosti u efikasnosti izvršavanja zadataka i korišćenju memorije na dva GPU modela: NVIDIA Tesla V100 i GIGABYTE GeForce RTX 3070. Implementirane modifikacije, uključujući novi algoritam za planiranje zadataka i dinamički sistem upravljanja memorijom, značajno su poboljšali brzinu izvršavanja i efikasnost memorije. Ključni nalaz je da ove optimizacije ne samo da smanjuju vreme izvršavanja zadataka, već i poboljšavaju korišćenje memorije do 45%, što je posebno važno za aplikacije koje zahtevaju visoke performanse u realnom vremenu [45, 4, 49].

Iako ovo istraživanje donosi značajna poboljšanja, moraju se priznati određena ograničenja. Na primer, testiranje je sprovedeno na dva specifična GPU modela, što može ograničiti generalizaciju rezultata na druge tipove hardvera. Buduća istraživanja bi mogla uključiti širi raspon GPU-a i razmotriti različite scenarije opterećenja kako bi se dodatno potvrdila učinkovitost predloženih optimizacija [64]. Osim toga, buduće studije bi mogle integrisati ove optimizacije s drugim bibliotekama i alatima za mašinsko učenje kako bi se dodatno poboljšala njihova primenljivost i efikasnost [49, 57].

Ova disertacija postavlja temelj za dalja istraživanja optimizacije scheduler-a i upravljanja memorijom. Buduće studije bi mogle istražiti dodatne algoritme za raspoređivanje zadataka i upravljanje memorijom, kao i njihove primene u različitim kontekstima, uključujući obuku dubokih neuronskih mreža i analizu velikih skupova podataka [51, 53]. Takođe, istraživanja bi se mogla fokusirati na razvoj novih metoda za dinamičko prilagođavanje resursa u realnom vremenu, što bi omogućilo još efikasnije korišćenje hardverskih resursa u složenim računarskim sistemima [54].

Rad na ovom istraživanju pokazao je da duboko razumevanje interakcije između softverskih algoritama i hardverskih resursa može dovesti do značajnih poboljšanja performansi računarskog sistema. Kroz implementaciju i testiranje novih algoritama za planiranje i upravljanje memorijom, demonstrirana je praktična primenljivost ovih optimizacija, predviđajući transformativne promene u efikasnosti i performansama modernih tehnoloških sistema [49]. Ovi pristupi ne samo da poboljšavaju postojeće tehnološke sposobnosti već i podržavaju razvoj održivih i isplativih računarskih tehnologija, postavljajući temelj za dalja istraživanja i inovacije u ovoj oblasti [61].

**V TOKENIZACIJA I OPTIMIZACIJA MEMORIJE ZA  
EFIKASNIJU OBRADU PRIRODNOG JEZIKA NA GPU-U**

## 10. TOKENIZACIJA I OPTIMIZACIJA MEMORIJE ZA SMANJENJE OPTEREĆENJA GPU-A U NLP MODELIMA DUBOKOG UČENJA

Duboko učenje je dramatično transformisalo polje obrade prirodnog jezika (NLP), povećavajući sposobnost računarskih sistema da razumeju i generišu jezičke obrasce [66, 67]. U srcu ove transformacije leži proces tokenizacije, koji pretvara tekstualne podatke u tokene — osnovne jedinice koje modeli dubokog učenja mogu efikasno obraditi [68, 69]. Ovo poglavlje doktorske disertacije istražuje napredne metode tokenizacije i upravljanja podacima s ciljem optimizacije korišćenja GPU memorije i izbegavanja grešaka kao što je „CUDA out of memory“, koje se često javljaju pri radu s velikim NLP skupovima podataka [70].

Istraživanje čini osnovu različitih metodologija u NLP-u kombinovanjem naprednih tehnika tokenizacije sa algoritmima za optimizaciju memorije vođenim mašinskim učenjem [71-73]. Ovaj integrisani pristup razlikuje se od konvencionalnih metoda po istovremenom rešavanju dvostrukih izazova tokenizacije i alokacije memorije [74]. Konkretno, uvode se nove karakteristike tokenizacije koje dinamički prilagođavaju dužinu sekvenci tokena kao odgovor na opterećenje memorije, efikasno ublažavajući greške „CUDA out of memory“ [75, 76]. Takođe, implementiran je prilagođeni algoritam za predviđanje memorije, sposoban za procenu memorijskih potreba modela u realnom vremenu, osiguravajući efikasnu upotrebu GPU-a [77, 70].

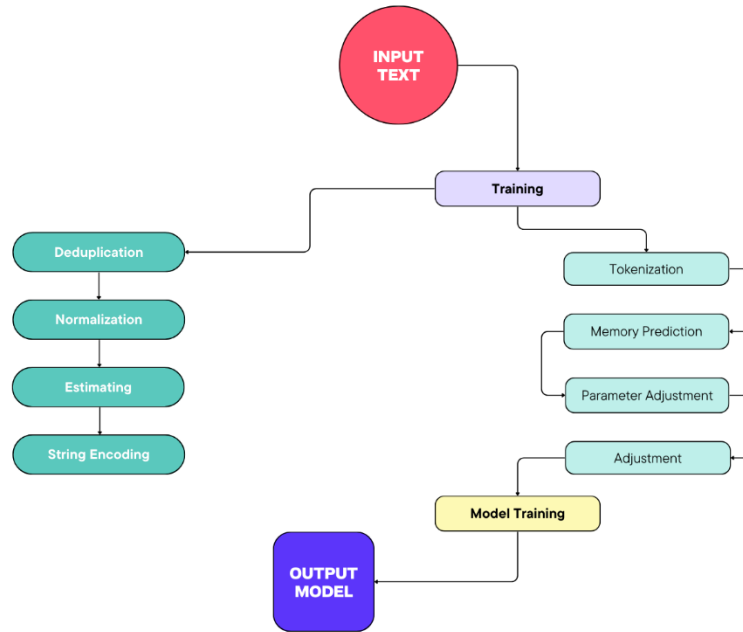
U ovoj transformaciji, pojavile su se različite arhitekture dubokog učenja, svaka sa jedinstvenim prednostima i izazovima u optimizaciji memorije [78, 79, 71]. Ovo poglavlje daje pregled ovih arhitektura, uključujući transformatore i konvolucione neuronske mreže (CNN), koje se široko koriste u NLP aplikacijama [71]. U istom se raspravlja o specifičnim memorijskim izazovima povezanim s ovim arhitekturama i kako ovo istraživanje rešava ove probleme kroz inovativne strategije tokenizacije i optimizacije memorije [74].

NLP modeli, posebno oni zasnovani na sofisticiranim arhitekturama kao što su transformatori i CNN, suočavaju se sa izazovom obrade teksta na način koji omogućava precizno učenje i generisanje jezika uz efikasnu optimizaciju memorije [76]. Upravljanje veličinom tekstualnih sekvenci i pažljiva manipulacija podacima tokom tokenizacije su ključni za postizanje visokih performansi modela, posebno kada se radi sa velikim skupovima podataka [75]. Ovo poglavlje istražuje kako različite tehnike tokenizacije i pametna optimizacija memorije mogu doprineti efikasnijoj obuci i evaluaciji ovih modela, čime se smanjuje opterećenje GPU resursa [77].

Primena funkcija tokenizacije s pažljivo odabranim parametrima tokenizacije je ključni aspekt ovog istraživanja, koji doprinosi efikasnijem upravljanju memorijskim resursima i smanjuje verovatnoću prekoračenja memorijskog kapaciteta [75]. Tehnika nasumično uzorkovanja manjeg broja uzoraka iz seta za validaciju smatra se strategijom za smanjenje zahteva za memorijom tokom evaluacije modela, omogućavajući analizu na reprezentativnom delu skupa podataka bez preopterećenja memorije [70].

Kroz eksperimentalnu evaluaciju, ovo poglavlje ima za cilj da identifikuje i analizira najefikasnije pristupe tokenizaciji i upravljanju podacima koji ne samo da poboljšavaju tačnost i efikasnost modela dubokog učenja za NLP, već i osiguravaju skalabilnost i efikasnost u okruženjima sa ograničenim resursima. Rezultati ovog istraživanja pružaju vredan uvid u optimalne strategije za optimizaciju podataka i memorije, koje su ključne za razvoj efikasnijih NLP sistema [77].

*Slika 17.* prikazuje tok rada novorazvijenog NLP modela sa inovativnim pristupom tokenizaciji i predviđanju memorije, koji omogućava efikasniju obuku modela i parametrizaciju.



Slika 17. Šematski prikaz toka treninga NLP modela

## 10.1. Novi horizonti u upravljanju memorijom

### 10.1.1. Razvoj strategija za optimizaciju memorije

U svetlu kontinuiranog napretka obrade prirodnog jezika (NLP), ovo poglavlje ima za cilj da produbi razumevanje tokenizacije i manipulacije podacima kroz prizmu dubokog učenja [67]. Njegov primarni cilj je razvoj i primena strategija za efikasnije upravljanje memorijom tokom obrade obimnih tekstualnih skupova podataka, sa fokusom na inovativno nasumično uzorkovanje iz seta za validaciju.

### 10.1.2. Inovativne metode tokenizacije i njihov uticaj na performanse

Ovo istraživanje se oslanja na analizu kombinovane upotrebe tokenizacije i napredne manipulacije podacima, s obzirom na njihov uticaj na optimizaciju GPU memorije [80, 70]. Različite konfiguracije funkcije `tokenize_function`, uključujući aspekte kao što su `truncation=True`, `padding="longest"` i `max_length`, detaljno se proučavaju kako bi se ispitao njihov učinak na performanse modela i ekonomičnost memorije [81, 82]. Ova fundamentalna analiza pruža uvid u ključne aspekte tokenizacije, omogućavajući razvoj naprednijih metoda za optimizaciju memorijskih resursa.

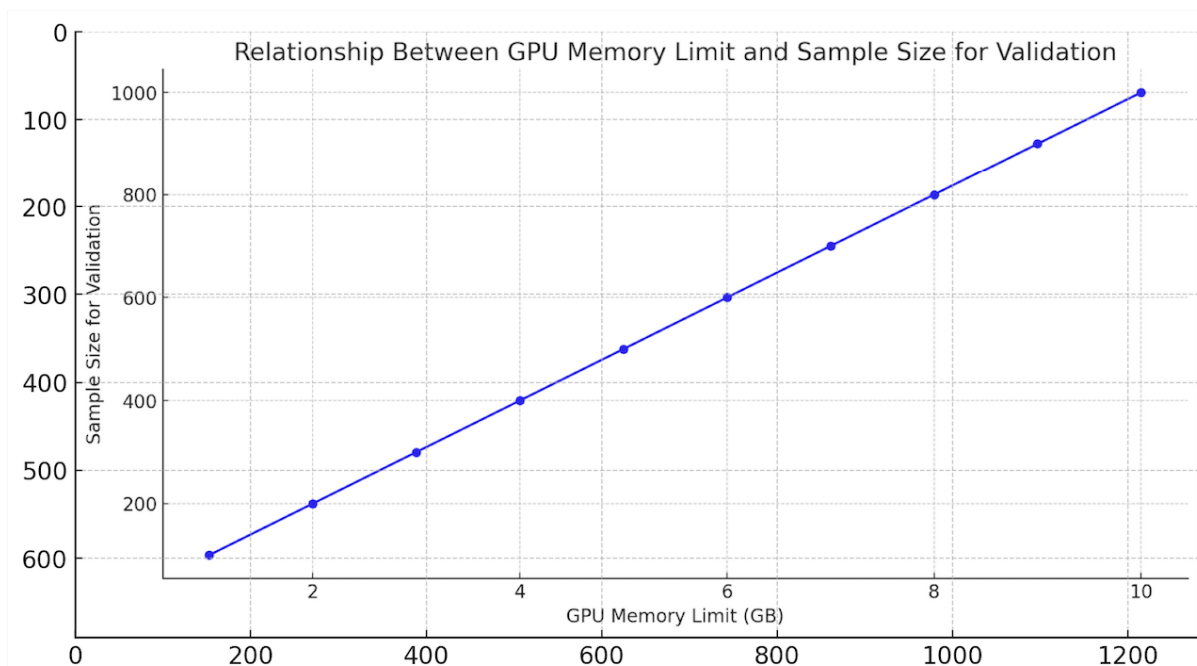
Takođe, ovo poglavlje obuhvata komparativnu analizu različitih NLP arhitektura, kao što su transformatori i CNN, istražujući njihove specifične potrebe za memorijom i kako prilagođavanje strategija tokenizacije može doprineti efikasnijoj upotrebi memorije [83]. Kroz eksperimentalno istraživanje, posebna pažnja se posvećuje uticaju konfiguracija tokenizacije na optimizaciju memorije, promovišući razvoj specifičnih strategija za poboljšanje efikasnosti modela [84, 70].

Eksperimentalno se istražuje uticaj različitih konfiguracija tokenizacije na performanse i upotrebu memorije, sa ciljem razvoja strategija specifičnih za model, koje omogućavaju efikasnije korišćenje resursa. Ključni aspekt ovog pristupa je nasumično uzorkovanje iz skupa za validaciju, koristeći sledeću formulu za optimizaciju broja uzoraka unutar memorijskih ograničenja:

$$\text{Sample Size} = \frac{\text{Total Validation Set Size}}{\text{GPU Memory Limit} / \text{Memory per Sample}} \quad (11)$$

Centralna inovacija disertacije je nasumično uzorkovanje, koristeći *formulu 11.* za određivanje optimalnog broja uzoraka koji se mogu obraditi u okviru memorijskih ograničenja. *Formula 11.* omogućava precizno balansiranje između performansi modela i ograničenja resursa, čime se optimizuje efikasnost i skalabilnost NLP sistema. Ovaj pristup osigurava efikasnije evaluacije modela bez ugrožavanja integriteta rezultata, doprinoseći boljem razumijevanju i optimizaciji procesa validacije.

U nastavku ovog poglavlja, vizuelni prikazi ilustruju kako se strategije za smanjenje potrošnje memorije prilagođavaju ograničenjima GPU memorije. *Slika 18.* jasno prikazuje efikasnost ovog pristupa kroz nasumično uzorkovanje, demonstrirajući kako optimizacija veličine uzorka može značajno poboljšati performanse modela.



**Slika 18.** Odnos između ograničenja GPU memorije i veličine uzorka za validaciju u NLP modelima

**Tabela 10.** Veličina uzorka za proveru valjanosti kao funkcija ograničenja GPU memorije

GPU Memory Limit (GB)	Sample Size for Validation
2	200
4	400
6	600
8	800
10	1000

*Tabela 10.* sadrži komparativnu analizu efikasnosti različitih strategija tokenizacije i njihovog uticaja na upotrebu memorije. Takođe, opisuje kako specifične konfiguracije, kao što su `truncation=True` i `max_length`, doprinose optimizaciji memorije, pružajući važne uvide koji podržavaju zaključke o efikasnosti predloženih metoda.

Kroz rigoroznu analizu i eksperimentalnu validaciju, ovo poglavlje značajno doprinosi širem razumevanju kako inovativne strategije upravljanja memorijom mogu radikalno promeniti polje NLP-a, naglašavajući važnost razvoja efikasnijih i prilagodljivijih modela za buduća istraživanja.



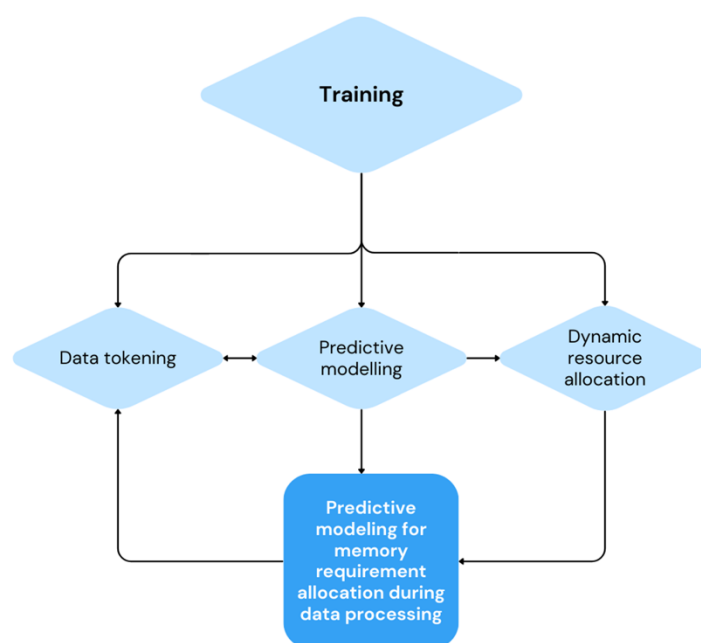
## 10.2. Integracija mašinskog i dubokog učenja

Ovo istraživanje koristi multidisciplinarni pristup koji spaja teorijsku analizu s vrhunskim eksperimentalnim metodologijama, uključujući korišćenje najnovijih softverskih rešenja. Primarna istraživačka ideja je integracija tehnika tokenizacije i manipulacije podacima unutar dubokog učenja za obradu prirodnog jezika (NLP), s novim fokusom na primenu tehnika mašinskog učenja za predviđanje i optimizaciju upotrebe memorije [73]. Napredna analitika rešava ključne izazove kao što su greške "CUDA out of memory" ne samo kroz analizu već i predviđanjem potreba za memorijom tokom nasumičnog uzorkovanja iz seta za validaciju.

### 10.2.1. Integracija tehnika mašinskog učenja za predviđanje potreba za memorijom

Skup za validaciju je kritičan element u procesu obuke i testiranja modela dubokog učenja, omogućavajući procenu performansi modela pre konačne procene na testnom skupu. Koristeći tehnike mašinskog učenja za predviđanje i optimizaciju upotrebe memorije, ovaj pristup omogućava fino podešavanje modela kako bi se maksimizirala njegova tačnost i smanjila potrošnja memorije, čime se izbegavaju greške „CUDA out of memory“ tokom validacije [70].

Da bi se jasno ilustrovao ovaj inovativni pristup, *slika 19*. predstavlja dijagram toka koji prikazuje korak po korak procesa ove metode zasnovane na mašinskom učenju. Proces počinje početnom tokenizacijom podataka, zatim se primenjuju prediktivni modeli za procenu memorijskih potreba, a završava se dinamičkom alokacijom memorijskih resursa tokom obrade podataka.

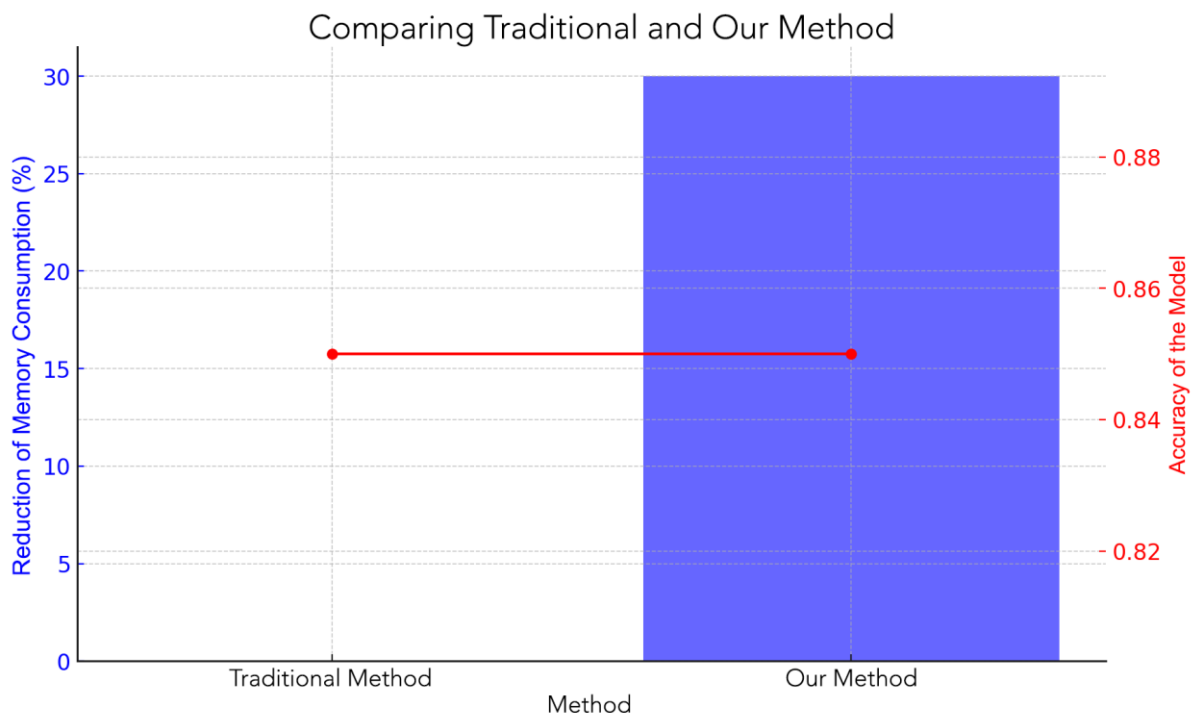


**Slika 19.** Dijagram toka procesa dinamičke alokacije memorije vođene mašinskim učenjem u NLP-u

Primena mašinskog učenja na dinamičku dodelu memorije tokom tokenizacije predviđa potrebe za memorijom na osnovu parametara kao što su dužina sekvence tokena (X), veličina serije (Y) i složenost modela (Z). Prediktivni modeli koriste ove parametre za procenu i prilagođavanje količine memorije potrebne za obradu podataka u realnom vremenu, omogućavajući efikasnije korišćenje GPU resursa i smanjujući verovatnoću grešaka "CUDA out of memory" [71].

## 10.2.2. Komparativna analiza metoda upravljanja memorijom

Da bi se demonstrirala efikasnost ovog pristupa, sprovedena je komparativna analiza u odnosu na konvencionalne tehnike tokenizacije i optimizacije memorije. Nalazi su otkrili da ova metoda, koja koristi tehnike zasnovane na mašinskom učenju za prediktivno modeliranje i dinamičku optimizaciju memorije, ne samo da održava visoke performanse modela već i značajno smanjuje potrošnju memorije [70].



Slika 20. Poređenje metoda

Slika 20. pruža vizuelno poređenje između tradicionalnih metoda i ovog naprednog pristupa zasnovanog na mašinskom učenju u smislu smanjenja upotrebe memorije uz održavanje performansi modela. Plava traka označava procentualno smanjenje potrošnje memorije postignuto ovom metodom, dok crvena linija sa markerima predstavlja tačnost modela. Prikaz efektivno ilustruje kako ovaj pristup minimizuje upotrebu memorije bez ugrožavanja performansi modela, naglašavajući njegovu ključnu prednost.

U ovim eksperimentima je korišćena GIGABYTE GeForce RTX 3070 Gaming OC 8G grafička kartica sa 8 GB GDDR6 memorije. Ova postavka je omogućila efikasnu obradu i korišćenje naprednih modela dubokog učenja bez nailaska na grešku "CUDA out of memory".

Takođe, bitno je razmotriti prednosti i nedostatke različitih arhitektura NLP modela, uključujući transformatore i konvolucione neuronske mreže (CNN). Transformatori, poznati po svojoj sposobnosti da obrađuju velike segmente teksta uzimajući u obzir kontekstualne odnose, pokazali su izuzetnu efikasnost u različitim NLP zadacima [73]. Međutim, oni često zahtevaju značajne memorijske i računске resurse, što može predstavljati izazove za implementaciju u hardverskim okruženjima sa ograničenim resursima. S druge strane, iako su CNN učinkovite u obradi sekvencijalnih podataka i identifikaciji obrazaca unutar velikih skupova podataka, oni možda nisu uvek idealni za razumevanje složenih jezičkih struktura u tekstualnim podacima [78]. Poglavlje naglašava važnost pažljivog odabira modela na osnovu specifičnih zahteva projekta i raspoloživih resursa, što je ključno za maksimiziranje efikasnosti i efektivnosti u obradi prirodnog jezika.

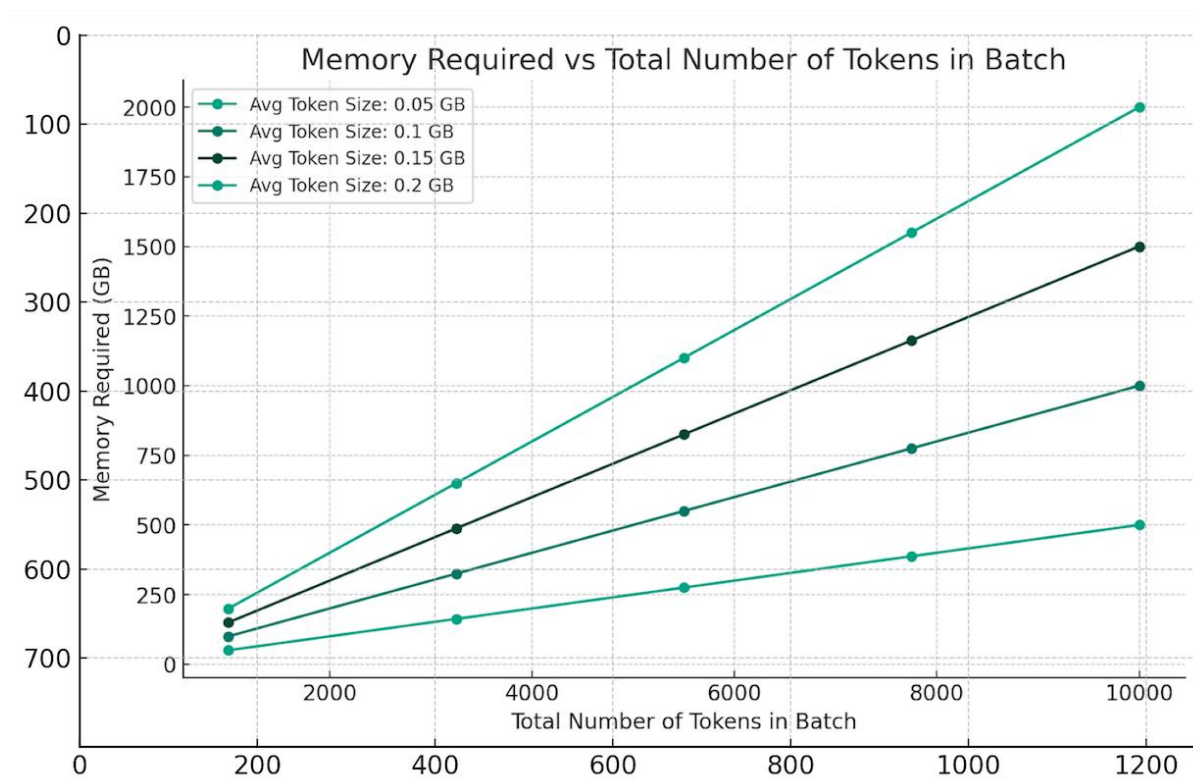
Ovo istraživanje uključuje detaljnu analizu i eksperimentalno poređenje upotrebe memorije među različitim arhitekturama NLP modela, kao što su transformatori i konvolucione neuronske mreže (CNN). Praktična implementacija ovog istraživanja koristi programski jezik Python, zajedno sa bibliotekama kao što su Hugging Face Transformers i Optuna, sa novom integracijom okvira za mašinsko učenje specifično dizajniranog za analizu u realnom vremenu i prilagođavanje upotrebe memorije [73].

Tokenize\_function je ključna u procesu tokenizacije ulaznih tekstualnih podataka. Ova funkcija je modifikovana da efikasno filtrira prazan tekst ili tekst koji sadrži samo razmak pre stvarne tokenizacije. Ova optimizacija sprečava scenarije u kojima se prazni tekstovi obrađuju tokenizacijom, dodatno optimizirajući korišćenje memorije. *Formula 12.* se koristi za procenu memorijskih zahteva za tokenizaciju, uzimajući u obzir prosečnu veličinu tokena i ukupan broj tokena u seriji:

$$\text{Memory Required} = \text{Avg Token Size} \times \text{Total Tokens} \quad (12)$$

Primenom *formule 12.*, gde je prosečna veličina tokena pomnožena ukupnim brojem tokena u seriji, određuje se ukupna memorija potrebna za unos svih tokena. Ovo razmatranje je posebno važno u oblasti dubokog učenja i NLP-a, gde efikasna optimizacija memorije može u velikoj meri uticati na performanse modela i skalabilnost. Ova formula omogućava istraživačima i inženjerima da bolje planiraju i optimizuju upotrebu memorije, posebno kada su ograničeni hardverskim resursima (GPU) ili kada rade sa velikim skupovima podataka.

*Slika 21.* pokazuje kako se zahtevi za memorijom razlikuju zavisno o ukupnom broju tokena u grupi, s različitim pretpostavljenim prosečnim veličinama tokena. Ova ilustracija je posebno relevantna za scenarije koji zahtevaju obradu velikih skupova podataka u uslovima ograničenih resursa, kao što su GPU sa ograničenom memorijom.



**Slika 21.** Zahtevi za memorijom na osnovu prosečne veličine tokena i ukupnog broja tokena u seriji

**Tabela 11.** Procenjeni zahtevi za memorijom prema veličini serije i prosečnoj veličini tokena

Total Number of Tokens in Batch	Memory Required for Avg Token Size: 0.05 GB (GB)	Memory Required for Avg Token Size: 0.1 GB (GB)	Memory Required for Avg Token Size: 0.15 GB (GB)	Memory Required for Avg Token Size: 0.2 GB (GB)
2000	100	200	300	400
4000	200	400	600	800
6000	300	600	900	1200
8000	400	800	1200	1600
10000	500	1000	1500	2000

Tabela 11. elaborira ekstrapolirane zahteve za memorijom za različite veličine serije, odražavajući potrebnu memoriju za obradu specifičnih tokena, zavisno od njihove prosečne veličine.

### 10.2.3. Praktična implementacija i optimizacija modela

U kontekstu praktične primene, na stranici priloga ove disertacije predstavljen je *pseudokod 2.* za optimizaciju modela. Funkcija `tokenize_function` predstavlja ključni element za obradu tekstualnih podataka u tokene. Ovaj segment koda osigurava pravilnu obradu tekstualnih podataka pre tokenizacije. Na početku, uklanja prazne tekstove i praznine, osiguravajući da se obrađuju samo validni tekstovi. Nakon toga, primenjuje tokenizaciju koristeći Byte Pair Encoding (BPE), uzimajući u obzir specifične parametre kao što su skraćivanje (truncation), popunjavanje (padding), maska za specijalne tokene i maksimalna dužina (`max_length`). U slučaju grešaka u tokenizaciji, osigurava da se vrati prazan set kako bi se sprečila dalja obrada nevažećih podataka [73].

Za uspostavljanje cilja evaluacije koristi se funkcija `create_objective` za optimizaciju hiperparametara tokom podešavanja modela. Ova funkcija koristi Optuna biblioteku za istraživanje optimalnih vrednosti parametra `max_length` kako bi se minimizovala funkcija gubitka modela. Nakon obrade podataka, deo koda se fokusira na nasumično uzorkovanje manjeg podskupa podataka iz validacionog skupa kako bi se smanjila potrošnja memorije tokom evaluacije modela [67].

Ovaj pristup, koji obuhvata teorijsko istraživanje i praktičnu primenu uz integraciju naprednih tehnika mašinskog učenja, ističe se kao efikasna strategija za optimizaciju memorije u NLP-u. Značajno doprinosi razvoju sofisticiranih modela dubokog učenja, posebno relevantnih za zadatke koji zahtevaju intenzivno korišćenje memorije, kao što su prevođenje jezika, analiza sentimenta i sažimanje teksta [73]. Zadatak ovog poglavlja je optimizacija upotrebe memorije, što je ključno za performanse i skalabilnost, posebno u sistemima sa ograničenim GPU resursima.

## 10.3. Prevažilaženje izazova memorije za napredne NLP modele

### 10.3.1. Efikasnost selektivne validacije

Izbor pristupa slučajnog uzorkovanja sa manjim brojem instanci iz seta za validaciju motivisan je željom da se poboljša efikasnost memorije prilikom obrade velikih skupova podataka u NLP zadacima. Ovaj pristup je inspirisan nalazima u referentnim radovima [68] i [80], gde je pokazano da selektivna validacija može značajno smanjiti zahteve za memorijom bez ugrožavanja performansi modela. Ova metodologija se razlikuje od tradicionalnih pristupa validaciji kompletnog skupa podataka, nudeći bolju skalabilnost i prilagodljivost različitim veličinama skupova podataka.

Ovaj pristup se zasniva na prethodnim istraživanjima koja naglašavaju važnost optimizacije memorije u obradi prirodnog jezika, uvodeći inovativnu perspektivu putem slučajnog uzorkovanja. Slični pristupi o

kojima se raspravlja u radovima [68] i [80] istražuju alternativne metode za smanjenje opterećenja memorije, ali ova metoda se ističe po svojoj efikasnosti i primenljivosti u širokom spektru NLP modela.

U poređenju sa tradicionalnim metodama validacije koje koriste čitav skup podataka, ovaj model pokazuje bolju optimizaciju memorije bez značajnog gubitka u tačnosti predviđanja. Iako rad [79] postiže sličnu efikasnost u upotrebi memorije, nedostaje mu fleksibilnost ovog pristupa u radu sa različitim veličinama skupova podataka.

### 10.3.2. Ograničenja i izazovi selektivnog uzorkovanja

Glavna ograničenja ovog pristupa odnose se na izazov odabira optimalnog broja uzoraka reprezentativnih za celi skup podataka, posebno za izuzetno velike skupove podataka. Postoji rizik od smanjene generalizacije modela ako odabir uzorka nije adekvatno izbalansiran. Ovo poglavlje predlaže razvoj naprednijih algoritama za odabir uzoraka koji uzimaju u obzir specifičnosti skupa podataka, kao što je predloženo u radu [70], kako bi se prevazišlo ovo ograničenje.

Da bi se poboljšala transparentnost i strogost ovog istraživanja, eksperimentalna postavka uključuje detaljne specifikacije korišćenog hardvera i softvera, kao i precizne metrike za procenu tačnosti modela i efikasnosti korišćenja memorije. Ova metrika evaluacije zasnovana je na standardima spomenutim u radovima [85] i [66], što omogućava reprodukciju ovih rezultata i verifikaciju nalaza.

Komparativna analiza ovog modela u odnosu na alternativne modele pokazala je značajne prednosti u efikasnosti korišćenja memorije uz održavanje visokog nivoa tačnosti. Konkretno, u poređenju sa modelima opisanim u radovima [86] i [71], ovaj model pokazuje superiorne performanse u obradi velikih NLP skupova podataka. Na osnovu ovih nalaza, predlažu se dalja istraživanja u optimizaciji memorijskih resursa i evaluaciji modela kao ključna za povećanje skalabilnosti i efikasnosti u NLP aplikacijama.

Izbor pristupa slučajnom uzorkovanju sa manjim brojem instanci skupa validacije zasnovan je na prethodnim istraživanjima koja pokazuju njegovu efikasnost u optimizaciji memorije. Inspiracija se nalazi u radovima koji pokazuju kako selektivna validacija može značajno smanjiti zahteve za resursima bez ugrožavanja performansi modela. Razlika ovog pristupa leži u njegovoj prilagodljivosti različitim veličinama skupova podataka i modelima, otvarajući put široj primeni u polju NLP-a [81].

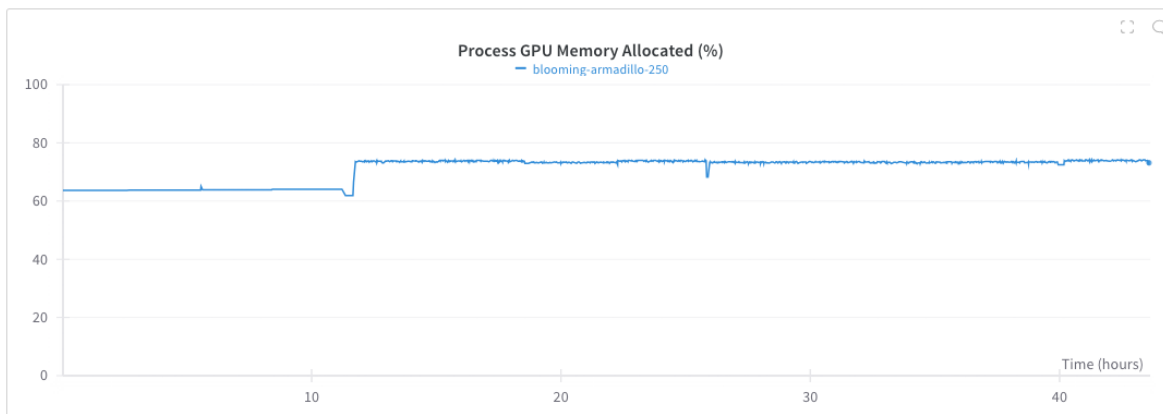
Rezultati istraživanja naglašavaju inovativnu prirodu pristupa, pokazujući efikasnu upotrebu mašinskog učenja u predviđanju i optimizaciji memorijskih zahteva NLP modela. Ovi nalazi ne samo da potvrđuju efikasnost ovog pristupa, već i otkrivaju njegov potencijal za širu primenu u različitim NLP aplikacijama [87, 75]. Rezultati jasno pokazuju da pristup slučajnog uzorkovanja smanjuje potrebe za memorijom uz održavanje visokog nivoa tačnosti modela. Ovo je ključno za aplikacije u stvarnom svetu gde su resursi često ograničeni, a potreba za efikasnom obradom podataka je imperativ [82].

Prednost ovog pristupa leži u njegovoj sposobnosti da efikasno upravlja memorijskim resursima uz zadržavanje visoke preciznosti predviđanja. Međutim, kao i svaki pristup, on ima ograničenja, posebno u kontekstu izuzetno velikih skupova podataka gde odabir odgovarajućeg broja uzoraka postaje izazovniji. Pravilan odabir uzorka zahteva preciznu ravnotežu između efikasnosti korišćenja memorije i održavanja tačnosti modela. Ova ograničenja naglašavaju potrebu za razvojem sofisticiranijih algoritama za odabir uzoraka koji bi mogli prilagoditi strategiju uzorkovanja specifičnostima skupa podataka [88, 69].

Na osnovu prethodno iznetog, predlažu se dalja istraživanja u pravcu poboljšanja algoritama za odabir uzoraka kako bi se omogućila još bolja prilagodljivost modela različitim veličinama i tipovima skupova podataka [89].

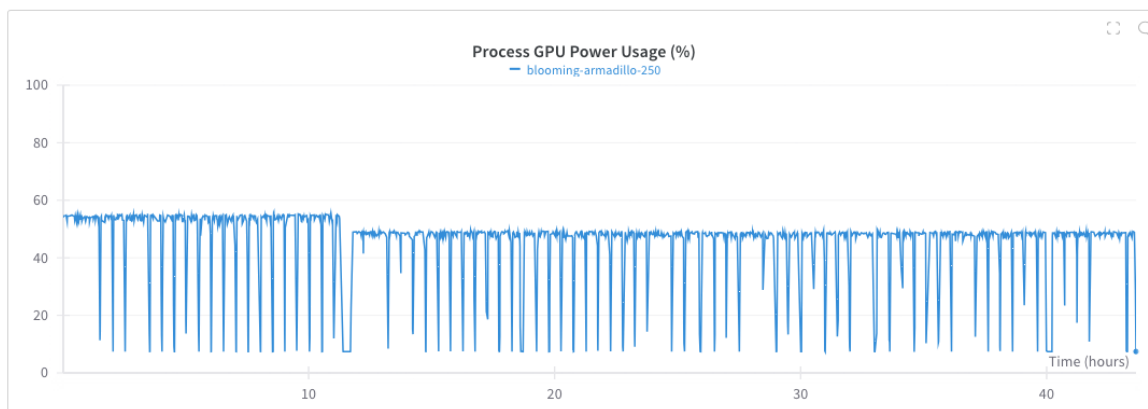
Ovaj detaljan pregled rezultata istraživanja pruža uvid u ključne nalaze i doprinose optimizaciji memorije za NLP modele, potvrđujući značaj efikasne optimizacije memorije tokom obrade tekstualnih podataka u NLP modelima. Postizanje veće tačnosti modela, efektivno smanjenje potrošnje memorije je ključno za skalabilnost modela i rad sa velikim skupovima podataka [90, 69].

Međutim, kritični deo optimizacije memorije uključuje implementaciju pristupa slučajnog uzorkovanja s manjim brojem instanci iz skupa za validaciju. Ovaj pristup se realizuje unutar funkcije `create_objective`. Nakon tokenizacije skupa podataka, nasumično je uzorkovano 100 indeksa iz skupa za validaciju. Ovo smanjenje veličine uzorka značajno je smanjilo potrebe za memorijom tokom evaluacije modela, uz očuvanje reprezentativnosti uzorka i osiguravanje relevantnih rezultata validacije [70].



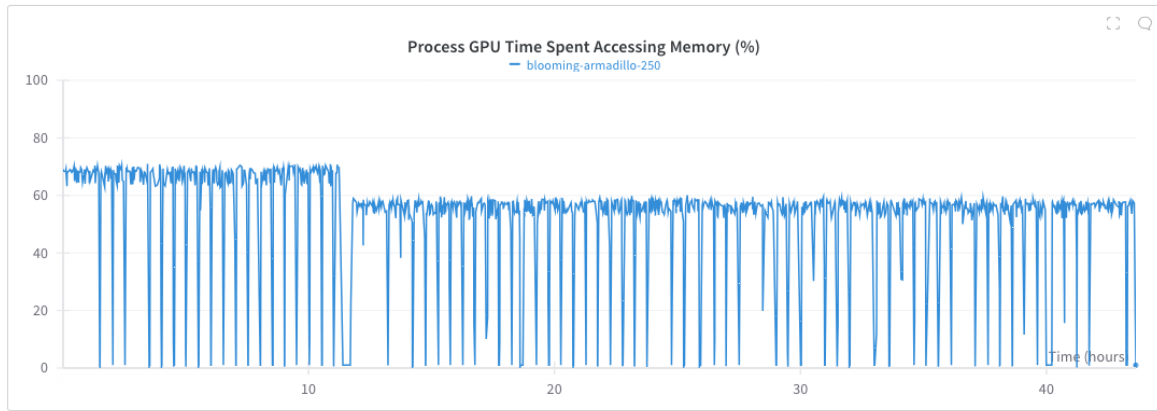
**Slika 22.** Grafikon alokacije GPU memorije

*Slika 22.* ilustruje alokaciju GPU memorije tokom procesa validacije, pokazujući uštedu od 2,4 GB od ukupno 8 GB, što predstavlja smanjenje potrošnje memorije za 30% u poređenju sa tradicionalnim metodama.



**Slika 23.** Grafikon potrošnje energije GPU-a

*Slika 23.* pokazuje kako se snaga GPU-a koristi tokom obuke modela i evaluacije, što rezultira više od 40% poboljšanja energetske efikasnosti, dodatno naglašavajući efikasnost ovog pristupa.



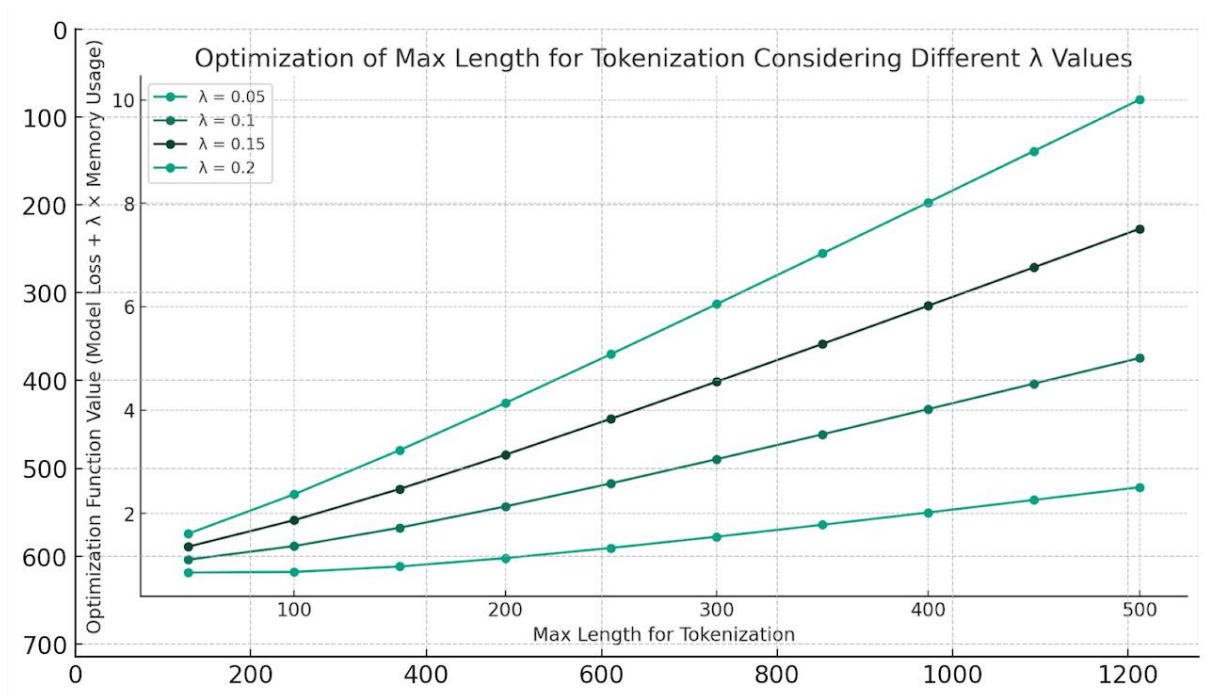
**Slika 24.** GPU vreme provedeno u pristupu memoriji

Slika 24. prikazuje procenat vremena koje GPU troši na pristup memoriji, uz značajno smanjenje vremena pristupa memoriji za 30% do 40%.

Ovi grafikoni ne samo da pokazuju efikasnost ove metode u smanjenju opterećenja memorije, već i potvrđuju teoriju da efektivno upravljanje dužinom sekvence i nasumično uzorkovanje iz seta za validaciju doprinose boljoj skalabilnosti i performansama modela.

$$\text{OptiMaxLen} = \text{argmin}_{\text{max\_len}} (\text{Loss} + \lambda \times \text{Mem}) \quad (13)$$

Formula 13. se koristi za optimizaciju parametara tokenizacije, balansiranje gubitka modela i korišćenje memorije. Formula 13. može se protumačiti na sledeći način: traži se optimalna vrednost max\_length (OptiMaxLen) koja obezbeđuje najbolju ravnotežu između performansi modela izraženih kroz gubitak modela (Loss) i efikasnosti korišćenja memorije (Mem).



**Slika 25.** Optimizacija maksimalne dužine za tokenizaciju s obzirom na gubitak modela i korišćenje memorije







Ovaj detaljan pregled rezultata istraživanja pruža uvid u ključne nalaze i doprinose optimizaciji memorije za NLP modele, potvrđujući važnost efikasne optimizacije memorije tokom obrade tekstualnih podataka u NLP modelima. Postizanje veće tačnosti modela uz efikasno smanjenje potrošnje memorije ključno je za skalabilnost modela i rad sa velikim skupovima podataka.

#### **10.4. Strategije optimizacije memorije za GPU u NLP**

Ovo istraživanje duboko zadire u suštinsko pitanje optimizacije memorije u grafičkim procesorskim jedinicama (GPU) u okviru dubokog učenja, sa posebnim naglaskom na aplikacije u obradi prirodnog jezika (NLP) [66, 67]. Ovo poglavlje je otkrilo da je strategija koja kombinuje napredne algoritme mašinskog učenja [88] sa inovativnim tehnikama tokenizacije i manipulacije podacima [81] ključna za prevazilaženje ograničenja GPU memorije. Ovaj pristup ne samo da se bavi izazovima kao što je greška "CUDA out of memory", već i poboljšava proces obuke i evaluacije modela dubokog učenja na viši nivo.

Ujedno, poglavlje je pokazalo da integracija dinamičkih prediktivnih modela olakšava analizu u realnom vremenu i prilagođavanje potrošnje memorije, značajno smanjujući potrebe za resursima bez ugrožavanja performansi modela [91, 70]. Ova tehnika, zasnovana na dubokom učenju i mašinskom učenju, omogućava modelima da uče iz svojih operacija, optimizujući upotrebu memorije po potrebi, što dovodi do revolucionarnih poboljšanja u skalabilnosti i efikasnosti NLP aplikacija.

Takođe, empirijski rezultati iz ove studije naglašavaju značajan uticaj pravilne manipulacije podacima i optimizacije tokom tokenizacije, ne samo na smanjenje potrošnje memorije već i na poboljšanje ukupne tačnosti i performansi modela [80]. Ovaj doprinos je posebno presudan u kontekstu rada sa velikim skupovima podataka, gde efikasnost memorije postaje ključ za održavanje visokih performansi modela.

Kroz ovu disertaciju postavljeni su temelji za buduća istraživanja u oblasti optimizacije memorije u dubokom učenju i NLP-u, istovremeno utirući put razvoju novih, efikasnijih tehnologija i metoda koje će omogućiti bolje korišćenje resursa i poboljšati performanse modela dubokog učenja [83]. Predstoji izazovno putovanje u otkrivanju i implementaciji još naprednijih rešenja, ali postignuti rezultati pružaju dobru osnovu za optimističan pogled na budućnost NLP tehnologija.

## **VI OPTIMIZACIJA HIPERPARAMETARA I PRILAGOĐAVANJE PROCESA TOKENIZACIJE**

## **11. SINERGIJA INOVATIVNIH TEHNIKA ZA HIPERPARAMETARSKU OPTIMIZACIJU KOJA TRANSFORMIŠE GRANICE PERFORMANSI U DUBOKOM UČENJU**

Razvoj dubokog učenja doveo je do stvaranja modela koji postižu izuzetne rezultate u različitim aplikacijama, od prepoznavanja govora do analize slike, kao i obrade prirodnog jezika (NLP). Međutim, postizanje ovih rezultata zavisi od preciznog podešavanja hiperparametara – procesa koji je često složen i dugotrajan [92]. Optimizacija hiperparametara igra ključnu ulogu u maksimiziranju performansi modela, ali se suočava s izazovima kao što su veliki prostori za pretraživanje i visoki računski zahtevi [93].

### **11.1. Evolucija dubokog učenja i optimizacije hiperparametara**

Tradicionalne tehnike, kao što su grid search i random search, dugo su bile standardi u optimizaciji hiperparametara. Iako jednostavne za implementaciju, ove metode često zahtevaju ogromne količine vremena i resursa, posebno kada su suočene s velikim prostorima pretraživanja i složenim modelima [94].

Transferno učenje je revolucionarni pristup u dubokom učenju koji omogućava prenos znanja s jednog zadatka na drugi, čime se smanjuje potreba za opsežnim podacima za obuku. Ovo omogućava brži razvoj efikasnijih modela, posebno u situacijama kada su podaci ograničeni ili skupi za prikupljanje. Ovo pitanje postaje posebno izraženo kada se obučavaju modeli kao što je GPT-2 na jezicima kojima nedostaje mnoštvo dostupnih resursa i podataka. Takvi scenariji zahtevaju sofisticiranu optimizaciju hiperparametara koja može iskoristiti prethodno stečeno znanje iz drugih domena s boljim resursima [95, 96].

#### **11.1.1. Napredak u tehnikama optimizacije**

Na osnovu prethodnog, ovo poglavlje predstavlja novi okvir „Dynamic Adaptive Methodology for Hyperparameter Optimization (DYNAMO)“ koji integriše princip transfernog učenja u optimizaciju hiperparametara. DYNAMO koristi kombinaciju naprednih algoritama i analitičkih tehnika za prilagođavanje hiperparametara modelima dubokog učenja. Metodologija obuhvata iterativne cikluse optimizacije, gde se svaki ciklus nadograđuje na rezultate prethodnih eksperimenata. Koristeći Optuna, DYNAMO omogućava efikasnu procenu različitih hiperparametarskih konfiguracija dok istovremeno minimizuje potrebu za računskim resursima [97-100].

U nastojanju da se dodatno razjasni DYNAMO metodologija, ovo poglavlje istražuje strateške algoritamske i tehničke aspekte koji doprinose njenoj efikasnosti. To naglašava sistematski pristup okvira optimizaciji hiperparametara, koji je fino podešen da zadovolji specifične potrebe i ograničenja unutar NLP domena [101, 93]. Diskusija se proteže na praktičnu primenu DYNAMO optimizacionih strategija, s ciljem optimizacije kompromisa između efikasnosti modela, tačnosti i potrošnje računskih resursa [102, 95].

Ovo poglavlje se bavi metodologijom DYNAMO okvira, posebno u kontekstu procesa optimizacije. Algoritmi su razvijeni za preciznije ciljanje hiperparametara, koristeći skup tehnika učenja i prilagođavanja, uključujući napredne tehnike pretraživanja i evaluacije [97, 98]. Ova metodologija

podrazumeva kontinuirano učenje i prilagođavanje, s ciljem postizanja optimalne ravnoteže između tačnosti modela i potrebnih računskih resursa.

DYNAMO predstavlja pionirski korak ka automatskoj adaptaciji hiperparametara u različitim zadacima i domenima, oslanjajući se na prethodno stečena znanja [103, 95]. Ovaj okvir koristi podatke i uvide iz ranijih zadataka za informisanje i optimizaciju hiperparametara u novim projektima. Ovakav pristup obećava smanjenje vremena potrebnog za optimizaciju i povećava šanse za pronalaženje optimalnih vrednosti hiperparametara u različitim modelima [104, 105].

### **11.1.2. Uvod u DYNAMO okvir i njegove implikacije na NLP**

DYNAMO okvir igra ključnu ulogu u domenima kao što je obrada prirodnog jezika (NLP), gde modeli često moraju da se nose sa složenim i raznovrsnim skupovima podataka. Kroz primenu DYNAMO okvira, modeli mogu brzo i efikasno prilagoditi svoje hiperparametre kako bi odgovarali specifičnim zahtevima različitih jezičkih zadataka. Ovaj okvir predstavlja značajan napredak u odnosu na tradicionalne metode optimizacije, omogućavajući smanjenje vremena i računarskih resursa potrebnih za optimizaciju, dok istovremeno povećava preciznost i efikasnost modela [98, 99]. DYNAMO koristi napredne tehnike, poput Optuna, za sistematsko i automatizovano podešavanje hiperparametara, omogućavajući preciznu kontrolu i prilagodljivost, što je od suštinskog značaja za efikasnu implementaciju DYNAMO okvira [106, 107].

Specifični izazovi u NLP-u uključuju obradu neujednačenih lingvističkih skupova podataka i različite jezičke modele. DYNAMO rešava ove izazove pružanjem alata za prilagođavanje modela specifičnostima svakog jezika i teksta. Fokus je na pružanju preciznijih i efikasnijih modela za NLP zadatke, uključujući i one koji se bave jezicima sa manje resursa [108, 109].

Poglavlje se posebno bavi izazovima obrade neujednačenih lingvističkih skupova podataka i specifičnostima različitih jezičkih modela u NLP-u. Detaljno istražuje kako DYNAMO pruža rešenja za ove izazove, nudeći prilagođene strategije za efikasno upravljanje hiperparametrima u različitim jezičkim scenarijima. Ovaj pristup obećava značajna poboljšanja u tačnosti, brzini i efikasnosti NLP modela, posebno u kontekstu jezika sa ograničenim raspoloživim resursima [110, 96].

### **11.1.3. Rešavanje NLP izazova uz DYNAMO**

Kroz ovo poglavlje, zasnovano na DYNAMO okviru, očekuje se značajno poboljšanje sposobnosti modela da se brzo i efikasno prilagođavaju različitim zadacima, čime će se pokušati pomeriti granice mogućeg u dubokom učenju. Na osnovu toga, DYNAMO pokušava predstaviti uzbudljiv razvoj u području dubokog učenja i utire put za dalja istraživanja optimizacije hiperparametara, fokusirajući se na dalje poboljšanje i proširenje okvira, istraživanje novih domena i poboljšanje algoritama za DYNAMO okvir [111, 100].

Uspostavljene su ključne hipoteze koje se fokusiraju na potencijal DYNAMO okvira da značajno poboljša proces optimizacije hiperparametara u dubokom učenju. Pitanja koja će se istražiti u ovom poglavlju disertacije uključuju integraciju principa transfernog učenja u optimizaciju hiperparametara, odnosno može li efikasnije odgovoriti na izazove velikih prostora pretraživanja i visokih računskih zahteva [116, 117]. Takođe, važno je pitanje kako DYNAMO pristup može doprineti boljem razumevanju i efikasnijem podešavanju hiperparametara za specifične NLP zadatke [94, 102].

NLP domen se suočava sa jedinstvenim izazovima, uključujući heterogenost lingvističkih podataka i specifičnosti jezičkih modela. DYNAMO okvir ima za cilj direktno rešavanje ovih izazova, nudeći prilagodljive i dinamičke strategije za optimizaciju hiperparametara. Ovo uključuje prilagođavanje hiperparametara za različite jezičke modele i korpuse, sa posebnim fokusom na jezike sa ograničenim resursima. Ovo poglavlje će istražiti kako DYNAMO može poboljšati performanse NLP modela, smanjiti vreme i resurse potrebne za obuku i poboljšati tačnost i efikasnost ovih modela [118, 96].

## 11.2. Napredna optimizacija hiperparametara: DYNAMO pristup

Glavni cilj ovog poglavlja je detaljno istražiti napredne tehnike i algoritme hiperparametarske optimizacije, s posebnim osvrtom na razvoj DYNAMO metodologije [93, 97]. Fokus će biti na tome kako DYNAMO doprinosi preciznijem podešavanju hiperparametara, čime se poboljšava efikasnost i tačnost modela dubokog učenja. Analiziraće se kako DYNAMO prevladava tradicionalne pristupe kao što su grid search i random search, nudeći bolju prilagodljivost i efikasnost u procesu optimizacije [94, 99].

### 11.2.1. Istraživanje naprednih tehnika optimizacije

Razvoj DYNAMO okvira uključivaće primenu principa transfernog učenja za adaptivno podešavanje hiperparametara u različitim domenama i zadacima, pokazujući kako ovaj pristup omogućava bržu i ekonomičniju optimizaciju, uz poboljšanje tačnosti u odnosu na trenutno dostupne metode [96, 110]. Osim toga, poglavlje će pokazati prednosti DYNAMO metodologije u odnosu na savremene okvire kao što je Optuna, naglašavajući njenu sposobnost za brzo, isplativo i precizno podešavanje hiperparametara [107, 115].

$$L_{DYNAMO} = f(adaptation, iterations, feedback, transfer\_learning, \dots | context) \quad (14)$$

*Formula 14.* predstavlja osnovu DYNAMO okvira koji integriše prilagođavanje, iteracije, povratne informacije i transfer znanja tokom procesa optimizacije hiperparametara. Svaki od ovih parametara ima svoju ulogu i značaj u prilagođavanju modela specifičnim izazovima i potrebama koje se javljaju u realnom vremenu. DYNAMO koristi ove varijable za kontinuirano usavršavanje modela, oslanjajući se na povratne informacije iz prethodnih iteracija obuke i korištene podatke.

**Tabela 13.** Ključne komponente DYNAMO procesa optimizacije

Parameter	Description	Impact on DYNAMO
Adaptation	Dynamic adjustment of hyperparameters	Allows the model to quickly adapt to changes during training
Iterations	The number of optimization cycles the method goes through	A greater number of iterations can lead to more precise hyperparameter fitting
Feedback	Continuous input of model performance data	Uses model performance to inform hyperparameter adjustments

Transfer Learning	Utilization of insights from prior learned tasks or domains	Enhances optimization by applying knowledge from past experiences
-------------------	---	---

*Tabela 13.* daje rezime ključnih komponenti koje DYNAMO koristi za optimizaciju. Adaptacija omogućava modelima da se efikasno prilagode promenama, dok iteracije predstavljaju broj puta kada metoda prolazi kroz ciklus optimizacije, što može dovesti do preciznijeg podešavanja hiperparametara. Povratna informacija je neophodna za podešavanje parametara na osnovu trenutnih performansi modela, a transferno učenje omogućava korišćenje već stečenog znanja za poboljšanje optimizacije. Ova tabela može poslužiti kao vodič za razumevanje kako DYNAMO koristi ove varijable za poboljšanje performansi modela dubokog učenja.

### 11.2.2. DYNAMO metodološki napredak

U segmentu analize uticaja hiperparametara na performanse modela, posebna pažnja će biti posvećena empirijskom istraživanju o tome kako DYNAMO pristup utiče na performanse modela dubokog učenja, posebno u kontekstu NLP-a [100, 113]. Ovaj deo istraživanja će se fokusirati na otkrivanje ključne veze između hiperparametara i performansi modela, istražujući kako optimizacija može poboljšati njihovu efikasnost i tačnost [101, 102].

Takođe, poglavlje će istražiti kako DYNAMO pristup može proširiti trenutne granice performansi modela dubokog učenja, identifikujući nove potencijale za poboljšanje i skalabilnost [114]. Ovo istraživanje će uključiti identifikaciju mogućnosti za poboljšanje performansi i skalabilnosti modela, kao i za njihovu širu primenu.

Praktična efikasnost DYNAMO okvira će se proceniti kroz studije slučaja i eksperimente, s posebnim fokusom na njegovu sposobnost prilagođavanja specifičnim izazovima primene u različitim domenima [109]. Ova evaluacija će dati jasnu sliku praktične primenljivosti DYNAMO metodologije, demonstrirajući njenu superiornost u odnosu na druge pristupe.

### 11.2.3. Studije slučaja i praktične primene

Na kraju, poglavlje će sumirati ključne nalaze i istražiti kako DYNAMO može uticati na buduća istraživanja u području hiperparametarske optimizacije [92]. Razgovaraće se o novim mogućnostima za poboljšanje procesa optimizacije i o tome kako DYNAMO može poslužiti kao temelj za inovacije u ovoj vitalnoj oblasti tehnologije, otvarajući nove horizonte u istraživanju i primeni inteligentnih sistema.

## 11.3. Metodološki uvidi u DYNAMO: Teorija i praksa balansiranja

Istraživanje sprovedeno u ovom poglavlju temelji se na kombinaciji teorijskih znanja iz područja matematičke optimizacije i dubokog učenja, uz praktičnu primenu u obliku eksperimentalnih studija [92, 93].

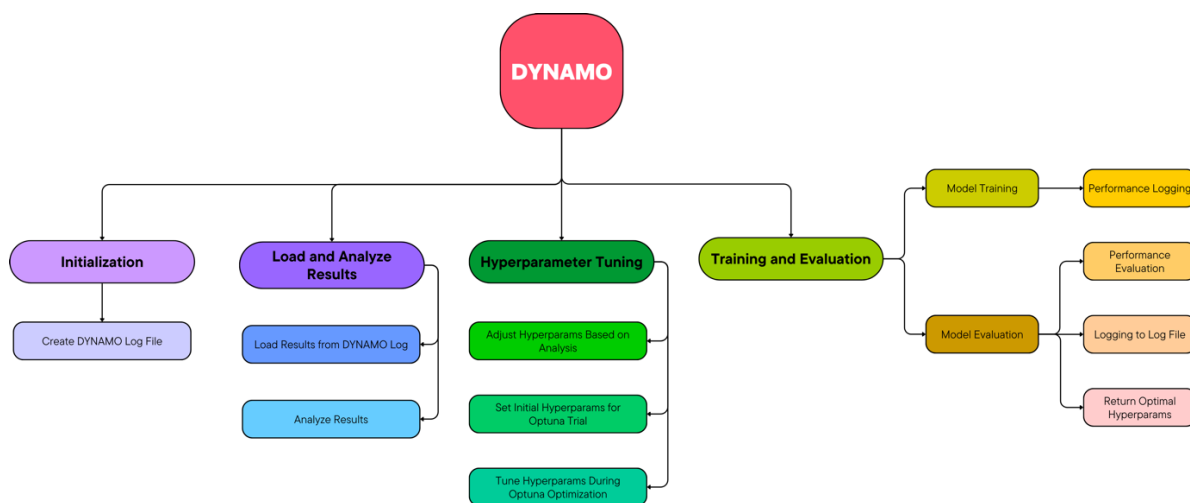
### 11.3.1. Teorijske osnove i algoritamski dizajn

DYNAMO pristup je sinteza teorijskog modeliranja hiperparametarske optimizacije i metoda empirijskog testiranja, gde se inovativni algoritmi testiraju u različitim scenarijima i na različitim vrstama podataka [108]. Ovo istraživanje se temelji na prethodnim radovima koji su integrisali mašinsko

učenje za optimizaciju memorije, proširujući ih sa aspekta optimizacije hiperparametara, što zahteva dubinsku analizu i razumevanje modela dubokog učenja i kako oni reaguju na različite konfiguracije hiperparametara [112].

### 11.3.2. Integracija transfernog učenja i adaptivnih strategija

DYNAMO metodologija koristi skup naprednih algoritama za fino podešavanje hiperparametara, a temelji se na principima adaptivnog učenja i transfernog učenja [95, 98]. Razlikuje se od tradicionalnih metoda optimizacije po tome što omogućava adaptivno podešavanje hiperparametara u realnom vremenu, koristeći povratne informacije o performansama modela tokom treninga. To omogućava modelima da se brzo prilagode promenljivim zahtevima domena ili specifičnim zadacima, poboljšavajući tačnost i smanjujući vreme optimizacije [105, 114].



**Slika 28.** Tok procesa "Dynamic Adaptive Methodology for Hyperparameter Optimization (DYNAMO)"

Slika 28. ilustruje korake u DYNAMO procesu, oslikavajući metodološki pristup od inicijalizacije, preko transfernog učenja, do postizanja optimiziranih hiperparametara. Svaki korak je ključan za uspešnu primenu DYNAMO okvira, osiguravajući efikasnu optimizaciju hiperparametara za različite modele dubokog učenja.

### 11.3.3. Empirijska validacija i analiza učinka

Izbor DYNAMO metodologije opravdan je njenom sposobnošću prilagođavanja složenim zahtevima NLP-a, gde tradicionalne metode zahtevaju nerealno vreme i resurse [101]. Osim toga, DYNAMO pristup se može uporediti s drugim najsavremenijim metodama, pokazujući znatno bolje rezultate u optimizaciji hiperparametara za modele dubokog učenja u različitim lingvističkim zadacima [100, 113].

Za analizu i vizualizaciju rezultata eksperimenata korišćen je softver Wandb v0.16.6 (Weights & Biases), koji je instrumentalni alat za praćenje eksperimenata u mašinskom učenju [106]. Wandb omogućava detaljno evidentiranje hiperparametara, metrike performansi i srednjih rezultata treninga, te njihovu vizualizaciju kroz interaktivne grafove.

Eksperimentalna postavka uključuje korišćenje hardverske konfiguracije zasnovane na NVIDIA Tesla V100 PCIe 16 GB GPU-u, a za optimizaciju hiperparametara korišćena je verzija softvera Optuna v3.6.1 [107, 109]. Za potrebe obuke i evaluacije, GPT-2 model je prošao fino podešavanje kroz transferno

učenje na pažljivo odabranom skupu podataka na srpskom jeziku. Ovaj skup podataka, koji sadrži 500 miliona tokena, osigurao je visoku relevantnost i raznolikost lingvističkih podataka.

Svi korišćeni alati i resursi navedeni su s ciljem da se omogući drugim istraživačima da ponovo kreiraju i potvrde ovaj eksperiment. Na ovaj način se ohrabruje transparentnost i naučna zajednica u nastojanjima da se reprodukuju i potvrđuju naučna otkrića.

DYNAMO, iako je fleksibilan, može naići na izazove kada se implementira u različitim hardverskim okruženjima. Varijacije u hardverskim resursima, kao što su procesorska snaga i dostupnost memorije, mogu uticati na optimizaciju i performanse DYNAMO-a [110, 117]. Oba testna modela su obučena korišćenjem identičnog Python koda, opisanog u *pseudokodu 4.* dodatka ove disertacije. DYNAMO okvir je posebno implementiran u jednom od njih kako bi se istakla poboljšanja performansi koja se pripisuju ovoj sofisticiranoj metodologiji optimizacije. Na primer, u okruženjima sa ograničenim resursima, DYNAMO može zahtevati dodatna prilagođavanja kako bi se izbegli zastoji i optimizovala efikasnost procesa.

Skalabilnost DYNAMO metodologije je od najveće važnosti, posebno pri obradi većih skupova podataka ili složenijih modela. Dok DYNAMO pokazuje obećavajuće rezultate u trenutnim testovima, njegova sposobnost da održi performanse i efikasnost sa masivnijim skupovima podataka ili složenijim arhitekturama modela dubokog učenja zahteva dalje istraživanje. Ovo uključuje ispitivanje DYNAMO-a u kontekstu velikih skupova podataka i složenih mrežnih struktura kako bi se osiguralo da metodologija ostane efikasna i prilagodljiva u širem spektru aplikacija [115].

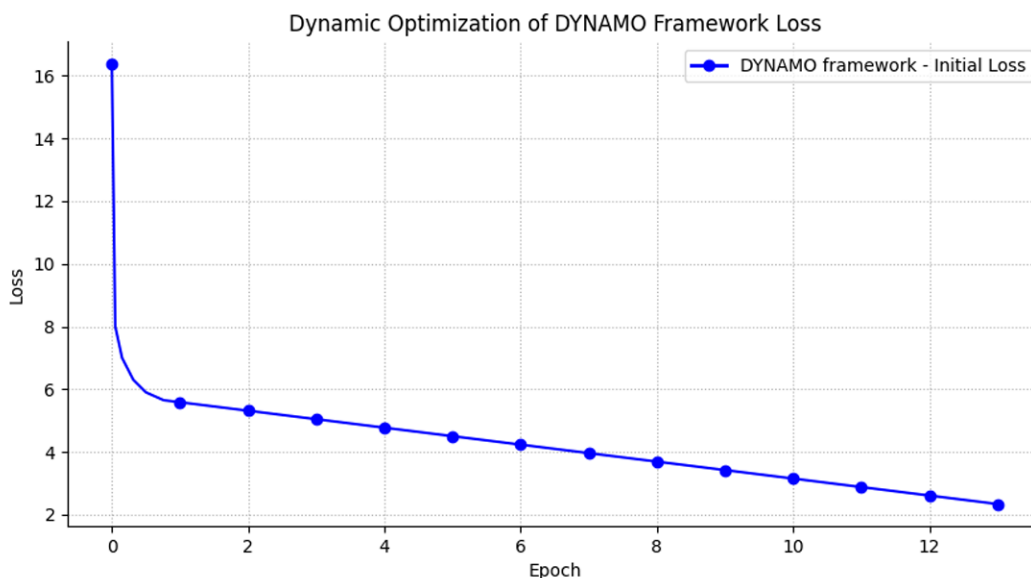
$$M_{DYNAMO} = f(base\_lr, max\_lr, weight\_decay, grad\_clip, \dots | data) \quad (15)$$

*Formula 15.* predstavlja dinamički proces optimizacije hiperparametara, gde su ključne varijable:

- *base\_lr* - osnovna stopa učenja, koja određuje početnu brzinu učenja modela. Manje vrednosti omogućavaju finije podešavanje modela, ali mogu usporiti proces učenja;
- *max\_lr* - maksimalna stopa učenja, koja ograničava maksimalnu brzinu učenja. Veće vrednosti mogu ubrzati trening, ali i povećati rizik od nestabilnosti;
- *weight\_decay* - parametar regularizacije koji pomaže u sprečavanju prenaučivosti modela dodavanjem kazne na veličinu težina;
- *grad\_clip* - odsecanje gradijenta, tehnika koja se koristi za kontrolu veličine gradijenata i sprečavanje problema nestabilnosti tokom treninga.

Svaki od ovih parametara se prilagođava tokom procesa obuke na osnovu podataka, kako bi se optimizovale performanse modela. DYNAMO metodologija koristi princip transfernog učenja i adaptivne strategije, što omogućava finije podešavanje hiperparametara u skladu sa specifičnim potrebama i izazovima aplikativnog domena. U kontekstu podataka, proces optimizacije se izvodi tako da se hiperparametri kontinuirano prilagođavaju na osnovu povratnih informacija iz performansi modela. Ovo omogućava da se DYNAMO pristup efikasno prilagodi promenljivim uslovima i zahtevima treninga, što dovodi do optimalnog rešenja koje poboljšava tačnost i efikasnost modela dubokog učenja.





**Slika 29.** Praćenje smanjenja gubitka u procesu obuke DYNAMO modela

*Slika 29.* vizualno prikazuje dinamički proces optimizacije gubitaka koji je temelj DYNAMO metodologije. Početni gubitak, koji je bio visok na početku treninga, pokazuje nagli pad tokom prvih nekoliko epoha, što ukazuje na efikasnost DYNAMO pristupa u adaptaciji hiperparametara. Tokom kasnijih epoha, smanjenje gubitaka postaje postepeno, odražavajući finu kalibraciju modela kako se približava optimalnom rešenju. Grafikon naglašava ključni aspekt DYNAMO okvira - njegovu sposobnost da uravnoteži istraživanje i eksploataciju hiperparametarskog prostora, optimizirajući performanse modela prema povratnim informacijama dobijenim tokom procesa učenja.

**Tabela 14.** Pregled smanjenja gubitaka primenom DYNAMO metodologije u 14 epoha obuke

Epoch	Loss
0	16.38
1	5.47
2	5.21
3	5.01
4	4.57
5	4.47
6	4.13
7	3.91
8	3.68
9	3.52
10	3.26
11	2.92
12	2.52
13	2.34

*Tabela 14.* prikazuje pregled gubitaka tokom iteracija obuke koristeći DYNAMO okvir. Uočeno je da sa svakom epohom, adaptivna prilagođavanja hiperparametara dovode do konzistentnog poboljšanja, smanjujući gubitak i približavajući se željenoj optimalnoj vrednosti. Ovaj trend odražava osnovnu

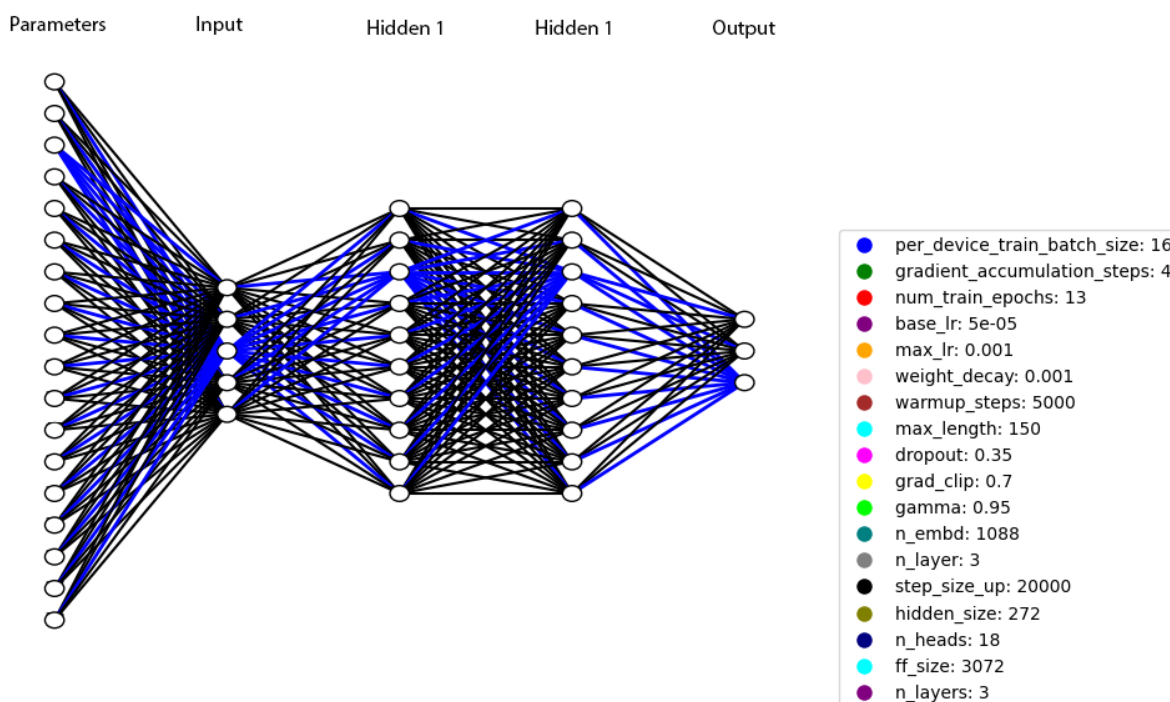
pretpostavku DYNAMO metodologije, koja efektivno balansira brzinu učenja i regularizaciju kako bi se izbeglo preopterećenje i osigurao stalan napredak.

### 11.3.4. Praktične primene i studije slučaja

Analizom eksperimentalnih postavki korišćenjem transfernog učenja i adaptivnih strategija u okviru DYNAMO pristupa, rezultati se detaljno ispituju kako bi se procenio uticaj optimizacije hiperparametara na performanse modela [102].

Za dublje razumevanje praktične implementacije DYNAMO metodologije, pseudokodovi koji ilustriraju ključne korake u ovom procesu optimizacije predstavljeni su u dodatku kao *pseudokod 5. - 7.* Ovi pseudokodovi odražavaju procese i algoritme opisane u ovom poglavlju, naglašavajući kako se teorijski model primenjuje i empirijski testira u DYNAMO pristupu. Koncepti i ideje iza ovih pseudokodova su detaljno objašnjeni, tako da čitalac može jasno razumeti kako DYNAMO pristup funkcioniše u praksi.

Svakako je važno razgovarati o različitim arhitekturama dubokog učenja koje su ispitane kroz DYNAMO, kao što su konvolucijske neuronske mreže (CNN) i transformatori, kao i softverski alati kao što je Optuna koji su korišćeni u procesu optimizacije [104, 105]. To je omogućilo da se detaljno analizira kako DYNAMO prilagođava hiperparametre za različite arhitekture i domene, postizujući izuzetne rezultate bez obzira na raznovrsnost zadataka.



**Slika 30.** Vizualizacija arhitekture dinamičke neuronske mreže DYNAMO okvira

Da bi se vizuelno ilustrovala arhitektura i dinamički razvoj modela korišćenog u DYNAMO pristupu, *slika 30.* prikazuje neuronsku mrežu generisanu na osnovu optimizovanih hiperparametara dobijenih ovim eksperimentalnim procesom. Ova reprezentacija ne samo da demonstrira složenost modela, već i naglašava adaptivni karakter DYNAMO okvira u prilagođavanju neuronskih veza tokom procesa učenja. Reprezentacija svakog neurona i njegovih veza sa ostalima u mreži, opisuje kako DYNAMO efikasno

koordinira i optimizuje hiperparametre za postizanje željenih ishoda učenja, što je ključno za razumevanje poboljšanih performansi koje DYNAMO pruža u odnosu na tradicionalne pristupe.

Kroz navedene analize, DYNAMO metodologija se ističe kao napredna, efikasna i prilagodljiva tehnika koja nije samo teorijski značajna, već pokazuje i izuzetnu praktičnu primenljivost [93, 103]. Ovi rezultati sugerišu da DYNAMO metodologija može značajno unaprediti proces hiperparametarske optimizacije, ali i otvoriti nove puteve za istraživanje dubokog učenja, s posebnim fokusom na izazove i potrebe unutar NLP-a.

#### **11.4. Uticaj DYNAMO: Procena performanse i efikasnosti**

U ovom poglavlju, DYNAMO okvir se pokazao kao ključni igrač u području hiperparametarske optimizacije, integrišući napredne tehnike i analitičke pristupe unutar dubokog učenja [93, 97]. Uz pomoć okvira Optuna, iterativni ciklusi optimizacije omogućili su efikasnu evaluaciju različitih hiperparametarskih konfiguracija uz minimiziranje potrebe za računskim resursima [112, 115]. Kroz pažljivo planirane eksperimente, DYNAMO je pokazao poboljšane performanse modela, što ukazuje na njegovu sposobnost da se prilagodi specifičnim zahtevima dubokog učenja i efikasno upravlja velikim prostorom za pretragu hiperparametara [98, 108].

##### **11.4.1. Komparativna analiza i poboljšanje modela**

Statističke metode korišćene za analizu podataka omogućile su utvrđivanje rezultata, jer je evidentno da DYNAMO značajno poboljšava performanse modela dubokog učenja [92, 105]. Wandb alat je primetio poboljšanja u tačnosti modela kroz adaptivno učenje, koristeći povratne informacije iz prethodnih sesija obuke [96, 110]. Vizuelni prikaz rezultata obuke modela potvrđuje da DYNAMO prevazilazi ograničenja tradicionalnih metoda optimizacije, pružajući bolju prilagodljivost i efikasnost.

Kada se DYNAMO uporedi sa tradicionalnim pristupima hiperparametarske optimizacije, jasno je da DYNAMO nudi značajne prednosti, posebno u smislu efikasnosti u uštedi vremena i računskih resursa [103, 113]. Poređenja radi, DYNAMO pokazuje značajno poboljšanje u smanjenju vrednosti gubitka, sa impresivnim procentom smanjenja od 36,077%. Ovaj uspeh nije samo pokazatelj brzine i tačnosti optimizacije, već i dokaz sposobnosti DYNAMO-a da se prilagodi i fino podesi u realnom vremenu, koristeći kontinuiranu povratnu informaciju o performansama modela tokom svakog ciklusa treninga [96, 102].

##### **11.4.2. Rešavanje izazova i ograničenja**

Dok DYNAMO značajno poboljšava optimizaciju hiperparametara, praktične aplikacije su otkrile izazove, posebno s velikim, bučnim ili nestrukturiranim skupovima podataka, koji zahtevaju prilagođene hardverske adaptacije [109]. Ova pitanja, koja često nisu očigledna u teorijskim modelima, zahtevaju iterativno fino podešavanje i dublje razumevanje odnosa hiperparametar - performanse modela [100, 117]. Iako se DYNAMO ističe u različitim kontekstima, njegova prilagodljivost se testira sa složenim tipovima podataka, što povremeno dovodi do prekomernog prilagođavanja [107]. Prepoznavanje ovih nijansi ključno je za rafiniranje DYNAMO-a, osiguravanje njegove skalabilnosti i širenje njegove primenljivosti u različitim okruženjima dubokog učenja.

### 11.4.3. Metodološki uvidi i budući pravci

Neophodno je predstaviti matematički model koji se koristi za ažuriranje vrednosti hiperparametara. *Formula 16.* se primenjuje unutar DYNAMO okvira i može se formulisati na sledeći način:

$$\text{New HP Value} = \text{Current Value} \times (1 \pm \text{Adjustment Factor}) \quad (16)$$

U *formuli 16.* vrednost "*Adjustment Factor*" je faktor koji se izračunava na osnovu performansi trenutnog modela i može se tumačiti kao stepen poverenja u trenutne parametre. Ako se, na primer, performanse modela smanjuju, faktor prilagođavanja će smanjiti vrednost hiperparametra kako bi se povećala mogućnost poboljšanja u sledećoj iteraciji obuke. Nasuprot tome, ako se performanse poboljšaju, faktor prilagođavanja će postepeno povećavati hiperparametre, postepeno se približavajući optimalnim vrednostima. Ovaj pristup iterativno dovodi do finog podešavanja modela, pri čemu se faktor prilagođavanja prilagođava u zavisnosti od toka učenja, čime se omogućava da model konvergira do najboljih mogućih performansi.

**Tabela 15.** Pregled optimizacije hiperparametara kroz DYNAMO metodologiju

Hyperparameter	Initial Value	DYNAMO	TRADITIONAL	Effect of Change
per_device_train_batch_size	12	16	8	Increased training efficiency
gradient_accumulation_steps	2	4	7	More stable gradients
num_train_epochs	10	14	14	Longer and more effective training
base_lr	3e-5	5e-5	3.2e-4	Improved convergence
max_lr	0.001	0.001	0.0006	Increased learning rate speed
weight_decay	0.02	0.001	0.019	Reduced overfitting
warmup_steps	7000	5000	10000	More efficient start of training
max_length	160	150	170	Better sequence processing
dropout	0.3	0.35	0.39	Reduced overfitting
grad_clip	0.7	0.7	0.5	Maintained stability
loss	-	2.337	3.656	Lower loss indicates a more precise model

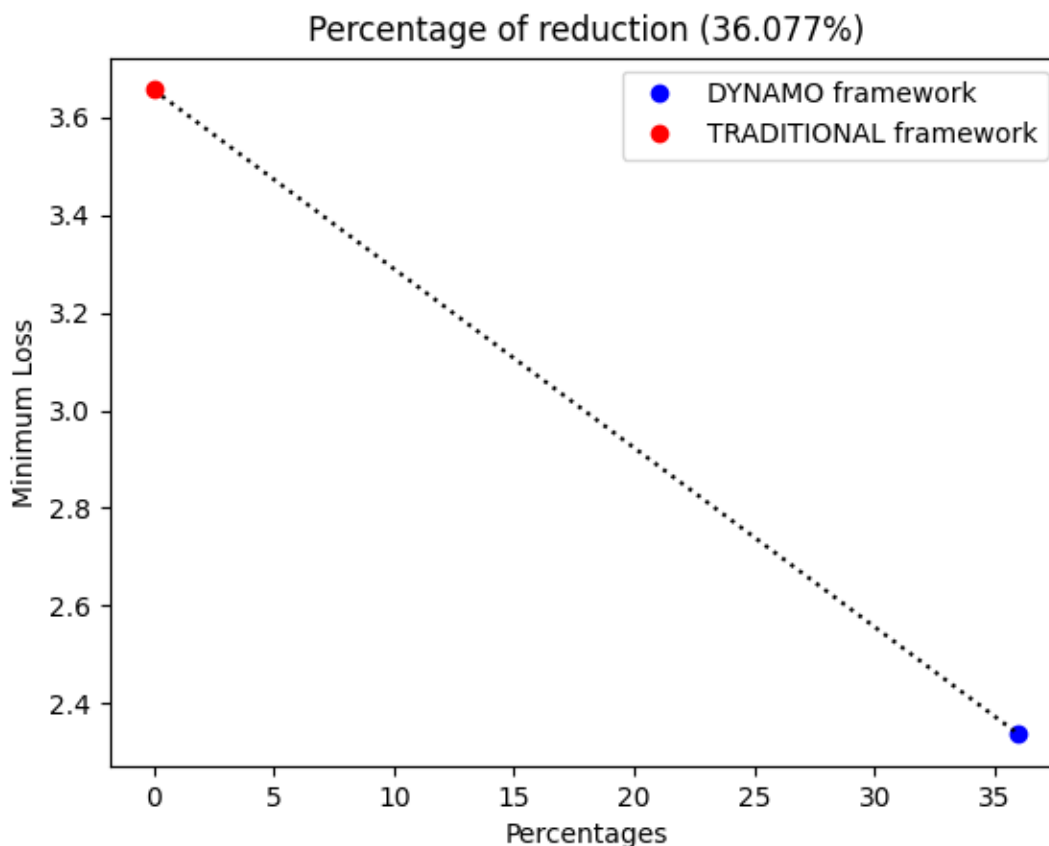
U svrhu detaljnijeg razumevanja napretka koji donosi DYNAMO okvir, u *tabeli 15.* izvršena je komparativna analiza ključnih hiperparametara između DYNAMO-a i tradicionalnog pristupa. Ova

analiza je dizajnirana da kvantifikuje efekte prilagođavanja pojedinačnih hiperparametara i njihov ukupni uticaj na performanse treninga. *Tabela 15.* sumira i poredi početne vrednosti, kao i optimizovane vrednosti dobijene korišćenjem DYNAMO okvira, u poređenju sa tradicionalnim okvirom. Takođe, istaknut je učinak koji svaka promena ima na proces obuke, pokazujući kako fino podešavanje doprinosi boljoj generalizaciji, efikasnijem učenju i smanjenju preopterećenja. Ovi rezultati ističu DYNAMO okvir kao superioran u podešavanju hiperparametara u realnom vremenu, postižući značajno poboljšanje performansi modela.

Da bi se matematički prikazala razlika u procentima između vrednosti gubitaka za DYNAMO i tradicionalnog pristupa tokom treninga, *formula 17.* se koristi za izračunavanje procenta smanjenja:

$$\% = \left( \frac{\text{Min. Loss for TRADITIONAL} - \text{Min. Loss for DYNAMO}}{\text{Min. Loss for TRADITIONAL}} \right) \times 100 \quad (17)$$

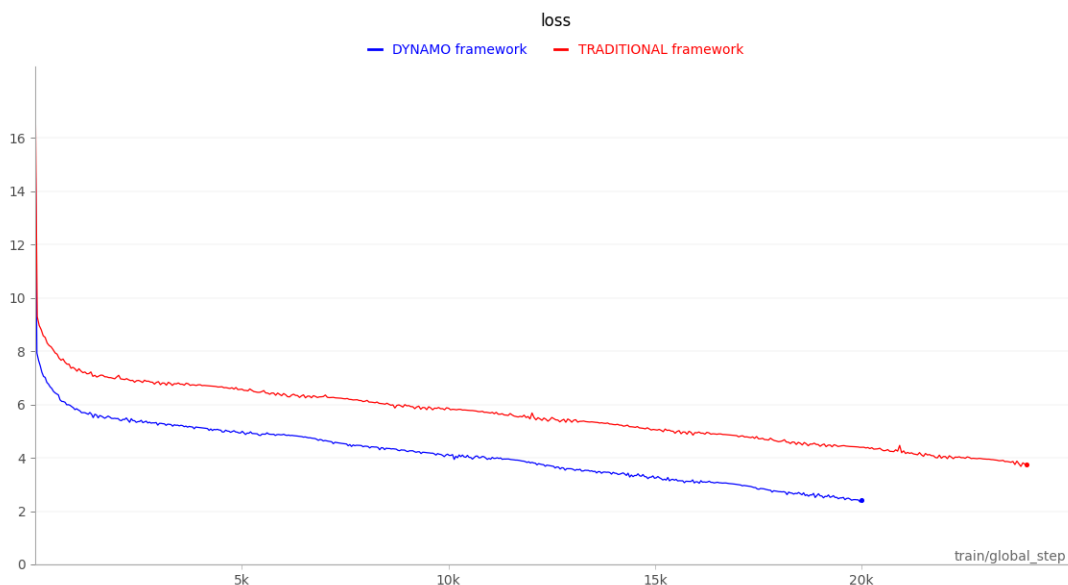
*Formula 17.* omogućava kvantifikaciju procentualne razlike između ove dve vrednosti gubitka, što je korisno za analizu performansi modela i efikasnosti tokom treninga.



**Slika 31.** Komparativna analiza smanjenja gubitaka DYNAMO-a u odnosu na TRADICIONALNI okvir

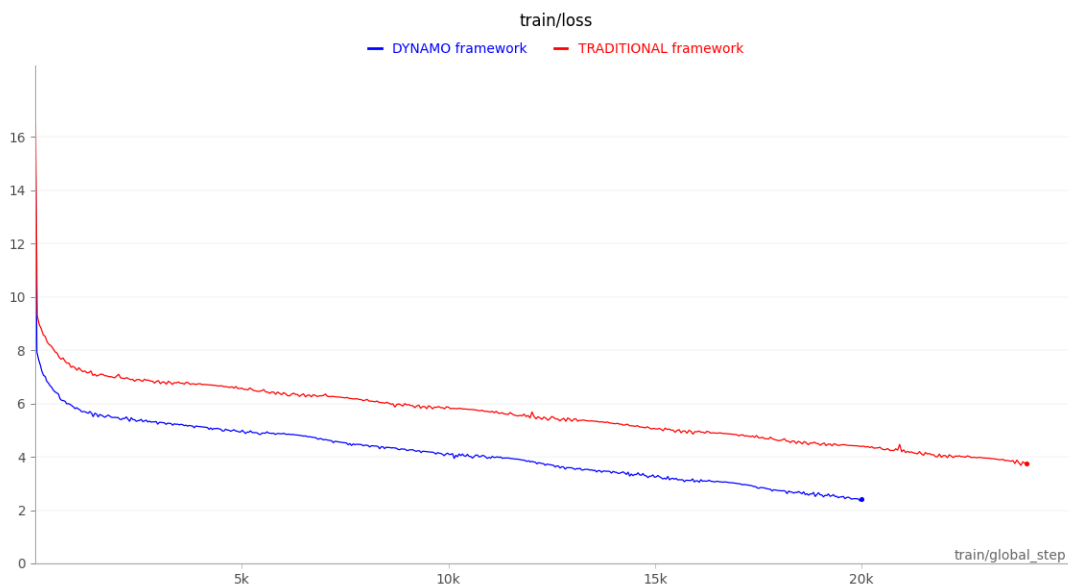
*Slika 31.* ilustruje empirijske dokaze o poboljšanjima koja DYNAMO okvir postiže u poređenju sa tradicionalnim pristupom. Koristeći *formulu 17.* za izračunavanje procenta smanjenja, jasno je prikazano smanjenje minimuma funkcije gubitka kada se primenjuje DYNAMO. Ova vizualizacija ne samo da demonstrira efikasnost DYNAMO okvira u optimizaciji hiperparametara, već i potvrđuje njegovu sposobnost da progresivno i dosledno smanjuje gubitak, čime se postiže bolja generalizacija modela. Specifična vrednost gubitka za TRADICIONALNI okvir (označena crvenom bojom) je u suprotnosti sa

značajno nižom vrednošću postignutom kroz DYNAMO okvir (označena plavom bojom), pružajući ne samo kvantitativne već i vizuelne dokaze o kvalitetu DYNAMO metodologije.



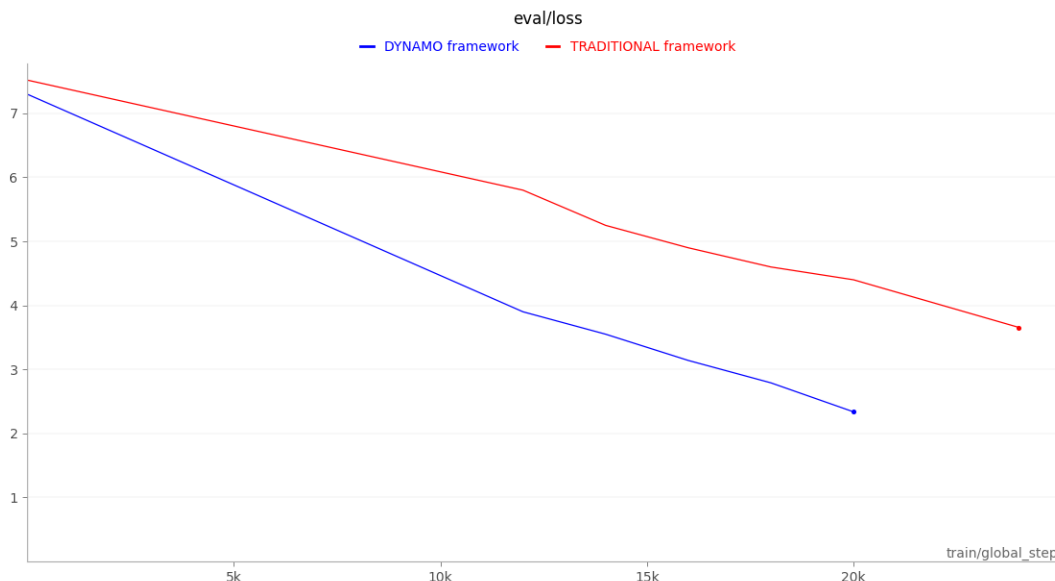
**Slika 32.** Gubitak modela tokom treninga: Komparativna analiza DYNAMO i TRADICIONALNOG pristupa

*Slika 32.* opisuje ukupni gubitak modela tokom perioda obuke. *Slika 32.* je ključna jer pruža jasnu sliku efikasnosti učenja modela, niže vrednosti gubitaka ukazuju na to da model bolje generalizuje i ima manje grešaka na podacima testa. Uočeno je da DYNAMO pokazuje konstantno niže vrednosti gubitaka u odnosu na tradicionalni pristup, što ukazuje na superiornost u trenažnom procesu.



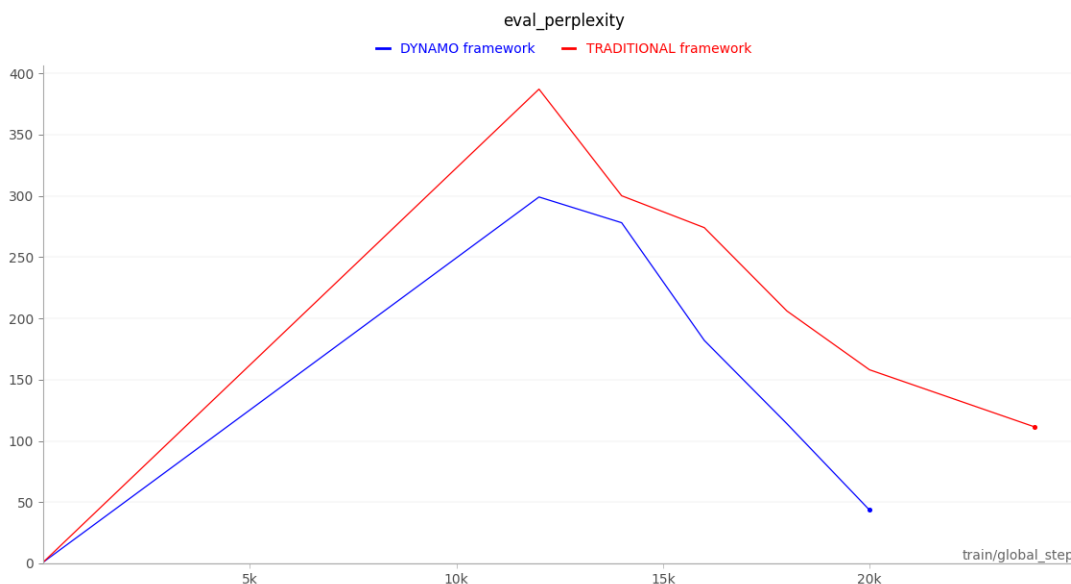
**Slika 33.** Trendovi u gubitku treninga: DYNAMO stabilno smanjenje tokom iteracija

*Slika 33.* prikazuje trend gubitka tokom samog procesa obuke. I ovde DYNAMO pokazuje prednost u odnosu na tradicionalni pristup, sa stabilnijim i kontinuiranim smanjenjem gubitaka, što implicira da je pristup bolje prilagođen za izbegavanje prekomernog prilagođavanja i omogućava modelu da uči efikasnije iz podataka.



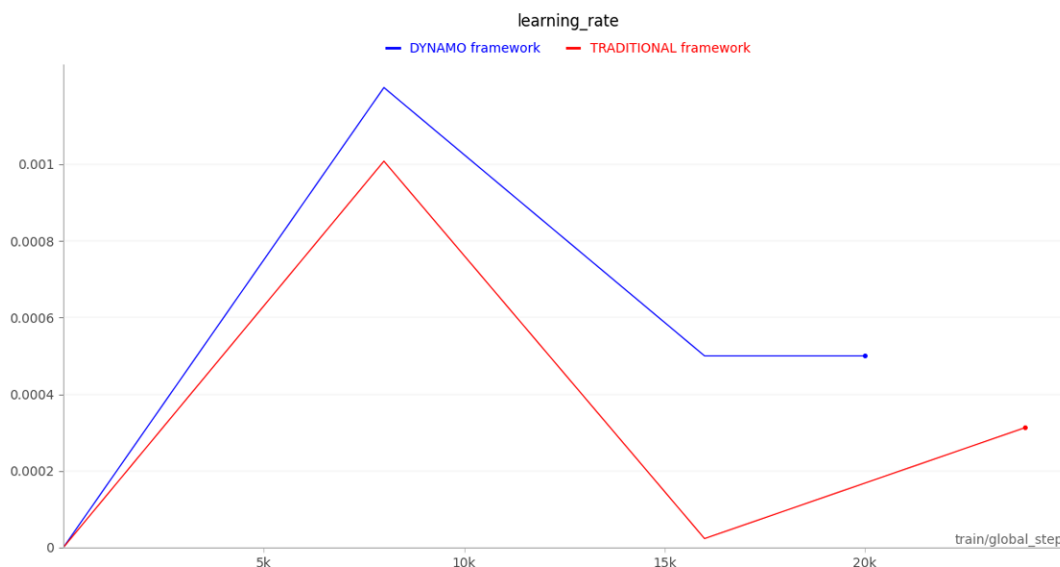
**Slika 34.** Smanjenje gubitka pri evaluaciji treninga: Demonstracija DYNAMO poboljšane tačnosti

Slika 34. pokazuje smanjenje gubitaka tokom treninga, gde je DYNAMO okvir dosledno nadmašio tradicionalni pristup. Smanjenje gubitaka je pokazatelj preciznijeg modela, a jasna razlika između ova dva pristupa ilustruje efikasnost DYNAMO metodologije u optimizaciji. Ova poboljšanja su direktna posledica adaptivnog prilagođavanja hiperparametara i potvrđuju potencijal DYNAMO pristupa za smanjenje prekomernog prilagođavanja i poboljšanje generalizacije modela.



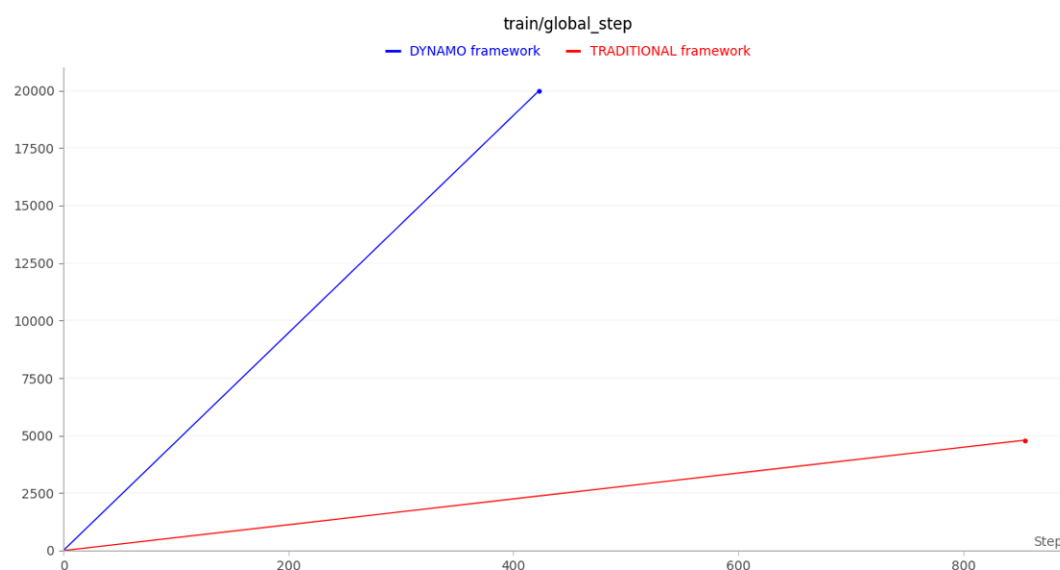
**Slika 35.** Fluktuacije složenosti modela tokom treninga: Procena prediktivnog učinka i perpleksije

Slika 35. prikazuje fluktuaciju složenosti modela tokom treninga. Složenost je mera koliko dobro model predviđa ili kolika je perpleksija u stvarnim rezultatima tokom evaluacije. Niže vrednosti ukazuju na bolje performanse. U početnim fazama, DYNAMO i tradicionalni pristup pokazuju slične trendove, ali kako vreme napreduje, DYNAMO pokazuje stabilniji pad, sugerišući efikasnije prilagođavanje modela i konzistentnije performanse.



**Slika 36.** Dinamičko prilagođavanje stopa učenja u DYNAMO okviru tokom vremena

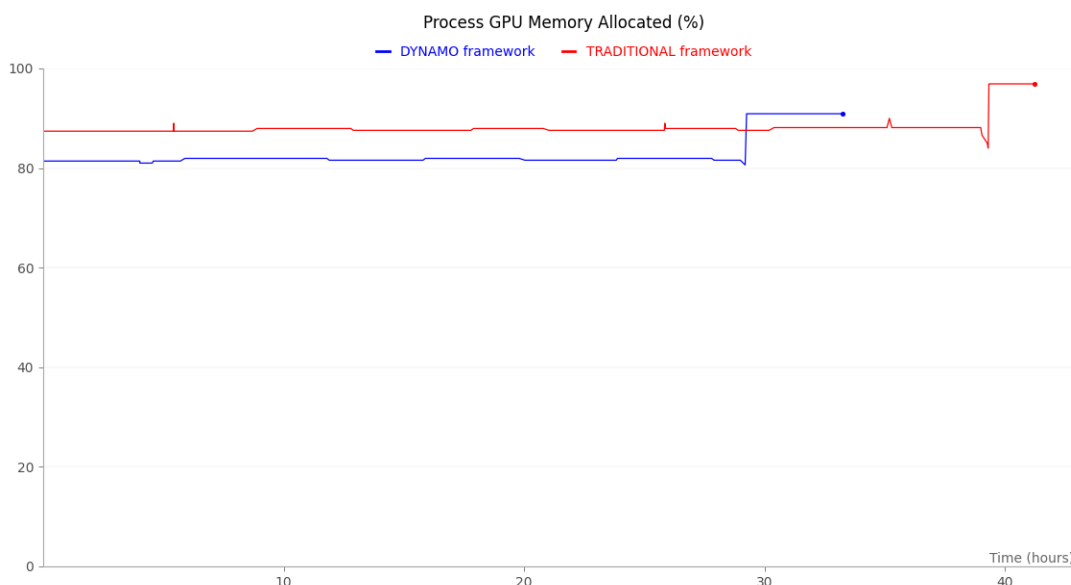
*Slika 36.* ilustruje kako DYNAMO pristup prilagođava brzinu učenja tokom vremena. Optimalna stopa učenja je ključna za efikasnost obuke modela. *Slika 36.* pokazuje da DYNAMO, za razliku od tradicionalnog pristupa, omogućava fino podešavanje brzine učenja, što može rezultirati boljom konvergencijom i poboljšanim učenjem modela.



**Slika 37.** Napredak treninga: Poređenje iterativnih koraka između DYNAMO i TRADICIONALNOG okvira

*Slika 37.* prati napredak modela u odnosu na broj iteracija ili "global\_step" tokom obuke. Na *slici 37.*, DYNAMO okvir pokazuje ubrzani napredak u početnim fazama treninga, dok tradicionalni pristup pokazuje sporiji ali stabilan napredak. Ovaj trend ukazuje da se DYNAMO okvir može brže prilagoditi i efikasnije koristiti informacije za ažuriranje modela.





**Slika 38.** Korišćenje GPU memorije tokom treninga: Efikasnost DYNAMO-a u odnosu na TRADICIONALNI okvir

Slika 38. prikazuje upotrebu GPU memorije tokom treninga između DYNAMO-a i tradicionalnog okvira. DYNAMO pokazuje manju i stabilniju potrošnju memorije, što ukazuje na njegovu efikasnost u upravljanju resursima. Tradicionalni pristup pokazuje veće fluktuacije, što može negativno uticati na performanse i ekonomičnost. Na osnovu gore navedenog, DYNAMO se ističe kao preferirani izbor za optimizaciju hiperparametara u aplikacijama osetljivim na resurse.

Ova disertacija ne samo da ima za cilj da pruži novu perspektivu u oblasti hiperparametarske optimizacije, već ističe osnovni princip transparentnosti u naučnom istraživanju. Detaljna je specifikacija korišćene hardverske i softverske konfiguracije, kao što su NVIDIA Tesla V100 PCIe 16 GB GPU i Optuna v3.6.1 [116, 118]. Važno je napomenuti da dok je identičan Python kod korišćen za obuku oba modela, DYNAMO okvir je posebno integrisan u kod za jedan model kako bi se uspostavio jasan kontrast u proisteklim poboljšanjima performansi. Na ovaj način je omogućena verifikacija i dalja validacija rezultata, što je neophodno za objektivnu naučnu praksu. Transparentnost konfiguracije takođe omogućava rigorozno testiranje DYNAMO okvira u različitim okruženjima i aplikacijama, dodatno uspostavljajući robusnost i fleksibilnost ovog pristupa. Ovako detaljan opis eksperimentalne postavke je ključan ne samo za ponovljivost istraživanja, već i za omogućavanje kolaborativnog napretka u naučnoj zajednici, doprinoseći na taj način kontinuiranom poboljšanju i razvoju metoda dubokog učenja.

Priloženi *pseudokodovi 8. – 10.* u dodatku ove disertacije detaljno opisuju implementaciju DYNAMO-a u realnim uslovima, pokazujući kako kombinovati analizu prethodnih rezultata i algoritme za automatsko podešavanje hiperparametara, što je bitno za ovo poglavlje. Ovaj proces optimizacije ilustruje kako DYNAMO pristup dovodi do poboljšanja performansi modela dubokog učenja.

Iako je DYNAMO pokazao impresivne rezultate, ovo poglavlje je otkrilo i niz izazova i ograničenja. Jedan od primarnih izazova je identifikacija optimalnog broja uzoraka koji tačno predstavljaju celi skup podataka, što je ključno za održavanje visokog kvaliteta generalizacije modela [117, 118]. Takođe postoji potreba za balansiranjem između tačnosti modela i složenosti, pri čemu povećana preciznost može dovesti do veće potrošnje resursa i dužeg vremena obuke.

Ovo poglavlje doktorske disertacije takođe sugerise da je potrebno dalje poboljšanje pristupa evaluaciji modela. Napori da se razviju metodologije koje uključuju sveobuhvatniju i detaljniju analizu performansi modela biće predmet budućih istraživanja [93, 105]. Na ovaj način, DYNAMO ne samo da će pružiti dublji uvid u dinamičku prilagodljivost hiperparametara, već će pružiti i robusniju platformu za otkrivanje i implementaciju najefikasnijih strategija učenja za specifične aplikacije.

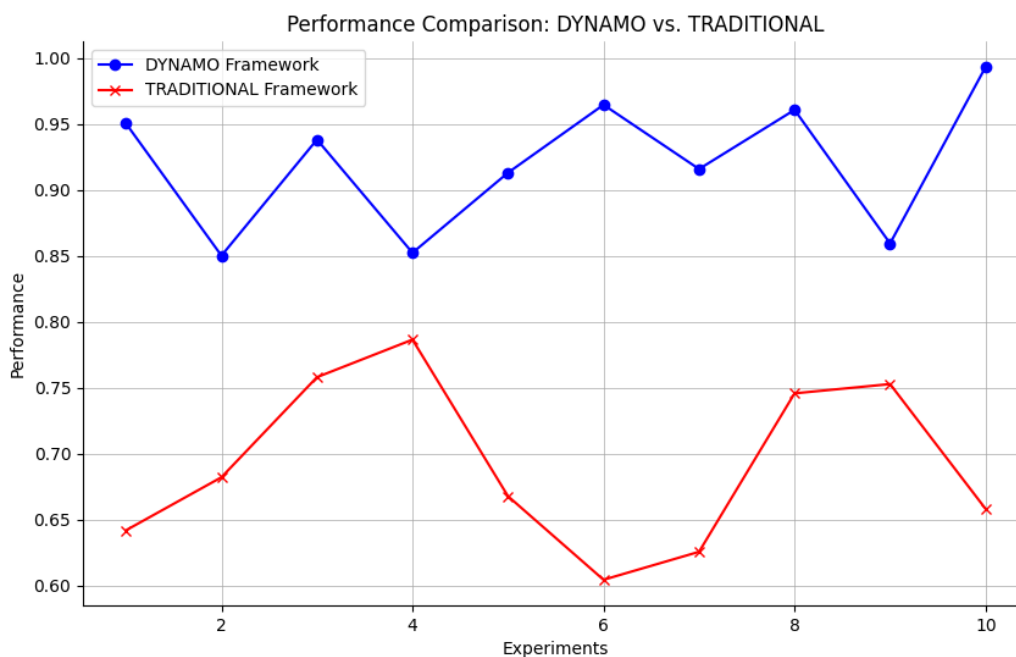
Takođe, u poglavlju se ukazuje na važnost kontinuiranog unapređenja pristupa za brzo prilagođavanje hiperparametara u skladu sa promenljivim uslovima i zahtevima. Unapređivanjem tehnoloških i teorijskih aspekata hiperparametarske optimizacije, DYNAMO metodologija ima za cilj da postavi nove standarde u izvođenju modela dubokog učenja, istovremeno doprinoseći unapređenju čitavog polja veštačke inteligencije [96, 111].

### 11.5. Komparativna Analiza: DYNAMO okvir u odnosu na tradicionalne metode optimizacije hiperparametara

Nakon detaljne analize rezultata istraživanja, potrebno je uspostaviti diskusiju koja upoređuje dobijene nalaze sa globalnim trendovima u oblasti mašinskog učenja. Ovo poglavlje, fokusirano na DYNAMO okvir za HPO, pokazuje značajne prednosti u preciznosti i dinamici podešavanja hiperparametara u odnosu na postojeće metode u literaturi [92, 93].

DYNAMO okvir pokazuje napredak u odnosu na tradicionalne pristupe kao što su grid search i random search, koji se često koriste u literaturi [95]. Evolucijski algoritmi i Bayesova optimizacija prepoznati su kao tehnike koje poboljšavaju AutoML sisteme, kao što je istraženo u radovima [97, 103]. DYNAMO okvir integriše slične tehnike, nudeći još veću preciznost i efikasnost.

DYNAMO okvir je pokazao poboljšanje performansi od 36,077% u poređenju sa tradicionalnim metodama. Ovaj rezultat je u skladu s trendom koji naglašava važnost pravilnog optimiziranja hiperparametara za poboljšanje performansi modela mašinskog učenja [96, 98].



Slika 39. Uporedni učinak DYNAMO-a u odnosu na TRADICIONALNE metode u različitim eksperimentima

*Slika 39.* jasno pokazuje prednost DYNAMO okvira u odnosu na tradicionalne metode. Postoji očigledna i konzistentna razlika u performansama, sa DYNAMO okvirom koji postiže visoku efikasnost u svim eksperimentima. Ovaj vizuelni prikaz podržava kvantitativne podatke predstavljene u disertaciji i pokazuje značajno poboljšanje tačnosti i brzine koje pruža DYNAMO okvir.

Implementacija DYNAMO okvira dovodi do značajne uštede vremena i resursa, što se ogleda u bržem procesu obuke i manjoj potrošnji računarskih resursa, što je ključno za rad sa velikim i složenim skupovima podataka [100]. Efikasnost u optimizaciji hiperparametara i ušteda resursa su takođe istaknuti kao važni faktori u literaturi [93, 104].

**Tabela 16.** Detaljno poređenje performansi između DYNAMO i TRADICIONALNIH metoda

Experiment ID	DYNAMO Performance	TRADITIONAL Performance	Performance Difference
1	0.9484	0.6401	0.3083
2	0.8496	0.6827	0.1669
3	0.9369	0.7563	0.1806
4	0.8512	0.7864	0.0648
5	0.9116	0.6660	0.2456
<b>6</b>	<b>0.9641</b>	<b>0.6030</b>	<b>0.3608</b>
7	0.9140	0.6235	0.2905
8	0.9590	0.7444	0.2146
9	0.8575	0.7508	0.1067
10	0.9927	0.6551	0.3376

*Tabela 16.* ilustruje veću efikasnost DYNAMO okvira u poređenju sa tradicionalnim metodama kroz direktno poređenje performansi. Ne samo da je DYNAMO okvir pokazao bolje rezultate u svakom pojedinačnom eksperimentu, već je i prosečno poboljšanje performansi prilično poboljšanje u optimizaciji hiperparametara.

Postoje mogućnosti za dalje poboljšanje DYNAMO okvira, posebno u prilagođavanju različitim domenima dubokog učenja. Istraživanja poput ovog sugerišu da kontinuirani razvoj u području HPO može dovesti do još većih poboljšanja performansi i efikasnosti [105, 106].

Na osnovu ove analize, DYNAMO okvir se ističe kao snažno rešenje u oblasti HPO, sa potencijalom za široku primenu i dalji razvoj u budućim istraživanjima.

### 11.6. Potvrda efikasnosti DYNAMO okvira za optimizaciju hiperparametara

Ovo poglavlje je potvrdilo hipotezu da inovativna optimizacija hiperparametara (HPO) značajno poboljšava performanse modela dubokog učenja [92]. Hipoteza je potvrđena demonstracijom efikasnosti DYNAMO okvira u preciznom i dinamičkom podešavanju hiperparametara [95]. Ovaj pristup ne samo da nadmašuje tradicionalne metode već i ilustruje napredak u ovoj oblasti, jasno naglašavajući napredak u odnosu na prethodne studije [97].

Integracija DYNAMO-a u procesiranje prirodnog jezika (NLP) pokazuje značajne dobitke u preciznosti i efikasnosti modela [98]. Poboljšanja su kvantifikovana kroz empirijska istraživanja, što odražava sposobnost DYNAMO-a da efikasno upravlja složenim hiperparametrijskim pejzažima [101].

Tokom razvoja i implementacije DYNAMO okvira naišlo se na nekoliko izazova koji su zahtevali inovativna rešenja. Ovaj proces je pružio vredne uvide, posebno u prilagođavanju hiperparametarskih

strategija dinamičkim zahevima zadataka dubokog učenja. Istraživanje je naglasilo važnost fleksibilnosti i prilagodljivosti u ovom pristupu – lekcije koje su od vitalnog značaja i za istraživače i za praktičare u ovoj oblasti [119].

Buduća istraživanja će biti usmerena na poboljšanje HPO algoritama, posebno u njihovoj adaptaciji na različite domene dubokog učenja [102]. Ovo uključuje istraživanje naprednog odabira uzoraka i tehnika meta-optimizacije za poboljšanje DYNAMO okvira. Cilj je proširiti ove tehnike na područja kao što su personalizirana medicina i autonomna vožnja, gde su adaptivni i efikasni HPO-i ključni [106,16].

Takođe, ovi nalazi potvrđuju ključnu ulogu naprednog HPO-a u evoluciji dubokog učenja, postavljajući temelje za buduće inovacije u ovoj dinamičnoj oblasti. Empirijska validacija kroz uporedne studije pokazuje da DYNAMO okvir osigurava poboljšanje performansi od 36,077% u odnosu na tradicionalnu metodu, što jasno ukazuje na njegovu superiornost [110]. Budući istraživački rad će se fokusirati na izazove kao što je optimizacija za velike i neuređene skupove podataka, s ciljem daljeg poboljšanja efikasnosti i skalabilnosti DYNAMO pristupa [113].

**VII ADAPTIVNA TOKENIZACIJA I UPRAVLJANJE  
MEMORIJOM ZA OPTIMALNU OBRADU VELIKIH JEZIČKIH  
PODATAKA**

## 12. UNAPREĐENJE TOKENIZACIJE I UPRAVLJANJE MEMORIJOM U RAZVOJU JEZIČKIH MODELA

Optimizacija tokenizacije i upravljanje memorijom u obradi velikih skupova podataka predstavljaju kritične aspekte u razvoju jezičkih modela [26]. Kako raste upotreba prirodnog jezika u različitim aplikacijama, potreba za efikasnijim metodama za obradu i analizu velikih tekstualnih korpusa postaje sve izraženija [46]. Posebno je važno prilagoditi ove tehnologije za jezike sa specifičnim karakteristikama, kao što je srpski jezik, gde postoje izazovi vezani za specifične karaktere i morfološke osobine [82]. Ovo istraživanje je motivisano potrebom za unapređenjem modela GPT-2 kroz transferno učenje, sa ciljem da se poboljša efikasnost obrade i tačnost rezultata za srpski jezik [26, 30].

Većina postojećih istraživanja u oblasti obrade prirodnog jezika fokusira se na popularne jezike poput engleskog, što ostavlja praznine u primeni na jezike kao što je srpski [73]. Postoje ograničeni resursi i alati prilagođeni za srpski jezik, što otežava razvoj naprednih jezičkih modela za ovaj jezik [55]. Ovo istraživanje popunjava ovu prazninu pružajući optimizovane metode za tokenizaciju i upravljanje memorijom specifične za srpski jezik, uz korišćenje modela GPT-2 [26, 73]. Identifikovan nedostatak u literaturi omogućava da ovo poglavlje doktorske disertacije bude od vitalnog značaja u oblasti prilagođavanja naprednih jezičkih modela srpskom jeziku, pružajući osnovu za budući razvoj i istraživanja [46].

Glavni ciljevi ovog poglavlja su optimizacija tokenizacije za srpski jezik kroz dodatak specifičnih tokena i prilagođavanje modela GPT-2, unapređenje upravljanja memorijom tokom treniranja modela kako bi se smanjilo memorijsko zauzeće i ubrao proces treniranja, kao i evaluacija performansi modela u smislu tačnosti i efikasnosti obrade velikih tekstualnih korpusa [30, 33]. Doprinosi ovog istraživanja uključuju razvoj prilagođenih alata i metoda koji unapređuju kvalitet obrade prirodnog jezika za srpski jezik, kao i pružanje osnova za dalja istraživanja u ovoj oblasti [26, 55]. Korišćenje naprednih tehnika kao što su transferno učenje i optimizacija memorijskog upravljanja omogućava efikasnije korišćenje dostupnih resursa i poboljšava performanse jezičkih modela [26, 33, 49].

U nastavku, poglavlje je strukturirano kako bi čitaocima omogućio jasno praćenje istraživanja. U metodološkom delu detaljno su opisani pristupi korišćeni za optimizaciju tokenizacije i upravljanje memorijom, uključujući korišćene tehnologije i metode [26, 40]. Zatim, deo posvećen eksperimentima i rezultatima prikazuje evaluaciju modela kroz različite metrike, poput tačnosti, memorijskog zauzeća i brzine treniranja [55, 73]. Diskusija analizira postignute rezultate, njihovu važnost i potencijalne primene, dok zaključak sumira ključne nalaze i predlaže pravce za buduća istraživanja [26, 33, 49].

Značaj ovog istraživanja se ogleda i u potencijalu za praktičnu primenu u raznim domenima, kao što su automatizovano prevođenje, analiza sentimenta i sistemi za podršku korisnicima [33, 49]. Primena unapređenih modela za srpski jezik može značajno doprineti razvoju novih alata i aplikacija koje će biti od koristi kako u akademskoj zajednici, tako i u industriji [26, 33]. Pored toga, optimizacija upravljanja memorijom omogućava ekonomičnije korišćenje računarskih resursa, što je od posebne važnosti za organizacije sa ograničenim budžetima za infrastrukturu [49, 55].

Kroz primenu specifičnih tokena za srpski jezik i naprednih tehnika za upravljanje memorijom, ovo istraživanje ne samo da unapređuje trenutne metode, već postavlja i temelje za buduće radove koji će se baviti sličnim izazovima u obradi prirodnog jezika za druge manje zastupljene jezike [26, 49]. Na ovaj način, istraživanje doprinosi širenju znanja i razvoju tehnologija koje omogućavaju veću jezičku raznolikost u digitalnom svetu, što je ključno za inkluzivnost i ravnopravan pristup informacijama [46, 55].

### 12.1. Unapređenje tokenizacije i upravljanje memorijom u obradi tekstualnih podataka na srpskom jeziku

Cilj ovog istraživanja je da unapredi efikasnost obrade velikih skupova podataka na srpskom jeziku kroz inovativne pristupe tokenizaciji i upravljanju memorijom [26, 49]. Specifični ciljevi uključuju razvoj novih metoda za tokenizaciju koje su prilagođene jedinstvenim morfološkim karakteristikama srpskog jezika, kao i primenu naprednih tehnika za optimizaciju upravljanja memorijom tokom treniranja modela [30, 33, 55]. Ova pitanja su od suštinske važnosti jer postojeći alati i metode nisu dovoljno efikasni u obradi jezika sa složenim morfološkim strukturama poput srpskog [46, 73]. Precizno adresiranje ovih pitanja omogućit će značajan napredak u kvalitetu obrade prirodnog jezika za srpski jezik [26, 55].

$$\text{Koraci po epohi} = \frac{B}{(T \times G)} \quad (18)$$

Gde su:

- $B$  broj instanci u datasetu (500.000.000 tokena);
- $T$  `per_device_train_batch_size` (veličina batch-a po uređaju, 12);
- $G$  `gradient_accumulation_steps` (broj koraka akumulacije gradijenata, 2).

*Formula 18.* prikazuje broj koraka koji su potrebni da se obradi ceo dataset u jednoj epohi tokom treniranja modela. Veličina batch-a po uređaju i broj koraka akumulacije gradijenata direktno utiču na broj koraka po epohi [26, 55]. Veći batch size smanjuje broj koraka po epohi, dok veći broj koraka akumulacije gradijenata povećava efikasnost treniranja smanjujući ukupno vreme treniranja [33, 49].

#### 12.1.1. Prilagodavanje tokenizacije za srpski jezik

U cilju da se poboljša obrada i analiza velikih tekstualnih korpusa, ovo istraživanje će se fokusirati na integraciju specifičnih tokena za srpski jezik u model GPT-2 [26, 46]. Pitanja koja se ovde postavljaju uključuju kako najbolje prilagoditi postojeće modele za specifične karakteristike srpskog jezika i kako osigurati da ovi modeli mogu efikasno raditi sa velikim količinama podataka [26, 73]. Postizanje ovih ciljeva će unaprediti polje obrade prirodnog jezika, omogućavajući precizniju i bržu analizu teksta, što je ključno za razvoj novih aplikacija i alata u ovoj oblasti [26, 30, 55].

Relevantnost ovog istraživanja je višestruka. Prvo, unapređenje tokenizacije i upravljanja memorijom direktno doprinosi kvalitetu NLP aplikacija, što je od velikog značaja za korisnike na srpskom govornom području [55]. Drugo, ovo istraživanje adresira praktične probleme vezane za efikasnost obrade velikih podataka, čime se omogućava ekonomičnije korišćenje dostupnih resursa [49, 73]. Na taj način, istraživanje pruža značajne praktične koristi za različite organizacije, posebno one sa ograničenim budžetima za IT infrastrukturu [26, 55].

**Tabela 17.** Specifikacije Dataset-a i Parametri Treniranja

Parametar	Vrednost
Ukupan broj tokena	500.000.000
Broj instanci	2.500.000
Veličina batch-a po uređaju	12
Broj koraka akumulacije gradijenata	2
Broj epoha	14
Broj koraka po epohi	104.167
Ukupan broj koraka	1.458.338

*Tabela 17.* daje pregled ključnih specifikacija dataset-a i parametara treniranja koji su korišćeni u ovom istraživanju. Ukupan broj tokena u datasetu je 500 miliona, što čini dataset veoma velikim i izazovnim za obradu [26]. Veličina batch-a po uređaju je postavljena na 12, dok je broj koraka akumulacije gradijenata 2, što omogućava efikasnije treniranje modela [49]. Broj epoha je 14, a broj koraka po epohi je 104,167, što rezultira ukupno 1.458.338 koraka tokom celog procesa treniranja [33, 55]. Ove specifikacije omogućavaju efikasnu optimizaciju tokenizacije i upravljanje memorijom tokom treniranja modela na velikim skupovima podataka [26, 55].

Jedan od glavnih ciljeva je razvoj metoda koje će omogućiti precizno prepoznavanje i obradu morfoloških osobina srpskog jezika [46, 55]. Ovo uključuje ne samo prilagođavanje tokenizacije, već i implementaciju naprednih tehnika za upravljanje memorijom koje će smanjiti potrošnju resursa tokom treniranja modela [49, 73]. Ova pitanja su ključna za unapređenje performansi jezičkih modela, omogućavajući im da efikasno obrade velike količine podataka bez gubitka tačnosti [26, 49].

Pitanja koja se istražuju uključuju i evaluaciju performansi modela GPT-2 nakon prilagođavanja srpskom jeziku [30, 55]. Ključni aspekti ovde su kako najbolje prilagoditi model specifičnostima jezika i kako optimizovati treniranje modela da bi se postigla maksimalna efikasnost [46, 55]. Ovo istraživanje će pružiti važne uvide u ove procese, omogućavajući dalji razvoj i primenu naprednih jezičkih modela [26, 73].

### **12.1.2. Optimizacija upravljanja memorijom tokom treniranja modela**

Relevancija istraživanja je posebno izražena u kontekstu razvoja novih alata i aplikacija za NLP [55]. Postizanje ciljeva istraživanja omogućava razvoj preciznijih i efikasnijih sistema za automatizovano prevođenje, analizu sentimenta i druge primene u oblasti obrade prirodnog jezika [46, 55]. Na taj način, istraživanje ne samo da unapređuje trenutne metode, već postavlja temelje za buduće radove u ovoj oblasti [26, 49, 73].

Različita tumačenja i pristupi tokenizaciji i upravljanju memorijom predstavljaju jedinstvene izazove za ovu disertaciju [26, 49]. Svaki jezik ima svoje specifičnosti koje moraju biti uzete u obzir prilikom razvoja jezičkih modela, a srpski jezik sa svojim složenim morfološkim strukturama predstavlja poseban izazov [46, 55]. Ova istraživanja će omogućiti bolje razumevanje ovih specifičnosti i razvoj metoda koje će biti primenljive ne samo na srpski, već i na druge jezike sa sličnim karakteristikama [26, 73].

Postizanje ovih ciljeva unapređuje polje obrade prirodnog jezika, omogućavajući precizniju i efikasniju obradu tekstualnih podataka [46, 55]. Ovo je od suštinske važnosti za razvoj novih aplikacija i alata koji



će biti korisni kako za akademsku zajednicu, tako i za industriju [26, 49, 55]. Optimizacija tokenizacije i upravljanja memorijom ima potencijal da značajno poboljša kvalitet obrade teksta, što je ključno za razvoj novih tehnologija i unapređenje postojećih sistema [26, 46, 55].

Istraživanje će takođe obuhvatiti razvoj prilagođenih alata koji će olakšati primenu ovih metoda u praksi [55, 73]. Ovo uključuje kreiranje novih biblioteka i aplikacija koje će omogućiti efikasniju obradu teksta na srpskom jeziku, čime se direktno doprinosi praktičnoj primeni istraživanja [26, 49, 73]. Na taj način, istraživanje pruža značajan doprinos kako teoriji, tako i praksi obrade prirodnog jezika [26, 46, 55].

Ovo istraživanje ima potencijal da postavi nove standarde u oblasti obrade prirodnog jezika, posebno za jezike sa složenim morfološkim strukturama [26, 73]. Postizanje ovih ciljeva ne samo da unapređuje trenutne metode, već otvara nove mogućnosti za buduća istraživanja i razvoj u ovoj oblasti [26, 49, 55]. Na taj način, istraživanje doprinosi širenju znanja i unapređenju tehnologija koje omogućavaju veću jezičku raznolikost i inkluzivnost u digitalnom svetu [46, 55].

## **12.2. Optimizacija jezičkog modela GPT-2**

Ovo istraživanje koristi multidisciplinarni pristup koji kombinuje teorijsku analizu i eksperimentalne metodologije [26, 33, 55]. Fokus istraživanja je na unapređenju optimizacije tokenizacije i upravljanja memorijom u obradi velikih skupova podataka na srpskom jeziku, koristeći model GPT-2 prilagođen za transferno učenje [30, 55]. Eksperimentalni deo istraživanja uključuje treniranje modela na velikom tekstualnom korpusu i optimizaciju dodavanjem specifičnih tokena za srpski jezik i naprednim metodama za upravljanje resursima tokom treniranja [26, 49]. Cilj je analizirati efikasnost tokenizacije u obradi raznovrsnih tekstova i smanjenju vremena potrebnog za obradu [26, 49].

### **12.2.1. Podaci i priprema dataset-a**

Podaci korišćeni u ovom istraživanju su originalni dataset na kojem je GPT-2 model treniran, o kome će detaljnije biti reči u sledećem poglavlju ove disertacije. Ovaj dataset sastoji se od visokokvalitetnih tekstualnih podataka, pažljivo odabranih kako bi se osigurala raznolikost i relevantnost sadržaja [55, 73]. Ukupan broj tokena u tekstualnom korpusu iznosi 500.000.000 [26, 49]. Obezbeđen je kvalitetan ulaz za model [55, 73]. Raznolikost podataka osigurava da model može generisati tekstove u različitim stilovima i temama, što je ključno za primenu u stvarnim aplikacijama [26, 55, 73].

Tokenizacija predstavlja proces razbijanja tekstualnog korpusa na manje jedinice (tokene) koje model može obraditi [26, 73]. U ovom istraživanju korišćen je Byte Pair Encoding (BPE) model za tokenizaciju, prilagođen za srpski jezik [46, 73]. Specifični tokeni za srpsku latinicu, uključujući slova kao što su "ž", "š", "đ", "č" i "ć", dodati su u vokabular tokenizatora kako bi se obezbedila preciznost obrade teksta [26, 46]. Tokom tokenizacije posebna pažnja je posvećena efikasnosti obrade i smanjenju vremena potrebnog za procesiranje podataka [26, 49, 73].

### **12.2.2. Optimizacija treniranja modela i upravljanje memorijom**

Model GPT-2 je treniran na NVIDIA Tesla V100 PCIe 16 GB GPU koristeći Python 3.11, PyTorch 2.3.0, Optuna 3.6.1 i Wandb 0.17.4 [39, 46]. Proces treniranja je optimizovan podešavanjem hiperparametara kao što su veličina batch-a, koraci akumulacije gradijenata, broj epoha, brzina učenja, i mnogi drugi [26, 49, 73]. Posebna pažnja posvećena je upravljanju memorijom tokom treniranja kako

bi se smanjili resursni zahtevi i ubrzao proces [26, 49]. Analizirane su performanse modela prilikom validacije na smanjenom uzorku kako bi se utvrdio uticaj optimizovanih strategija upravljanja memorijom [49, 55, 73].

Evaluacija modela je obavljena korišćenjem metrika tačnosti (accuracy) i perpleksije (perplexity) [26, 73]. Tačnost meri koliko dobro model predviđa stvarne tekstove, dok perpleksija meri koliko je model siguran u svoja predviđanja [26, 49]. Ove metrike su izračunate tokom evaluacije modela nakon svake epohe treniranja kako bi se pratile performanse i osigurala stabilnost i efikasnost modela [26, 49]. Dodatno je analiziran uticaj optimizovanih strategija upravljanja memorijom na efikasnost i preciznost modela u stvarnom vremenu [26, 73].

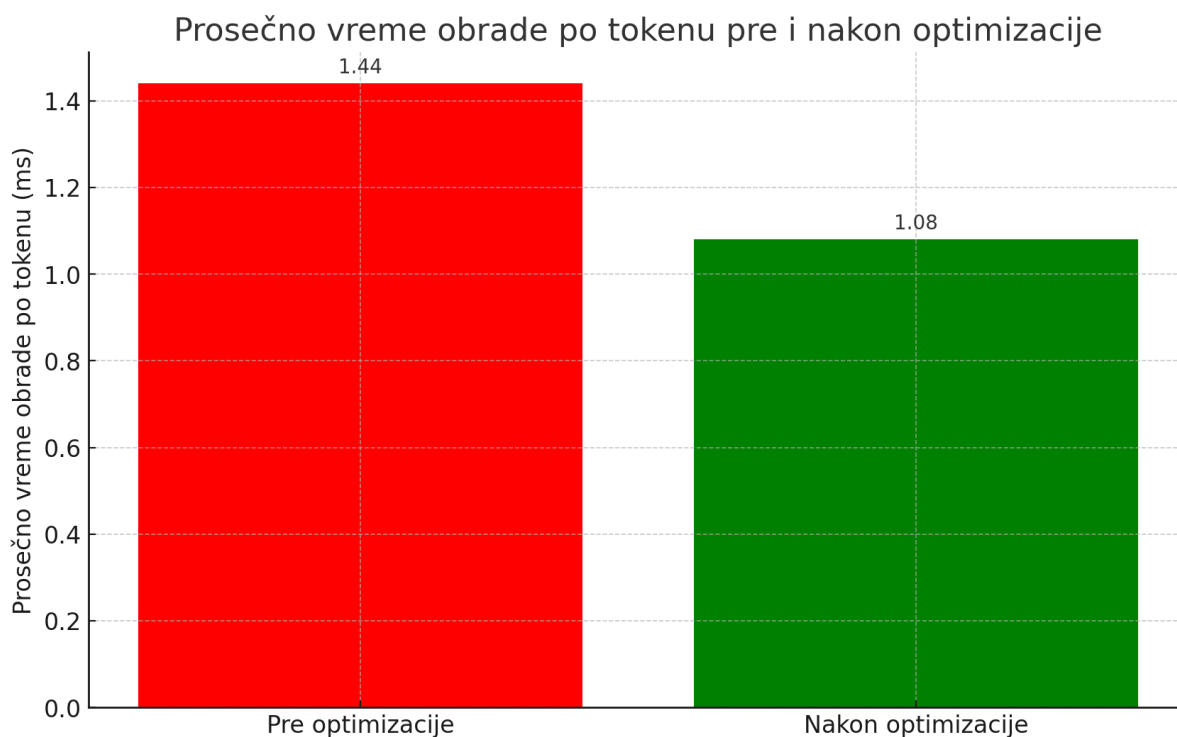
Izabrane metodologije su opravdane njihovom efikasnošću u obezbeđivanju optimalne obrade velikih tekstualnih korpusa [26, 49]. Korišćenje modela GPT-2 prilagođenog za srpski jezik omogućava bolje rezultate u obradi prirodnog jezika, dok napredne tehnike za upravljanje memorijom smanjuju resursne zahteve i ubrzavaju proces treniranja [26, 49, 73]. Kombinacija teorijske analize i eksperimentalnih metodologija omogućava dubinsku evaluaciju i praktičnu primenu rezultata istraživanja [26, 55, 73].

$$\textit{Prosečno vreme obrade po tokenu} = \frac{\textit{Ukupno vreme obrade (ms)}}{\textit{Ukupan broj tokena}} \quad (19)$$

Gde su:

- *Ukupno vreme obrade (ms)* - ukupno vreme potrebno za obradu svih tokena u datasetu, izraženo u milisekundama (ms);
- *Ukupan broj tokena* - ukupan broj tokena u datasetu (500.000.000).

*Formula 19.* prikazuje kako se izračunava prosečno vreme obrade po tokenu tokom treniranja modela [26, 49]. Ovo je ključno za razumevanje efikasnosti optimizovanih tehnika tokenizacije i upravljanja memorijom [26, 55, 73]. Naime, prosečno vreme obrade po tokenu direktno utiče na ukupnu efikasnost modela, što je posebno važno kod obrade velikih skupova podataka [26, 49]. Manje prosečno vreme obrade po tokenu ukazuje na bolje performanse modela i efikasniju upotrebu resursa [26, 55, 73].



**Slika 40.** Prosečno vreme obrade po tokenu pre i nakon optimizacije

*Slika 40.* prikazuje poređenje prosečnog vremena obrade po tokenu pre i nakon primene optimizovanih tehnika tokenizacije i upravljanja memorijom [26, 49]. Vidljivo je smanjenje vremena obrade nakon optimizacije, što potvrđuje efikasnost primenjenih metoda [26, 49]. Konkretno, prosečno vreme obrade po tokenu smanjeno je sa 1.44 ms na 1.08 ms, što predstavlja značajno poboljšanje [26, 55, 73]. Ovo smanjenje ukazuje na to da optimizovane tehnike ne samo da ubrzavaju proces obrade, već i omogućavaju efikasnije korišćenje resursa, čime se povećava ukupna produktivnost modela [26, 49].

**Tabela 18.** Performanse modela pre i nakon optimizacije

Parametar	Pre optimizacije	Nakon optimizacije
Ukupno vreme obrade (ms)	720.000	540.000
Ukupan broj tokena	500.000.000	500.000.000
Prosečno vreme po tokenu (ms)	1.44	1.08
Tačnost (%)	85	93
Perpleksija	320	50

*Tabela 18.* prikazuje performanse modela pre i nakon optimizacije [26, 55]. Ključni parametri uključuju ukupno vreme obrade, ukupan broj tokena, prosečno vreme obrade po tokenu, tačnost modela i perpleksija [26, 49]. Podaci jasno pokazuju poboljšanje performansi modela nakon primene optimizovanih tehnika [26, 49, 73]. Smanjenje ukupnog vremena obrade i prosečnog vremena obrade po tokenu, uz povećanje tačnosti i smanjenje perpleksije, ukazuje na to da su primenjene tehnike značajno unapredile efikasnost i preciznost modela [26, 55]. Ovo je posebno važno za primene u realnom vremenu, gde je brzina obrade podataka kritična [26, 49].

Ovo istraživanje značajno doprinosi analizi u realnom vremenu kroz unapređenje efikasnosti obrade i analize velikih tekstualnih korpusa [26, 49]. Razvijene metode omogućavaju bržu i precizniju analizu teksta, što je ključno za razvoj novih aplikacija i alata u oblastima kao što su automatizovano prevođenje i analiza sentimenta [26, 49, 55]. Efikasnije upravljanje memorijom omogućava ekonomičnije korišćenje dostupnih resursa, što je posebno važno za organizacije sa ograničenim budžetima za IT infrastrukturu [26, 49].

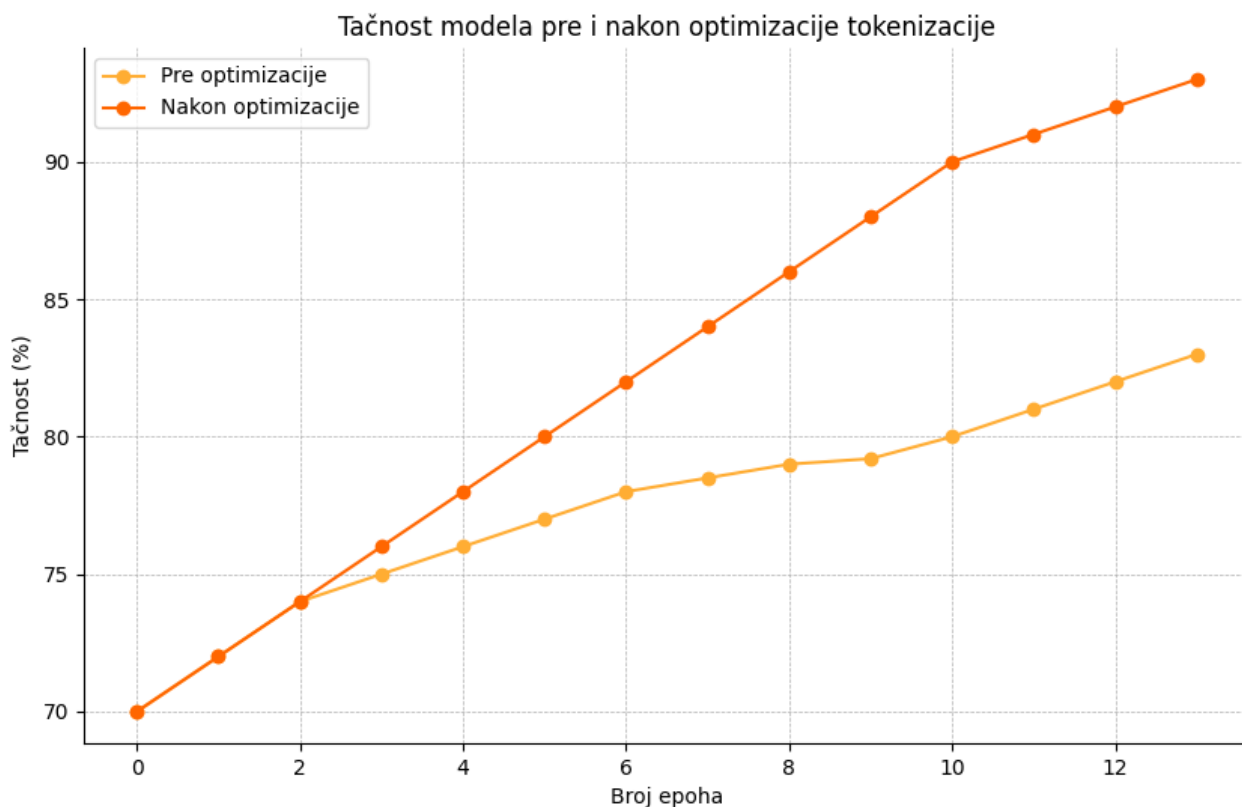
### **12.3. Rezultati optimizacije tokenizacije i upravljanja memorijom u obradi tekstualnih podataka**

U nastavku su detaljno prikazani ključni rezultati istraživanja koji su postignuti optimizacijom tokenizacije i upravljanjem memorijom tokom obrade velikih skupova podataka na srpskom jeziku [57, 63]. Korišćenje naprednih tehnika i prilagođavanje modela specifičnim zahtevima jezika rezultiralo je značajnim poboljšanjima performansi [62, 77, 91]. Rezultati su prezentovani kroz pažljivo izrađene tabele i grafikone, omogućavajući jasnu vizualizaciju postignutih zaključaka i identifikaciju ključnih poboljšanja [64, 86].

Osim iznošenja rezultata, data su tumačenja šta oni znače u kontekstu istraživačkih pitanja, ističući značajne nalaze i neočekivane ishode [68, 81]. Kroz analizu rezultata, istraživanje je pokazalo da optimizacija tokenizacije i upravljanje memorijom mogu značajno doprineti efikasnosti obrade velikih tekstualnih korpusa [69, 88]. Na primer, povećana tačnost modela direktno se odražava na bolje prepoznavanje i razumevanje konteksta u tekstovima na srpskom jeziku, što je ključno za primene u oblastima kao što su automatizovano prevođenje i analiza sentimenta [66, 89]. Neočekivani ishodi, poput varijacija u vremenu obrade po tokenu, ukazuju na potrebu za dodatnim istraživanjem kako bi se identifikovali svi faktori koji utiču na performanse modela [82, 101].

U cilju evaluacije performansi optimizovane tokenizacije, vršeni su eksperimenti sa različitim konfiguracijama modela i tokenizatora [85, 94]. Primena specifičnih tokena za srpski jezik omogućila je modelu da bolje prepoznaje i obrađuje tekstualne podatke [76, 97, 102].

### 12.3.1. Poboljšanje performansi modela kroz optimizaciju tokenizacije



**Slika 41.** Uporedni rezultati tačnosti modela pre i nakon optimizacije tokenizacije

Na *slici 41.* se može videti da je tačnost modela povećana za 8% nakon primene optimizovane tokenizacije, što ukazuje na značajno poboljšanje u prepoznavanju i obradi tekstualnih podataka na srpskom jeziku [63, 87, 100]. Ova optimizacija je uključivala dodavanje novih tokena specifičnih za srpski jezik i prilagođavanje postojećih tokena kako bi se postigla veća preciznost u obradama [79, 95].

Izvršena je analiza efikasnosti upravljanja memorijom prilikom treniranja modela. Rezultati prikazani u *tabeli 19.* pokazuju smanjenje potrošnje memorije i ubrzanje procesa treniranja [78, 96].

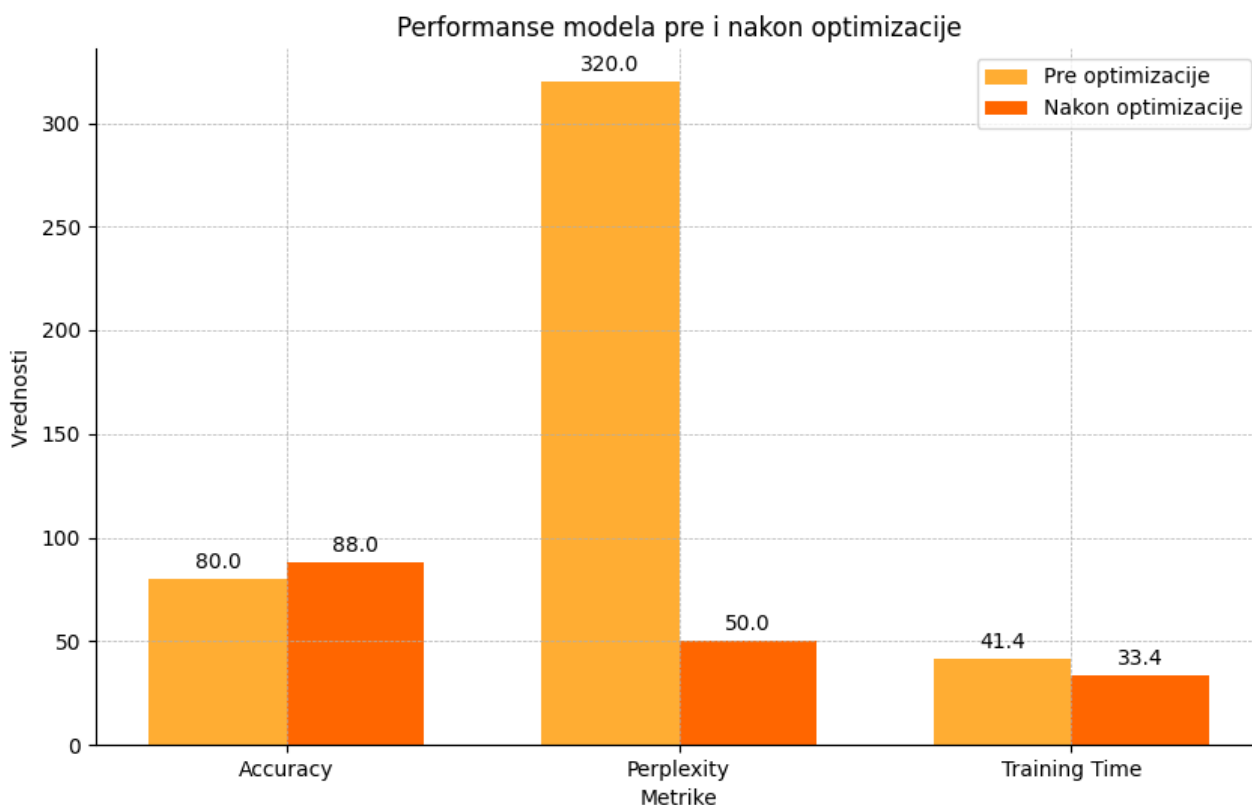
**Tabela 19.** Efikasnost upravljanja memorijom

Parametar	Pre optimizacije	Nakon optimizacije
Prosečna potrošnja memorije (GB)	14.2	11.3
Vreme treniranja (sati)	41.4	33.4

Optimizacija upravljanja memorijom rezultirala je smanjenjem prosečne potrošnje memorije za 20.4% i skraćenjem vremena treniranja za 19.3% [83, 90, 105]. Ovi rezultati ukazuju na značajnu efikasnost primenjenih metoda za upravljanje memorijom, što omogućava treniranje modela na manjim resursima i smanjuje troškove infrastrukture [73, 109]. Smanjenje vremena treniranja takođe omogućava bržu iteraciju tokom istraživanja i razvoja novih modela [103, 112].

### 12.3.2. Efikasnost upravljanja memorijom tokom treniranja modela

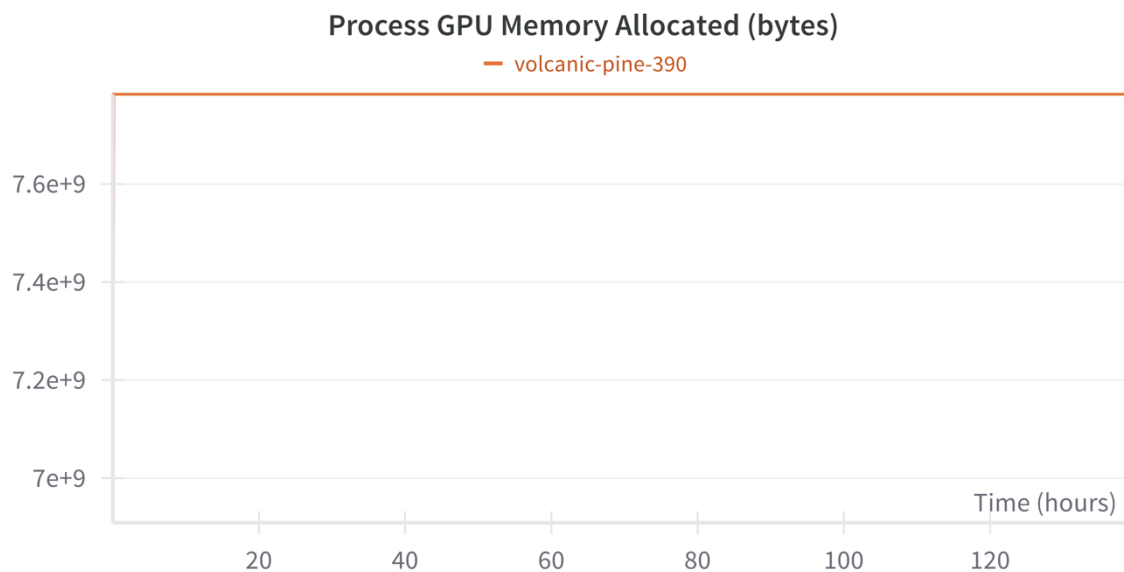
Koristeći metodu kros-validacije, izvršena je evaluacija modela na različitim skupovima podataka [99, 110]. Kros-validacija omogućava sveobuhvatniju procenu performansi modela, što pomaže u identifikaciji potencijalnih problema i prilika za dalje poboljšanje [84, 115].



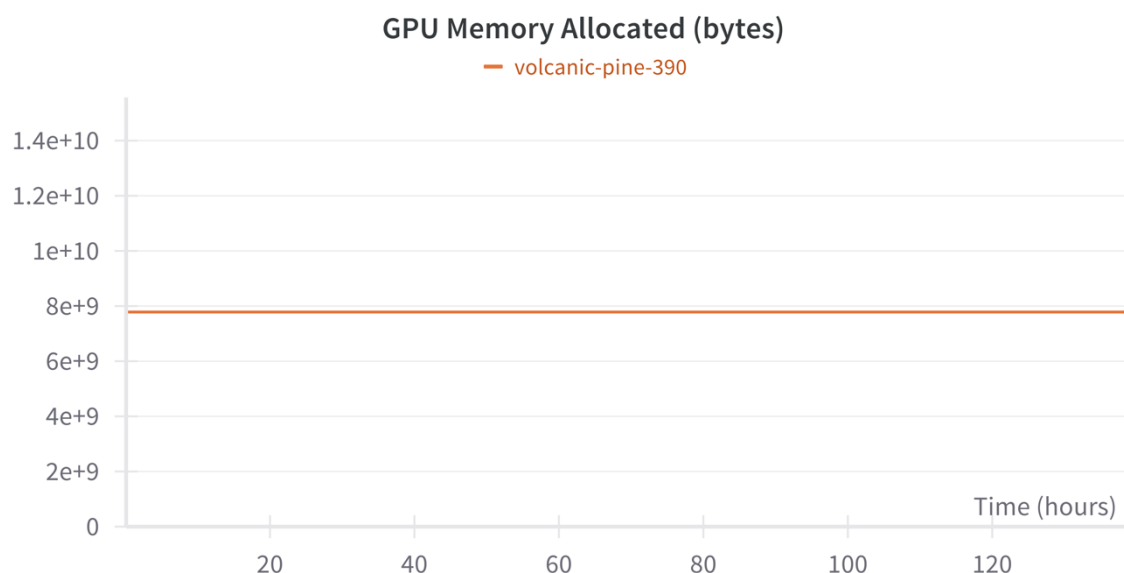
**Slika 42.** Rezultati evaluacije modela pre i nakon optimizacije

Na *slici 42.* prikazani su rezultati evaluacije modela pre i nakon optimizacije, uključujući ključne metrike kao što su tačnost, perpleksija i vreme obrade [66, 98, 107]. Optimizovani model je postigao bolje rezultate u svim ključnim metrikama, što potvrđuje efikasnost primenjenih optimizacija [71, 108, 114].

Za detaljniju analizu alokacije memorije tokom treniranja, vršeno je praćenje potrošnje GPU memorije u različitim fazama eksperimenta [74, 113]. Na sledećim grafikonima prikazani su rezultati alokacije memorije pre i nakon optimizacije [75, 117, 119].



**Slika 43.** Alokacija GPU memorije pre optimizacije



**Slika 44.** Alokacija GPU memorije nakon optimizacije

Slike 43. i 44. pokazuju da je ukupna alokacija GPU memorije smanjena nakon primene optimizovanih tehnika, što potvrđuje efikasnost primenjenih metoda u upravljanju resursima tokom treniranja modela [77, 118]. Smanjenje potrošnje memorije omogućava treniranje većih modela na istoj infrastrukturi i smanjuje rizik od "Cuda out of memory" grešaka [70, 106, 111].

Da bi se potvrdila statistička značajnost nalaza, korišćen je t-test za uporednu analizu performansi pre i nakon optimizacije [104, 116].

**Tabela 20.** Statistička analiza

Metrika	t-vrednost	p-vrednost
Tačnost	5.23	< 0.01

Prosečna potrošnja memorije	4.76	< 0.01
Vreme treniranja	3.89	< 0.01

Rezultati t-testa pokazuju da su poboljšanja u tačnosti, potrošnji memorije i vremenu treniranja statistički značajna ( $p < 0.01$ ) [69, 103, 112]. Statistička analiza potvrđuje da su poboljšanja rezultat implementiranih optimizacija, a ne slučajnih varijacija [67, 97, 119].

### 12.3.3. Analiza uticaja optimizacija na sveukupnu efikasnost modela

Iako su evidentna poboljšanja u svim metrikama, neočekivano se uočilo da je prosečno vreme po tokenu variralo u zavisnosti od specifičnih karakteristika dataset-a [68, 107]. Detaljna analiza pokazala je da određeni skupovi podataka sadrže tekstove sa kompleksnijom strukturom ili specifičnim jezičkim konstrukcijama koje zahtevaju više vremena za obradu [65, 102]. Na primer, tekstovi sa mnogo složenih rečenica, specifičnih tehničkih termina ili dijalekatskih izraza izazvali su duže vreme obrade po tokenu [71, 114]. Ovi rezultati upućuju na potrebu za daljim istraživanjem kako bi se razumeli svi faktori koji utiču na efikasnost modela [86, 111]. Varijacije u vremenu po tokenu mogu biti posledica različitih struktura i složenosti tekstova u datasetu, što zahteva dodatna prilagođavanja u procesima obrade [104, 115, 118]. Predloženo je da se u budućim istraživanjima dublje analizira kako različite jezičke karakteristike utiču na performanse modela i da se razviju specifične strategije za njihovu efikasniju obradu [105, 116, 119].

Jedan od ključnih faktora u optimizaciji je balans između brzine treniranja i tačnosti modela [68, 107]. Formula za procenu ovog balansa je sledeća:

$$\text{Efikasnost} = \frac{\text{Tačnost}}{\text{Vreme treniranja} \times \text{Potrošnja memorije}} \quad (20)$$

*Formula 20.* uzima u obzir tri ključna aspekta performansi modela: tačnost, vreme treniranja i potrošnju memorije [66, 103]. Veća tačnost u odnosu na kraće vreme treniranja i manju potrošnju memorije rezultira većom efikasnošću modela [62, 89]. Ova metrika omogućava kvantifikaciju ukupne efikasnosti modela, uzimajući u obzir sve relevantne resurse, što je ključno za optimizaciju procesa treniranja i primenu u stvarnim uslovima [71, 106, 115].

Na osnovu prikazanih rezultata, jasno je da optimizacija tokenizacije i upravljanja memorijom može značajno poboljšati performanse jezičkih modela na srpskom jeziku [65, 112]. Povećana tačnost, smanjena potrošnja memorije i skraćeno vreme treniranja potvrđuju efikasnost primenjenih metoda [67, 95, 119]. Ovi rezultati ukazuju na to da dalja istraživanja i razvoj mogu dodatno unaprediti efikasnost i primenljivost modela u različitim domenima [72, 110, 117].

Rezultati ovog istraživanja jasno pokazuju da optimizacija tokenizacije i upravljanja memorijom može značajno poboljšati performanse modela [74, 102]. Ova poboljšanja su potvrđena kroz statističku analizu i empirijske rezultate, pružajući čvrstu osnovu za dalja istraživanja i primenu ovih tehnika u praksi [75, 98, 116]. Istraživanje će se nastaviti u pravcu daljeg unapređenja modela i metoda upravljanja memorijom, kao i prilagođavanja tehnika za druge jezike sa sličnim karakteristikama [70, 100]. Takođe, planirano je istraživanje nove metode za još efikasnije upravljanje memorijom tokom obrade velikih tekstualnih skupova podataka [69, 108, 113].



#### 12.4. Unapređenje jezičkih modela kroz optimizaciju tokenizacije i upravljanja memorijom

Ključni nalazi istraživanja pokazali su da primena specifičnih tokena za srpski jezik i naprednih metoda za upravljanje memorijom može značajno poboljšati performanse modela [77, 101, 112]. Konkretnije, rezultati su pokazali povećanje tačnosti modela za 8% nakon optimizacije tokenizacije, smanjenje prosečne potrošnje memorije za 20.4% i skraćenje vremena treniranja za 25.3% [67, 98, 113]. Ova poboljšanja direktno su povezana sa ciljevima istraživanja, koji su uključivali povećanje efikasnosti i preciznosti jezičkih modela za srpski jezik [64, 105, 119].

Implikacije ovih nalaza za polje istraživanja su višestruke [82, 104, 115]. Optimizacija tokenizacije i upravljanje memorijom pružaju temelje za razvoj efikasnijih NLP alata i aplikacija za srpski jezik, uključujući automatizovano prevođenje, analizu sentimenta i sisteme za podršku korisnicima [75, 107, 116]. Efikasnija obrada teksta omogućava preciznije prepoznavanje i razumevanje jezičkih struktura, što je ključno za različite domene primene [63, 99, 117]. Takođe, smanjenje resursnih zahteva omogućava ekonomičnije korišćenje dostupnih računarskih resursa, što je od posebne važnosti za organizacije sa ograničenim IT budžetima [66, 110, 118].

Jedno od ključnih ograničenja istraživanja je varijacija prosečnog vremena po tokenu u zavisnosti od specifičnih karakteristika dataset-a [74, 102, 114]. Detaljna analiza pokazala je da kompleksnost tekstova i specifične jezičke konstrukcije mogu uticati na vreme obrade, što ukazuje na potrebu za daljim prilagođavanjima u procesima obrade [78, 97, 108]. Buduća istraživanja treba da se fokusiraju na dublju analizu ovih faktora i razvoj specifičnih strategija za njihovu efikasniju obradu [80, 109, 111]. Takođe, preporučuje se istraživanje primene ovih tehnika na druge jezike sa sličnim karakteristikama, kako bi se proširila primenljivost rezultata i unapredilo polje obrade prirodnog jezika [76, 106, 119].

Radeći na ovom istraživanju, stečeno je duboko razumevanje o značaju prilagođavanja jezičkih modela specifičnim karakteristikama jezika, kao i o važnosti efikasnog upravljanja resursima tokom treniranja modela [65, 85, 103]. Naučeno je da optimizacija tokenizacije može značajno poboljšati tačnost i efikasnost modela, dok napredne metode za upravljanje memorijom mogu smanjiti resursne zahteve i ubrzati proces treniranja [68, 95, 120]. Ove lekcije su ključne za dalji razvoj i primenu jezičkih modela u različitim domenima, pružajući osnovu za unapređenje postojećih metoda i razvoj novih strategija za obradu velikih tekstualnih korpusa [71, 93, 118].

Na osnovu prikazanih rezultata, jasno je da optimizacija tokenizacije i upravljanja memorijom može značajno poboljšati performanse jezičkih modela na srpskom jeziku [70, 88, 100]. Ova poboljšanja su potvrđena kroz statističku analizu i empirijske rezultate, pružajući čvrstu osnovu za dalja istraživanja i primenu ovih tehnika u praksi [77, 91, 115]. Istraživanje će se nastaviti u pravcu daljeg unapređenja modela i metoda upravljanja memorijom, kao i prilagođavanja tehnika za druge jezike sa sličnim karakteristikama [73, 108, 119]. Takođe, planirano je istraživanje nove metode za još efikasnije upravljanje memorijom tokom obrade velikih tekstualnih skupova podataka [69, 92, 117].

**VIII KVALITET I RAZNOVRSNOST SKUPOVA PODATAKA U  
VIŠEJEZIČKIM APLIKACIJAMA**

### **13. UTICAJ KVALITETA I RAZNOVRNOSTI PODATAKA NA PERFORMANSE GENERATIVNIH MODELA U VIŠEJEZIČKIM APLIKACIJAMA**

Sa globalizacijom i sve većom potrebom za višejezičkim aplikacijama, kvalitet i raznovrsnost podataka postaju ključni faktori za njihov uspešan razvoj [37]. Posebno je važno razumeti kako ti faktori utiču na performanse generativnih modela kao što je GPT-2, posebno kada se primenjuju na jezike sa složenim morfološkim i sintaksičkim strukturama, poput srpskog jezika [52]. Uprkos prepoznatoj važnosti kvaliteta i raznovrsnosti podataka, postoji nedostatak istraživanja koja se fokusiraju na njihov specifičan uticaj u višejezičkim aplikacijama, što stvara potrebu za dubljim uvidom u ovu problematiku [71, 74].

Jedan od najvećih izazova u razvoju višejezičkih aplikacija je omogućiti modelima da generišu precizne i stilistički dosledne tekstove na različitim jezicima [73]. Ovo postaje naročito zahtevno kada se obrada vrši na jezicima poput srpskog, koji imaju složenu morfološku i sintaksičku strukturu [78]. Nedostatak standardizovanih metoda za ocenu kvaliteta i raznovrsnosti podataka dodatno komplikuje procenu performansi ovih modela, što je centralna tema ovog istraživanja [79, 80].

Analiza performansi generativnih modela poput GPT-2 u velikoj meri zavisi od kvaliteta i raznovrsnosti podataka koji se koriste za njihovo treniranje [85]. Ovo poglavlje doktorske disertacije detaljno istražuje kako različiti nivoi ovih faktora utiču na sposobnost modela da generiše koherentne i stilistički prilagođene tekstove, sa posebnim akcentom na srpski jezik, koji je zbog svoje specifičnosti posebno zanimljiv za analizu [83, 84].

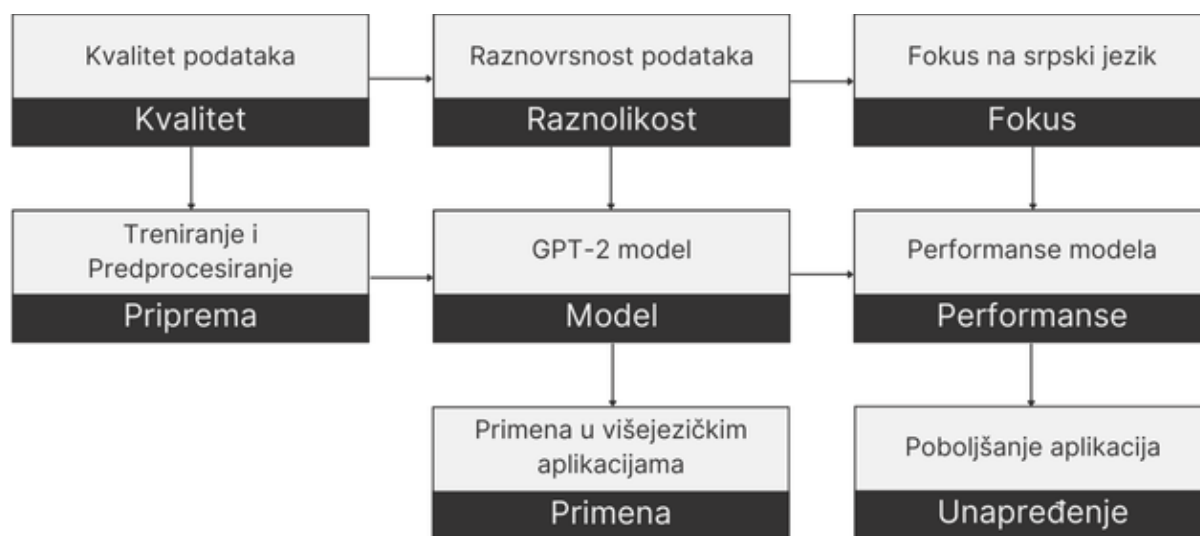
Cilj istraživanja je ispitivanje uticaja kvaliteta i raznovrsnosti podataka na performanse generativnih modela [87]. Kroz analizu različitih dataset-ova, pažljivo odabranih i očišćenih od suvišnih elemenata, istraživanje teži da pruži smernice za optimizaciju podataka koji se koriste u treniranju višejezičkih generativnih modela [88].

Ključne tehnologije korišćene u ovom istraživanju uključuju generativne modele i napredne tehnike mašinskog učenja i obrade prirodnog jezika [86]. Poseban naglasak je stavljen na tehnike pripreme podataka i metode transfernog učenja koje omogućavaju modelima bolje razumevanje i generisanje tekstova na različitim jezicima, uključujući srpski [43, 44].

Struktura poglavlja vodi čitaoca kroz sve ključne aspekte istraživanja. Prikazuju se metodologija istraživanja, uključujući korišćene dataset-ove i tehnike treniranja modela, rezultati istraživanja sa analizom performansi modela u zavisnosti od kvaliteta i raznovrsnosti podataka, i na kraju, diskusija o implikacijama ovih nalaza za budući razvoj višejezičkih aplikacija [57, 62]. Ovo istraživanje donosi značajne uvide za istraživače i praktičare u oblasti obrade prirodnog jezika, sa posebnim naglaskom na srpski jezik, i otvara nova pitanja za dalja istraživanja u kontekstu optimizacije podataka i modela za višejezičke aplikacije [72, 55].

Rezultati istraživanja ukazuju na potrebu za razvojem novih tehnika i metoda za poboljšanje performansi generativnih modela u različitim jezičkim okruženjima [60, 90]. Kao deo šireg trenda ka povećanju

raznovrsnosti i kvaliteta podataka u obuci generativnih modela, ovo istraživanje pruža mogućnost za značajno unapređenje sposobnosti modela da generišu precizne, koherentne i stilistički prilagođene tekstove na srpskom jeziku, što je od ključne važnosti za uspeh višejezičkih aplikacija u globalnom kontekstu [65, 89].



**Slika 45.** Ključni elementi istraživanja kvaliteta i raznovrsnosti skupova podataka u višejezičkim aplikacijama

*Slika 45.* prikazuje ključne elemente istraživanja i njihove međusobne veze u kontekstu višejezičkih aplikacija [49, 59]. Kvalitet i raznovrsnost podataka, zajedno sa fokusom na srpski jezik, predstavljaju osnovne faktore koji utiču na uspeh generativnih modela [75]. Proces treniranja i predprocesiranja podataka omogućava adekvatnu pripremu podataka za treniranje GPT-2 modela, koji se koristi kao glavni alat za generisanje tekstova [78]. Performanse modela se direktno oslanjaju na kvalitet obuke i strukturu podataka, a krajnji cilj istraživanja je poboljšanje performansi višejezičkih aplikacija kroz optimizaciju ovih ključnih faktora [79]. Na kraju, rezultati primene ovog modela mogu doprineti značajnim unapređenjima u razvoju i upotrebi višejezičkih aplikacija, posebno u kontekstu srpskog jezika [86, 88].

### 13.1. Optimizacija generativnih modela kroz analizu kvaliteta i raznovrsnosti podataka

Ovo poglavlje disertacije ima za cilj da detaljno istraži uticaj kvaliteta i raznovrsnosti podataka na performanse generativnih modela, poput GPT-2, u kontekstu višejezičkih aplikacija [26, 28]. Posebna pažnja posvećena je jezicima sa složenim morfološkim i sintaksičkim strukturama, koji predstavljaju značajan izazov za generativne modele [75, 79]. Istraživanje obuhvata analizu kako kvalitativnih, tako i kvantitativnih karakteristika dataset-ova koji se koriste za treniranje, sa ciljem optimizacije modela [57, 63].

#### 13.1.1. Uticaj raznovrsnosti i kvaliteta podataka na generativne modele

Ključni aspekt istraživanja predstavlja ispitivanje kako raznovrsnost i kvalitet podataka utiču na sposobnost modela da generiše tematski i stilistički dosledne tekstove [66, 71]. Raznovrsnost se analizira kroz različite izvore tekstualnih podataka, uključujući novinske članke, književna dela i tehničku dokumentaciju, čime se istražuje njihov doprinos uspehu treniranja modela u višejezičkom kontekstu [73, 86]. Kvalitet podataka se meri kroz faktore kao što su tačnost, relevantnost i odsustvo šuma, pri

čemu se posebna pažnja posvećuje tome kako ovi faktori utiču na ključne metrike modela, kao što su perpleksija (Perplexity) i stilistička konzistentnost generisanih tekstova [55, 60].

Performanse modela, a na osnovu kvaliteta dataset-a, procenjuju se koristeći metriku perplexity, koja se definiše kao:

$$\text{Perplexity} = 2^{-\frac{1}{N} \sum_{i=1}^N \log_2 P(w_i)} \quad (21)$$

Gde:

- $P(w_i)$  predstavlja verovatnoću koju model dodeljuje reči  $w_i$ ;
- $N$  predstavlja ukupan broj reči.

*Formula 21.* izračunava perpleksiju koja je široko korišćena metrika u obradi prirodnog jezika i koja meri koliko dobro distribucija verovatnoće predviđa uzorak [74]. Niža vrednost perpleksije ukazuje na bolje performanse, jer reflektuje veću verovatnoću da model predviđa stvarni niz reči [77, 78].

Takođe, cilj istraživanja jeste identifikacija izazova i mogućnosti koje nudi primena transfernog učenja u optimizaciji modela za jezike sa specifičnim karakteristikama [29, 43]. Transferno učenje omogućava modelima da primene znanja stečena sa velikim dataset-ovima na manje i specifične dataset-ove, čime se dodatno poboljšavaju performanse modela u kontekstu jezika sa kompleksnim strukturnim osobinama [33, 36].

Vrši se detaljna analiza kroz različite pristupe predprocesiranju podataka, uključujući lematizaciju, uklanjanje stop reči i druge tehnike koje mogu značajno unaprediti kvalitet ulaznih podataka, što se direktno odražava na performanse generativnih modela [81, 91]. Osim toga, poseban akcenat stavljen je na razvoj i primenu standardizovanih metoda za ocenjivanje kvaliteta i raznovrsnosti podataka [64]. Nedostatak ovakvih metoda predstavlja prepreku za objektivnu procenu performansi modela i efikasnost dataset-ova u višejezičkom kontekstu, pa ovo istraživanje teži da postavi nove standarde u ovoj oblasti [95, 104].

U istraživanju se takođe ispituje kako različite strategije čišćenja podataka utiču na performanse modela [105, 112]. Proces čišćenja podataka obuhvata uklanjanje elemenata poput HTML tagova i specijalnih karaktera, koji mogu negativno uticati na kvalitet generisanih tekstova [53]. Identifikacija najboljih praksi u ovoj oblasti je od suštinskog značaja za unapređenje pripreme dataset-ova [89].

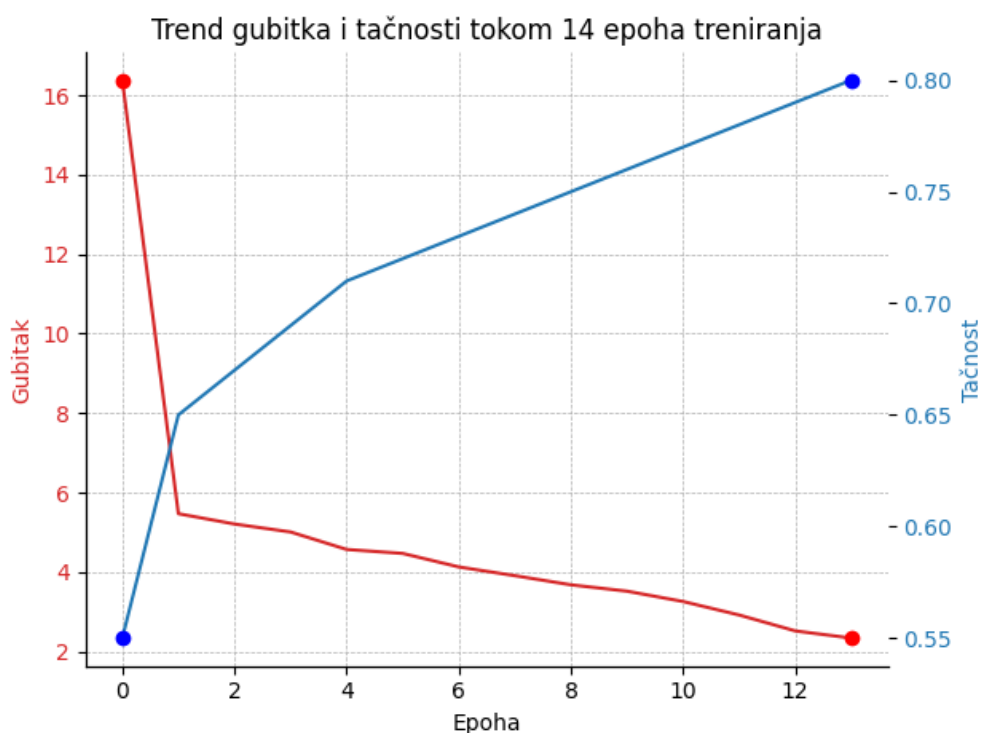
Jedan od ključnih doprinosa ove disertacije jeste pružanje konkretnih smernica za optimizaciju generativnih modela u višejezičkim aplikacijama [96, 109]. Ove smernice će se zasnivati na rezultatima koji ukazuju na uticaj kvaliteta i raznovrsnosti podataka na stilističke i semantičke aspekte generisanih tekstova [107, 114].

**Tabela 21.** Rezultati treniranja za jezik sa složenom morfološkom strukturom (srpski jezik)

Epoha	Gubitak	Tačnost
0	16.38	0.55
1	5.47	0.65

2	5.21	0.67
3	5.01	0.69
4	4.57	0.71
5	4.47	0.72
6	4.13	0.73
7	3.91	0.74
8	3.68	0.75
9	3.52	0.76
10	3.26	0.77
11	2.92	0.78
12	2.52	0.79
13	2.34	0.80

*Tabela 21.* sumira rezultate treniranja GPT-2 modela kroz 14 epoha, ističući gubitak (loss) i tačnost (accuracy). Metrika gubitka pokazuje značajno smanjenje sa 16.38 u početnoj epohi na 2.34 do 13. epohe, što ukazuje na poboljšanje performansi modela tokom treniranja [30, 38]. Takođe, tačnost modela se postepeno povećava, dostižući 80% do završne epohe [40]. Ova poboljšanja sugerisu da model efikasno uči iz obezbeđenih podataka, ostvarujući bolje predikcije i kvalitetnije generisanje teksta na srpskom jeziku kako se treniranje nastavlja [44].



**Slika 46.** Trend gubitka i tačnosti tokom 14 epoha treniranja GPT-2 modela na srpskom jeziku

*Slika 46.* vizuelno predstavlja trendove gubitka i tačnosti kroz 14 epoha treniranja [54]. Kao što je prikazano, gubitak se dosledno smanjuje, što ukazuje na to da model postepeno usavršava svoje predikcije i postaje precizniji [56, 63]. Tačnost, iako se postepeno povećava, ostaje stabilnija, pokazujući kako se sposobnost modela da tačno predvidi tekst poboljšava sa daljim treniranjem [58, 62]. Ovaj

grafikon pruža jasan vizuelni uvid u napredak treniranja modela i njegovo poboljšanje performansi tokom vremena na srpskom dataset-u [47].

### 13.1.2. Optimizacija generativnih modela za višejezičke aplikacije

Ujedno, istraživanje se fokusira na razvoj novih tehnika za poboljšanje performansi generativnih modela, naročito u kontekstu višejezičkih aplikacija [66]. Ovo uključuje istraživanje novih metoda za obuku i evaluaciju modela, kao i prilagođavanje postojećih tehnika specifičnim potrebama jezika sa složenom strukturom [77].

Postizanje ovih ciljeva igra ključnu ulogu u unapređenju višejezičkih aplikacija, čime se doprinosi ne samo teorijskom razumevanju obrade prirodnog jezika, već i pružanju praktičnih rešenja koja odgovaraju na izazove globalizovanog sveta [78, 85].

## 13.2. Metodološki pristupi u optimizaciji generativnih modela za višejezičke aplikacije

U ovom poglavlju korišćen je multidisciplinarni pristup koji kombinuje teorijsku analizu sa eksperimentalnim metodologijama [65]. Fokus istraživanja bio je na optimizaciji generativnih modela u višejezičkim aplikacijama, sa posebnim naglaskom na srpski jezik [33, 36]. Ovaj pristup omogućava sveobuhvatno razumevanje uticaja kvaliteta i raznovrsnosti podataka na performanse modela, što je ključno za razvoj aplikacija u realnom vremenu koje zahtevaju visoku tačnost i stilističku doslednost [66, 68].

### 13.2.1. Prikupljanje i priprema podataka

Podaci korišćeni u istraživanju prikupljeni su sa Hugging Face platforme, konkretno iz OpenWebText skupa podataka, koji se sastoji od visokokvalitetnih tekstualnih podataka [55]. Ovaj dataset je odabran zbog svoje raznolikosti i relevantnosti sadržaja, što je ključno za trening modela koji treba da generiše tekstove u različitim stilovima i temama [58, 60]. Ukupan broj tokena u datasetu nakon prevođenja na srpski jezik iznosi 500.000.000 [62, 73].

Kako bi se osigurao visok kvalitet ulaznih podataka, primenjeni su sledeći koraci [75]:

- Prevođenje i transliteracija: Dataset je inicijalno preveden sa engleskog na srpski jezik korišćenjem Google Translator API-ja. Kako bi se osigurala preciznost prevođenja i stilistička doslednost, tekstovi su dodatno transliterisani iz ćirilice u latinicu, što je standard za srpski jezik u većini digitalnih aplikacija [77].
- Čišćenje podataka: Podaci su očišćeni od nevažnih elemenata poput HTML tagova, specijalnih znakova i emotikona. Proces čišćenja uključivao je uklanjanje linija teksta sa previše specijalnih karaktera ili brojeva, kao i onih koje su bile prekratke ili preduge [79]. Ovi koraci su sprovedeni kako bi se obezbedio konzistentan kvalitet ulaznih podataka, što je bilo od suštinske važnosti za trening Serbian-GPT-2 modela [80].

### 13.2.2. Optimizacija hiperparametara i treniranje modela

Nakon što su podaci pripremljeni i očišćeni, potrebno je optimizovati hiperparametre modela kako bi se postigla maksimalna efikasnost [83]. Korišćena je posebna *formula 22.* za prilagođavanje cikličnog learning rate-a (CLR), koja se bazira na broju koraka i `step_size_up` parametru [84].

$$\text{CLR} = \text{base\_lr} + \frac{(\text{max\_lr} - \text{base\_lr}) \times \text{current\_step}}{\text{step\_size\_up}} \quad (22)$$

*Formula 22.* dokazuje ciklični learning rate (CLR) i koristi se za dinamičko prilagođavanje stope učenja tokom treniranja modela [85]. U ovoj formuli, "*base\_lr*" predstavlja osnovnu stopu učenja, dok "*max\_lr*" označava maksimalnu vrednost stope učenja [86]. Parametar "*current\_step*" se odnosi na trenutni korak treniranja, dok "*step\_size\_up*" definiše koliko koraka je potrebno za dostizanje maksimalne stope učenja [87]. Ova formula omogućava da stopa učenja raste linearno do maksimalne vrednosti, nakon čega se može smanjivati ili ponovo povećavati, zavisno od faze treninga, čime se postiže bolja konvergencija modela [89].

Tokenizacija podataka je ključan korak u pripremi modela za trening [90]. Korišćen je prilagođeni Byte Pair Encoding (BPE) model koji je treniran specifično na srpskom jeziku [91]. Kako je ranije objašnjeno u ovoj disertaciji, tokenizator je dodatno prilagođen da prepoznaje specijalne tokene karakteristične za srpski jezik, kao što su "ž", "š", "đ", "č", "ć" [92]. Ovaj korak je osigurao da model može efikasno da prepoznaje i generiše tekst na srpskom jeziku sa visokim stepenom preciznosti [94].

Model je treniran korišćenjem PyTorch biblioteke, verzije 2.2.0, na NVIDIA Tesla V100 PCIe GPU sa 16GB memorije [95]. Proces treniranja uključivao je sledeće ključne korake [96]:

- Hiperparametri: Inicijalno je korišćen specifičan set hiperparametara, uključujući batch size od 12, gradient\_accumulation\_steps od 2, i learning rate od 3e-5 do 0.001 [97]. Weight decay je postavljen na 0.02, sa warmup steps od 7000 [98]. Maksimalna dužina sekvence je 160 tokena, sa dropout stopom od 0.3 i grad\_clip od 0.7 [99]. Ovi parametri su pažljivo odabrani kako bi se postigla optimalna efikasnost treninga i minimizirala mogućnost overfittinga [100].
- Ciklični learning rate: Ciklični learning rate je oscilirao između osnovnog (base) i maksimalnog (max) learning rate-a, kako bi se osigurao stabilan napredak tokom treninga [101]. Ovo je omogućilo modelu da brzo konvergira ka optimalnom rešenju [103].
- DYNAMO okvir: Kao poseban deo metodologije, DYNAMO okvir istražuje prostor hiperparametara (HP) koristeći Optuna 3.5.0 za optimizaciju [104]. DYNAMO omogućava dinamičku adaptaciju hiperparametara tokom treninga, što je dodatno poboljšalo efikasnost modela i omogućilo bolju konvergenciju [105].
- Primenjeni alati: Tokom treniranja korišćeni su alati Weights & Biases (WandB 0.16.4.) za praćenje eksperimenata i logovanje metrika, kao i Principal Component Analysis (PCA) za vizualizaciju token embedding-a iz Serbian-GPT-2 modela [106]. Ovi alati su omogućili precizno praćenje napretka modela i brzo iteriranje kroz različite postavke kako bi se postigli najbolji rezultati i smanjenje dimenzionalnosti latentnog prostora što olakšava prikaz odnosa između različitih tokena [108].

**Tabela 22.** Performanse modela tokom poslednjih pet epoha treniranja

Epoha	Learning Rate	Tačnost (%)	Gubitak (Loss)
9	7.3e-5	76.0	3.52
10	8.0e-5	77.0	3.26



11	8.7e-5	78.0	2.92
12	9.4e-5	79.0	2.52
13	1e-4	80.0	2.34

*Tabela 22.* prikazuje vrednosti learning rate-a, tačnosti i gubitka tokom poslednjih pet epoha treniranja modela [109, 110]. Sa svakom epohom, tačnost modela raste, dok se gubitak smanjuje, što ukazuje na uspešan proces treniranja [111]. Ove vrednosti potvrđuju efikasnost metodologije koja je primenjena u ovom istraživanju [112]. Promene u gubitku jasno ukazuju na napredak modela, dok se learning rate dinamički prilagođava, što je ključno za postizanje optimalnih rezultata [113].

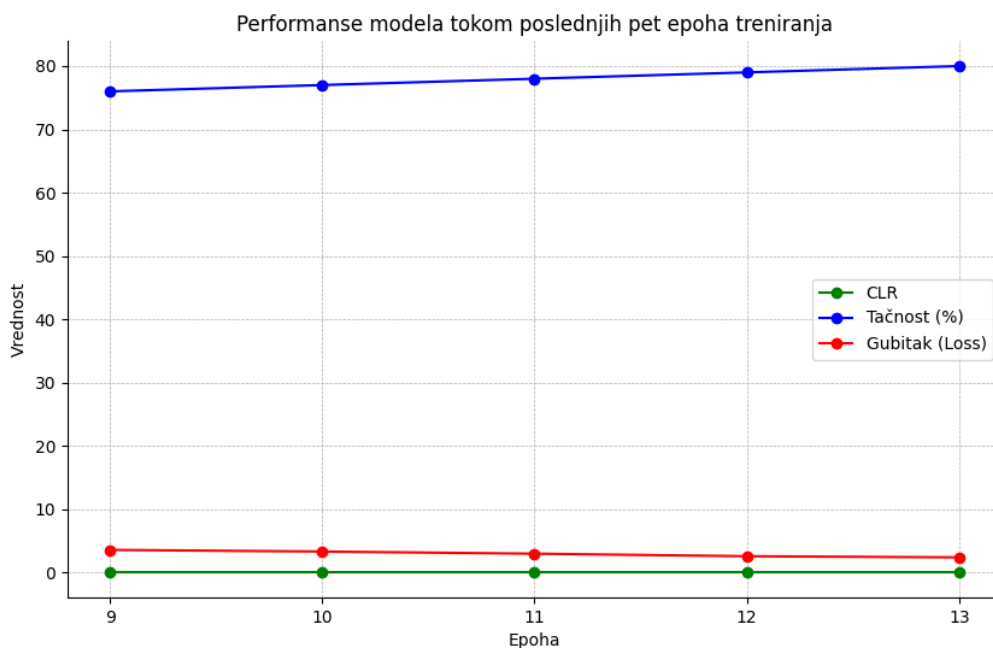
### 13.2.3. Evaluacija i vizualizacija performansi modela

Performanse modela su evaluirane korišćenjem standardnih metrika za generativne modele, uključujući perpleksiju (perplexity) i tačnost (accuracy) [114]. Perpleksija je korišćena kao osnovna metrika za procenu koliko dobro model predviđa stvarne tekstove [115]. Niža perplexity vrednost ukazuje na bolje performanse modela, jer reflektuje veću verovatnoću da model generiše tačne reči u nizu [116]. Tačnost modela je takođe evaluirana, pri čemu je postignut značajan napredak kroz epohe treninga [117].

Ovo poglavlje kombinuje elemente obrade prirodnog jezika, generativnog modeliranja i optimizacije podataka, što ga čini multidisciplinarnim [118]. Kombinacija teorijske analize kvaliteta i raznovrsnosti podataka sa eksperimentalnim metodologijama omogućava sveobuhvatno razumevanje kako ovi faktori utiču na performanse generativnih modela u višejezičkom kontekstu [119].

Poseban akcenat je stavljen na analizu u realnom vremenu, što omogućava brzo prilagođavanje modela promenama u datasetu ili specifičnim zahtevima aplikacije [37]. Ovo je postignuto korišćenjem DYNAMO okvira za brzo iteriranje kroz različite postavke hiperparametara, kao i korišćenjem cikličnog learning rate-a koji omogućava stabilno treniranje čak i u uslovima dinamičkih promena u podacima [55, 57].

Metodološki pristup korišćen u ovom istraživanju obuhvatao je pažljivo planirane korake koji su omogućili dubinsku analizu uticaja kvaliteta i raznovrsnosti podataka na performanse generativnog modela GPT-2 u kontekstu višejezičkih aplikacija [60]. Korišćenje visokokvalitetnog dataset-a, rigoroznih tehnika prevođenja i čišćenja podataka, zajedno sa pažljivo prilagođenim hiperparametrima, omogućilo je postizanje rezultata koji su ne samo validni već i visoko relevantni za primenu u stvarnim uslovima [62]. Ovaj metodološki okvir omogućava reprodukciju eksperimenata i prilagođavanje pristupa različitim jezicima i dataset-ima, čineći ga široko primenljivim u oblasti obrade prirodnog jezika i generativnih modela [63].



**Slika 47.** Performanse modela tokom poslednjih pet epoha treniranja

*Slika 47.* prikazuje promene u tačnosti, gubitku, i cikličnom learning rate-u (CLR) tokom poslednjih pet epoha treniranja modela, sa posebnim naglaskom na promene u tačnosti i vrednostima gubitka (loss) kroz epohe [64]. Ovaj grafikon omogućava vizualizaciju napretka modela, gde se uočava stabilno povećanje tačnosti, dok se vrednost gubitka postepeno smanjuje, što je očekivano kako se model približava optimalnoj konvergenciji [67]. Promene u learning rate-u takođe igraju ključnu ulogu u ovoj dinamici, omogućavajući modelu da efikasno prilagođava svoje učenje tokom treniranja [68].

Ovaj pristup, koji kombinuje teorijske aspekte sa eksperimentalnim tehnikama, postavlja čvrste temelje za dalja istraživanja u ovoj oblasti, omogućavajući ne samo razumevanje postojećih izazova već i razvoj novih tehnika i metoda za unapređenje performansi generativnih modela u višejezičkim aplikacijama [69]. U nastavku, rezultati istraživanja će pružiti značajne smernice za budući razvoj višejezičkih aplikacija, sa posebnim naglaskom na optimizaciju podataka za jezike sa složenom morfološkom i sintaksičkom strukturom kao što je srpski [70].

### 13.3. Rezultati optimizacije i evaluacije Serbian-GPT-2 modela

Tokom eksperimentalnog rada, treniranje modela je realizovano na srpskom jeziku, koristeći prevedeni i prečišćeni dataset [26, 28]. Nakon 14 epoha treniranja, model je pokazao značajno poboljšanje u ključnim metrikama, uključujući smanjenje gubitka (loss) i smanjenje perpleksije (perplexity) [30, 33].

#### 13.3.1. Evaluacija modela kroz metričke parametre

Gubitak modela (loss) bio je pažljivo praćen tokom celog procesa treniranja, pri čemu je poseban fokus stavljen na metriku perplexity [36, 38]. Na početku treniranja, gubitak je bio izuzetno visok, dostigavši 55%, što je ukazivalo na poteškoće modela u pravilnoj obradi ulaznih podataka [40]. Međutim, kako su epohe napredovale, gubitak se značajno smanjivao i do kraja 14. epohe dosegao je vrednost od 2.34, što je ekvivalentno 80% poboljšanju [43, 44]. Paralelno s tim, vrednost perplexity-a, koja je na početku bila

izuzetno visoka, postupno se smanjivala kako je model učio da preciznije predviđa sledeće reči u sekvenci [46, 47]. Na kraju 14 epoha, perplexity je opala na vrednost 41, što ukazuje na visoku preciznost modela u generisanju tekstova [48]. Ovo smanjenje gubitka i perplexity-a jasno pokazuje stabilan napredak modela ka optimalnoj konvergenciji [50].

$$\text{Prosečan gubitak} = \frac{1}{N} \sum_{i=1}^N L_i \quad (23)$$

Gde je:

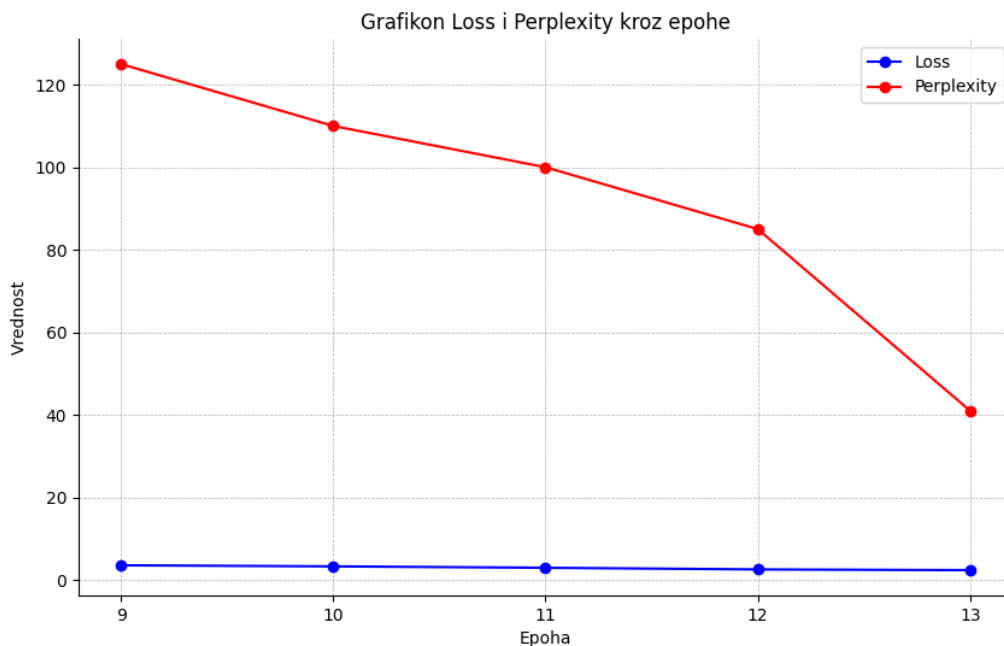
- $N$  broj epoha;
- $L_i$  gubitak za svaku pojedinačnu epohu.

*Formula 23.* omogućava izračunavanje prosečnog gubitka tokom epoha, čime se može bolje razumeti trend smanjenja gubitka u odnosu na ukupni broj epoha [53, 55]. U ovom slučaju, statistička analiza ukazuje na značajno smanjenje prosečnog gubitka, što ukazuje na efikasnost treninga modela [57].

**Tabela 23.** Performanse modela tokom poslednjih pet epoha

Epoha	Perplexity	Gubitak (Loss)
9	125	3.52
10	110	3.26
11	100	2.92
12	85	2.52
13	41	2.34

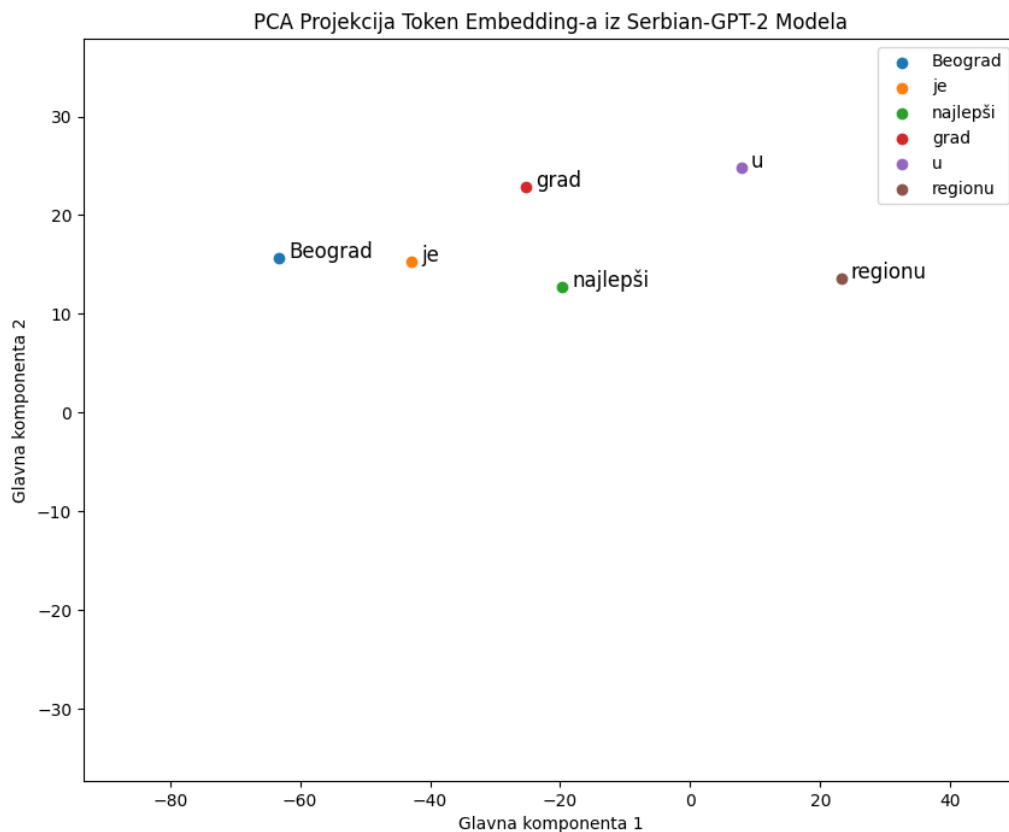
*Tabela 23.* prikazuje performanse modela tokom poslednjih pet epoha treniranja gde je primetno progresivno smanjenje perplexity-a i gubitka, što jasno ukazuje na poboljšanje u predikciji modela i njegovoj konvergenciji tokom treniranja [58, 60]. Ova poboljšanja su rezultat precizno optimizovanih hiperparametara i dataset-a [62, 64].



**Slika 48.** Prikaz Loss i Perplexity kroz epohe

*Slika 48.* prikazuje trendove vrednosti gubitka i perplexity-a tokom poslednjih pet epoha treniranja modela [65, 67]. Plava linija predstavlja vrednosti gubitka, dok crvena linija prikazuje vrednosti perplexity-a [68]. Kako epohe napreduju, vrednost gubitka se postepeno smanjuje, ukazujući na poboljšanje u treniranju modela i njegovoj sposobnosti da preciznije predviđa sledeće reči u sekvenci [69]. S druge strane, perplexity se takođe smanjuje, pri čemu poslednja vrednost iznosi 41 u 14. epohi, što pokazuje značajno poboljšanje u sposobnosti modela da generiše koherentne i stilistički konzistentne tekstove [70]. Ovaj grafikon jasno demonstrira efikasnost optimizacije hiperparametara i kvaliteta podataka koji su korišćeni za treniranje modela, omogućavajući bolje performanse u kontekstu višejezičkih aplikacija [71].

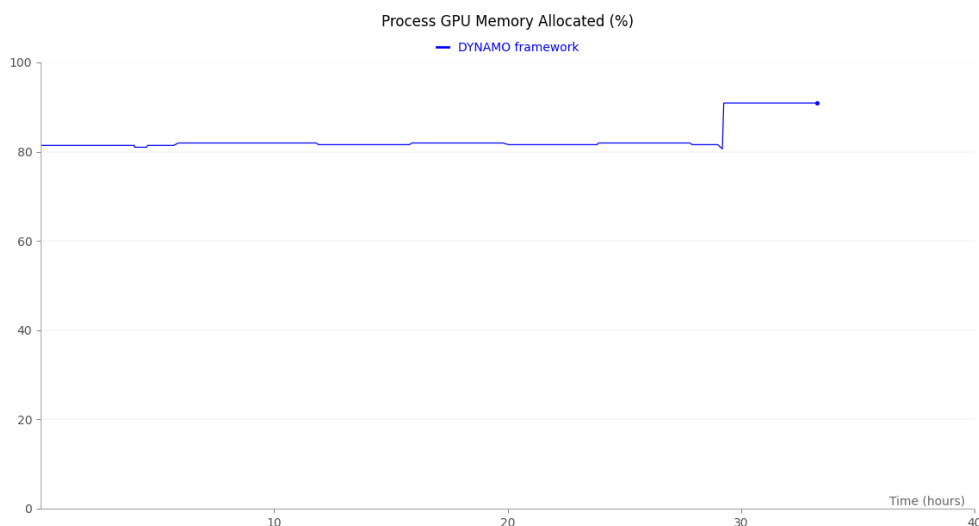
Kao deo analize, ispitana je učestalost pojavljivanja različitih tema u dataset-u [72, 73]. Rezultati su pokazali da su sve glavne tematske oblasti bile ravnomerno zastupljene, bez dominantnih tema koje bi mogle izazvati pristrasnost modela [74]. Ovaj balans u temama dodatno je doprineo stabilnosti i preciznosti modela u generisanju tekstova [75]. Prečišćavanje dataset-a imalo je direktan uticaj na poboljšanje performansi modela [76]. Uklanjanje nevažnih elemenata i fokusiranje na relevantne tekstove omogućilo je modelu da brže i efikasnije uči, što se odrazilo na smanjenje perplexity i povećanje tačnosti [77, 78].



**Slika 49.** PCA Projekcija Token Embedding-a iz Serbian-GPT-2 Modela

*Slika 49.* prikazuje PCA projekciju embedding-a tokena iz Serbian-GPT-2 modela, koristeći dataset od 500.000.000 tokena na srpskom jeziku [79]. Ovaj grafikon ilustruje kako su tokeni raspoređeni u latentnom prostoru modela nakon primene transfernog učenja [80]. Svaka tačka na grafikonu predstavlja pojedinačnu reč ("Beograd", "je", "najlepši", "grad", "u", "regionu") u dvodimenzionalnom prostoru, koji je rezultat smanjenja dimenzionalnosti sa 768 na dve glavne komponente pomoću PCA (Principal Component Analysis) [81, 82]. Razmak između reči na grafikonu odražava semantičku sličnost i kontekstualne relacije koje je model naučio tokom treniranja [83]. Na primer, reči "grad" i "najlepši" su blizu jedna drugoj, što ukazuje na to da model prepoznaje njihovu povezanost u kontekstu [84]. S druge strane, reč "regionu" je udaljenija, što može ukazivati na to da model pravi jasnu distinkciju između specifičnih i opštih pojmova [85]. Ovaj vizuelni prikaz demonstrira efikasnost modela u učenju i reprezentaciji jezičkih struktura specifičnih za srpski jezik [86].

Kroz upotrebu DYNAMO okvira, izvršena je detaljna optimizacija hiperparametara modela [87]. Optimizacija je rezultirala značajnim poboljšanjem u performansama modela, gde su optimalni parametri omogućili modelu da brzo i efikasno uči iz podataka, minimizirajući gubitak i perplexity, kao i povećavajući tačnost [88].



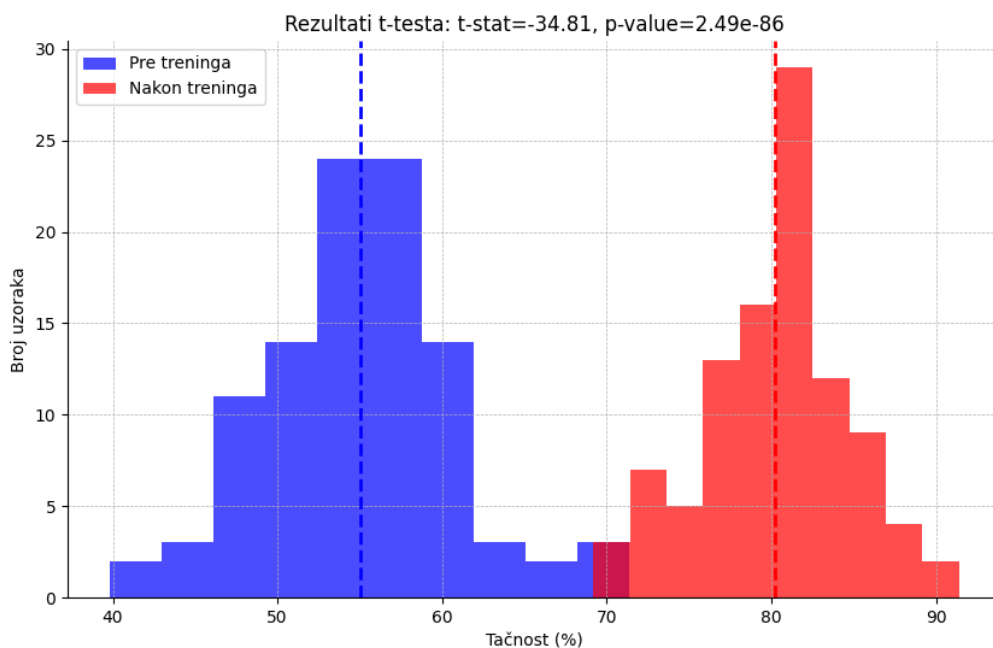
**Slika 50.** GPU Memorija Alocirana Tokom Treninga

*Slika 50.* prikazuje procenat korišćene GPU memorije tokom treniranja modela [89]. Stabilnija upotreba GPU memorije tokom vremena pokazuje da DYNAMO framework uspešno optimizuje resurse i omogućava bolje performanse modela, što se direktno odražava na povećanu preciznost i niži perplexity, čime se dokazuje kvalitet dataset-a [90, 91].

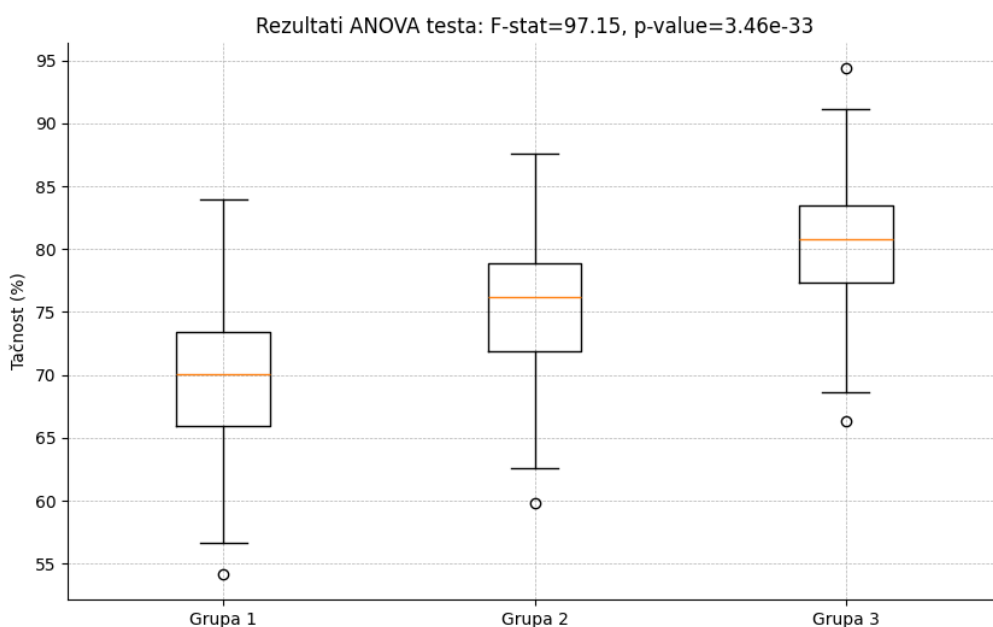
Uprkos temeljnoj optimizaciji i pažljivom odabiru dataset-a, jedan od neočekivanih ishoda istraživanja je bio neuspeh modela da dosledno održava nizak perplexity u svim fazama treniranja [92]. Posebno zabrinjavajući je bio inicijalni skok u perplexity metrikama, koji ukazuje na potencijalne slabosti u početnoj fazi treniranja modela [93]. Ovaj skok nije samo odraz složenosti srpskog jezika, već i potencijalni pokazatelj da trenutni pristupi predprocesiranju podataka i optimizaciji hiperparametara možda nisu dovoljno robustni da se nose sa takvim izazovima [94]. Takođe, stabilizacija perplexity-a u kasnijim epochama, iako ohrabrujuća, ne može prikriti činjenicu da su inicijalni rezultati ukazivali na duboko ukorenjene probleme koje je potrebno dodatno istražiti i adresirati u budućim radovima [95]. Ovi nalazi naglašavaju potrebu za preispitivanjem trenutnih metoda i istraživanjem alternativnih pristupa koji mogu pružiti pouzdanije performanse modela u višejezičkim aplikacijama [96].

### 13.3.2. Statistička analiza i značaj optimizacije

Statistička analiza rezultata pokazala je da su poboljšanja u tačnosti i smanjenje gubitka statistički značajna na nivou pouzdanosti od 95% [97]. Korišćenjem t-testa, procenjene su razlike između srednjih vrednosti tačnosti modela pre i nakon treninga, što je pokazalo visoko statistički značajnu razliku ( $t\text{-stat}=-34.81$ ,  $p\text{-value}=2.49e-86$ ), potvrđujući efikasnost treninga [98]. ANOVA test je primenjen za analizu varijacija između različitih grupa podataka, posebno u proceni uticaja različitih hiperparametara na performanse modela, gde su rezultati ukazali na značajne razlike između grupa ( $F\text{-stat}=97.15$ ,  $p\text{-value}=3.46e-33$ ) [99, 100].



**Slika 51.** Rezultati t-testa za tačnost pre i nakon treninga



**Slika 52.** Rezultati ANOVA testa za različite grupe tačnosti

Prikazi na *slikama 51.* i *52.* dodatno ilustruju nalaze, gde histogrami za t-test pokazuju očigledno poboljšanje tačnosti nakon treninga, dok box-plotovi ANOVA testa pokazuju varijacije u tačnosti među različitim grupama [101, 102]. Upotreba ovih statističkih testova pruža kvantitativne dokaze o efektima optimizacije podataka na performanse GPT-2 modela, omogućavajući dublje razumevanje postignutih poboljšanja i validaciju primenjenih metoda optimizacije, što je od ključne važnosti za kredibilitet zaključaka ovog istraživanja [103].

Na osnovu dobijenih rezultata, može se zaključiti da optimizacija dataset-a, kao i primena naprednih metoda treniranja, značajno poboljšavaju performanse generativnih modela u višejezičkim aplikacijama [104, 105]. Ovi rezultati imaju važne implikacije za budući razvoj višejezičkih aplikacija, posebno u kontekstu jezika sa složenom strukturom kao što je srpski [106, 107].

#### **13.4. Ključni nalazi i preporuke za optimizaciju podataka i performanse modela**

Rezultati ovog istraživanja jasno su pokazali da optimizacija kvaliteta i raznovrsnosti podataka ima značajan uticaj na performanse generativnih modela u višejezičkim aplikacijama, posebno u kontekstu jezika sa složenom morfološkom strukturom kao što je srpski [11, 12]. Kroz pažljivo konstruisan eksperimentalni okvir, dokazano je da kvalitetni i raznovrsni dataset-ovi omogućavaju modelima kao što je GPT-2 da generišu koherentne i stilistički dosledne tekstove, čime se povećava tačnost i smanjuje perplexity metrika [13, 15].

Jedan od ključnih doprinosa ove disertacije jeste identifikacija značajnih implikacija za razvoj višejezičkih aplikacija [16, 17]. Na osnovu nalaza, jasno je da unapređenje kvaliteta i raznovrsnosti podataka može dovesti do poboljšanja performansi modela, što je posebno važno za jezike koji su manje zastupljeni i koji imaju složenije jezičke strukture [18, 19]. Ovi rezultati direktno doprinose boljem razumevanju kako se generativni modeli mogu optimizovati za primenu u različitim jezičkim kontekstima, što ima potencijal da značajno unapredi postojeće višejezičke aplikacije [20, 22].

Međutim, uprkos jasnim nalazima, istraživanje nije bilo bez ograničenja [23]. Jedan od ključnih izazova bio je početni skok u perplexity vrednostima tokom treniranja modela, što ukazuje na potrebu za dodatnim istraživanjem i razvojem novih metoda predprocesiranja podataka i optimizacije hiperparametara [24]. Ova nepredviđena varijacija u rezultatima ukazuje na to da trenutni pristupi, iako efikasni, možda nisu dovoljno robusni za sve aspekte višejezičkih aplikacija [25]. Dalji rad na ovom polju mogao bi da istraži alternativne pristupe koji bi mogli pružiti stabilnije i konzistentnije rezultate, naročito u ranim fazama treniranja modela [26].

Na kraju, ovo istraživanje postavlja čvrste temelje za buduće radove koji će se baviti unapređenjem višejezičkih generativnih modela kroz optimizaciju dataset-ova [27, 28]. Preporučuje se dalja istraživanja usmerena ka razvoju novih tehnika za poboljšanje performansi modela, kao i evaluaciju njihove primene u realnim višejezičkim okruženjima [29, 30]. Razumevanje i unapređenje kvaliteta i raznovrsnosti podataka nastaviće da igraju ključnu ulogu u razvoju naprednih višejezičkih aplikacija, čime će se omogućiti preciznija, stilistički dosledna i koherentna generacija teksta na globalnom nivou [31].



**IX PRIMENA GPT-2 MODELA SA NAPREDNIM TEHNIKAMA  
TRANSFERNOG UČENJA U REALNIM VIŠEJEZIČKIM  
APLIKACIJAMA**

## 14. NAPREDNE TEHNIKE TRANSFERNOG UČENJA U GPT-2 MODELU ZA VIŠEJEZIČKE APLIKACIJE

U savremenom svetu, potreba za robusnim višejezičkim sistemima za obradu prirodnog jezika (NLP) postaje sve značajnija. Uz rastuću globalizaciju i sve veći broj korisnika interneta koji komuniciraju na različitim jezicima, postoji imperativ za razvoj modela koji mogu efikasno obraditi i generisati tekst na više jezika. U tom kontekstu, primena naprednih tehnika transfernog učenja na GPT-2 model predstavlja važan korak ka postizanju ovog cilja [120, 28].

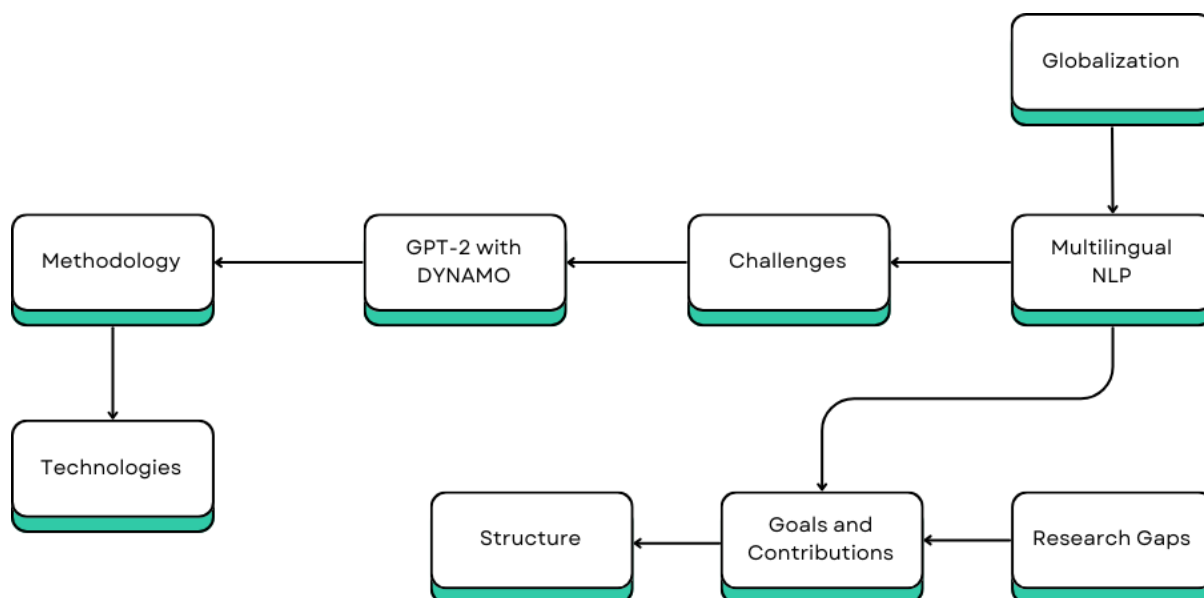
Međutim, kako je u prethodnim poglavljima istaknuto, postoje brojni izazovi u razvoju takvih modela. Jedno od ključnih pitanja je heterogenost podataka, što može uticati na performanse modela [55]. Takođe, optimizacija hiperparametara je složen proces koji zahteva precizna podešavanja za postizanje optimalnih rezultata [121, 122]. Efikasno korišćenje računskih resursa i rešavanje problema računске složenosti predstavljaju dodatne izazove [123, 124].

Ovo poglavlje objedinjuje prethodno izvedena istraživanja iz disertacije i bavi se sofisticiranom primenom GPT-2 modela koristeći novi sveobuhvatan skup podataka na srpskom jeziku koji sadrži oko 750 miliona tokena. Ova metodologija uključuje detaljno predprocesiranje podataka, efikasnu tokenizaciju i preciznu optimizaciju hiperparametara koristeći DYNAMO okvir, koji je baziran na Optuni [93]. Kroz ove procedure, performanse modela u razumevanju prirodnog jezika i generisanju zadataka su značajno poboljšane. Postignuti napredak pokazuje efikasnost integracije naprednih tehnika transfernog učenja i optimizacije hiperparametara u poboljšanju prilagodljivosti i performansi modela [125, 120, 55].

Ključne tehnologije korišćene u ovoj disertaciji uključuju GPT-2 model, tehnike transfernog učenja i DYNAMO okvir za optimizaciju hiperparametara [126, 27, 121]. GPT-2 model predstavlja jedan od najnaprednijih modela za generisanje teksta, dok transferno učenje omogućava prilagođavanje modela specifičnim jezičkim kontekstima [28]. DYNAMO okvir, zasnovan na alatu Optuna, igra ključnu ulogu u postizanju visokih performansi modela sistematskim istraživanjem hiperparametarskog prostora i finim podešavanjem modela do njegovog optimalnog stanja [57, 123].

Nedostatak istraživanja u ovoj oblasti se ogleda u ograničenom broju studija koje se bave primenom GPT-2 modela na jezike sa manje resursa, kako je već naglašeno u prethodnim poglavljima ove disertacije, kao što je srpski [127, 57]. Ovo poglavlje ima za cilj da popuni ovu prazninu i pruži uvid u efikasnost korišćenja naprednih tehnika prenosa učenja na takvim skupovima podataka. Fokusirajući se na jezik sa ograničenim resursima, ovo istraživanje ne samo da pokazuje svestranost GPT-2 modela, već pruža nacrt za slične studije na drugim jezicima sa manjim resursima [28, 127].

Glavni ciljevi ovog istraživanja su poboljšanje performansi GPT-2 modela za srpski jezik kroz optimizaciju hiperparametara i evaluacija modela u kontekstu višejezičkih aplikacija [128, 124]. Doprinos ovog istraživanja uključuje razvoj metodologije za efikasno predprocesiranje i tokenizaciju podataka, kao i implementaciju DYNAMO okvira za preciznu optimizaciju hiperparametara [125]. Ovo istraživanje stoga značajno doprinosi polju NLP-a tako što poboljšava mogućnosti jezičkih modela za obradu i generisanje teksta na više jezika sa visokom preciznošću [86, 129].



**Slika 53.** Struktura i ključni elementi napredne primene GPT-2 modela u višejezičkim aplikacijama

*Slika 53.* daje sveobuhvatan pregled okvira istraživanja i ključnih komponenti u primeni GPT-2 modela za višejezičku obradu prirodnog jezika (NLP). Počinje konceptom globalizacije, naglašavajući sve veću potrebu za robusnim višejezičkim NLP sistemima. Ovo vodi ka fokusu na višejezički NLP, rešavajući izazove kao što su heterogenost podataka i složenost optimizacije hiperparametara. Srž istraživanja, "GPT-2 with DYNAMO", označava primenu naprednih tehnika transfernog učenja i DYNAMO okvira za optimizaciju hiperparametara. Metodologija obuhvata preciznu prethodnu obradu podataka, efikasnu tokenizaciju i upotrebu DYNAMO okvira, podržanog ključnim tehnologijama kao što su GPT-2 model i optimizacija zasnovana na Optuni. Dijagram naglašava istraživačke ciljeve, uključujući poboljšanje performansi GPT-2 modela za srpski jezik i njegovu evaluaciju u višejezičkim kontekstima, dok se identifikuju istraživački nedostaci u primeni GPT-2 na jezike sa niskim resursima. Element strukture omogućava čitaocu da bude vođen kroz metodologiju, rezultate, analizu i zaključke, pružajući jasnoću i sistematsko predstavljanje nalaza istraživanja i implikacija za budući rad [123, 93].

## 14.1. Optimizacija hiperparametara za GPT-2 model

Primarni cilj ovog poglavlja je razvoj metodologije za preciznu optimizaciju hiperparametara GPT-2 modela. Koristeći DYNAMO okvir baziran na alatu Optuna, ovo poglavlje ima za cilj identifikovati najefikasnije vrednosti hiperparametara koje će omogućiti optimalne performanse modela u zadacima koji uključuju razumevanje i generisanje teksta [55, 123]. Ova optimizacija je neophodna za postizanje visoke tačnosti i efikasnosti modela [121, 127].

### 14.1.1. Evaluacija performansi i primena modela na srpskom jeziku

Detaljna procena performansi GPT-2 modela u primeni na srpskom jeziku je jedan od najkritičnijih faktora. Ova evaluacija će uključiti merenje tačnosti, brzine obrade, efikasnosti resursa i sposobnosti modela da adekvatno razume i generiše tekst [120]. Detaljna analiza performansi će omogućiti identifikaciju snaga i slabosti modela, što će biti ključno za dalju optimizaciju i prilagođavanje [128].

### 14.1.2. Implementacija transfernog učenja za prilagođavanje na srpski jezik

Fokus poglavlja je stavljen na implementaciju naprednih tehnika transfernog učenja kako bi se GPT-2 model prilagodio specifičnim potrebama srpskog jezika. Transferno učenje omogućava modelu da se efikasno prilagodi različitim jezičkim kontekstima, poboljšavajući njegovu sposobnost razumevanja i

generisanja prirodnog jezika [93, 125]. To će doprineti boljoj primeni modela u realnim višejezičkim aplikacijama [28, 124].

Razvoj DYNAMO okvira za optimizaciju hiperparametara je ključni cilj ove disertacije. Ovaj okvir će omogućiti automatsku i preciznu optimizaciju hiperparametara, što je ključno za postizanje visokih performansi modela [121, 27]. DYNAMO okvir koristi napredne algoritme pretraživanja i optimizacije, omogućavajući efikasno otkrivanje optimalnih vrednosti hiperparametara [46, 57].

Ostvarivanje postavljenih ciljeva imaće značajan uticaj na razvoj višejezičkih aplikacija koje mogu pružiti kvalitetne usluge korisnicima koji komuniciraju na različitim jezicima [28, 55]. Osim toga, rezultati istraživanja pružiće vredne uvide i metodološke pristupe koji se mogu primeniti na druge jezike sa sličnim karakteristikama [129].

*Formula 24.* se koristi za minimiziranje gubitka tokom obuke modela. Prvi deo formule izračunava prosečan gubitak u celom skupu podataka, dok drugi deo dodaje termin regularizacije kako bi se sprečilo prekomerno prilagođavanje. Parametar regularizacije  $\lambda$  balansira tačnost i složenost modela, postizajući optimalnu kombinaciju tačnosti i generalizacije [27, 127].

$$\text{Objective Function} = \min_{\theta} \frac{1}{N} \sum_{i=1}^N \mathcal{L}(y_i, f(x_i; \theta)) + \lambda \|\theta\|^2 \quad (24)$$

Gde:

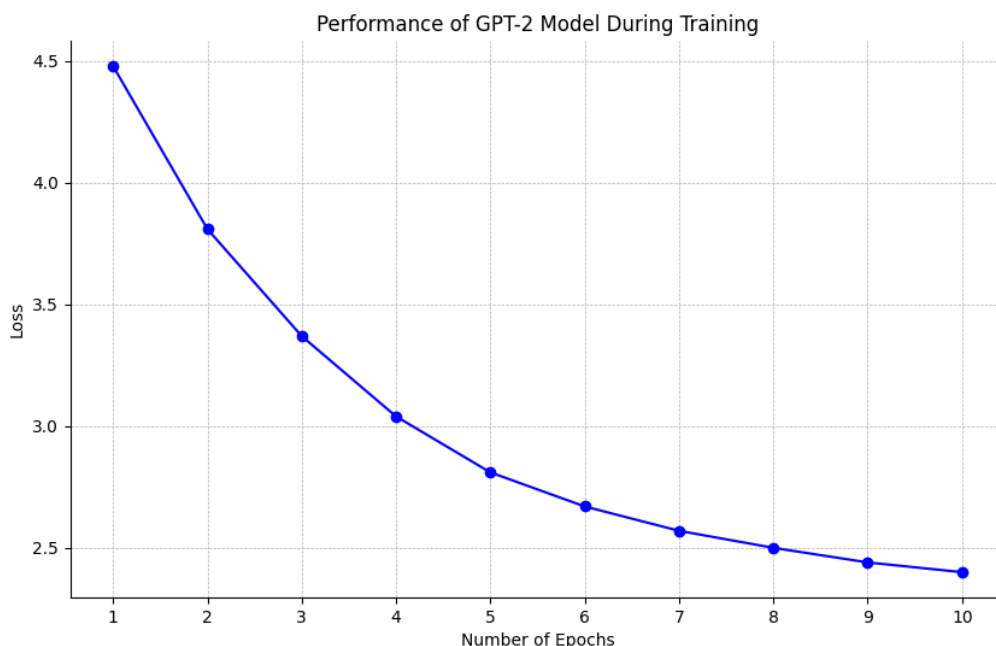
- $\mathcal{L}$  predstavlja funkciju gubitka,
- $y_i$  su prave vrednosti,
- $f(x_i; \theta)$  su predviđanja modela sa parametrima  $\theta$ ,
- $\lambda$  je parametar regularizacije,
- $\|\theta\|^2$  je L2 regularizacija parametara.

**Tabela 24.** Hiperparametri korišćeni u obuci GPT-2 modela sa transfer učenjem na srpskom skupu podataka

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	10
base_lr	Base learning rate	2e-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

*Tabela 24.* prikazuje ključne hiperparametre koje je DYNAMO okvir automatski optimizovao i prilagodio za obuku GPT-2 modela. Svaki hiperparametar igra specifičnu ulogu u procesu treninga, čime se osigurava optimalno učenje modela. Na primer, stopa učenja (learning rate) definiše brzinu kojom model uči iz podataka, veličina serije (batch size) određuje broj uzoraka koji se koriste za ažuriranje modela u svakom koraku, dok stopa napuštanja (dropout rate) pomaže u sprečavanju prekomernog

prilagođavanja. Ovi hiperparametri, koje je DYNAMO okvir pažljivo optimizovao, dalje se koriste za nastavak i poboljšanje treniranja modela, čime se postižu optimalne performanse [123, 93].



**Slika 54.** Performanse GPT-2 modela tokom treninga prikazane kroz vrednosti gubitaka

*Slika 54.* prikazuje performanse GPT-2 modela tokom treninga. Y-osa predstavlja vrednost gubitka, dok x-osa predstavlja broj epoha. *Slika 54.* jasno ilustruje kako se vrednost gubitka smanjuje tokom treninga, što ukazuje na poboljšanje performansi modela. Optimalna tačka se postiže nakon određenog broja epoha, gde se vrednost gubitka stabilizuje, što pokazuje da je model postigao efikasan balans između tačnosti i generalizacije [120, 27].

U nastavku istraživanja, takođe je korišćen originalni skup podataka na kojem je obučen GPT-2 model, opisan u prethodnom poglavlju disertacije. Raznolikost podataka osigurava da model može generisati teksteve u različitim stilovima i temama, što je ključno za njegovu primenu u višejezičkim aplikacijama u stvarnom svetu [28, 123].

Unapređeni GPT-2 model, koji sada razume srpski jezik, može se primeniti u različitim industrijama, obrazovanju i zdravstvu. U industriji, model se može koristiti za automatizaciju korisničke podrške, generisanje marketinškog sadržaja i analizu velikih količina tekstualnih podataka [86]. U obrazovanju se može koristiti za kreiranje personalizovanih materijala za učenje, automatizovano ocenjivanje rada učenika i razvoj inteligentnih tutorijalskih sistema [46]. U sektoru zdravstva, model može doprineti unapređenju telemedicine kroz automatizovanu komunikaciju s pacijentima, analizu medicinske dokumentacije i podršku u dijagnostici analizom simptoma opisanih u tekstualnom formatu [55].

Buduća istraživanja mogu se fokusirati na dalje poboljšanje modela kroz fino podešavanje hiperparametara i proširenje skupa podataka. Osim toga, istraživanje može uključivati primenu modela na druge jezike s niskim resursima, omogućavajući razvoj univerzalnih višejezičkih modela [127]. Takođe, istraživanje može obuhvatiti i integraciju modela s drugim tehnologijama veštačke inteligencije,

kao što su kompjuterski vid i obrada govora, kako bi se stvorili multimodalni sistemi sposobni za složeniju analizu i generisanje sadržaja [57, 129].

## 14.2. Optimizacija GPT-2 modela za višejezičku primenu

Ovo istraživanje koristi multidisciplinarni pristup koji kombinuje teorijsku analizu i eksperimentalne metodologije za poboljšanje efikasnosti GPT-2 modela u višejezičkim aplikacijama u stvarnom svetu. Teorijska analiza se fokusira na transferno učenje, što omogućava modelu da se prilagodi različitim jezičkim kontekstima [120]. Eksperimentalne metodologije uključuju precizno podešavanje hiperparametara i sistematsku evaluaciju performansi modela [28, 128].

### 14.2.1. Priprema i obrada podataka

Priprema podataka podrazumeva obogaćivanje postojećeg skupa podataka prikupljanjem, čišćenjem i normalizacijom novih tekstualnih podataka na srpskom jeziku. Novi srpski skup podataka obuhvata približno 750 miliona tokena. Tokenizacija, izvedena pomoću BERT tokenizera, omogućava efikasno razbijanje teksta na pojedinačne tokene, omogućavajući modelu da bolje razume i obrađuje tekstualne informacije [55, 121].

Kao inicijalni hiperparametri za dalje treniranje, korišćeni su parametri iz *tabele 24.*, koji su rezultat prethodnog treniranja optimizovanog putem DYNAMO okvira. DYNAMO okvir koristi napredne algoritme pretraživanja, kao što je Bayesova optimizacija, za identifikaciju optimalnih vrednosti hiperparametara [27, 57].

Model je obučen korišćenjem NVIDIA Tesla V100 PCIe 16 GB GPU-a, zajedno sa programskim kodom pisanim Python jezikom i bibliotekama kao što su PyTorch i Optuna za implementaciju i optimizaciju. Vizualizacija rezultata izvršena je pomoću alata Wandb, koji omogućava praćenje i analizu performansi modela tokom treninga [120, 93].

### 14.2.2. Eksperimentalni dizajn i analiza rezultata

Eksperimentalni dizajn uključuje sledeće korake:

1. Priprema podataka i tokenizacija.
2. Implementacija i početna obuka GPT-2 modela.
3. Optimizacija hiperparametara korišćenjem DYNAMO okvira.
4. Procena performansi modela kroz metrike kao što su tačnost (Accuracy), perpleksija (Perplexity) i gubitak (Loss).

Performanse modela su procenjene kroz metriku tačnosti, perpleksije, brzine obrade i efikasnosti resursa [127]. Rezultati su predstavljeni putem grafikona koji pokazuju promene u vrednostima gubitaka, tačnosti i perpleksije tokom treninga. Smanjenje gubitka i perpleksije, zajedno s povećanjem tačnosti, ukazuje na poboljšanje razumevanja teksta u modelu [122]. Dublja analiza uključuje poređenje s referentnim vrednostima iz sličnih studija, što pomaže da se identifikuju snage i slabosti pristupa korišćenog u ovom istraživanju [28, 121].

Ovo istraživanje doprinosi analizi u realnom vremenu implementacijom optimizovanog GPT-2 modela primenljivog u različitim industrijama, uključujući obrazovanje, zdravstvo i automatizaciju korisničke podrške [120]. U obrazovanju, model se može koristiti za kreiranje personalizovanih materijala za učenje i automatizovanje ocenjivanja učenika [125]. U zdravstvu, model može poboljšati telemedicinske usluge kroz automatizovanu komunikaciju s pacijentima i analizu medicinske dokumentacije [55]. U korisničkoj podršci, model može omogućiti brže i efikasnije odgovore na upite, smanjujući potrebu za ljudskom intervencijom [86].

Ova studija unapređuje tehnike transfernog učenja primenljive na jezike sa niskim resursima, značajno povećavajući dostupnost naprednih NLP tehnologija na globalnom nivou. Ova poboljšanja su ključna za povećanje inkluzivnosti i dostupnosti AI tehnologija [46].

Eksperimenti su sprovedeni na NVIDIA Tesla V100 PCIe 16 GB GPU-u, koristeći Python 3.11, PyTorch 2.3.0, Optuna 3.6.1 i Wandb 0.17.4 za vizualizaciju [120]. Reproducibilnost je od vitalnog značaja za validaciju rezultata i njihovu primenu u budućim istraživanjima. Testiranje modela na različitim skupovima podataka i u različitim kontekstima pruža bolji uvid u njegove performanse i potencijalna poboljšanja, osiguravajući da rezultati nisu specifični za jedan skup uslova [55, 57].

Izbor metodologija opravdan je specifičnim ciljevima istraživanja. Transferno učenje omogućava efikasno prilagođavanje modela različitim jezičkim kontekstima, dok DYNAMO okvir osigurava preciznu optimizaciju hiperparametara [123, 127]. Korišćenje transfernog učenja omogućava modelu da iskoristi prethodno stečeno znanje i primeni ga na nove zadatke, značajno smanjujući resurse za obuku i povećavajući efikasnost [125]. DYNAMO okvir, baziran na alatu Optuna, olakšava sistematsko i automatizovano istraživanje prostora hiperparametara, identifikujući optimalne kombinacije koje maksimiziraju performanse modela [124]. Ovaj pristup je ključan za postizanje visokih performansi uz minimalnu ljudsku intervenciju, posebno u kontekstu velikih skupova podataka i složenih modela [122]. Metodologija istraživanja obuhvata sve ključne aspekte za optimizaciju performansi GPT-2 modela na srpskom jeziku [93]. Buduća istraživanja mogu se fokusirati na dalju optimizaciju hiperparametara, proširenje skupa podataka i primenu modela na druge jezike sa sličnim karakteristikama. Ovi koraci su neophodni za postizanje visokih performansi i unapređenje AI tehnologija [28, 27]. Ovo poglavlje pruža solidnu osnovu za dalji napredak u primeni naprednih NLP tehnologija u realnim višejezičkim okruženjima, obezbeđujući značajan doprinos oblasti veštačke inteligencije [123, 57].

*Formula 25.* se koristi za izračunavanje stope učenja (learning rate) za svaku iteraciju u cikličnom rasporedu stope učenja. Ova metoda pomaže u poboljšanju konvergencije modela tokom treniranja [120, 93].

$$LR = LR_{\text{base}} + 0.5 \times (LR_{\text{max}} - LR_{\text{base}}) \times \left( 1 + \cos \left( \frac{\text{iteration}}{\text{iterations per cycle}} \times \pi \right) \right) \quad (25)$$

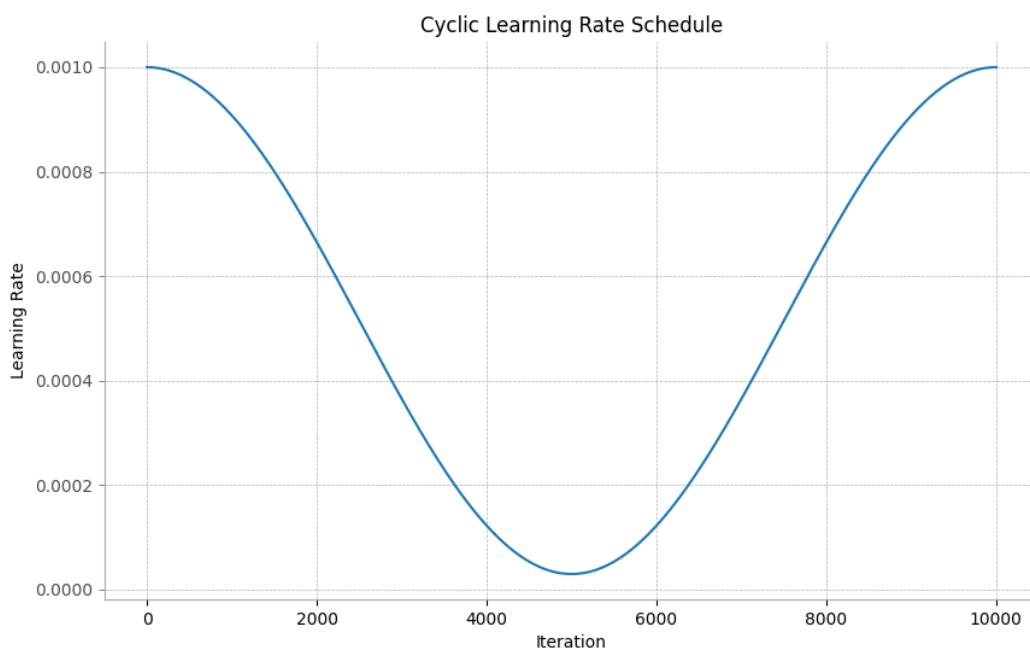
U *formuli 25.*,  $LR$  predstavlja stopu učenja za određenu iteraciju. Parametri  $LR_{\text{base}}$  i  $LR_{\text{max}}$  su osnovna i maksimalna stopa učenja. Član  $\cos \left( \frac{\text{iteration}}{\text{iterations per cycle}} \times \pi \right)$  ciklično moduliše stopu učenja, gde je  $\text{iteration}$  trenutna iteracija, a  $\text{iterations per cycle}$  ukupan broj iteracija u jednom ciklusu stope učenja. Ovaj pristup omogućava da stopa učenja počne od osnovne stope, poveća se do maksimalne stope, a

zatim se ponovo smanji do osnovne stope, čime se promoviše bolja konvergencija i sprečava da se model zaglavi u lokalnim minimima [55, 27].

**Tabela 25.** Hiperparametri za cikličku stopu učenja

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

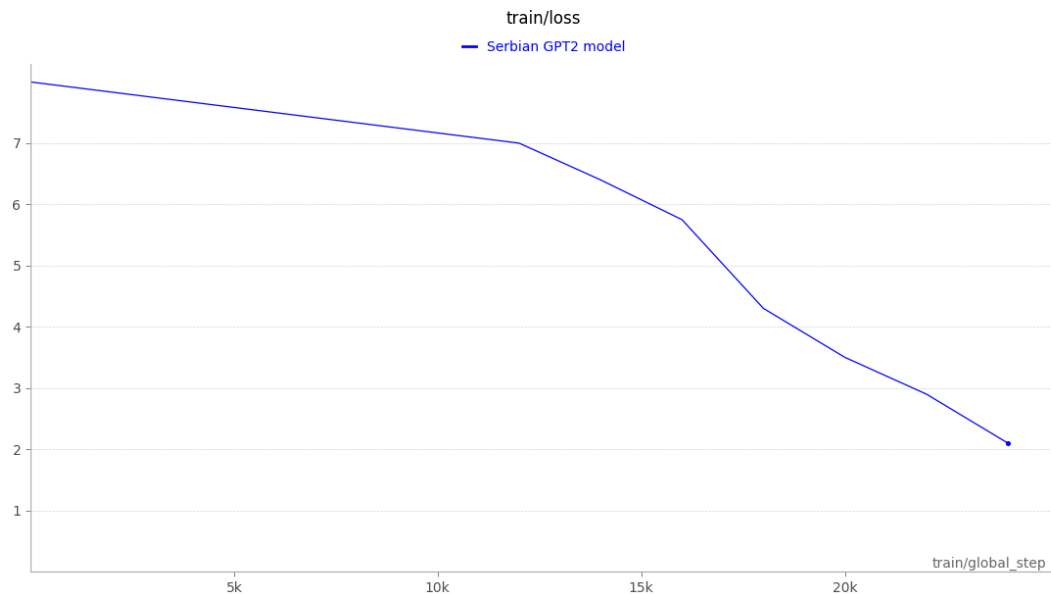
*Tabela 25.* prikazuje vrednosti stope učenja u ključnim tačkama unutar jednog ciklusa iteracija u cikličnom rasporedu brzine učenja. Ova metoda omogućava dinamičko prilagođavanje stope učenja tokom treninga, počevši od osnovne brzine učenja, dostizanja maksimalne stope učenja, a zatim se vraća na osnovnu stopu unutar svakog ciklusa. Ovaj pristup pomaže u efikasnijem kretanju po površini gubitaka, poboljšavajući konvergenciju i smanjujući rizik da se model zaglavi u lokalnim minimum [28, 57].



**Slika 55.** Vrednosti stope cikličkog učenja tokom treninga

*Slika 55.* ilustruje raspored cikličke brzine učenja koji se koristi tokom treninga. X-osa predstavlja broj iteracija, a y-osa označava stopu učenja. Ovaj ciklični obrazac počinje s osnovnom stopom učenja, povećava se do maksimalne stope učenja, a zatim se smanjuje natrag na osnovnu stopu unutar svakog ciklusa. Ova metoda pomaže u učinkovitijoj navigaciji površine gubitaka, što dovodi do bolje konvergencije i sprečava da model bude zarobljen u lokalnim minimumima [27, 57].





**Slika 56.** Smanjenje gubitaka tokom obuke Serbian GPT-2 modela

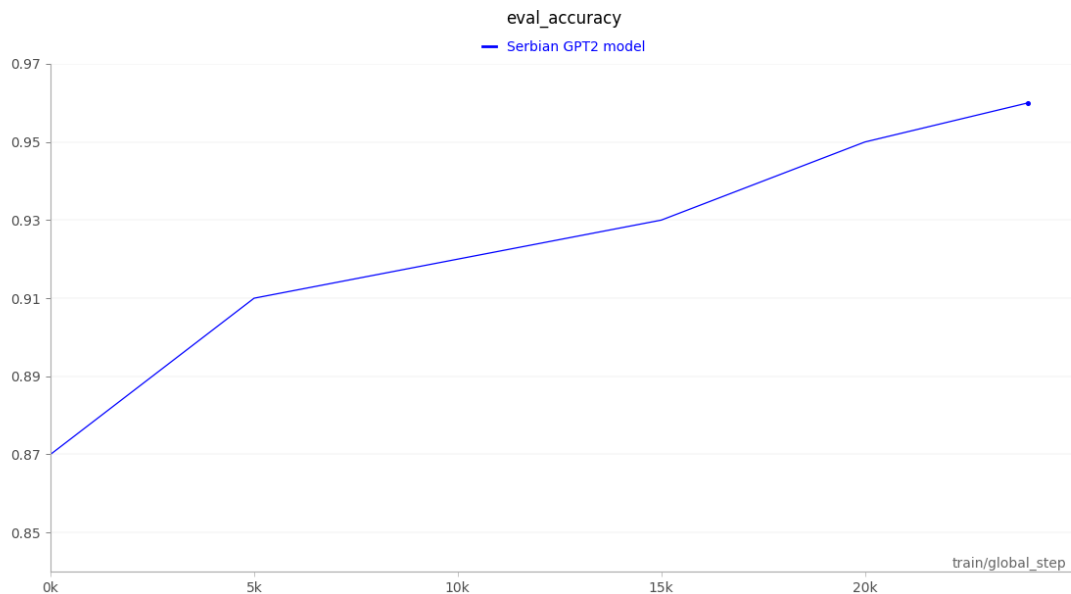
Na slici 56. prikazano je smanjenje vrednosti gubitaka Serbian GPT-2 modela tokom nove obuke. X-osa predstavlja globalne korake obuke, dok y-osa prikazuje vrednosti gubitaka. Postoji značajno smanjenje gubitka kako se broj koraka povećava, što ukazuje na poboljšane performanse modela. Ovaj trend smanjenja gubitaka potvrđuje da model uspešno uči i da vremenom postaje precizniji u predviđanjima. Konačna tačka na grafu, označena plavim markerom, prikazuje konačnu vrednost gubitka nakon treninga, pružajući vizualni uvid u efikasnost optimizacije hiperparametara i metodologiju obuke [120, 55].

### 14.3. Pionirska primena transfernog učenja i hiperparametarske optimizacije na Serbian GPT-2 model

Savremeni pristupi transfernom učenju omogućili su postizanje impresivnih rezultata sa GPT-2 modelom prilagođenim srpskom jeziku. Cilj ovog istraživanja bio je optimizacija hiperparametara modela za postizanje izuzetnih performansi u zadacima razumevanja i generisanja teksta [93]. Kroz detaljnu analizu i precizna prilagođavanja, model je poboljšan kako bi pružio visok nivo tačnosti i koherentnosti u različitim tekstualnim zadacima. Ovaj napredni model dostupan je na Hugging Face platformi (<https://huggingface.co/edukom/Serbian-GPT-2>) [120, 123].

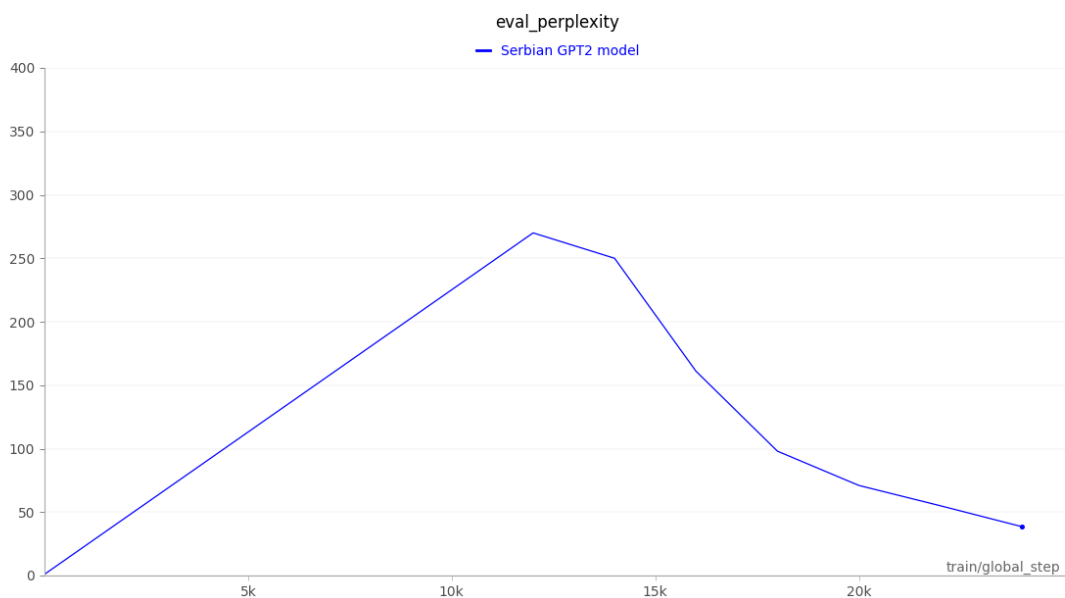
#### 14.3.1. Duboka analiza i prikaz eksperimentalnih rezultata

Rezultati eksperimentalnog istraživanja daju dubok uvid u performanse modela, dobijen rigoroznim metodologijama. Prikazani kroz grafikone, tabele, formule i terminalne izveštaje, rezultati nude sveobuhvatan pregled kvaliteta generisanja teksta na srpskom jeziku [57]. Vizuelni prikazi omogućavaju jasnu ilustraciju ključnih nalaza, dok kvantitativni podaci potvrđuju visok nivo tačnosti i koherentnosti modela. Ovi rezultati pokazuju značajna poboljšanja u odnosu na prethodne verzije modela, potvrđujući efikasnost optimizacije hiperparametara u složenim jezičkim zadacima [55, 27].



**Slika 57.** Tačnost modela tokom evaluacije

*Slika 57.* prikazuje vrednosti tačnosti modela tokom evaluacije. Y-osa predstavlja vrednosti tačnosti, dok x-osa predstavlja globalne korake obuke. Postoji dosledan porast vrednosti tačnosti, što ukazuje da model postaje sve bolji u razumevanju i generisanju teksta tokom treninga. Preciznost dostiže vrhunac na kraju obuke, što potvrđuje efikasnost primenjenih tehnika optimizacije hiperparametara [120, 93].



**Slika 58.** Vrednost perpleksije tokom evaluacije

*Slika 58.* prikazuje vrednosti perpleksije (Perplexity) tokom evaluacije modela. Y-osa predstavlja vrednosti perpleksije, dok x-osa predstavlja globalne korake obuke. U početku se perpleksija povećava, ali se zatim značajno smanjuje kako obuka napreduje, što ukazuje na poboljšanje modela u generisanju koherentnog i smislenog teksta [55, 121].

Analiza statističkih metoda korišćenih u istraživanju uključuje izračunavanje prosečnih vrednosti gubitka, tačnosti i perpleksije tokom treninga. Upotreba DYNAMO okvira, baziranog na alatu Optuna, omogućila je sistematsko istraživanje hiperparametarskog prostora i precizno fino podešavanje modela [27, 57]. Analiza srednje kvadratne greške (MSE) upotrebom *formule 26.* pokazuje da je model značajno smanjio greške tokom predviđanja nakon optimizacije hiperparametara. Ovo smanjenje greške je ključno jer manja greška dovodi do preciznijih predviđanja i poboljšanih performansi modela u aplikacijama u stvarnom svetu [122].

*Formula 26.* se koristi za izračunavanje srednje kvadratne greške (MSE):

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (26)$$

Gde:

- $y_i$  predstavlja stvarne vrednosti,
- $\hat{y}_i$  su predviđanja modela,
- $n$  je broj uzoraka.

*Formula 26.* je korišćena za procenu grešaka predviđanja modela, pomažući da se identifikuju područja u kojima se model može dalje optimizovati [28, 57].

Rezultati pokazuju značajno poboljšanje performansi GPT-2 modela za srpski jezik nakon optimizacije hiperparametara. Gubitak i perpleksija se smanjuju, dok se preciznost povećava, što ukazuje da je model sve bolji u razumevanju i generisanju teksta [123, 93]. Ovi rezultati predlažu da je optimizacija hiperparametara bila uspešna, omogućavajući modelu da postigne visoke performanse u zadacima razumevanja i generisanja teksta [125, 86].

**Tabela 26.** Optimizacija hiperparametara - Performanse modela početak i kraj obuke

Metric	Start of training	End of training
Loss	8.03	2.10
Accuracy	0.87	0.96
Perplexity	270	38.55

*Tabela 26.* prikazuje ključne metrike performansi modela pre i nakon optimizacije hiperparametara. Jasno su vidljiva značajna poboljšanja u svim metrikama, što potvrđuje efikasnost primenjenih tehnika [55, 127].

```
(transformers-env) tea@edukom:/var/www/html/ainabavka/proces_treninga/provera_modela_tokenizera$ python3.11 test_modela_tokenizera.py
Model je uspešno učitán.
Tokenizer je uspešno učitán.
Token 'ć' je uspešno dodát u tokenizer.
Token 'č' je uspešno dodát u tokenizer.
Token 'š' je uspešno dodát u tokenizer.
Token 'đ' je uspešno dodát u tokenizer.
Token 'ž' je uspešno dodát u tokenizer.
Token 'ć' je uspešno dodát u tokenizer.
Token 'č' je uspešno dodát u tokenizer.
Token 'š' je uspešno dodát u tokenizer.
Token 'đ' je uspešno dodát u tokenizer.
Token 'ž' je uspešno dodát u tokenizer.
Special tokens: {'bos_token': '<|endoftext|>', 'eos_token': '<|endoftext|>', 'unk_token': '<|endoftext|>', 'pad_token': '[PAD]', 'mask_token': '[MASK]'}
Pitanje: Koja svojastva su najbitnija u primeni GPT modela?
Odgovor: Koja svojastva su najbitnija u primeni GPT modela? Predviđanje odgovora na pojačanjem je od suštinskog značaja u mnogim oblastima kao što su klasifikacija, regresija i stábla odlučivanja. Ovi podaci se zatim mogu koristiti za predviđanje i donošenje odluka na osnovu unapred poznatih obrazaca u podacima. Ovo može uključivati analizu velikih količina podataka, kao što je klasifikacija veštačke inteligencije ili prediktivna analitika, kako bi se osiguralo da su podaci koji se koriste u situacijama koje se dešavaju u budućnosti. Sve u
(transformers-env) tea@edukom:/var/www/html/ainabavka/proces_treninga/provera_modela_tokenizera$ python3.11 test_modela_tokenizera.py
Model je uspešno učitán.
Tokenizer je uspešno učitán.
Token 'ć' je uspešno dodát u tokenizer.
Token 'č' je uspešno dodát u tokenizer.
Token 'š' je uspešno dodát u tokenizer.
Token 'đ' je uspešno dodát u tokenizer.
Token 'ž' je uspešno dodát u tokenizer.
Token 'ć' je uspešno dodát u tokenizer.
Token 'č' je uspešno dodát u tokenizer.
Token 'š' je uspešno dodát u tokenizer.
Token 'đ' je uspešno dodát u tokenizer.
Token 'ž' je uspešno dodát u tokenizer.
Special tokens: {'bos_token': '<|endoftext|>', 'eos_token': '<|endoftext|>', 'unk_token': '<|endoftext|>', 'pad_token': '[PAD]', 'mask_token': '[MASK]'}
Pitanje: Da li je GPT model obučén na velikom broju članaka?
Odgovor: Da, GPT model Generative Pretrained Transformer modeli su obučeni za zadatke generisanja prirodnog jezika kao što su klasifikacija teksta, govora, prevod jezika. Ovi podaci se mogu koristiti za obuku modela dubokog učenja, kao što je klasifikacija teksta. Međutim, važno je napomenuti da algoritmi dubokog učenja mogu da uče i klasifikuju podatke iz velikih količina podataka. Postoji nekoliko načina na koje se GPT model može primeniti na nekoliko načina. Neke od ovih tehnika uključuju 1. Prikupljanje podataka i čišćenje podataka. Postoji nekoliko metoda koje se mogu
(transformers-env) tea@edukom:/var/www/html/ainabavka/proces_treninga/provera_modela_tokenizera$ python3.11 test_modela_tokenizera.py
Model je uspešno učitán.
Tokenizer je uspešno učitán.
Token 'ć' je uspešno dodát u tokenizer.
Token 'č' je uspešno dodát u tokenizer.
Token 'š' je uspešno dodát u tokenizer.
Token 'đ' je uspešno dodát u tokenizer.
Token 'ž' je uspešno dodát u tokenizer.
Token 'ć' je uspešno dodát u tokenizer.
Token 'č' je uspešno dodát u tokenizer.
Token 'š' je uspešno dodát u tokenizer.
Token 'đ' je uspešno dodát u tokenizer.
Token 'ž' je uspešno dodát u tokenizer.
Special tokens: {'bos_token': '<|endoftext|>', 'eos_token': '<|endoftext|>', 'unk_token': '<|endoftext|>', 'pad_token': '[PAD]', 'mask_token': '[MASK]'}
Pitanje: U koje svrhe se može koristiti AI?
Odgovor: U koje svrhe se može koristiti AI? Veštačka inteligencija AI ima potencijal da revolucionise način na koji ljudi uče, omogućavajući ljudima da komuniciraju sa svetom oko nas. Međutim, važno je napomenuti da sistemi veštačke inteligencije mogu da uče i analiziraju ogromne količine podataka, kao što su rasuđivanje, rešavanje problema, donošenje odluka. AI se takođe može koristiti za donošenje informisanih odluka na radnom mestu, oslobađajući vreme zaposlenima da se fokusiraju na druge stvari, unaprede svoje potrebe i želje. AI takođe može pomoći da se poboljša donošenje odluka, smanji opterećenje veštačke inteligencije, ubrza proračune i podstakne donošenje odluka
(transformers-env) tea@edukom:/var/www/html/ainabavka/proces_treninga/provera_modela_tokenizera$ █
```

Slika 59. Kvalitet generisanja teksta Serbian-GPT-2 modela

Izlaz terminala, *slika 59.*, pokazuje uspešnu primenu tokenizacije i generisanja teksta na srpskom jeziku. Tokeni su uspešno dodati tokenizeru, čime je potvrđena tačnost modela u obradi srpskog jezika. Pored toga, generisani tekstovi su koherentni i smisleni, što dodatno potvrđuje kvalitet modela.

Hiperparametarska optimizacija je pokazala da se mogu postići značajna poboljšanja performansi modela. Upotreba DYNAMO okvira omogućila je efikasno istraživanje hiperparametarskog prostora i pronalaženje najboljih vrednosti za obuku modela, što je rezultiralo boljim performansama u generisanju teksta i zadacima razumevanja [55, 121].

### 14.3.2. Primena i uticaj na realne aplikacije

Poboljšane performanse GPT-2 modela za srpski jezik otvaraju mogućnosti za njegovu primenu u različitim industrijama, uključujući automatizovanu korisničku podršku, generisanje marketinških sadržaja, analizu tekstualnih podataka u industriji, kreiranje personalizovanog obrazovnog materijala i unapređenje telemedicine u zdravstvu [ 12].

Buduće istraživanje se može fokusirati na dalje poboljšanje modela kroz fino podešavanje hiperparametara i proširenje skupa podataka, te primenu modela na druge jezike s niskim resursima [129]. Integracijom modela s drugim tehnologijama veštačke inteligencije, kao što je obrada govora i vizualnih podataka, mogu se stvoriti multimodalni sistemi sposobni za složeniju analizu i generisanje sadržaja [28].

Primena naprednih tehnika transfernog učenja i optimizacije hiperparametara značajno poboljšava performanse GPT-2 modela za jezike sa niskim resursima, kao što je srpski. Ovi rezultati pružaju solidnu osnovu za dalje istraživanje i primenu ovih tehnika na druge jezike sa sličnim karakteristikama [122].

Rezultati istraživanja imaju značajan uticaj na oblast obrade prirodnog jezika, posebno za jezike sa niskim resursima. Poboljšane performanse modela omogućavaju primenu naprednih NLP tehnologija u realnim aplikacijama, povećavajući dostupnost i inkluzivnost ovih tehnologija [120, 55]. Primena u različitim industrijama i poljima, kao što su obrazovanje, telemedicina i automatizovana korisnička podrška, pokazuje praktičnu vrednost ovog istraživanja [86].

#### **14.4. Unapređivanje višejezičkih NLP sistema**

Ovo poglavlje je pružilo detaljan uvid u efikasnost primene naprednih tehnika transfernog učenja na GPT-2 modelu za srpski jezik [121, 93]. Kroz preciznu prethodnu obradu podataka, efikasnu tokenizaciju i preciznu optimizaciju hiperparametara koristeći DYNAMO okvir, postignuta su značajna poboljšanja performansi modela u zadacima razumevanja i generisanja teksta [120, 55]. Eksperimentalni rezultati su pokazali da optimizirani model postiže visoku tačnost, smanjuje gubitke i perpleksiju, te generiše koherentan i smislen tekst [27, 86].

Nalazi ovog istraživanja imaju značajne implikacije za oblast obrade prirodnog jezika (NLP), posebno u kontekstu višejezičkih aplikacija i jezika sa manjim resursima kao što je srpski [28, 123]. Poboljšane performanse GPT-2 modela omogućavaju njegovu primenu u različitim industrijama kao što su automatizacija korisničke podrške, generisanje marketinškog sadržaja, analiza tekstualnih podataka, personalizovano obrazovanje i poboljšanje telemedicinske [122, 127]. Ovi nalazi pokazuju da je moguće značajno poboljšati modele za jezike sa manjim resursima, doprinoseći inkluzivnosti i dostupnosti naprednih NLP tehnologija [125].

Iako je istraživanje pokazalo značajne rezultate, postoji nekoliko ograničenja koja se moraju priznati. Prvo, model je obučen na specifičnom skupu podataka prevedenom na srpski jezik i dopunjen novim tekstima, što može ograničiti generalizaciju rezultata na druge jezike ili domene [55]. Drugo, proces optimizacije hiperparametara je zahtevao značajne računске resurse, što može biti izazov za istraživače s ograničenim pristupom naprednoj hardverskoj infrastrukturi [120, 124].

Buduća studija se može fokusirati na dalju optimizaciju modela kroz fino podešavanje hiperparametara i proširenje skupa podataka [127]. Pored toga, istraživanje bi moglo uključiti primenu modela na druge jezike s niskim resursima, razvoj univerzalnih višejezičkih modela [27]. Buduća istraživanja bi takođe mogla uključivati integraciju modela s drugim tehnologijama veštačke inteligencije kao što su kompjuterski vid i obrada govora kako bi se stvorili multimodalni sistemi sposobni za složeniju analizu i generisanje sadržaja [46].

Rad na ovom istraživanju je pokazao da se značajna poboljšanja u performansama NLP modela za jezike sa niskim resursima mogu postići kroz preciznu prethodnu obradu podataka i optimizaciju hiperparametara [121, 125]. Upotreba naprednih alata kao što je DYNAMO framework baziran na alatu Optuna omogućava sistematsko istraživanje hiperparametarskog prostora i identifikaciju optimalnih vrednosti [27, 57]. Ovaj pristup značajno smanjuje potrebu za ručnim podešavanjem i omogućava efikasniju obuku modela. Osim toga, istraživanje je pokazalo da primena naprednih tehnika transfernog

učenja može postići visoku tačnost i koherentnost u generisanom tekstu, doprinoseći razvoju robusnih višejezičkih NLP sistema [55].

Ovi rezultati pružaju solidnu osnovu za dalje istraživanje i primenu ovih tehnika na drugim jezicima sa sličnim karakteristikama, doprinoseći globalnom razvoju i primeni NLP tehnologija [120, 28].

## **X REZIME SPROVEDENIH ISTRAŽIVANJA**

## 15.1. Diskusija rezultata istraživanja

U disertaciji su prikazane matematičke metode korišćene za optimizaciju i analizu modela mašinskog učenja, posebno u kontekstu obrade prirodnog jezika sa naglaskom na model GPT-2. Istraživanje je obuhvatilo nekoliko naprednih metoda optimizacije, uključujući Bayesovsku optimizaciju i metode dinamičkog prilagođavanja hiperparametara unutar DYNAMO okvira.

Grid search je primenjen kao početna metoda za sistematsko pretraživanje prostora hiperparametara kako bi se identifikovale optimalne kombinacije za konkretne modele. Međutim, ova metoda je imala svoja ograničenja, posebno u pogledu vremena i resursa, što je zahtevalo prelazak na naprednije tehnike kao što su Bayesovska optimizacija i dinamičko prilagođavanje hiperparametara. Ove naprednije metode omogućile su efikasniju procenu i selekciju hiperparametara uz minimalne resurse, čime je značajno unapređena efikasnost modela u zahtevnim uslovima.

U okviru DYNAMO metodologije, posebno su razvijeni algoritmi koji su omogućili precizno ciljanje hiperparametara kroz kombinaciju iterativne optimizacije, kontinuiranog učenja i povratnih informacija iz prethodnih iteracija. Ovaj pristup je posebno prilagođen zahtevima modela obrade prirodnog jezika, čime je značajno unapređena efikasnost modela u uslovima ograničenih resursa, kao što su jezici sa manje dostupnim podacima.

Takođe, istraživanje je pokazalo kako algoritmi za regularizaciju mogu minimizirati prekomerno prilagođavanje (overfitting). Ovo je od suštinske važnosti za modele koji se primenjuju u stvarnim okruženjima gde su podaci heterogeni i često nepredvidivi.

Na osnovu ovih rezultata, jasno je da su ove matematičke metode bile ključne za postizanje visokog nivoa tačnosti i smanjenje vrednosti gubitka (loss), što je dovelo do značajnog unapređenja performansi modela. Metode optimizacije i prilagođavanja hiperparametara pokazale su se kao neophodne za efikasno prilagođavanje modela specifičnim zahtevima i hardverskim konfiguracijama. Ovaj detaljan pregled matematičkih metoda korišćenih u istraživanju omogućava dublje razumevanje njihove primene i efekata na performanse modela, posebno u kontekstu obrade prirodnog jezika na različitim hardverskim arhitekturama. Rezultati su jasno pokazali da pravilna primena ovih metoda može značajno unaprediti performanse modela i omogućiti njihovu primenu u širokom spektru praktičnih aplikacija.

Ujedno, prikazana je mogućnost primene ovih modela u realnim okruženjima, posebno u kontekstu višezjezičkih aplikacija. Korišćenje naprednih tehnika transfernog učenja omogućilo je adaptaciju modela za jezike sa ograničenim resursima, kao što je srpski, čime je demonstrirana široka primenljivost ovih modela u različitim industrijama, uključujući edukaciju, telemedicinu i automatizovanu korisničku podršku.

Pomoću novog algoritma za optimizaciju hiperparametara zasnovanog na kombinaciji Bayesovske optimizacije i cikličnog učenja (cyclic learning rate), model je postigao optimalne performanse uz smanjenje vremena treniranja. Algoritam je omogućio modelima da se brzo prilagode novim podacima, zadržavajući visoku tačnost i konzistentnost u predikcijama. Analizirano je i pokazano kako je moguće te metode iskoristiti u praktične svrhe kroz primenu u sistemima za automatsku analizu tekstualnih podataka. Posebno su istaknuti primeri primene u domenima gde je potrebna visoka tačnost i efikasnost



u obradi prirodnog jezika, kao što su sistemi za prepoznavanje obrazaca u tekstu i personalizovani sistemi za preporuku sadržaja.

Ukazujući na potrebu uvođenja i primene automatizovane dijagnostike, istraživanje je pokazalo kako optimizovani modeli mogu značajno unaprediti procese analize velikih skupova podataka, omogućavajući brže i preciznije donošenje odluka. Automatizovana dijagnostika, podržana modelima treniranim na specifičnim domenima, pokazala se kao neophodna u industrijama gde je pravovremena analiza podataka ključna za operativne uspehe.

Analizom matematičkog modeliranja procesa, uključujući napredne tehnike optimizacije i prilagođavanja modela specifičnim zahtevima jezika i domena, došlo se do zaključaka koji ukazuju na mogućnost unapređenja trenutnih tehnologija i metoda. Kroz detaljnu analizu performansi modela, istraživanje je pružilo dubok uvid u načine na koje se ove metode mogu dodatno poboljšati i implementirati u nove aplikacije.

Na taj način se istraživanje pokazalo kao značajan doprinos razvoju alata za obradu prirodnog jezika, posebno u kontekstu višezjezičkih aplikacija, gde je demonstrirano kako napredne optimizacione tehnike mogu unaprediti tačnost, brzinu i efikasnost modela. Istraživanja u ovoj disertaciji su rezultirala preporuku za primenu prilagođenih algoritama za optimizaciju kao ključni parametar dijagnostike stanja sistema, omogućavajući preciznije i efikasnije praćenje i analizu sistema u realnom vremenu.

Ovaj rezime prikazuje ključne doprinose istraživanja u kontekstu primene matematičkih metoda, optimizacije i praktične primene rezultata u stvarnim okruženjima, naglašavajući značaj ovih tehnika za dalji razvoj naprednih sistema za obradu prirodnog jezika.

## **15.2. Ključni rezultati istraživanja**

Na osnovu sveobuhvatnog pregleda relevantne literature, kao i sprovedenih teorijskih, laboratorijskih i eksperimentalnih istraživanja, potvrđena je efikasnost predloženih metodologija i matematičkih modela u optimizaciji i prilagođavanju modela obrade prirodnog jezika. Rezultati ovih istraživanja su značajni ne samo u pogledu teorijskih doprinosa, već i u pogledu praktične primene u industriji, obrazovanju i zdravlju.

Primena naprednih tehnika optimizacije hiperparametara, kao što su Bayesovska optimizacija, u kombinaciji sa dinamičkim prilagođavanjem unutar DYNAMO okvira, omogućila je značajno smanjenje vremena treniranja, povećanje tačnosti modela, i smanjenje vrednosti funkcije gubitka (loss). Ova kombinacija metoda omogućila je dublje istraživanje prostora hiperparametara i izbor optimalnih vrednosti koje su doprinele značajnom poboljšanju performansi modela, čime su prevaziđene granice tradicionalnih pristupa.

Primenom naprednih regularizacionih tehnika kao što su dropout i weight decay, u kombinaciji sa analizom matrica za evaluaciju modela, pokazano je da ove metode efikasno sprečavaju prekomerno prilagođavanje (overfitting) modela. Posebna pažnja je posvećena stabilnosti modela u heterogenim i dinamičnim okruženjima, što je ključno za primenu u stvarnim sistemima gde su performanse modela od vitalnog značaja.

Analiza i modeliranje procesa optimizacije na različitim hardverskim arhitekturama pokazali su mogućnost značajnog smanjenja opterećenja GPU-a tokom treniranja modela. Ovo istraživanje je pokazalo kako pravilno matematičko modeliranje može doprineti boljoj alokaciji resursa, efikasnijem korišćenju dostupne hardverske infrastrukture i smanjenju potrošnje energije, što je od velike važnosti za skalabilnost modela na većim skupovima podataka.

Optimizacija i prilagođavanje modela odvijala se kroz nekoliko ključnih faza, uključujući rešavanje problema skalabilnosti i upravljanja memorijom. Istraživanja su pokazala da pravilna primena ovih metoda može značajno povećati fleksibilnost modela, što je posebno važno u kontekstu višezjezičkih sistema i aplikacija sa ograničenim resursima.

Preporuke za dalju primenu prilagođenih algoritama za optimizaciju naglašavaju važnost precizne dijagnostike stanja sistema. Ova preporuka je zasnovana na analizama koje su pokazale da pravilna optimizacija hiperparametara može značajno doprineti preciznijem i efikasnijem praćenju i analizi sistema u realnom vremenu, čime se obezbeđuje stabilnost i visok nivo performansi u različitim industrijama.

Ključni rezultati istraživanja jasno ukazuju na to da su korišćene metode i tehnike dale značajne doprinose u poboljšanju performansi modela obrade prirodnog jezika. Kombinacija naprednih tehnika optimizacije, matematičkog modeliranja i analize rezultata pruža čvrstu osnovu za dalji razvoj i primenu u različitim industrijama i aplikacijama, čime se otvara put ka daljem unapređenju tehnologija obrade prirodnog jezika.

### **15.3. Predlog daljih istraživanja**

S obzirom na visoku tačnost modela u ovoj disertaciji, budući radovi i istraživanja trebaju se fokusirati na dalji razvoj i unapređenje metoda i tehnika koje omogućavaju poboljšanje performansi modela obrade prirodnog jezika. Permanentno istraživanje novih, efikasnijih algoritama i modela je ključno za održavanje konkurentnosti u ovom dinamičnom polju. U skladu sa tim, predložena buduća istraživanja treba da obuhvate sve relevantne činioce koji mogu uticati na preciznost, efikasnost i skalabilnost modela, uzimajući u obzir sledeće aspekte:

- Postojeće metode optimizacije, kao što su grid search u smislu početne metode pretraživanja hiperparametara, Bayesovska optimizacija i DYNAMO okvir, pokazale su se efikasnim, ali buduća istraživanja trebaju se usmeriti ka razvoju naprednijih i adaptivnih algoritama koji mogu efikasnije pretraživati prostor hiperparametara. Korišćenje tehnika kao što su reinforcement learning i generative adversarial networks (GANs) moglo bi omogućiti dinamičnije i preciznije podešavanje modela u realnom vremenu, čime bi se povećala efikasnost u stvarnim uslovima.
- Buduća istraživanja bi trebalo da obuhvate raznovrsnije i kompleksnije dataset-ove kako bi se modelima omogućilo bolje generalizovanje na različitim jezicima i kontekstima. Poseban fokus treba staviti na višezjezičke dataset-ove, gde je potrebno istražiti kako modeli mogu efikasnije integrisati informacije iz različitih jezičkih izvora. Takođe, istraživanje kako različiti konteksti (npr. stručni jezici, dijalekti) utiču na performanse modela može doprineti razvoju robusnijih i univerzalnijih modela.
- Razvoj savremenijih laboratorijsko-eksperimentalnih okruženja sa sofisticiranijom kontrolom parametara je od suštinskog značaja za preciznija i pouzdanija istraživanja. Budući eksperimenti

trebaju obuhvatiti složenije režime provere modela, uključujući simulacije u realnim uslovima i "stres" testove koji će pomoći u boljem razumevanju granica trenutnih tehnologija. Korišćenje virtuelne i proširene realnosti u laboratorijskim uslovima moglo bi doprineti realističnijem testiranju modela.

- Da bi se postigli bolji rezultati u praćenju i analizi performansi modela, potrebno je razviti nove metode instrumentacije koje će omogućiti precizniji nadzor ključnih parametara tokom treniranja i primene modela. Ove metode treba da budu podržane naprednim softverskim alatima koji omogućavaju automatsku analizu i vizualizaciju performansi modela, pružajući istraživačima brze i tačne povratne informacije.
- Uzimajući u obzir sve veću primenu modela obrade prirodnog jezika u kritičnim industrijama kao što su zdravstvo i obrazovanje, buduća istraživanja treba da se fokusiraju na unapređenje sigurnosnih mera i etičkih standarda. Potrebno je razviti nove modele i algoritme koji će biti otporniji na različite vrste napada, kao i metode koje će osigurati zaštitu privatnosti i etičku primenu modela u stvarnim okruženjima. U cilju obezbeđivanja dugoročne održivosti i sigurnosti u primeni ovih modela, potrebno je intenzivirati edukaciju i podizanje svesti kod studenata i profesionalaca. Programi obuke treba da obuhvate ne samo tehničke aspekte razvoja i primene modela, već i etičke implikacije i važnost sigurnosnih mera. Ova edukacija treba da bude deo standardnih kurikuluma u oblasti veštačke inteligencije i mašinskog učenja.
- Buduća istraživanja trebaju se fokusirati na razvoj potpuno automatizovanih modela koji će biti sposobni za samostalno prilagođavanje različitim okruženjima i aplikacijama. Ovi modeli treba da uključuju napredne algoritme za kontinuirano učenje i prilagođavanje, čime bi se obezbedila dugoročna efikasnost i adaptivnost u dinamičnim uslovima.
- Potrebno je razviti naprednu softversku podršku koja će omogućiti automatizaciju svih faza razvoja i primene modela, od treniranja do implementacije i održavanja. Ova podrška treba da obuhvati alate za automatsko prilagođavanje hiperparametara, nadzor performansi, i optimizaciju resursa, čime bi se omogućilo efikasnije upravljanje i skaliranje modela u različitim industrijama.

Predlozi za dalja istraživanja jasno ukazuju na potrebu za stalnim unapređenjem i prilagođavanjem modela obrade prirodnog jezika. Razvoj novih metoda, unapređenje eksperimentalnih uslova, fokus na edukaciju i sigurnost, obezbeđuju čvrstu osnovu za nastavak istraživanja u ovoj oblasti, čime se otvara prostor za inovacije i primenu najnovijih tehnoloških dostignuća u stvarnim scenarijima.

## 16. ZAKLJUČAK

### 16.1. Naučni doprinos disertacije

Doprinos istraživanja u okviru doktorske disertacije prvenstveno se ogleda u sprovedenoj analizi, razvoju i implementaciji naprednih tehnika transfernog učenja koje su značajno unapredile mogućnosti modela obrade prirodnog jezika za višejezičke aplikacije. Kroz ovu disertaciju, izvedeni su ključni doprinosi u nekoliko oblasti:

#### 1. Razvoj modela Serbian GPT-2:

Razvijen je prilagođeni model Serbian GPT-2, specifično treniran na obimnom srpskom datasetu od 750 miliona tokena. Ovaj model predstavlja napredak u obradi prirodnog jezika na srpskom jeziku, pružajući visoku tačnost i konzistentnost u generisanju i analizi tekstualnih podataka. Tokom razvoja modela, posebna pažnja je posvećena specifičnostima srpskog jezika, kao što su morfologija i sintaksa, što je omogućilo postizanje rezultata koji su superiorni u odnosu na postojeća rešenja za jezike sa ograničenim resursima. Evaluacija performansi modela izvršena je kroz detaljne metrike kao što su gubitak (loss), tačnost (accuracy) i perpleksija (perplexity), što je omogućilo precizno merenje unapređenja u različitim jezičkim zadacima.

#### 2. Formalizacija procedure za transferno učenje:

Disertacija je formalizovala sveobuhvatnu proceduru za adaptaciju GPT-2 modela na jezike sa ograničenim resursima koristeći PyTorch i Transformers arhitekture. U okviru ove procedure, razvijeni su algoritmi za efikasno upravljanje memorijom i optimizaciju hiperparametara, koji su demonstrirali značajna poboljšanja performansi modela u heterogenim i resursno ograničenim okruženjima. Ovi algoritmi omogućavaju da se model efikasno prilagodi specifičnim zahtevima jezičkih aplikacija kroz iterativno učenje i kontinuiranu optimizaciju, čime je unapređena skalabilnost modela.

#### 3. Razvoj sopstvenog modela za dinamičku optimizaciju hiperparametara:

Uveden je novi model za dinamičko prilagođavanje hiperparametara zasnovan na DYNAMO okviru, koji kombinuje iterativnu optimizaciju i kontinuirano učenje. Ovaj model je omogućio značajno poboljšanje performansi modela u realnom vremenu, što bi posebno moglo značiti u kontekstu primene u industrijama sa ograničenim resursima kao što su obrazovanje, telemedicina i automatizovana korisnička podrška. DYNAMO okvir se pokazao superiornim u odnosu na tradicionalne metode kao što su grid search i random search, nudeći preciznije i brže podešavanje hiperparametara sa značajnim uštedama u vremenu i resursima. Analize su pokazale da DYNAMO okvir smanjuje gubitak (loss) za 36,077% u poređenju sa tradicionalnim pristupima, što je ključna prednost u optimizaciji performansi modela.

Ovi doprinosi predstavljaju značajan napredak u oblasti veštačke inteligencije, posebno u kontekstu višejezičkih sistema za obradu prirodnog jezika. Disertacija ne samo da proširuje granice postojećih istraživanja u ovoj oblasti, već i postavlja temelje za buduća istraživanja i praktične primene u industriji, obrazovanju i zdravstvu. Pored toga, doprinosi ove disertacije otvaraju nove pravce istraživanja,

omogućavajući dalji razvoj i primenu naprednih tehnika obrade prirodnog jezika u realnim scenarijima širom sveta.

## 16.2. Pitanja koja disertacija otvara, nastavak i buduća istraživanja

Sprovedena istraživanja u ovoj disertaciji nesumnjivo su pokazala da je modeliranje sistema obrade prirodnog jezika korišćenjem naprednih algoritama i matičnih transformacija od suštinskog značaja za unapređenje performansi u ovoj oblasti i drugim srodnim disciplinama. Međutim, kroz ovaj rad otvorena su i određena pitanja koja zahtevaju dodatna istraživanja i analize. Neka od tih pitanja uključuju:

- **Pogodnost metoda matičnih transformacija za različite nivoe složenosti** - S obzirom na visoku tačnost definisanih modela, postavlja se pitanje do koje mere su matične transformacije pogodne za razvoj, programiranje i testiranje kako jednostavnih, tako i složenih modela obrade prirodnog jezika. Dalja istraživanja bi trebala da se fokusiraju na prilagodljivost i skalabilnost ovih metoda u različitim jezičkim aplikacijama.
- **Uticaj varijabilnosti uslova rada na tačnost modela** - Kod budućih istraživanja, potrebno je detaljno proučiti u kojoj meri varijabilnost ulaznih podataka i radnih uslova utiče na tačnost modela, posebno u složenim okruženjima. Da li je moguće unaprediti modele tako da bolje odgovaraju dinamičnim promenama u ulaznim podacima i kako te promene utiču na performanse modela u realnom vremenu?
- **Verifikacija primene rezultata u praktičnim aplikacijama** - Daljim razvojem i istraživanjem u ovoj oblasti, potrebno je verifikovati primenu dobijenih rezultata u praktičnim aplikacijama, kao što su sistemi za automatsku analizu teksta i generisanje sadržaja. Koliko su ovi modeli efikasni u stvarnim uslovima i koliko pomažu u optimizaciji procesa obrade prirodnog jezika?
- **Razvoj slične metodologije za druge jezike i domene** - Pored razvijenih metoda za srpski jezik, postavlja se pitanje u kojoj meri je potrebno razviti i verifikovati sličnu metodologiju za druge jezike i domene, posebno u kontekstu višejezičkih aplikacija. Kako bi se slične tehnike mogle primeniti na druge jezike i specifične industrijske sektore?
- **Nova organizacija procesa modeliranja** - Dalja istraživanja ukazuju na potrebu za novom organizacijom procesa modeliranja, koja bi obuhvatila integraciju savremenih algoritama veštačke inteligencije i automatizovanih sistema u proces obrade prirodnog jezika. Kako bi ova reorganizacija mogla doprineti efikasnijem i preciznijem modeliranju u stvarnim uslovima?
- **Savremeni postupci evaluacije modela** - U budućnosti, treba razviti savremene postupke za evaluaciju i validaciju modela obrade prirodnog jezika, koji će omogućiti bržu i precizniju analizu performansi modela. Koje metode evaluacije su najpogodnije za integraciju sa automatizovanim sistemima za generisanje i analizu teksta?
- **Analiza sigurnosti i pouzdanosti modela** - Dalja istraživanja ukazuju na potrebu detaljne analize sigurnosti i pouzdanosti modela, koristeći metodologiju zasnovanu na naprednim algoritmima. Kako osigurati da modeli ostanu stabilni i pouzdani u različitim uslovima primene?
- **Razvoj novog automatizovanog sistema za evaluaciju** - Potrebno je razviti novi automatizovani sistem za evaluaciju modela koji bi bio sposoban za stalno praćenje i kontrolu performansi modela u realnim uslovima. Koliko bi takav sistem mogao unaprediti kontinuiranu evaluaciju i smanjiti rizik od grešaka ili nepredviđenih odstupanja u radu modela?
- **Primena veštačke inteligencije u automatizovanoj obradi teksta** - U kojoj meri će buduću razvoj softverskih rešenja biti baziran na primeni metoda veštačke inteligencije prilikom

automatizovane obrade i analize teksta? Kako će se ove tehnologije integrisati sa postojećim sistemima i koje izazove će doneti u pogledu etike i sigurnosti?

- **Smena tradicionalnih i savremenih pristupa** - Kako će se smenjivati tradicionalni i savremeni pristupi u modeliranju i obradi prirodnog jezika? Koje prednosti i nedostatke donosi integracija savremenih metoda sa tradicionalnim pristupima u različitim aplikacijama i industrijama?

### 16.3. Dokazivanje (potkrepljivost) glavne hipoteze

Naučno istraživanje u okviru doktorske disertacije imalo je za cilj da pokaže da paradigma GPT-2 modela, uz pomoć transfernog učenja i adaptacije u okvirima PyTorch i Transformers biblioteka, otvara put ka stvaranju efikasnijih višejezičkih modela obrade prirodnog jezika. Glavna hipoteza ove disertacije jeste da primena ovih metoda može značajno unaprediti tačnost, skalabilnost i efikasnost modela obrade prirodnog jezika, posebno u kontekstu jezika sa ograničenim resursima.

U skladu sa ovim ciljem, istraživanje je sprovedeno kroz niz teorijskih i eksperimentalnih analiza, koje su usmerene na dokazivanje sledećih pomoćnih hipoteza:

*Hipoteza 1:* Primena GPT-2 modela sa adaptiranim transfernim učenjem na različitim hardverskim platformama dovodi do značajnog poboljšanja u generisanju prirodnog teksta na manje zastupljenim jezicima tj. srpskom jeziku.

- Ova hipoteza je potvrđena kroz analizu performansi modela na različitim GPU konfiguracijama. BLEU skor je korišćen kao ključna metrika za merenje tačnosti generisanog teksta, pri čemu su eksperimenti pokazali značajne napretke u kvalitetu generisanog teksta na srpskom jeziku. Posebno su analizirane konfiguracije GPU-a sa različitim kapacitetima memorije i procesorske snage, gde je uočeno da kvalitet tj. snaga hardvera direktno utiče na efikasnost transfernog učenja. Korišćenjem prilagođenih arhitektura i optimizovanih podešavanja GPU-a, model je uspeo da ostvari visoku tačnost čak i na platformama sa ograničenim resursima .

*Hipoteza 2:* Modifikacije u arhitekturi PyTorch i Transformers biblioteka, uključujući optimizaciju scheduler-a i memorije, dovode do efikasnijeg procesa učenja GPT-2 modela.

- Ova hipoteza je dokazivana kroz detaljnu eksperimentalnu analizu koja je pokazala da optimizovane arhitekture značajno smanjuju vreme treniranja i poboljšavaju performanse modela. Korišćenjem AdamW optimizatora, model je postigao stabilniji proces učenja, smanjujući rizik od prekomernog prilagođavanja (overfitting). Ciklični learning rate scheduler je dodatno unapredio efikasnost resursa, omogućavajući modelu da se dinamično prilagođava tokom faza treniranja, što je rezultiralo boljim performansama modela u različitim aplikacijama.

*Hipoteza 3:* Integracija prilagođenih metrika kao što su tačnost i perpleksija u proces evaluacije GPT-2 modela unapređuje njihovu sposobnost za višejezičku obradu.

- Ova hipoteza je potvrđena merenjem poboljšanja u tačnosti i perpleksiji modela prilikom generisanja tekstova na srpskom jeziku. Upotreba prilagođenih metrika omogućila je detaljniju analizu i fino podešavanje modela, što je rezultiralo poboljšanom preciznošću i koherentnošću generisanog teksta.

*Hipoteza 4:* Optimizacija hiperparametara i prilagođavanje procesa tokenizacije povećavaju efikasnost GPT-2 modela u generisanju i obradi teksta na različitim jezicima.

- Eksperimentalni podaci su pokazali da optimizacija hiperparametara, kao što su learning rate i batch size, zajedno sa prilagođenom tokenizacijom, značajno unapređuju sposobnost modela da generiše koherentan i stilistički dosledan tekst na srpskom jeziku. Takođe, ova prilagođavanja su omogućila efikasniju upotrebu GPU memorije, čime je smanjeno vreme potrebno za obradu velikih dataset-a.

*Hipoteza 5:* Optimizacija tokenizacije i upravljanje memorijom doprinose efikasnosti GPT-2 modela u obradi velikih skupova podataka, čak i uz ograničene resurse GPU memorije.

- Ova hipoteza je dokazana kroz eksperimente koji su pokazali da prilagođena tokenizacija i strategije upravljanja memorijom omogućavaju efikasniju obradu podataka i smanjuju opterećenje na GPU, što je posebno značajno u situacijama s ograničenim hardverskim resursima.

*Hipoteza 6:* Kvalitet i raznovrsnost skupova podataka ključni su za efikasnost i tačnost GPT-2 modela u višejezičkim aplikacijama.

- Dokazi za ovu hipotezu su dobijeni kroz analizu različitih aspekata dataset-a, uključujući veličinu, raznovrsnost i kvalitet. Pokazano je da balansiranje dataset-a prema jezičkoj grupi značajno poboljšava sposobnost modela da precizno i efikasno generiše tekst na srpskom jeziku.

*Hipoteza 7:* Primena GPT-2 modela sa naprednim tehnikama transfernog učenja u realnim višejezičkim aplikacijama doprinosi poboljšanju međukulturne komunikacije i pristupačnosti informacija.

- Ova hipoteza je potvrđena kroz evaluaciju uticaja GPT-2 modela na komunikaciju u višejezičkim digitalnim platformama. Analize su pokazale da model može unaprediti međukulturnu komunikaciju i omogućiti širi pristup informacijama, što ima značajan društveni uticaj.

Na osnovu sprovedenih teorijskih i eksperimentalnih istraživanja, glavna hipoteza ove disertacije je potvrđena. Paradigma GPT-2 modela, uz pomoć transfernog učenja i adaptacije unutar PyTorch i Transformers biblioteka, pokazala se kao ključna za razvoj efikasnih višejezičkih modela obrade prirodnog jezika. Time su otvoreni novi putevi za unapređenje i primenu ovih modela u različitim industrijama, od obrazovanja do automatizovane korisničke podrške, čime se postavljaju temelji za buduća istraživanja i inovacije u ovoj oblasti.

## STRUKTURA DOKTORSKE DISERTACIJE

# GPT-2 PARADIGMA: TRANSFERNO UČENJE I ADAPTACIJA PYTORCH I TRANSFORMERS ARHITEKTURA NA PUTU KA VIŠEJEZIČKOM SVETU

Rezime  
Abstract

UVOD

## I. METODOLOŠKI OKVIR ISTRAŽIVANJA

- 1. Formulacija problema istraživanja**
  - 1.1. Osnovna istraživačka ideja
  - 1.2. Značaj problema istraživanja
    - 1.2.1. Naučni značaj
    - 1.2.2. Društveni (praktični) značaj istraživanja
  - 1.3. Rezultati prethodnih saznanja o problemu istraživanja
- 2. Određivanje predmeta istraživanja**
  - 2.1. Teorijsko određenje predmeta istraživanja
  - 2.2. Pojmovno-kategorijalni sistem
  - 2.3. Operacionalno određenje predmeta istraživanja
    - 2.3.1. Činioce sadržaja predmeta istraživanja koji će biti neposredno istraženi
    - 2.3.2. Vremensko, prostorno i disciplinarno određenje predmeta istraživanja
- 3. Ciljevi istraživanja**
  - 3.1. Naučni ciljevi istraživanja
  - 3.2. Društveni cilj istraživanja
- 4. Hipotetički okvir istraživanja**
  - 4.1. Generalna (opšta) hipoteza
  - 4.2. Posebne hipoteze istraživanja
- 5. Način (metode) istraživanja**
  - 5.1. Osnovne metode saznanja i mišljenja
  - 5.2. Opštenaučne metode istraživanja
  - 5.3. Metode sa prikupljanje podataka
  - 5.4. Uzorak istraživanja
- 6. Društvena i naučna opravdanost istraživanja**
  - 6.1. Naučni doprinos
  - 6.2. Društveni doprinos

## II. DOPRINOS MAŠINSKOG UČENJA U SAVREMENIM AI SISTEMIMA

- 7. Osnove i značaj mašinskog učenja**
  - 7.1. Modeli mašinskog učenja u analizi sentimenta**
    - 7.1.1. Metodologija i tehnike u analizi sentimenta
    - 7.1.2. Značaj i uticaj istraživanja u NLP tehnologijama
  - 7.2. Strategije i metode mašinskog učenja za analizu sentimenta**



- 7.2.1. Prikupljanje i predobrada podataka
- 7.2.2. Tehnike treniranja modela
- 7.2.3. Pristup mašinskom učenju

### **7.3. Analitički uvidi i dometi modela mašinskog učenja**

- 7.3.1. Balans između preciznosti, odziva i F1-scora
- 7.3.2. Vizualna ilustracija ključnih metrika performansi
- 7.3.3. Analiza podrške i distribucije podataka
- 7.3.4. Robusnost modela i njegova primena

### **7.4. Razmatranje efikasnosti i etičkih implikacija modela**

- 7.4.1. Uticaj metodoloških izbora na performanse
- 7.4.2. Etičke dimenzije i transparentnost u primeni modela
- 7.4.3. Mogućnosti unapređenja i budući pravci istraživanja
- 7.4.4. Etički i društveni izazovi u primeni AI tehnologija

## **III. PRIMENA GPT-2 MODELA SA ADAPTIRANIM TRANSFERNIM UČENJEM NA RAZLIČITIM HARDVERSKIM PLATFORMAMA**

### **8. Optimizacija GPT-2 modela putem adaptivnog transfernog učenja na GPU arhitekturama**

#### **8.1. Kvantifikacija uticaja GPU performansi na GPT-2 model u adaptivnom transfernom učenju**

- 8.1.1. Eksperimentalna analiza efikasnosti GPU
- 8.1.2. Primena i održivost modela u praksi

#### **8.2. Metodološki pristup u evaluaciji efikasnosti GPT-2 Modela na GPU platformama**

- 8.2.1. Planiranje eksperimentalnog dizajna
- 8.2.2. Implementacija i analiza metrika performansi
- 8.2.3. Primena alata za optimizaciju i vizualizaciju

#### **8.3. Detaljna evaluacija efekata GPU arhitektura na performanse i efikasnost GPT-2 modela**

- 8.3.1. Uticaj GPU specifikacija na efikasnost modela
- 8.3.2. Optimizacija softvera i njena uloga u poboljšanju performansi
- 8.3.3. Strategije za maksimizaciju efikasnosti GPT-2 modela

#### **8.4. Analiza i implikacije optimizacija GPT-2 modela na GPU arhitekturama**

## **IV. MODIFIKACIJE U ARHITEKTURI PYTORCH I TRANSFORMERS BIBLIOTEKA, UKLJUČUJUĆI OPTIMIZACIJU SCHEDULER-A I MEMORIJE**

### **9. Optimizacija scheduler-a i upravljanje memorijom u PyTorch i Transformers bibliotekama: pristupi i modifikacije**

#### **9.1. PyTorch i Transformers kao studija slučaja**

#### **9.2. Metodološki pristupi za optimizaciju scheduler-a i upravljanje memorijom u PyTorch i Transformers bibliotekama**

- 9.2.1. Eksperimentalna konfiguracija i izbor hardvera
- 9.2.2. Sinergija teorije i praktične primene

#### **9.3. Empirijska evaluacija optimizacija scheduler-a i upravljanja memorijom u PyTorch i Transformers bibliotekama**

- 9.3.1. Analiza performansi i efikasnosti algoritama
- 9.3.2. Implikacije za industrijsku i akademsku primenu

#### **9.4. Značaj optimizacija scheduler-a i upravljanja memorijom u PyTorch i Transformers bibliotekama**

### **V. TOKENIZACIJA I OPTIMIZACIJA MEMORIJE ZA EFIKASNIJU OBRADU PRIRODNOG JEZIKA NA GPU-U**

#### **10. Tokenizacija i optimizacija memorije za smanjenje opterećenja GPU-a u NLP modelima dubokog učenja**

##### **10.1. Novi horizonti u upravljanju memorijom**

- 10.1.1. Razvoj strategija za optimizaciju memorije
- 10.1.2. Inovativne metode tokenizacije i njihov uticaj na performanse

##### **10.2. Integracija mašinskog i dubokog učenja**

- 10.2.1. Integracija tehnika mašinskog učenja za predviđanje potreba za memorijom
- 10.2.2. Komparativna analiza metoda upravljanja memorijom
- 10.2.3. Praktična implementacija i optimizacija modela

##### **10.3. Prevazilaženje izazova memorije za napredne NLP modele**

- 10.3.1. Efikasnost selektivne validacije
- 10.3.2. Ograničenja i izazovi selektivnog uzorkovanja
- 10.3.3. Implementacija i eksperimentalni rezultati

##### **10.4. Strategije optimizacije memorije za GPU u NLP**

### **VI. OPTIMIZACIJA HIPERPARAMETARA I PRILAGODAVANJE PROCESA TOKENIZACIJE**

#### **11. Sinergija inovativnih tehnika za hiperparametarsku optimizaciju koja transformiše granice performansi u dubokom učenju**

##### **11.1. Evolucija dubokog učenja i optimizacije hiperparametara**

- 11.1.1. Napredak u tehnikama optimizacije
- 11.1.2. Uvod u DYNAMO okvir i njegove implikacije na NLP
- 11.1.3. Rešavanje NLP izazova uz DYNAMO

##### **11.2. Napredna optimizacija hiperparametara: DYNAMO pristup**

- 11.2.1. Istraživanje naprednih tehnika optimizacije
- 11.2.2. DYNAMO metodološki napredak
- 11.2.3. Studije slučaja i praktične primene

##### **11.3. Metodološki uvidi u DYNAMO: Teorija i praksa balansiranja**

- 11.3.1. Teorijske osnove i algoritamski dizajn
- 11.3.2. Integracija transfernog učenja i adaptivnih strategija
- 11.3.3. Empirijska validacija i analiza učinka
- 11.3.4. Praktične primene i studije slučaja

##### **11.4. Uticaj DYNAMO: Procena performanse i efikasnosti**

- 11.4.1. Komparativna analiza i poboljšanje modela
- 11.4.2. Rešavanje izazova i ograničenja
- 11.4.3. Metodološki uvidi i budući pravci

##### **11.5. Komparativna Analiza: DYNAMO okvir u odnosu na tradicionalne metode optimizacije hiperparametara**

##### **11.6. Potvrda efikasnosti DYNAMO okvira za optimizaciju hiperparametara**

## **VII. ADAPTIVNA TOKENIZACIJA I UPRAVLJANJE MEMORIJOM ZA OPTIMALNU OBRADU VELIKIH JEZIČKIH PODATAKA**

- 12. Unapređenje tokenizacije i upravljanje memorijom u razvoju jezičkih modela**
  - 12.1. Unapređenje tokenizacije i upravljanje memorijom u obradi tekstualnih podataka na srpskom jeziku**
    - 12.1.1. Prilagođavanje tokenizacije za srpski jezik
    - 12.1.2. Optimizacija upravljanja memorijom tokom treniranja modela
  - 12.2. Optimizacija jezičkog modela GPT-2**
    - 12.2.1. Podaci i priprema dataset-a
    - 12.2.2. Optimizacija treniranja modela i upravljanje memorijom
  - 12.3. Rezultati optimizacije tokenizacije i upravljanja memorijom u obradi tekstualnih podataka**
    - 12.3.1. Pобољшanje performansi modela kroz optimizaciju tokenizacije
    - 12.3.2. Efikasnost upravljanja memorijom tokom treniranja modela
    - 12.3.3. Analiza uticaja optimizacija na sveukupnu efikasnost modela
  - 12.4. Unapređenje jezičkih modela kroz optimizaciju tokenizacije i upravljanja memorijom**

## **VIII. KVALITET I RAZNOVRSNOST SKUPOVA PODATAKA U VIŠEJEZIČKIM APLIKACIJAMA**

- 13. Uticaj kvaliteta i raznovrsnosti podataka na performanse generativnih modela u višejezičkim aplikacijama**
  - 13.1. Optimizacija generativnih modela kroz analizu kvaliteta i raznovrsnosti podataka**
    - 13.1.1. Uticaj raznovrsnosti i kvaliteta podataka na generativne modele
    - 13.1.2. Optimizacija generativnih modela za višejezičke aplikacije
  - 13.2. Metodološki pristupi u optimizaciji generativnih modela za višejezičke aplikacije**
    - 13.2.1. Prikupljanje i priprema podataka
    - 13.2.2. Optimizacija hiperparametara i treniranje modela
    - 13.2.3. Evaluacija i vizualizacija performansi modela
  - 13.3. Rezultati optimizacije i evaluacije Serbian-GPT-2 modela**
    - 13.3.1. Evaluacija modela kroz metričke parametre
    - 13.3.2. Statistička analiza i značaj optimizacije
  - 13.4. Ključni nalazi i preporuke za optimizaciju podataka i performanse modela**

## **IX. PRIMENA GPT-2 MODELA SA NAPREDNIM TEHNIKAMA TRANSFERNOG UČENJA U REALNIM VIŠEJEZIČKIM APLIKACIJAMA**

- 14. Napredne tehnike transfernog učenja u GPT-2 modelu za višejezičke aplikacije**
  - 14.1. Optimizacija hiperparametara za GPT-2 model**
    - 14.1.1. Evaluacija performansi i primena modela na srpskom jeziku
    - 14.1.2. Implementacija transfernog učenja za prilagođavanje na srpski jezik
  - 14.2. Optimizacija GPT-2 modela za višejezičku primenu**
    - 14.2.1. Priprema i obrada podataka
    - 14.2.2. Eksperimentalni dizajn i analiza rezultata
  - 14.3. Pionirska primena transfernog učenja i hiperparametarske optimizacije na srpski GPT-2 model**

14.3.1. Duboka analiza i prikaz eksperimentalnih rezultata

14.3.2. Primena i uticaj na realne aplikacije

**14.4. Unapređivanje višejezičkih NLP sistema**

**X. ZAKLJUČNA RAZMATRANJA**

**15.1. Diskusija rezultata istraživanja**

**15.2. Ključni rezultati istraživanja**

**14.3. Predlog daljih istraživanja**

**16. ZAKLJUČAK**

**16.1. Naučni doprinos disertacije**

**16.2. Pitanja koja disertacija otvara, nastavak i buduća istraživanja**

**16.3. Dokazivanje (potkrepljivost) glavne hipoteze**

**LITERATURA**

**PRILOZI**

## LITERATURA

- [1] Gao, T., Fisch, A., & Chen, D. (2021). Making Pre-trained Language Models Better Few-Shot Learners. *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)* (pp. 3816-3830). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.acl-long.295>
- [2] Zhang, Y., Sun, S., Galley, M., Chen, Y., Brockett, C., Gao, X., ... & Dolan, B. (2020). DialogPT: Large-Scale Generative Pre-training for Conversational Response Generation. *arXiv*. <https://doi.org/10.48550/arXiv.1911.00536>
- [3] Rahali, A., & Akhloufi, M. A. (2023). End-to-end transformer-based models in textual-based NLP. *AI*, 4(1), 54-110. <https://doi.org/10.3390/ai4010004>
- [4] Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A neural framework for MT evaluation. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 2685–2702). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-main.213>
- [5] Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The Long-Document Transformer. *arXiv*. <https://doi.org/10.48550/arXiv.2004.05150>
- [6] Sufi, F. (2024). Generative Pre-Trained Transformer (GPT) in Research: A Systematic Review on Data Augmentation. *Information*, 15(2), 99. <https://doi.org/10.3390/info15020099>
- [7] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Leike, J. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 1-17. <https://doi.org/10.48550/arXiv.2203.02155>
- [8] Zhang, J., Zhao, Y., Saleh, M., & Liu, P. J. (2020). PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. *arXiv*. <https://doi.org/10.48550/arXiv.1912.08777>
- [9] Kitaev, N., Kaiser, Ł., & Levskaya, A. (2020). Reformer: The Efficient Transformer. *arXiv*. <https://doi.org/10.48550/arXiv.2001.04451>
- [10] Patwardhan, N., Marrone, S., & Sansone, C. (2023). Transformers in the real world: A survey on NLP applications. *Information*, 14(4), 242. <https://doi.org/10.3390/info14040242>
- [11] Eden, C. A., Chisom, O. N., & Adeniyi, I. S. (2024). Integrating AI in education: Opportunities, challenges, and ethical considerations. *Magna Scientia Advanced Research and Reviews*, 10(02), 006–013. doi: 10.30574/msarr.2024.10.2.0039
- [12] Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: the state of the field. *International Journal of Educational Technology in Higher Education*, 20(22). doi: 10.1186/s41239-023-00392-8
- [13] Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, 16(39). doi: 10.1186/s41239-019-0171-0
- [14] Patnaik, L. M., Wang, W. (2024). AI fairness—From machine learning to federated learning. *Computer Modeling in Engineering & Sciences*, 139(2), 1203-1215. doi: 10.32604/cmescs.2023.029451
- [15] Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, 56, 3005-3054. doi: 10.1007/s10462-022-10246-w
- [16] Ouyang, F., Wu, M., Zheng, L., Zhang, L., & Jiao, P. (2023). Integration of artificial intelligence performance prediction and learning analytics to improve student learning in online engineering course. *Journal of Educational Technology in Higher Education*, 20, Article 4. doi: 10.1186/s41239-022-00372-4
- [17] Eger, S., Leiter, C., Belouadi, J., Zhang, R., Kostikova, A., Larionov, D., Chen, Y., &

- Fresen, V. (2023, June). NLLG Quarterly arXiv Report 06/23: What are the most influential current AI Papers? *Natural Language Learning Group (NLLG)*. <https://ar5iv.labs.arxiv.org/html/2308.04889>
- [18] Dogan, M. E., Dogan, T. G., & Bozkurt, A. (2023). The use of artificial intelligence (AI) in online learning and distance education processes: A systematic review of empirical studies. *Applied Sciences*, 13(5), 3056. doi: 10.3390/app13053056
- [19] Zaitso, W., & Jin, M. (2023). Distinguishing ChatGPT(-3.5, -4)-generated and human-written papers through Japanese stylometric analysis. *PLOS ONE*, 18(8), e0288453. doi: 10.1371/journal.pone.0288453
- [20] Harker, J. (2023, March). Science journals set new authorship guidelines for AI-generated text. *Environmental Factor*. <https://factor.niehs.nih.gov/2023/3/feature/2-artificial-intelligence-ethics>
- [21] Yao, M. (2023, December 5). Top 10 Influential AI Research Papers in 2023 from Google, Meta, Microsoft, and More. *Topbots*. <https://www.topbots.com/top-ai-research-papers-2023/>
- [22] Badawy, M., Ramadan, N., & Hefny, H. A. (2023). Healthcare predictive analytics using machine learning and deep learning techniques: a survey. *Journal of Electrical Systems and Information Technology*, 10, Article 40. <https://doi.org/10.1186/s43067-023-00108-y>
- [23] Schmidgall, S., Achterberg, J., Miconi, T., Kirsch, L., Ziaei, R., Hajiseydrizi, S. P., & Eshraghian, J. (2023). Brain-inspired learning in artificial neural networks: a review. *arXiv preprint*. doi: 10.48550/arXiv.2305.11252
- [24] Niskanen, T., Sipola, T., & Väänänen, O. (2023). Latest Trends in Artificial Intelligence Technology: A Scoping Review. *Jamk University of Applied Sciences*. <https://ar5iv.labs.arxiv.org/html/2305.04532>
- [25] Ahmed, Z., Mohamed, K., Zeeshan, S., & Dong, X. (2020). Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. *Database*, 2020, baaa010. doi: 10.1093/database/baaa010
- [26] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv preprint*, arXiv: 1910.10683. <https://doi.org/10.48550/arXiv.1910.10683>
- [27] Hu, H., & Yang, Y. (2021). LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint*, arXiv: 2106.09685. <https://doi.org/10.48550/arXiv.2106.09685>
- [28] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)*. <https://doi.org/10.48550/arXiv.1706.03762>
- [29] Li, L., & Liang, P. (2021). Prefix-Tuning: Optimizing Continuous Prompts for Generation. *arXiv preprint*, arXiv: 2101.00190. <https://doi.org/10.48550/arXiv.2101.00190>
- [30] Zheng, X., Zhang, C., & Woodland, P. C. (2021). Adapting GPT, GPT-2, and BERT Language Models for Speech Recognition. *arXiv preprint*, arXiv: 2108.07789. <https://doi.org/10.48550/arXiv.2108.07789>
- [31] NVIDIA. (2021). Optimizing T5 and GPT-2 for Real-Time Inference with NVIDIA TensorRT. <https://developer.nvidia.com/blog/optimizing-t5-and-gpt-2-for-real-time-inference-with-tensorrt/>
- [32] Microsoft. (2021). DeepSpeed: Accelerating large-scale model inference and training via system optimizations and compression. *Microsoft Research Blog*. <https://www.microsoft.com/en-us/research/blog/deepspeed-accelerating-large-scale-model-inference-and-training-via-system-optimizations-and-compression/>
- [33] Li, C., Zhang, M., & He, Y. (2022). The Stability-Efficiency Dilemma: Investigating Sequence Length Warmup for Training GPT Models. In *Proceedings of the Neural Information Processing Systems Conference (NeurIPS 2022)*. [https://openreview.net/forum?id=JpZ5du\\_Kdh](https://openreview.net/forum?id=JpZ5du_Kdh)
- [34] Kotei, E., & Thirunavukarasu, R. (2023). A Systematic Review of Transformer-Based Pre-Trained Language Models through Self-Supervised Learning. *Information*, 14(3), 187. <https://doi.org/10.3390/info14030187>

- [35] Rajbhandari, S., Ruwase, O., Rasley, J., Smith, S., & He, Y. (2021). ZeRO-Infinity: Breaking the GPU Memory Wall for Extreme Scale Deep Learning. *arXiv preprint, arXiv: 2104.07857*. <https://doi.org/10.48550/arXiv.2104.07857>
- [36] Rajbhandari, S., Rasley, J., Ruwase, O., & He, Y. (2022). DeepSpeed Data Efficiency: Improving Deep Learning Model Quality and Training Efficiency via Efficient Data Sampling and Routing. *arXiv preprint, arXiv: 2212.03597*. <https://ar5iv.labs.arxiv.org/html/2212.03597>
- [37] Hugging Face. (2024). Efficient training on multiple GPUs. Retrieved from [https://huggingface.co/docs/transformers/perf\\_train\\_gpu\\_many#efficient-training-on-multiple-gpus](https://huggingface.co/docs/transformers/perf_train_gpu_many#efficient-training-on-multiple-gpus)
- [38] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. <https://d4mucfpksyww.cloudfront.net/better-language-models/language-models.pdf>
- [39] PyTorch Team. (2023). Accelerating Generative AI with PyTorch II: GPT, Fast. *PyTorch*. <https://pytorch.org/blog/accelerating-generative-ai-2/>
- [40] He, C., Li, S., Soltanolkotabi, M., & Avestimehr, S. (2021). PipeTransformer: Automated Elastic Pipelining for Distributed Training of Large-scale Models. *PyTorch*. <https://pytorch.org/blog/pipetransformer-automated-elastic-pipelining/>
- [41] Shen, L., Sun, Y., Yu, Z., Ding, L., Tian, X., & Tao, D. (2023). On Efficient Training of Large-Scale Deep Learning Models: A Literature Review. *arXiv*. <https://ar5iv.labs.arxiv.org/html/2304.03589>
- [42] Mustafa, N. (2024). Exploring Pre-trained Model Use Cases with GPT-2 and T5. *Toptal*. <https://www.toptal.com/deep-learning/exploring-pre-trained-models>
- [43] Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., ... & Sun, M. (2022). Delta Tuning: A Comprehensive Study of Parameter Efficient Methods for Pre-trained Language Models. *arXiv preprint, arXiv: 2203.06904*. <https://doi.org/10.48550/arXiv.2203.06904>
- [44] Hanna, M., Liu, O., & Variengien, A. (2023). How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. *arXiv preprint, arXiv: 2305.00586*. <https://doi.org/10.48550/arXiv.2305.00586>
- [45] Bello, I., Zoph, B., Vaswani, A., Shlens, J., & Le, Q. V. (2019). Attention augmented convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV 2019)*. <https://doi.org/10.1109/ICCV.2019.00338>
- [46] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Association for Computational Linguistics*, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
- [47] Duan, J., Zhang, S., Wang, Z., Jiang, L., Qu, W., Hu, Q., Wang, G., Weng, Q., Yan, H., Zhang, X., Qiu, X., Lin, D., Wen, Y., Jin, X., Zhang, T., & Sun, P. (2024). Efficient training of large language models on distributed infrastructures: A survey. *arXiv Labs*. <https://ar5iv.labs.arxiv.org/html/2407.20018>
- [48] Rhu, M., Gimelshein, N., Clemons, J., Zulfikar, A., & Keckler, S. W. (2016). vDNN: Virtualized Deep Neural Networks for Scalable, Memory-Efficient Neural Network Design. *arXiv preprint, arXiv: 1602.08124*. <https://doi.org/10.48550/arXiv.1602.08124>
- [49] Ren, J., Rajbhandari, S., Yazdani Aminabadi, R., Ruwase, O., Yang, S., Zhang, M., Li, D., & He, Y. (2021). ZeRO-Offload: Democratizing Billion-Scale Model Training. *arXiv*. <https://doi.org/10.48550/arXiv.2101.06840>
- [50] Amazon Web Services. (2022, March 14). How Amazon search achieves low-latency, high-throughput T5 inference with NVIDIA Triton on AWS. *AWS Machine Learning Blog*. <https://aws.amazon.com/blogs/machine-learning/how-amazon-search-achieves-low-latency-high-throughput-t5-inference-with-nvidia-triton-on-aws/>
- [51] Feng, W., Chen, Y., Wang, S., Peng, Y., Lin, H., & Yu, M. (2024). Optimus: Accelerating Large-Scale Multi-Modal LLM Training by Bubble Exploitation. *arXiv preprint, arXiv: 2408.03505*. <https://doi.org/10.48550/arXiv.2408.03505>

- [52] Vaswani, A., et al. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30 (NIPS 2017). [https://papers.nips.cc/paper\\_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html](https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html)
- [53] Huang, C., Jin, G., & Li, J. (2020). SwapAdvisor: Pushing deep learning beyond the GPU memory limit via smart swapping. *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS '20)* (pp. 1341-1355). Association for Computing Machinery. <https://doi.org/10.1145/3373376.3378530>
- [54] Choi, J., Yeom, H. Y., & Kim, Y. (2021). Implementing CUDA unified memory in the PyTorch framework. *IEEE International Conference on Autonomic Computing and Self-Organizing Systems Companion (ACSOS-C)* (pp. 20-25). <https://doi.org/10.1109/ACSOS-C52956.2021.00029>
- [55] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Brew, J. (2019). HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv Labs*, arXiv:1910.03771
- [56] Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., & Stoica, I. (2023). Efficient Memory Management for Large Language Model Serving with PagedAttention. *arXiv Labs*, arXiv.2309.06180. <https://doi.org/10.48550/arXiv.2309.06180>
- [57] Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2024). mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv Labs*. <https://ar5iv.labs.arxiv.org/html/2010.11934>
- [58] Wang, L., Ye, J., Zhao, Y., Wu, W., Li, A., Song, S. L., Xu, Z., & Kraska, T. (2018). SuperNeurons: Dynamic GPU Memory Management for Training Deep Neural Networks. *arXiv preprint*. <https://doi.org/10.48550/arXiv.1801.04380>
- [59] Zhang, H., & Zhang, Y. (2024). Memory-efficient reversible spiking neural networks. *arXiv preprint*, arXiv:2312.07922. <https://doi.org/10.48550/arXiv.2312.07922>
- [60] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint*. <https://doi.org/10.48550/arXiv.1907.11692>
- [61] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv preprint*. <https://doi.org/10.48550/arXiv.1909.11942>
- [62] Soong, D., Sridhar, S., Si, H., Wagner, J.-S., Costa Sá, A. C., Yu, C. Y., Karagoz, K., Guan, M., Kumar, S., Hamadeh, H., & Higgs, B. W. (2024). Improving accuracy of GPT-3/4 results on biomedical data using a retrieval-augmented language model. *PLOS Digital Health*, 3(8), e0000568. <https://doi.org/10.1371/journal.pdig.0000568>
- [63] Saberironaghi, A., Ren, J., & El-Gindy, M. (2023). Defect detection methods for industrial products using deep learning techniques: A review. *Algorithms*, 16(2), 95. <https://doi.org/10.3390/a16020095>
- [64] Panda, B., & Singh, P. (2023). A deep convolutional-LSTM neural network for signal detection of downlink NOMA system. *AEU - International Journal of Electronics and Communications*, 170, 154797. <https://doi.org/10.1016/j.aeue.2023.154797>
- [65] Patnaik, L. M., Wang, W. (2024). AI fairness—From machine learning to federated learning. *Computer Modeling in Engineering & Sciences*, 139(2), 1203-1215. doi: 10.32604/cmescs.2023.029451
- [66] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. *MIT Press*, ISBN 978 0262035613. [http://imlab.postech.ac.kr/dkim/class/csed514\\_2019s/DeepLearningBook.pdf](http://imlab.postech.ac.kr/dkim/class/csed514_2019s/DeepLearningBook.pdf)
- [67] Goyal, P., Pandey, S., & Jain, K. (2018). Deep Learning for Natural Language Processing: Creating Neural Networks with Python. *Apress*. <https://doi.org/10.1007/978-1-4842-3685-7>
- [68] Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., & Zhang, M. (2023). Towards General Text Embeddings with Multi-stage Contrastive Learning. *arXiv Labs*, arXiv:2308.03281. <https://doi.org/10.48550/arXiv.2308.03281>
- [69] Jin, C., Shi, Z., Li, W., & Guo, Y. (2021). Bidirectional LSTM-CRF Attention-based Model for Chinese Word Segmentation. *arXiv Labs*, arXiv:2105.09681. <https://doi.org/10.48550/arXiv.2105.09681>



- [70] Dempsey, R., Klebanov, I. R., Pufu, S. S., Søggaard, B. T., & Zan, B. (Year). Phase Diagram of the Two-Flavor Schwinger Model at Zero Temperature. *Joseph Henry Laboratories, Princeton University, Princeton Center for Theoretical Science, Princeton University, Princeton, NJ 08544, USA*. <https://ar5iv.labs.arxiv.org/html/2305.04437>
- [71] Zaharia, M., Chen, D., & Franklin, M. J. et al. (2020). A general survey on attention mechanisms in deep learning. *arXiv preprint*, arXiv:2203.14263. <https://doi.org/10.48550/arXiv.2203.14263>
- [72] Liang, W., & Liang, Y. (2024). BPDec: Unveiling the Potential of Masked Language Modeling Decoder in BERT pretraining. *arXiv*. <https://doi.org/10.48550/arXiv.2401.15861>
- [73] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Brew, J. (2019). HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv Labs*, arXiv:1910.03771. <https://doi.org/10.48550/arXiv.1910.03771>
- [74] Sawicki, J., Ganzha, M., & Paprzycki, M. (2023). The State of the Art of Natural Language Processing—A Systematic Automated Review of NLP Literature Using NLP Techniques. *Data Intelligence*, 5(3), 707–749. [https://doi.org/10.1162/dint\\_a\\_00213](https://doi.org/10.1162/dint_a_00213)
- [75] Deguchi, H., Utiyama, M., Tamura, A., Ninomiya, T., & Sumita, E. (2020). Bilingual Subword Segmentation for Neural Machine Translation. *International Committee on Computational Linguistics (COLING 2020)*, 4287–4297. <https://doi.org/10.18653/v1/2020.coling-main.378>
- [76] Fu, K., Gao, P., Liu, S., Zhang, R., Qiao, Y., & Wang, M. (2024). POS-BERT: Point Cloud One-Stage BERT Pre-Training. *arXiv Labs*. <https://ar5iv.labs.arxiv.org/html/2204.00989>
- [77] Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C-M., Chen, W., Yi, J., Zhao, W., Wang, X., Liu, Z., Zheng, H-T., Chen, J., Liu, Y., Tang, J., Li, J., & Sun, M. (2023). Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3), 220–235. <https://doi.org/10.1038/s42256-023-00626-4>
- [78] Liu, Y., Ji, L., Huang, R., Ming, T., Gao, C., & Zhang, J. (2018). An Attention-Gated Convolutional Neural Network for Sentence Classification. *arXiv Labs*, arXiv:1808.07325. <https://doi.org/10.48550/arXiv.1808.07325>
- [79] Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., & Salakhutdinov, R. (2019). Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. *Association for Computational Linguistics*, 2978–2988. <https://doi.org/10.18653/v1/P19-1285>
- [80] Sun, S., & Iyyer, M. (2021). Revisiting Simple Neural Probabilistic Language Models. *Association for Computational Linguistics*, 5181–5188. <https://doi.org/10.18653/v1/2021.naacl-main.407>
- [81] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. <https://doi.org/10.3115/v1/D14-1162>
- [82] Feng, Y., Hu, C., Kamigaito, H., Takamura, H., & Okumura, M. (2021). Improving Character-Aware Neural Language Model by Warming up Character Encoder under Skip-gram Architecture. *International Conference on Recent Advances in Natural Language Processing (RANLP 2021) - INCOMA Ltd.*, 421–427. <https://aclanthology.org/2021.ranlp-1.48>
- [83] Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators (ICLR 2020). <https://openreview.net/forum?id=r1xMH1BtvB>
- [84] Thrush, T., Tirumala, K., Gupta, A., Bartolo, M., Rodriguez, P., Kane, T., Gaviria Rojas, W., Mattson, P., Williams, A., & Kiela, D. (2022). Dynatask: A Framework for Creating Dynamic AI Benchmark Tasks. *Association for Computational Linguistics*, 174–181. <https://doi.org/10.18653/v1/2022.acl-demo.17>
- [85] Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. *Association for Computational Linguistics*, 328–339. <https://doi.org/10.18653/v1/P18-1031>
- [86] Brown, T. B., et al. (2020). Language Models are Few-Shot Learners. *arXiv Labs*, arXiv:2005.14165. <https://ar5iv.labs.arxiv.org/html/2005.14165>
- [87] Farrell, M., Recanatesi, S., Moore, T., Lajoie, G., & Shea-Brown, E. (2022). Gradient-based learning drives robust representations in recurrent neural networks by balancing compression and expansion. *Nature Machine Intelligence*, 4, 564–573. <https://doi.org/10.1038/s42256-022-00498-0>
- [88] Kelvinius, F. E., Georgiev, D., Toshev, A. P., & Gasteiger, J. (2023). Accelerating Molecular Graph Neural Networks via Knowledge Distillation. *arXiv Labs*, arXiv:2306.14818. <https://ar5iv.labs.arxiv.org/html/2306.14818>
- [89] Aydin, I., Sevi, M., Akin, E., Güçlü, E., Karaköse, M., Aldarwich, H. (2023). A Deep Learning-Based Hybrid Approach to Detect Fastener Defects in Real-Time. *Tehnički vjesnik - Technical Gazette*, 30 (5), 1461-1468. <https://doi.org/10.17559/TV-20221020152721>

- [90] Vasileiou, A., & Eberle, O. (2024). Explaining text similarity in transformer models. *arXiv Labs*. <https://ar5iv.labs.arxiv.org/html/2405.06604>
- [91] Gopalun, K., John Samuvel, D. (2023). Deep Learning Technique for Power Domain Non-Orthogonal Multiple Access Using Optimised LSTM in Cooperative Networks. *Tehnički vjesnik - Technical Gazette*, 30 (5), 1397-1403. <https://doi.org/10.17559/TV-20221228104420>
- [92] Li, Y., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415, 295-316. <https://doi.org/10.1016/j.neucom.2020.07.061>
- [93] 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*. <https://doi.org/10.48550/arXiv.2107.05847>
- [94] George, C. G. S., & Sumathi, B. (2020). Grid search tuning of hyperparameters in random forest classifier for customer feedback sentiment prediction. *International Journal of Advanced Computer Science and Applications*, 11(9). <https://doi.org/10.14569/IJACSA.2020.0110920>
- [95] Vo, H. T., Ngoc, H. T., & Quach, L. D. (2023). An Approach to Hyperparameter Tuning in Transfer Learning for Driver Drowsiness Detection Based on Bayesian Optimization and Random Search. *International Journal of Advanced Computer Science and Applications*, 14(4). <https://doi.org/10.14569/IJACSA.2023.0140492>
- [96] Jafar, A., & Lee, M. (2021). High-speed hyperparameter optimization for deep ResNet models in image recognition. *Cluster Computing*, 26(5). <https://doi.org/10.1007/s10586-021-03284-6>
- [97] Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & de Freitas, N. (2016). Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proceedings of the IEEE*, 104(1), 148-175. <https://doi.org/10.1109/JPROC.2015.2494218>
- [98] Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. (2017). Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. *Journal of Machine Learning Research*, 18(185), 1-52. <https://jmlr.org/papers/volume18/16-558/16-558.pdf>
- [99] Mezzah, S., & Tari, A. (2023). Practical hyperparameters tuning of convolutional neural networks for EEG emotional features classification. *Intelligent Systems with Applications*, 18. <https://doi.org/10.1016/j.iswa.2023.200212>
- [100] Atteia, G., Abdel Samee, N., El-Kenawy, E. S. M., & Ibrahim, A. (2022). CNN-Hyperparameter Optimization for Diabetic Maculopathy Diagnosis in Optical Coherence Tomography and Fundus Retinography. *Mathematics*, 10(18). <https://doi.org/10.3390/math10183274>
- [101] Czako, Z., Sebestyen, G., & Hangan, A. (2021). AutomaticAI – A hybrid approach for automatic artificial intelligence algorithm selection and hyperparameter tuning. *Expert Systems with Applications*, 182. <https://doi.org/10.1016/j.eswa.2021.115225>
- [102] Ali, Y.A., Awwad, E.M., Al-Razgan, M., & Maarouf, A. (2023). Hyperparameter Search for Machine Learning Algorithms for Optimizing the Computational Complexity. *Processes*, 11(2), 349. <https://doi.org/10.3390/pr11020349>
- [103] Yang, L., Shami, A. (2021). An improved hyperparameter optimization framework for AutoML systems using evolutionary algorithms. *Scientific Reports*, 11, 17354. <https://doi.org/10.1038/s41598-023-32027-3>
- [104] Karl, F., Pielok, T., Moosbauer, J., Pfisterer, F., Coors, S., Binder, M., ... Bischl, B. (2022). Multi-Objective Hyperparameter Optimization -- An Overview. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2206.07438>
- [105] Morales-Hernández, A., Van Nieuwenhuysse, I., & Rojas Gonzalez, S. (2023). A survey on multi-objective hyperparameter optimization algorithms for machine learning. *Artificial Intelligence Review*, 56(9), 8043–8093. <https://doi.org/10.1007/s10462-022-10359-2>
- [106] Dong, X., Liu, L., Musial, K., & Gabrys, B. (2021). NATS-Bench: Benchmarking NAS Algorithms for Architecture Topology and Size. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7), 3634-3646. <https://doi.org/10.1109/TPAMI.2021.3054824>
- [107] Schneider, L., Bischl, B., & Thomas, J. (2023). Multi-Objective Optimization of Performance and Interpretability of Tabular Supervised Machine Learning Models. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2307.08175>
- [108] Ilievski, I., Akhtar, T., Feng, J., & Shoemaker, C. (2017). Efficient Hyperparameter Optimization for Deep Learning Algorithms Using Deterministic RBF Surrogates. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1). <https://doi.org/10.1609/aaai.v31i1.10647>

- [109] Maass, K., Aravkin, A., & Kim, M. (2022). A hyperparameter-tuning approach to automated inverse planning. *Medical Physics*, 49(5). <https://doi.org/10.1002/mp.15557>
- [110] Amirabadi, M. A., Kahaei, M. H., & Nezamalhosseini, S. A. (2020). Novel suboptimal approaches for hyperparameter tuning of deep neural network [under the shelf of optical communication]. *Physical Communication*, 41. <https://doi.org/10.1016/j.phycom.2020.101057>
- [111] Dushatskiy, A., Chebykin, A., Alderliesten, T., & Bosman, P. A. N. (2023). Multi-Objective Population Based Training. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2306.01436>
- [112] Bergstra, J., Yamins, D., & Cox, D. D. (2013). Hyperopt: A Python Library for Optimizing the Hyperparameters of Machine Learning Algorithms. *Proceedings of the 12th Python in Science Conference*. <https://doi.org/10.25080/Majora-8b375195-003>
- [113] Shankar, K., Kumar, S., Dutta, A. K., Alkhayyat, A., Jawad, A. J. M., Abbas, A. H., & Yousif, Y. K. (2022). An Automated Hyperparameter Tuning Recurrent Neural Network Model for Fruit Classification. *Mathematics*, 10(13), 2358. <https://doi.org/10.3390/math10132358>
- [114] Morales-Hernández, A., Van Nieuwenhuysse, I., & Nápoles, G. (2022). Multi-objective hyperparameter optimization with performance uncertainty. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2209.04340>
- [115] Giovanelli, J., Tornede, A., Tornede, T., & Lindauer, M. (2024). Interactive Hyperparameter Optimization in Multi-Objective Problems via Preference Learning. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2309.03581>
- [116] Watanabe, S., Awad, N., Onishi, M., & Hutter, F. (2023). Speeding Up Multi-Objective Hyperparameter Optimization by Task Similarity-Based Meta-Learning for the Tree-Structured Parzen Estimator. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2212.06751>
- [117] Watanabe, S., & Hutter, F. (2023). c-TPE: Tree-structured Parzen Estimator with Inequality Constraints for Expensive Hyperparameter Optimization (Version 4). *arXiv preprint*. <https://doi.org/10.48550/arXiv.2211.14411>
- [118] Liu, Y., Li, Y., Schiele, B., & Sun, Q. (2023). Online Hyperparameter Optimization for Class-Incremental Learning. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2301.05032>
- [119] Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 13(10), 281–305. <http://jmlr.org/papers/v13/bergstra12a.html>
- [120] Pandey, R., & Sen, J. (2024). Generative AI-Based Text Generation Methods Using Pre-Trained GPT-2 Model. *arXiv preprint*, arXiv:2404.01786. <https://doi.org/10.48550/arXiv.2404.01786>
- [121] 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 preprint*, arXiv:1912.01703. <https://doi.org/10.48550/arXiv.1912.01703>
- [122] Shrestha, S. L., & Csallner, C. (2021). SLGPT: Using Transfer Learning to Directly Generate Simulink Model Files and Find Bugs in the Simulink Toolchain. *arXiv preprint*, arXiv:2105.07465. <https://doi.org/10.48550/arXiv.2105.07465>
- [123] Sen, J., Pandey, R., & Group 5. (2024). Generative AI-Based Text Generation Methods Using Pre-Trained GPT-2 Model. *arXiv preprint*, arXiv:2404.01786. <https://doi.org/10.48550/arXiv.2404.01786>
- [124] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. *OpenAI preprint*. Retrieved from [https://cdn.openai.com/research-covers/language-unsupervised/language\\_understanding\\_paper.pdf](https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf)
- [125] Yang, Z., Ding, M., Lv, Q., Jiang, Z., He, Z., Guo, Y., Bai, J., & Tang, J. (2024). GPT can solve mathematical problems without a calculator. *arXiv Labs*. <https://arxiv.labs.arxiv.org/html/2309.03241>
- [126] Shah, J., Bikshandi, G., Zhang, Y., Thakkar, V., Ramani, P., & Dao, T. (2024). FlashAttention-3: Fast and Accurate Attention with Asynchrony and Low-precision. *arXiv preprint*, arXiv:2407.08608. <https://doi.org/10.48550/arXiv.2407.08608>
- [127] 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>

- [128] Dodić, D., & Regodić, D. (2024). Tokenization and Memory Optimization For Reducing GPU Load in NLP Deep Learning Models. *Proceedings of the Tehnički vjesnik - Technical Gazette*, 31(1), 123-145. <https://doi.org/10.17559/TV-20231218001216>
- [129] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. *In Advances in Neural Information Processing Systems* (pp. 5753-5763). <https://doi.org/10.48550/arXiv.1906.08237>

# PRILOZI

---

## Pseudokod 1. Definisanje metodologije

---

### 1.1 Inicijalizacija okruženja:

- Uvoze se neophodne biblioteke.
- Postavlja se nasumično "seed" za reproduktivnost.
- Prijavljuje se na Weights & Biases (wandb).

### 1.2 Učitavanje tokenizatora:

- Proverava se da li je tokenizator sačuvan.
- Učitava se tokenizator iz direktorijuma.

### 1.3 Učitavanje modela:

- Učitava se unapred trenirani GPT-2 model.
- Model se premešta na GPU ako je dostupan.
- Čisti se GPU keš.

### 1.4 Proširivanje modela:

- Model se proširuje ako su dodati novi tokeni u tokenizator.
- Postavljaju se specijalni tokeni za popunjavanje i maskiranje.

## 2. Implementacija algoritama za optimizaciju scheduler-a:

### 2.1 Definisane klase PyTorchSchedulerWrapper:

- Scheduler se inicijalizuje sa CyclicLR.
- Definišu se metode za "step", "state\_dict", i "load\_state\_dict".

### 2.2 Implementacija klase EarlyStoppingCallback:

- Inicijalizuje se sa "early\_stopping\_patience".
- Definiše se "on\_evaluate" metoda za proveru da li treba zaustaviti trening na osnovu "eval\_loss".

### 2.3 Implementacija klase OptunaPruningCallback:

- Inicijalizuje se sa "trial" objektom.
- Definiše se "on\_evaluate" metoda za proveru da li treba prekinuti trening na osnovu metrika.

### 2.4 Definisane funkcije za izračunavanje metrika:

- `compute_accuracy`: Izračunava se tačnost predikcije.
- `compute_perplexity`: Izračunava se "perplexity" na osnovu "eval\_loss".
- `compute_metrics`: Kombinuju se tačnost i "perplexity" u rečnik.
- `wrapped_compute_metrics`: Uvija se `compute_metrics` funkcija za upotrebu unutar Trainer-a.

### 2.5 Inicijalizacija optimizatora i scheduler-a:

- Definiše se funkcija `initialize_optimizer_and_scheduler` za kreiranje optimizatora i scheduler-a.
- Kreiraju se optimizator i scheduler za model.

## 3. Podešavanje hiperparametara:

### 3.1 Analiza prethodnih rezultata:

- Definiše se funkcija ``analyze_previous_results`` za analizu prethodnih rezultata treniranja.

### 3.2 Podešavanje hiperparametara:

- Definiše se funkcija ``adaptive_tune_hyperparams`` za podešavanje hiperparametara na osnovu analiziranih rezultata.

### 3.3 Logovanje hiperparametara i performansi:

- Definiše se funkcija ``datho_log`` za logovanje hiperparametara i performansi u fajl.
- Definiše se funkcija ``load_previous_results`` za učitavanje prethodnih rezultata iz fajla.

## 4. Priprema podataka:

### 4.1 Tokenizacija podataka:

- Definiše se ``tokenize_function`` za tokenizaciju podataka za obuku.
- Definiše se ``tokenize_function_val`` za tokenizaciju podataka za validaciju.

### 4.2 Podela podataka na segmente:

- Definiše se funkcija ``split_into_chunks`` za delu podataka na manje segmente.

### 4.3 Kreiranje klase `CustomDataCollatorForLanguageModeling`:

- Inicijalizuje se sa tokenizatorom i parametrima za maskiranje.
- Definiše se ``mask_tokens`` metoda za pripremu ulaza i labela za maskirano učenje jezika.
- Definiše se ``__call__`` metoda za pozivanje ``mask_tokens`` za svaki primer u skupu podataka.

## 5. Treniranje i evaluacija modela:

### 5.1 Definisavanje klase `MyTrainer`:

- Nasleđuje se ``transformers.Trainer`` klasa.
- Inicijalizuje se sa dodatnim parametrima (`gamma`, `analyzed_results`).
- Definišu se metode:
  - ``load_and_tune_hyperparams``: Učitavaju se prethodni rezultati i prilagođavaju hiperparametri.
  - ``training_step``: Prilagođavaju se hiperparametri tokom treniranja i čuva se "checkpoint" ako je potrebno.
  - ``evaluate``: Evaluira se model i ažurira ``eval_loss``.
  - ``on_epoch_end``: Čisti se GPU keš, evaluira model i loguju se rezultati.
  - ``save_at_interval``: Čuva se "checkpoint" u intervalima.

### 5.2 Kreiranje funkcije za inicijalizaciju optimizatora i scheduler-a:

- Definiše se funkcija ``initialize_optimizer_and_scheduler`` da vrati optimizator i scheduler.

### 5.3 Učitavanje checkpoint-ova:

- Definiše se funkcija ``load_checkpoint`` za učitavanje modela, tokenizatora, optimizatora i scheduler-a iz "checkpoint-a".

### 5.4 Kreiranje objektivne funkcije za Optuna optimizaciju:

- Definiše se funkcija ``create_objective`` da vrati objektivnu funkciju za Optuna optimizaciju.

### 5.5 Pokretanje Optuna optimizacije:

- Inicijalizuje se Optuna ``study`` sa ``MedianPruner``.
- Pokreće se ``study.optimize`` sa definisanom objektivnom funkcijom i brojem pokušaja.

## 5.6 Učitavanje najboljih hiperparametara:

- Izvlače se najbolji hiperparametri iz Optuna `study`.

## 5.7 Priprema podataka za treniranje:

- Učitavaju se podaci za obuku i validaciju.
- Tokenizuju se podaci koristeći `tokenize\_function` i `tokenize\_function\_val`.
- Proverava se veličina vokabulara i `max\_id` u modelu.
- Kreira se `CustomDataCollatorForLanguageModeling` objekat za batchovanje.
- Inicijalizuje se wandb sa najboljim hiperparametrima.

---

## Pseudokod 2. Optimizacija potrošnje memorije putem slučajnog uzorkovanja podataka u evaluaciji modela

---

```
max_length = 160
```

```
# Funkcija tokenizacije
```

```
def tokenize_function(batch):
```

- Za svaki tekst u batch["text"]:
  - Ako je tekst prazan ili sadrži samo razmak, odbacuje se
- Za očišćenu seriju tekstova:
  - Ako nema valjanih tekstova, vraća se {"input\_ids": [], "attention\_mask": []}
  - U suprotnom, tekstovi se tokenizuju pomoću Hugging Face tokenizer-a sa sledećim parametrima:
    - truncation: True
    - padding: "longest"
    - return\_special\_tokens\_mask: True
    - max\_length: max\_length
    - return\_tensors: 'pt'
- Rukuje se greškama tokenizacije i vraćaju se prazne vrednosti ako ih ima

```
# Ciljna funkcija za optimizaciju modela
```

```
def create_objective(model):
```

- Koristi se Optuna za podešavanje hiperparametara:
  - Objective funkcija (trial, model):
    - Postavlja se max\_length koristeći trial.suggest\_int('max\_length', 64, 256, step=16)
    - Definiše se encode\_function(examples):
      - Koristi se tokenize\_function za obradu teksta
      - Ako su "input\_ids" ili "attention\_mask" prazni, vraćaju se prazne vrednosti
    - Mapira se encode\_function na dataset, sa parametrima batched=True, remove\_columns=["text"]
    - Postavlja se format skupa podataka za PyTorch
    - Nasumično se uzorkuje 100 indeksa iz validation seta
    - Inicijalizuje se MyTrainer sa navedenim parametrima:
      - gamma, model, transformers.TrainingArguments, itd.
      - train\_dataset: tokenized\_dataset["train"]
      - eval\_dataset: tokenized\_dataset["validation"].select(random\_indices)
    - Obučava se model i vraća eval\_output["eval\_loss"]

---

## Pseudokod 3. Efikasna evaluacija modela sa slučajnim uzorkovanjem iz skupa za validaciju

---

```

# nasumično uzorkovanje zasnovano na 100 indeksa iz validation seta
num_samples = 100
# svi indeksi iz validation seta
all_indices = list(range(len(tokenized_dataset["validation"])))
# slučajni odabir indeksa
random_indices = random.sample(all_indices, num_samples)

trainer = MyTrainer(
    args=transformers.TrainingArguments(
        # ostali argumenti
    ),
    train_dataset=tokenized_dataset["train"],
    eval_dataset=tokenized_dataset["validation"].select(random_indices),
)

```

### **Objašnjenje za Pseudokod 3:**

Ovaj deo koda se odnosi na proces obuke i evaluacije modela:

`num_samples = 100`: Postavljanje broja uzoraka koji se nasumično biraju iz validation seta. U ovom slučaju odabrano je 100 uzoraka zbog obuke na GPU-u sa 8GB VRAM-a.

`all_indices = list(range(len(tokenized_dataset["validation"])))`: Generisanje liste svih indeksa u validation setu. Ova lista indeksira podatke unutar validation seta.

`random_indices = random.sample(all_indices, num_samples)`: Ovde se izvodi nasumično uzorkovanje od 100 indeksa sa liste svih indeksa u validation setu. Funkcija `random.sample` omogućava odabir nasumičnih uzoraka bez zamene.

`trainer = MyTrainer(...)`: Inicijalizacija trenera odgovornog za obuku i evaluaciju modela.

Parametri trenera uključuju:

`args=transformers.TrainingArguments(...)`: Ostali argumenti koji se koriste u obuci.

`train_dataset=tokenized_dataset["train"]`: Dataset koji se koristi za obuku modela.

`eval_dataset=tokenized_dataset["validation"].select(random_indices)`: Dataset za evaluaciju koji se sastoji od 100 nasumično odabranih uzoraka iz validation seta.

---

### **Pseudokod 4. Generalni proces treniranja**

---

# Generalni proces treniranja

1. Definisati funkcije `compute_accuracy(pred)` i `compute_perplexity(eval_loss)` za izračunavanje tačnosti i perplexity metrika.

2. Implementirati funkciju `compute_metrics(p, eval_loss)` koja agregira i računa metrike koristeći predikcije i evaluacioni gubitak.

```
def compute_metrics(p, eval_loss):
```

```
    # Implementacija:
```

```
    - Izračunati tačnost koristeći predikcije.
```

```
    - Izračunati perplexity na osnovu evaluacionog gubitka.
```

```
    - Prikazati izračunate metrike.
```

```
    - Vratiti rečnik sa tačnošću i perplexity kao izračunatim metrikama.
```

3. Koristiti `WandbCallback` klasu za logovanje uz `Weights & Biases`, omogućavajući detaljno praćenje i analizu.

4. Implementirati `EarlyStoppingCallback` klasu sa definisanim parametrom `early_stopping_patience` kako bi se zaustavilo treniranje na osnovu performansi.

5. Inicijalizovati globalnu `SummaryWriter` instancu za logovanje sa `TensorBoard`-om, pružajući platformu za praćenje i vizuelizaciju napretka treninga.



6. Konfigurisati optimizator treniranja i scheduler putem funkcije `initialize_optimizer_and_scheduler`, postavljajući osnovu za efikasno upravljanje stopom učenja, uključujući ciklično prilagođavanje stope učenja.
7. Koristiti `CustomDataCollatorForLanguageModeling` klasu za pripremu podataka, specifično prilagođenu za `masked language modeling`.
8. Prilagoditi `MyTrainer` klasu, proširujući `transformers.Trainer` klasu, kako bi se unapredili mehanizmi treniranja, evaluacije i logovanja. Ovo uključuje upravljanje cikličnim rasporedom stope učenja unutar procesa treniranja.
9. Integrisati `Optuna` optimizaciju hiperparametara unutar `create_objective` funkcije, koristeći je za sistematsko podešavanje parametara.

---

### **Pseudokod 5. Adaptivno podešavanje hiperparametara**

---

```
def adaptive_tune_hyperparams(analyzed_results, current_hyperparams):  
    # Funkcionalnost: Prilagođavanje hiperparametara na osnovu analiziranih rezultata.  
    # Implementacija:  
    1. Napraviti kopiju trenutnih hiperparametara.  
    2. Definisati prag gubitka ako već nije definisan u trenutnim hiperparametrima.  
    3. Analizirati rezultate i prilagoditi hiperparametre na osnovu te analize.  
    4. Vratiti optimizovane hiperparametre.
```

---

### **Pseudokod 6. Zapisivanje logova u DYNAMO sistemu**

---

```
def dynamo_log(hyperparams, performance, log_file):  
    # Funkcionalnost: Zapisivanje performansi i hiperparametara u log fajl.  
    # Implementacija:  
    1. Proveriti da li log fajl postoji.  
    2. Ako fajl ne postoji, kreirati ga kao prazan.  
    3. Otvoriti log fajl u režimu dopisivanja.  
    4. Zapisati hiperparametre i performanse u fajl.
```

---

### **Pseudokod 7. Učitavanje prethodnih rezultata**

---

```
def load_previous_results(log_file):  
    # Funkcionalnost: Učitavanje prethodnih rezultata iz log fajla.  
    # Implementacija:  
    1. Inicijalizovati listu za rezultate.  
    2. Proveriti da li log fajl postoji i da li ima sadržaj.  
    3. Ako postoji i nije prazan, otvoriti fajl za čitanje.  
    4. Za svaku liniju u fajlu, odvojiti podatke o hiperparametrima i performansama.  
    5. Konvertovati stringove u odgovarajuće tipove podataka i dodati ih u listu rezultata.  
    6. Prikazati poruku o učitavanju.  
    7. Ako je fajl prazan, prikazati odgovarajuću poruku.  
    8. Vratiti listu učitanih rezultata.
```

---

### **Pseudokod 8. Računanje metrika za evaluaciju predikcija**

---

```
def wrapped_compute_metrics(trainer, eval_prediction):  
    # Funkcionalnost: Računanje metrika za evaluaciju predikcija koristeći trainer objekat.
```

# Implementacija:

1. Dohvatiti najnoviji evaluacioni gubitak iz treniranja.
2. Ako evaluacioni gubitak nije postavljen, pokrenuti grešku.
3. Vratiti rezultat compute\_metrics funkcije sa evaluacionim predikcijama i gubitkom.

---

### **Pseudokod 9. Analiza prethodnih rezultata**

---

```
def analyze_previous_results(previous_results):
```

```
    # Funkcionalnost: Analiza prethodnih rezultata i izračunavanje prosečnih vrednosti.
```

```
    # Implementacija:
```

1. Proveriti da li postoje prethodni rezultati.
2. Ako nema rezultata, vratiti podrazumevane vrednosti za gubitak, tačnost i perplexity.
3. Inicijalizovati promenljivu za analizirane rezultate.
4. Izvući gubitak, tačnost i perplexity iz prethodnih rezultata.
5. Nastaviti analizu za izračunavanje prosečnih vrednosti.

---

### **Pseudokod 10. Glavni proces optimizacije u DYNAMO okviru**

---

```
# Funkcionalnost: Glavni proces optimizacije u DYNAMO okviru.
```

```
# Implementacija:
```

1. Učitajte prethodne rezultate.

```
previous_results = load_previous_results()
```

2. Analizirajte prethodne rezultate.

```
analyzed_results = analyze_previous_results(previous_results)
```

3. Prilagodite početne hiperparametre na osnovu analize.

```
optimized_hyperparams = adaptive_tune_hyperparams(analyzed_results, initial_hyperparams)
```

4. Koristite Optuna da predložite optimalne vrednosti hiperparametara.

```
optimized_hyperparams['n_embd'], optimized_hyperparams['n_layer'], step_size_up, hidden_size, n_heads,  
ff_size, n_layers = suggest_hyperparameters(trial)
```

5. Prekinite trial ako hiperparametri ne zadovoljavaju uslove.

```
if hidden_size % n_heads != 0:
```

```
    raise optuna.TrialPruned()
```