

TRANSFORMERI U PRAKSI

Goran Savić¹

Rezime: Ovaj rad pruža pregled primene transformera u različitim zadacima obrade prirodnog jezika, pokazujući njihovu efikasnost, fleksibilnost i potencijal u rešavanju složenih NLP (Natural Language Processing) problema. Kroz analizu sentimenta, prepoznavanje imenovanih entiteta, odgovaranje na pitanja, sažimanje teksta, prevod i generisanje teksta, prezentovane su konkretne primere koje ilustruju kako transformeri, kao što su BERT, T5, BART i drugi, mogu da vrše složene zadatke analize i generisanja teksta.

Ključne reči: NLP, BERT, imenovani entiteti, odgovori na pitanja, sumariizacija, prevod

TRANSFORMERS IN PRACTICE

Abstract: This paper provides an overview of the application of transformers in various natural language processing tasks, showing their effectiveness, flexibility and potential in solving complex NLP (Natural Language Processing) problems. Through sentiment analysis, named entity recognition, question answering, text summarization, translation, and text generation, concrete examples are presented that illustrate how transformers, such as BERT, T5, BART, and others, can perform complex text analysis and generation tasks..

Key words: NLP, BERT, named entity recognition, question answering, summarization, translation

1. UVOD

Transformeri su doprineli neverovatnom brzom razvoju obrade prirodnog jezika i širokoj upotrebi velikih jezičkih modela (Large Language Models - LLM). Sve do 2017. godine i objave rada *Attention is all you need* [1], gde su transformeri predstavljeni, zadaci obrade jezika su se vršili najčešće pomoću povratnih neuronskih mreža (recurrent neural networks). Najveći izazov povratnih neuronskih mreža je bila obrada dugačkih nizova, gde se informacija sa početka niza mogla izgubiti prilikom obrade dok se dođe do kraja niza. Ovaj problem je prevaziđen primenom mehanizma *attention*, koji određuje prioritet (težinski koeficijent) skrivenih stanja koja će doći na ulaz u dekodir [2]. Ovo značajno otkriće doprinelo je velikoj primeni u različitim oblastima obrade jezika kao što su analiza sentimenta, prepoznavanje imenovanih entiteta, odgovaranje na pitanja, sažimanje teksta, prevođenje i generisanje teksta.

Transformeri mogu da koriste različitu arhitekturu u zavisnosti od vrste zadatka [2]:

- Enkoder – klasifikacija teksta, prepoznavanje imenovanih entiteta. Neki od modela koji koriste enkoder su BERT, DistilBERT, ModernBERT
- Dekoder – generisanje teksta, konverzacioni AI, kreativno pisanje. Neki od modela koji koriste dekodir su GPT, LLaMa, Gemma, SmolLM
- Enkoder-dekoder – sumariizacija, prevođenje, generativno odgovaranje na pitanja. Neki od modela koji koriste enkoder-dekoder arhitekturu su BART, T5, Marian, mBART.

Cilj ovog rada je da se prikaže kako radi algoritam transformers na nekoliko praktičnih primera, uz prikaz koda i rezultata izvršenja.

2. MODEL

Za potrebe ovog rada koristiće se biblioteka Hugging face transformers koja podržava tri najveća radna okruženja mašinskog učenja (PyTorch, Tensorflow i JAX). Ova biblioteka može da se koristi za fino podešavanje transformera za potrebe specifičnih zadataka u okviru NLP-a, a može i direktno da se

¹ msc, Goran Savić, Visoka tehnička škola strukovnih studija, Novi Sad, savic.goran@vtsns.edu.rs

koristi za zadatke kao što su analiza sentimenta, prepoznavanje imenovanih entiteta, odgovaranje na pitanja, sumariizacija teksta, prevod i generisanje teksta.

Ovi zadaci će se izvršavati u Tensorflow radnom okruženju, koristeći uslugu Google colab.

Svaki NLP zadatak započinje nekim tekstom. Za ovaj rad izabrana je jedna nasumična recenzija filma "Star Wars Episode IV – A new hope" sa sajta www.imdb.com na kojoj će biti primenjeni zadaci NLP-a.

text = "Filled with great characters and a fun story, Star Wars is well deserving of its reputation as a classic. John Williams score alone makes this a film worth seeing. The characters are unforgettable and the special effects impressive for its time, but very outdated now. The different alien species are one of the best things about the film. The Cantina scene showing an array of Lucas' creations is particularly fun."

I give this film a 7/10. Looking at it subjectively, it isn't a really good film. A lot of clichés, bad dialogue, cracker jack philosophy, and unimpressive acting. Yet, for some reason, I find it hard to say anything bad about this movie. My favorite film as a child, and one I still enjoy, mostly for nostalgia."

2.1. Analiza sentimenta (Sentiment Analysis)

Prema [3] transformeri koriste predtrenirane jezičke modele, poput BERT-a (Bidirectional Encoder Representations from Transformers), za analizu sentimenta u tekstu. Tokom fine-tuninga, model uči da prepozna emotivne tonove (npr. pozitivan, neutralan ili negativan) analizirajući kontekst reči unutar rečenica. Ovo se postiže zahvaljujući sposobnosti transformera da uzimaju u obzir globalni kontekst rečenice i na taj način mogu da razumeju kontekst na dubljem nivou.

Neke od prednosti transformera u analizi sentimenta su da transformeri mogu prepoznati složene emocionalne nijanse, uključujući sarkazam i ironične izraze [3].

Kako bi smo ilustrovali na koji način može da se uradi klasifikacija teksta, u Tensorflow framework su unete sledeće linije koda:

```
from transformers import pipeline
import pandas as pd
klasifikator = pipeline("text-classification")
rezultat = klasifikator(text)
pd.DataFrame(rezultat)
```

Koristeći Hugging face Transformers i biblioteku pipeline za zadatak klasifikacije, odnosno analizu sentimenta komentara dobijen je sledeći rezultat:

Tabela 1 – Rezultat klasifikacije teksta

	label	score
0	NEGATIVE	0.628249

Na osnovu rezultatata može da se vidi da je klasifikator imao težak zadatak jer komentar nije isključivo ni pozitivan ni negativan. U ovom primeru je korišćena python biblioteka pandas da bi se rezultat prikazao u dataframe obliku - kao tabela.

2.2. Prepoznavanje imenovanih entiteta (Named Entity Recognition - NER)

Prema [4] Named Entity Recognition (NER) je zadatak obrade prirodnog jezika (Natural Language Processing - NLP) koji se bavi identifikacijom i klasifikacijom imenovanih entiteta u tekstu. Imenovani entiteti obuhvataju kategorije kao što su imena ljudi, mesta, organizacija, datumi, vremenski intervali, i druge semantičke klase. Transformeri, kao što su BERT i RoBERTa, pružaju napredne mogućnosti za rešavanje ovog zadatka.

Neke od prednosti transformera za NER su da modeli prepoznaju entitete čak i u komplikovanim sintaktičkim strukturama, fino podešeni modeli mogu biti optimizovani za specifične jezike i domene i transformeri pružaju dosledno dobre rezultate čak i na velikim korpusima podataka [3].

Kako bi smo ilustrovali na koji način može da se uradi prepoznavanje imenovanih entiteta u Tensorflow framework su unete sledeće linije koda:

```
from transformers import pipeline
import pandas as pd
prepoznavanje_entiteta = pipeline("ner", aggregation_strategy="simple")
rezultat = prepoznavanje_entiteta(text)
pd.DataFrame(rezultat)
```

Koristeći Hugging face Transformers i biblioteku pipeline za zadatak prepoznavanja imenovanih entiteta, dobijen je sledeći rezultat:

Tabela 2 – Rezultat prepoznavanja imenovanih entiteta

	entity_group	score	word	start	end
0	MISC	0.994748	Star Wars	46	55
1	PER	0.994308	John Williams	106	119
2	MISC	0.840158	Cantina	344	351
3	PER	0.997373	Lucas	378	383

Ovde može da se vidi da je pipeline ispravno detektovao sve entitete i svrstao ih u dve kategorije: PER (osobe) i MISC (razno). Identifikovanje entiteta i svrstavanje u kategorije je uradio sa visokim stepenom sigurnosti.

2.3. Odgovaranje na pitanja (Question Answering - QA)

Transformeri, poput T5 (Text-to-Text Transfer Transformer) i BERT-a, mogu direktno odgovarati na pitanja analizirajući ulazni tekst. Model se trenira tako što razume pitanje i koristi pažnju (attention) za pronalaženje tačnog odgovora u odgovarajućem delu teksta [5].

Kako bi ilustrovali da li algoritam može da odgovara na pitanja i na koji način to radi u Tensorflow su unete sledeće linije koda:

```
from transformers import pipeline
import pandas as pd
odgovaranje = pipeline("question-answering")
pitanje = "Are there any special effects in the film?"
rezultat = odgovaranje(question=pitanje, context=text)
pd.DataFrame([rezultat])
```

Koristeći Hugging face Transformers i biblioteku pipeline za zadatak odgovaranja na pitanja, dobijen je sledeći rezultat:

Tabela 3 – Rezultat odgovora na pitanje

	score	start	end	answer
0	0.133152	201	267	the special effects impressive for its time, but very outdated now

Transformer je pronašao ispravan odgovor. Uz odgovor je prikazao integer koji predstavljaju indekse početka i kraja dela teksta koji je vezan za odgovor.

2.4. Sažimanje teksta (Summarization)

Prema [2] cilj sumarizacije je da algoritam na osnovu dugačkog teksta koji predstavlja ulaz u transformer prikaže skraćenu verziju sa svim relevantnim činjenicama.

Za sažimanje, transformeri kao što su BART (Bidirectional and Auto-Regressive Transformers) koriste sekvencijalnu analizu. Modeli prepoznaju važne delove teksta i generišu sažetak koji čuva značenje originala, dok izostavlja suvišne informacije [6].

Kako bi smo ilustrovali na koji način može da se uradi sumarizacija teksta, u Tensorflow framework su unete sledeće linije koda:

```
from transformers import pipeline
sumiranje = pipeline("summarization")
rezultat = sumiranje(text, max_length=60, clean_up_tokenization_spaces=True)
print(rezultat[0]['summary_text'])
```

Dobijeni rezultat:

```
Filled with great characters and a fun story, Star Wars is well deserving of its
reputation as a classic. The Cantina scene showing an array of Lucas' creations is
particularly fun. The characters are unforgettable and the special effects are
impressive for its time, but very outdated now.
```

U funkciji sumiranje() može da se unese proizvoljna maksimalna dužina odgovora.

2.5. Prevod (Translation)

Prema [2] kao i sumarizacija, prevod je zadatak gde izlazni rezultat sadrži generisani tekst.

Modeli poput mBART-a i mT5-a transformeri su specijalizovani za prevođenje među više jezika. Oni koriste sekvence tokena koje modeliraju kontekst kako bi generisali precizan prevod [7].

Kako bi smo ilustrovali na koji način može da se uradi prevod teksta u italijanski jezik, u Tensorflow framework su unete sledeće linije kôda:

```
from transformers import pipeline
prevodilac = pipeline("translation",
                      model="Helsinki-NLP/opus-mt-en-it")
rezultat = prevodilac(text, clean_up_tokenization_spaces=True, min_length=100)
print(rezultat[0]['translation_text'])
```

Dobijeni rezultat:

```
Ricca di grandi personaggi e di una storia divertente, Star Wars
merita molto la sua reputazione di classico. La sola partitura di
John Williams rende questo film un film particolarmente divertente.
I personaggi sono indimenticabili e gli effetti speciali
impressionanti per il suo tempo, ma molto obsoleti ora. Le diverse
specie aliene sono una delle cose migliori del film. La scena della
Cantina che mostra una serie di creazioni di Lucas è particolarmente
divertente. Io do a questo film un 7/10. Guardandolo
soggettivamente, non è un buon film. Un sacco di cliché, cattivo
dialogo, filosofia cracker jack, e recitazione non impressionante.
Eppure, per qualche motivo, trovo difficile dire qualcosa di brutto
su questo film. Il mio film preferito da bambino, e uno che ancora
mi piace, soprattutto per nostalgia.
```

Ako želimo prevod na neki drugi jezik potrebno je promeniti samo treću liniju kôda, odnosno model:

```
model="Helsinki-NLP/opus-mt-en-de")
```

Dobijeni rezultat:

Gefüllt mit großartigen Charakteren und einer lustigen Geschichte, verdient Star Wars seinen Ruf als Klassiker. John Williams Partitur allein macht diesen Film sehenswert. Die Charaktere sind unvergesslich und die Special Effects beeindruckend für seine Zeit, aber sehr veraltet. Die verschiedenen Alien-Arten sind eines der besten Dinge über den Film. Die Cantina Szene zeigt eine Reihe von Lucas' Kreationen ist besonders Spaß. Ich gebe diesem Film eine 7/10. Betrachtet man ihn subjektiv, es ist kein wirklich guter Film. Eine Menge Klischees, schlechter Dialog, Cracker Jack Philosophie und unimpressives Schauspiel. Doch aus irgendeinem Grund finde ich es schwer, etwas Schlechtes über diesen Film zu sagen. Mein Lieblingsfilm als Kind, und einer, den ich immer noch genieße, vor allem wegen Nostalgie.

Za potrebe ovog rada dovoljno je da se prikaže da model kreira tekst u odgovorajućem jeziku, a za tačnost prevoda bi ipak trebalo da se angažuju stručnjaci koji poznaju ova dva jezika, kako bi smo mogli da kvantifikujemo uspešnost prevodenja.

2.6. Generisanje teksta (Text Generation)

Za generisanje teksta, modeli poput GPT-3 (Generative Pretrained Transformer 3) koriste dekoderski pristup transformera kako bi kreirali tečan i koherentan tekst. Generisanje se oslanja na uzastopnu predikciju tokena na osnovu datog ulaza [8].

Generisanje teksta može da se koristi u slučajevima kada je potrebno odgovoriti na primedbe i zahteve kupaca ili korisnika nekih usluga. Za potrebe ovog rada, prikazano je generisanje teksta koje predstavlja odgovor režisera filma Star Wars našem nezadovoljnom recenzentu. Kôd za generisanje teksta u Hugging face transformeru:

```
from transformers import pipeline
generator = pipeline("text-generation")
odgovor = "Dear viewer, I am sorry to hear that you find that Star wars movies isn't a really good film."
prompt = "\n\nDirector's response:\n" + odgovor
rezultat = generator(prompt, max_length=400)
print(outputs[0]['generated_text'])
```

Dobijeni rezultat se sastoji od teksta recenzije i odgovora generatora:

Director's response:

Dear viewer, I am sorry to hear that you find that Star wars movies isn't a really good film. In fact, I think that I am quite frustrated by most of Star Wars movies. I was very pleasantly surprised and glad to see so many of them rereleased as films. While I wish for a more diverse and more mature Star Wars, I do think that the movies as a whole can still continue making wonderful movies that do not quite get off the ground. Although, Star Wars is a long season in itself, so it might be a good idea to watch more movies about this sci-fi fad.

In my opinion, Star Wars doesn't deserve an 8/10. For example, Star Wars: The Clone Wars didn't receive a 7/10. Even more, Star Wars: Return of the Jedi, though not universally regarded, didn't

get a 9/10. *Star Wars: The Force Awakens* doesn't even manage to get a 6/10 in my mind's eye. And even *Star Wars: The Lord of the Rings* is just not worth its 8/10. Even *Star Wars: The Force Awakens*, however, is

Odgovor smo ograničili na 400 karaktera, ali ovaj parametar može da se menja i prilagođava potrebama. Na osnovu dobijenog rezultata možemo da vidimo da generisani odgovor sadrži još podataka iz serijala *Star Wars* koji se ne nalaze u ulaznoj recenziji. U odgovoru se pojavila i jedna greška, naime kao jedan od filmova iz ovog serijala se spominje i *Star Wars: The Lord of the Rings* što naravno nije tačno. Model je bio “zbunjen” jer postoji knjiga *Star Wars: Lords of the Sith*.

3. ZAKLJUČAK

Kroz praktične implementacije u sklopu framework-a TensorFlow i biblioteka Hugging Face Transformers, prikazano je da transformeri obezbeđuju visoke performanse u svim navedenim zadacima. Na primer, model za analizu sentimenta je mogao da identifikuje složene emocionalne nijanse, dok je NER postigao visok nivo preciznosti u prepoznavanju entiteta. Odgovaranje na pitanja je pokazalo sposobnost modela da pronađe relevantne delove teksta, dok su sažimanje i prevod pokazali učinkovitost u redukciji sadržaja i reprezentaciji jezika. Generisanje teksta, iako pokazalo ograničenja (npr. generisanje neispravnih informacija o filmovima), ukazuje na potencijal transformera za kreativne i interaktivne primene.

Buduće istraživanje bi trebalo da se fokusira na poboljšanje modela za specifične domene, povećanje tačnosti u generisanju teksta kako bi se smanjile greške i povećala pouzdanost rešenja. Ovaj rad takođe pokazuje da nezavisno od tehnoloških ograničenja, transformeri imaju ogroman potencijal za primenu u širokom spektru industrijskih, akademskih i komercijalnih scenarija.

4. LITERATURA

- [1] Vaswani, A; Shazeer, N; Parmar, N; Uszkoreit, J; Jones, L; Gomez, A; Kaiser, L; Polosukhin, I.: *Attention is all you need*, Advances in neural information processing systems 30 (2017)
- [2] Tunstall, V; Werra, L; Wolf, T: *Natural language processing with transformers*, O'Reilly Media, 2022, isbn: 9781098103248
- [3] Devli, J; Chang, M; Lee, K; Toutanova, K: *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2019, arxiv: 1810.04805
- [4] Liu, Y; Ott, M; Goyal, N; Du, J; Joshi, M; Chen, D; Levy, O; Lewis, M; Zettlemoyer, L, Stoyanov, V: *RoBERTa: A Robustly Optimized BERT Pretraining Approach*, 2019, arxiv: 1907.11692
- [5] Raffel, C; Shazeer, N; Roberts, A; Lee, K; Narang, S; Matena, M; Zhou, Y; Li, W; Liu, P: *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*, 2020, arxiv: 1910.10683
- [6] Lewis, M; Liu, Y, Goyal, N; Ghazvininejad, M; Mohamed, A; Levy, O; Stojanov, V; Zettlemoyer, L: *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*, 2020, arxiv: 1910.13461
- [7] Tang, Y; Tran, C; Li, X; Chen, P; Goyal, N; Chaudhary, V; Gu, J; Fan, A: *Multilingual Translation with Extensible Multilingual Pretraining and Fine-Tuning*, 2020, arxiv: 2008.00401
- [8] Brown, T; Mann, B; Ryder, N; Subbiah, M: *Language Models are Few-Shot Learners*, 2020, arxiv: 2005.14165