Obrada prirodnog jezika korištenjem velikih jezičnih modela

Putić, Marko

SVEUČILIŠTE U RIJECI
TEHNIČKI FAKULTET
Prijediplomski studij računarstva

Završni rad

Obrada prirodnog jezika korištenjem velikih jezičnih modela

Rijeka, rujan 2024.

Marko Putić
0069088085

SVEUČILIŠTE U RIJECI
TEHNIČKI FAKULTET
Prijediplomski studij računarstva

Završni rad

Obrada prirodnog jezika korištenjem velikih jezičnih modela

Mentor: izv.prof.dr.sc. Goran Mauša

Rijeka, rujan 2024.

Marko Putić
0069088085

SVEUČILIŠTE U RIJECI
TEHNIČKI FAKULTET
POVJERENSTVO ZA ZAVRŠNE ISPITE

Rijeka, 17.03.2024.

Zavod: Zavod za računarstvo

Predmet: Programsko inženjerstvo

ZADATAK ZA ZAVRŠNI RAD

Pristupnik: **Marko Putić (0069088085)**

Studij: Sveučilišni prijediplomski studij računarstva (1035)

Zadatak: **Obrada prirodnog jezika korištenjem velikih jezičnih modela / Natural language processing using large language models**

Opis zadatka:

Istražiti modele dubokog učenja koji se koriste za obradu prirodnog jezika. Analizirati arhitekturu i principe rada postojećih modela te opisati postupak kojim jezični modeli razlučuju važne podatke. Proučiti način funkcioniranja postojećih aplikacija koje u svom radu koriste prethodno istrenirane jezične modele. Istražiti hardverska i softverska rješenja razvijena za potrebe jezičnih modela te povećanje njihove učinkovitosti.

Rad mora biti napisan prema Uputama za pisanja diplomskih / završnih radova koje su objavljene na mrežnim stranicama studija.

Zadatak uručen pristupniku: 20.03.2024.

Mentor:
izv. prof. Goran Mauša

Predsjednik povjerenstva za
završni ispit:
prof. dr. sc. Miroslav Joler

Student: Marko Putić

Studijski program: Prijediplomski studij računarstva

JMBAG: 0069088085

IZJAVA O SAMOSTALNOJ IZRADI ZAVRŠNOG RADA

kojom izjavljujem da sam završni rad s naslovom Obrada prirodnog jezika korištenjem velikih jezičnih modela

izradio samostalno pod mentorstvom izv. prof. dr. sc. Goran Mauša

U radu sam primijenio metodologiju izrade stručnog/znanstvenog rada i koristio literaturu koja je navedena na kraju završnog rada. Tuđe spoznaje, stavove, zaključke, teorije i zakonitosti koje sam izravno ili parafrazirajući naveo u završnom radu na uobičajen, standardan način citirao sam i povezao s fusnotama i korištenim bibliografskim jedinicama te nijedan dio rada ne krši bilo čija autorska prava. Rad je pisan u duhu hrvatskoga jezika.

Student  (potpis)

Marko Putić

Rijeka, srpanj 2024.

Sadržaj

Popis slika

Poglavlje 1

Uvod

Računalna obrada jezika kojim se svakodnevno služimo u govoru i pismu zastupljena je u brojnim tehnologijama s kojima se susrećemo koristeći pametne telefone, računala i ostale pametne uređaje. Ova vrsta računalne obrade jezika funkcionira u oba smjera. Moguće je da računalo bilježi, interpretira i obrađuje ljudski jezik, ali isto tako računalo mora biti u mogućnosti korisniku pružiti prikladan odgovor ili rješenje u obliku u kojem će ga čovjek moći razumjeti. Ovaj kompleksan koncept obrade prirodnog jezika veliki uzlet i poboljšanje dobio je upravo korištenjem velike količine dostupnih tekstualnih podataka u kombinaciji s naprednim arhitekturama dubokog učenja. Ovaj rad će obrađivati razvoj, arhitekturu i primjenu jezičnih modela uz analizu i proučavanje postojećih alata koji se služe istima. Konkretno, obrađivat će se povratne neuronske mreže i model Transformera koji predstavlja temelj velikih jezičnih modela. Razradit će se i dva modela obrade prirodnog jezika - ChatGPT[3] i BERT[6]. Zbog velike raširenosti i potrebe za što pravilnijim radom modela, razvijena su brojna hardverska i softverska rješenja koja pospješuju brzinu izvršavanja zadataka te energetsku učinkovitost sustava u cjelini. Ispitat će se prednosti grafičkih procesora (eng. Graphics Processing Unit) i jedinica za obradu tenzora (eng. Tensor Processing Unit) te njihov način rada i primjena. Sa sigurnošću se može tvrditi kako se razvojem tehnologija koje se bave obradom prirodnog jezika otvorila potpuno nova mogućnost i iskustvo komunikacije čovjeka i računala.

Poglavlje 2

Osnovni pojmovi i arhitektura

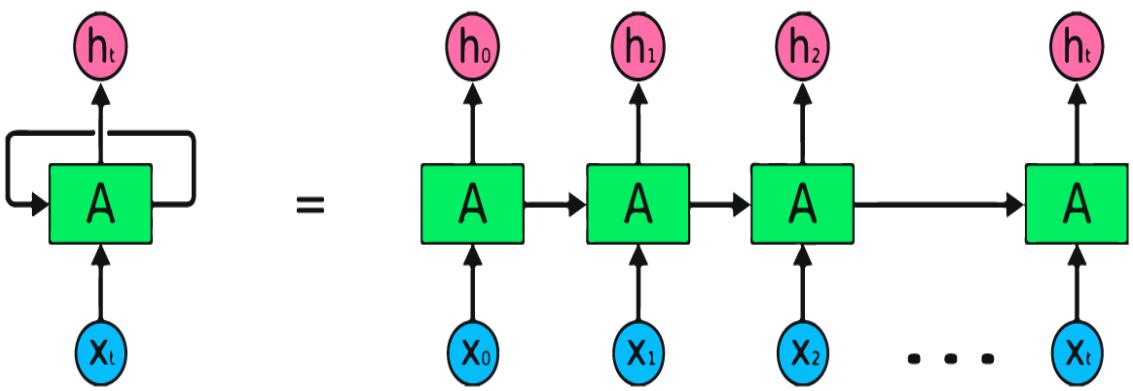
Veliki jezični modeli su tipovi strojnog učenja istrenirani tako da razumiju, generiraju i analiziraju ljudski jezik. Ovi modeli su većinom bazirani na arhitekturi neuronskih mreža i metodama dubokog učenja, a uče se na velikom broju tekstualnih podataka. Jedna od uloga velikih jezičnih modela je generiranje međuzavisnosti u riječima prirodnog jezika što im omogućava razumijevanje i stvaranje teksta u semantičkom i sintaktičkom duhu jezika traženog od strane korisnika[1]. Ovo će poglavlje obrađivati temeljne modele dubokog učenja koji čine bazu velikih jezičnih modela. Bit će objašnjene povratne neuronske mreže i modeli Transformera te će se analizirati radi čega su upravo modeli Transformera prikladniji za proces obrade prirodnog jezika.

2.1 Uvod

Transformeri su relativno novi modeli neuronskih mreža koji su se pokazali optimalnim u procesima obrade prirodnog jezika. Predstavljaju značajan napredak u učinkovitosti ovih procesa u usporedbi s povratnim neuronskim mrežama (eng. Recurrent Neural Networks). Za lakše shvaćanje Transformera valja opisati način rada povratnih neuronskih mreža kako bi se razlučile temeljne razlike u arhitekturi ovih dvaju modela.

2.1.1 Povratne neuronske mreže

Povratne neuronske mreže su tip umjetnih neuronskih mreža (eng. Artificial Neural Network) koja se sastoji od ulaznog, izlaznog sloja te od n skrivenih slojeva u kojima je svaka sljedeća vrijednost ovisna o elementima niza koji joj prethode. Ove se mreže koriste za obradu sekvencijalnih podataka, što znači da svaki čvor u skrivenom sloju ne prima isključivo vrijednost prethodnog sloja već vrijednosti svih prošlih skrivenih stanja[5]. Slika 2.1 jednostavan je prikaz povratnih neuronskih mreža. Neka je dan primjer predviđanja cijena nafte korištenjem jednostavnih podataka u obliku cijelog broja (cijene barela nafte). Tako će svaki unos od X_0 do X_t sadržavati prošlu vrijednost (cijenu), gdje će na primjer X_0 imati vrijednost 82, a X_1 85 te će se obje vrijednosti koristiti za predviđanje svakog sljedećeg broja u nizu. Ulazni sloj X obrađuje početni ulaz i prosljeđuje ga srednjem sloju A. Srednji sloj se sastoji od više skrivenih slojeva od kojeg svaki sadrži aktivacijske funkcije, težinske vrijednosti i pristranosti. Ti se parametri standardiziraju kroz skriveni sloj tako da se umjesto kreiranja više skrivenih slojeva, kreira jedan koji ima svrhu petlje u toku podataka[5].



Slika 2.1 RNN (preuzeto iz [5])

2.2 Transformeri

Ideja modela Transformera prvi put je predstavljena 2017. godine od strane Googlea u znanstvenom članku *Attention Is All You Need*[6]. Ovaj koncept je ubrzo odjeknuo i postao temelj suvremenih velikih jezičnih modela poput ChatGPTa[3], BERTa[6], Claudea[3][6] i drugih. Pored toga što je ovo relativno nov i inovativan način, postavlja se pitanje zašto implementirati drukčiji model ako već postoje prethodno spomenute povratne neuronske mreže. Razlog tome je što one nisu u potpunosti prikladne za razumijevanje i obradu prirodnog jezika zbog neoptimalnog detektiranja međuzavisnosti podataka, npr. riječi unutar rečenice. Do toga dolazi radi izostanka korištenja dostupnih resursa grafičkih procesora koji su dizajnirani za paralelnu obradu podataka te neučinkovitosti u slučajevima kada su dva elementa unutar teksta međusobno udaljena. Tada se gubi veza tih dvaju elemenata te ju model u daljnjem računanju više ne uzima u obzir. Ovaj problem rješavaju i slojevi LSTM (eng. Long Short-Term Memory) unutar povratnih neuronskih mreža, no u usporedbi s modelima Transformera, imaju slabiji učinak. To se očituje u činjenici da se kod LTSM tipa podaci obrađuju sekvencijalno, a kod Transformera paralelno što rezultira bržim vremenom izvršavanja. Nadalje, ključnu ulogu u brzini i točnosti imaju upravo mehanizmi pažnje (eng. attention mechanisms), koji u slučaju povratnih neuronskih mreža u potpunosti izostaju. Pojava Transformera riješila je dva ključna problema u metodama obrade prirodnog jezika eliminiravši utjecaj udaljenosti pojedinih elemenata na kontekst te sveukupno je doprinijela povećanju brzine obrade velikih količina podataka.

2.2.1 Arhitektura Transformera

Arhitektura modela Transformera može se podijeliti na 4 ključna dijela ili koraka[7]:

- tokenizacija (eng. tokenization),
- ugrađivanje (eng. embedding),
- pozicijsko kodiranje (eng. positional encoding),
- predviđanje sljedeće riječi.

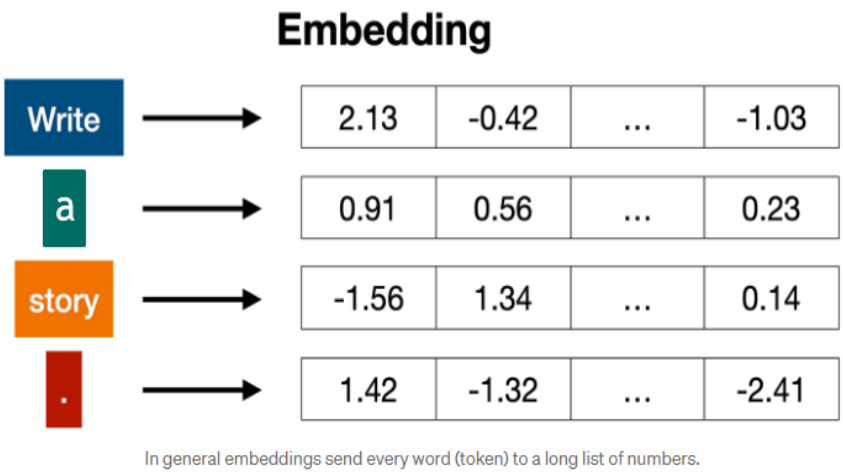
Tokenizacija

Tokenizacija je primarni korak u kojem dolazi do podjele teksta u manje jedinice, tzv. tokene koji mogu predstavljati riječi, slova, interpunkcijske znakove, razmake, itd. Korak tokenizacije uzorkuje svaku riječ, njene elemente i okolne znakove (npr. interpunkcijske ili razmake) te ih šalje k poznatom tokenu unutar dostupne knjižnice. Slika 2.2 prikazuje prethodno opisanu podjelu rečenice na manje jedinice. Radi jasnog razlučivanja, svaki zasebni element rečenice smješten je u pravokutnike različitih boja. Kao takvi će se pojavljivati i u idućim koracima.

Slika 2.2 Tokenizacija (preuzeto iz [7])

Ugrađivanje

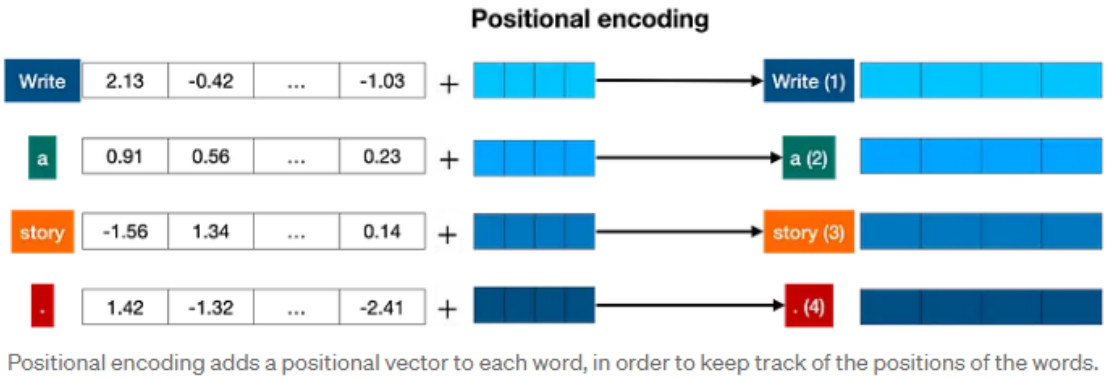
Nakon što je ulaz tokeniziran, riječi treba pretvoriti u brojeve. U procesu ugrađivanja svaki dio teksta šalje se u polje brojeva, što je vidljivo na slici 2.3. Ako su dva dijela teksta slična, tada su i brojevi u njihovim odgovarajućim poljima podudarni (vrijedi po komponentama rečenice, što znači da svaki par brojeva na istoj poziciji je podudaran)[7]. Analogno tome, ako se dva dijela teksta razlikuju, tada se brojčane vrijednosti unutar odgovarajućih polja razlikuju.



Slika 2.3 Ugrađivanje (preuzeto iz [7])

Pozicijsko kodiranje

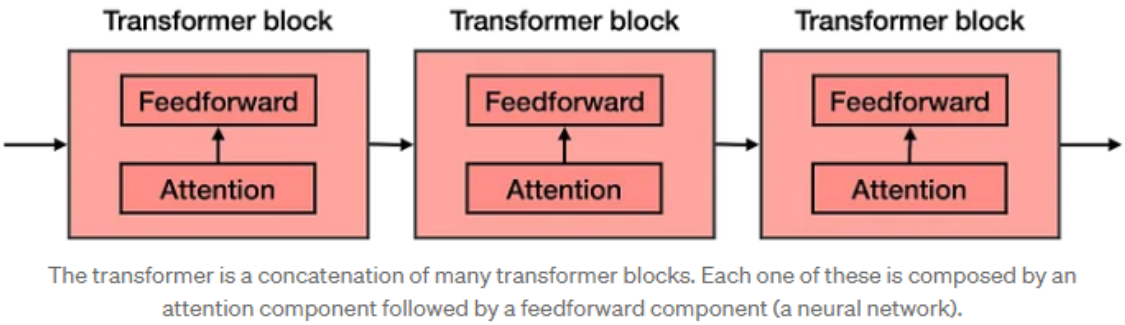
U ovome koraku potrebno je prethodno kreirana polja koja sadrže brojčane vrijednosti pretvoriti u jedinstveno polje spremno za obradu. Najčešći i najjednostavniji način pretvaranja mnoštva vektora u jedan je njihovo zbrajanje. Međutim, u nekim jezicima primjenom navedenog u ovom slučaju može se naići na problem kod rečenica u kojima zamjena dviju riječi istoj daju potpuno drugo značenje, dok zbog komutativnosti zbrajanja dobivamo identičan rezultantni vektor. Jedno od mogućih rješenja ovog problema je upravo pozicijsko kodiranje. Pozicijsko kodiranje sastoji se od dodavanja niza unaprijed definiranih vektora vektorima dobivenim u koraku ugrađivanja. Potonje osigurava da svako ugrađivanje riječi razumije njeno značenje, ali i da ima informaciju o njenom položaj u nizu. Upravo to omogućuje modelu da razlikuje rečenice koje sadrže iste riječi napisane različitim redoslijedom pošto će izlazne vrijednosti pozicijskog kodiranja varirati s obzirom na opisane različitosti unutar rečenice. U primjeru prikazanom na Slici 2.4 vektori koji odgovaraju riječima *"Write"*, *"a"*, *"story"* i *"."* postaju vektori koji u sebi nose podatak o njihovom položaju u rečenici te budu označeni kao *"Write (1)"*, *"a (2)"*, *"story (3)"* i *". (4)"*. Naposljetku, dobiveno je ono što je zapravo i cilj ovog koraka, a to je da ovakvi modificirani vektori nose informaciju i o samoj riječi i o njenoj poziciji unutar rečenice.



Slika 2.4 Pozicijsko kodiranje (preuzeto iz [7])

Predviđanje sljedeće riječi

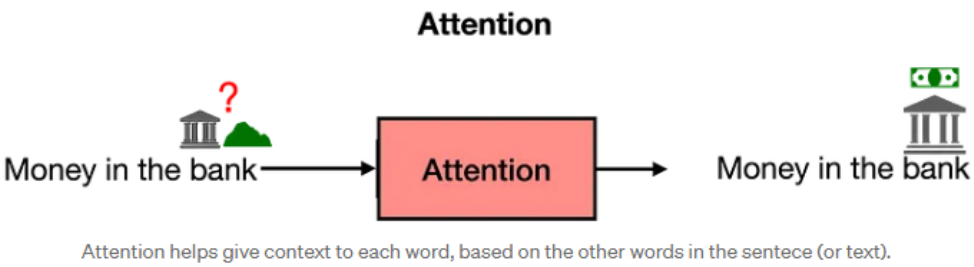
Idući korak u obrađenom slučaju je predviđanje sljedeće riječi u rečenici. To se radi uz pomoć velikih neuronskih mreža koje su istrenirane specijalno za željenu namjenu, a u ovom slučaju je to upravo predviđanje sljedećih riječi u rečenici. Takva neuronska mreža se može naučiti, no njena učinkovitost se znatno može unaprijediti dodavanjem mehanizma pažnje. Taj mehanizam, je glavni razlog zašto su Transformeri toliko uspješna i prihvaćena tehnologija modernog vremena[6].
Komponenta mehanizma pažnje ugrađuje se u svaki blok unaprijedne mreže. Dakle, ako se radi o prethodno spomenutom slučaju predviđanja riječi velika unaprijedna neuronska mreža sastojat će se od više blokova manjih neuronskih mreža. Komponenta pažnje dodavat će se svakom od tih manjih blokova. Naposljetku, svaki će se blok Transformera sastojati od 2 ključne komponente: Komponenta pažnje i unaprijedna (eng. feedforward) komponenta, što je prikazano na slici 2.5.



Slika 2.5 Blok Transformera (preuzeto iz [7])

2.3 Mehanizmi pažnje

Ključna komponenta modela Transformera su mehanizmi pažnje (eng. attention mechanisms). Ovi mehanizmi rješavaju glavni problem u razumijevanju ljudskog jezika, a to je kontekst riječi unutar rečenice. Ponekad jedna te ista riječ može imati različito značenje s obzirom na način uporabe u rečenici. Razlikovanje značenja te riječi model ne stječe u koraku ugrađivanja, već mu je za ovu vrstu optimizacije potreban upravo mehanizam pažnje. Neka je dan primjer na engleskom jeziku kako bi se zorno mogao prikazati problem kojeg rješavaju mehanizmi pažnje. Obrađuju se dvije rečenice: *"The bank of the river"* i *"Money in the bank"*. U obje rečenice koristi se riječ *"bank"*, no u svakoj od njih ona poprima različito značenje. U prvoj rečenici ta se riječ odnosi na nasip uz obalu rijeke, dok u drugoj ima značenje institucije u kojoj se čuva novac. Ono što je korisno, i pomaže u stvaranju konteksta, su upravo preostale riječi unutar rečenice. Tako na primjer u prvoj rečenici stavljanju u kontekst riječi *"bank"* najviše će koristiti riječ *"river"* jer ta riječ indicira kako se tu radi o nasipu/obali rijeke. U drugoj rečenici je ta *najkorisnija* riječ *"money"* koja u ovom slučaju ukazuje kako bi se moglo raditi o banci. Slika 2.6 prikazuje objašnjeni postupak kojim model putem mehanizma pažnje stvara kontekst i omogućuje da rečenica ima semantički smisao. Zaključno, mehanizmi pažnje zapravo približavaju ključne riječi u koraku ugrađivanja. Isto će se dogoditi i u danom primjeru. Riječ *"river"* će dati značenje riječi *"bank"*, dok će u drugoj rečenici riječ *"money"* biti ključna za stavljanje željene riječi u kontekst.



Slika 2.6 Pažnja (eng. attention) (preuzeto iz [7])

Poglavlje 3

Aplikacije zasnovane na jezičnim modelima

Nakon modela učenja obrađenih u prethodnom poglavlju, slijede konkretni primjeri alata koji se koriste u obradi prirodnog jezika. Bit će obrađeni ChatGPT i BERT, od kojih svaki koristi već spomenutu arhitekturu Transformera. Analizirat će se njihov način rada, primjena te prednosti i nedostaci. OpenAI je ChatGPT predstavio kao funkcionalnu aplikaciju koja pruža korisniku informacije u gotovo svim područjima ljudske djelatnosti te je primjenjiv kod najrazličitijih zahtjeva korisnika. S druge strane, BERT ne nudi konkretnu aplikaciju kojom se prosječni korisnik može služiti, no odličan je alat za korisnika koji zna što želi i na koji specifičan način želi prilagoditi model sebi i svojim potrebama. Tako Google tvrdi da korisnici mogu istrenirati funkcionalan sustav pitanja i odgovora u samo 30 minuta na *cloud TPU* (eng. Tensor Processing Unit) te u nekoliko sati koristeći grafičku karticu (eng. Graphics Processing Unit)[15, 16]. Detaljnije tehničke specifičnosti ovih dvaju modela bit će obrađene u nastavku.

3.1 ChatGPT

ChatGPT (Generative Pre-trained Transformer) je generativni program umjetne inteligencije koji koristi modele obrade prirodnog jezika za razumijevanje i kreiranje

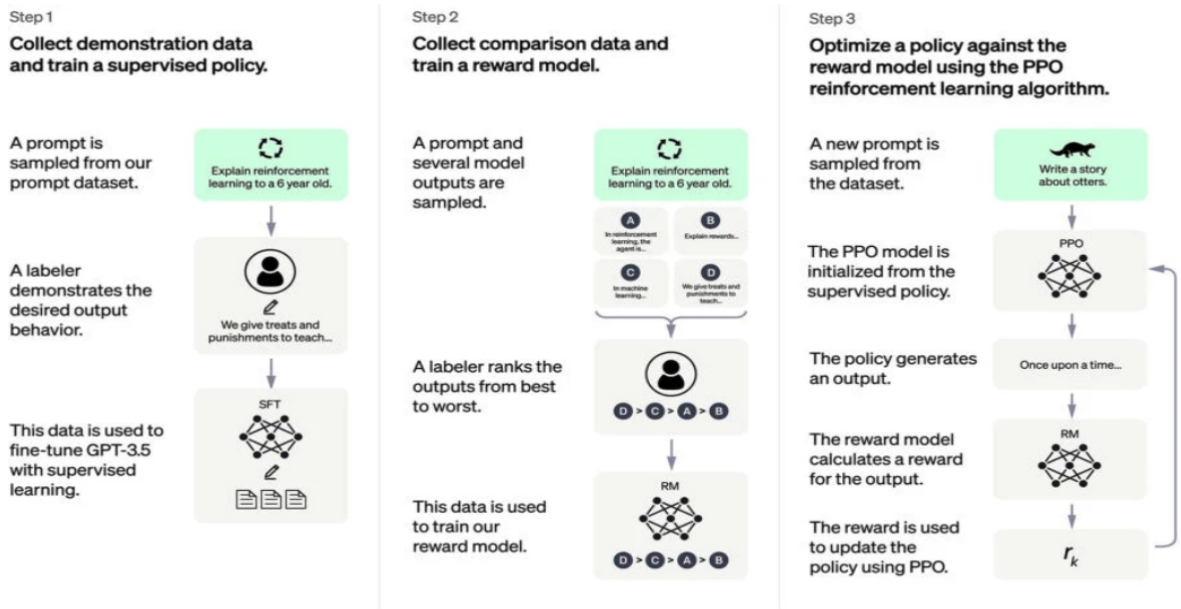
teksta razumljivog čovjeku. Način korištenja ovog programa nalik je automatiziranim sustavima podrške raznih tvrtki. Program je predstavljen široj javnosti u studenom 2022. godine od strane organizacije OpenAI koja se bavi istraživanjem umjetne inteligencije. Sama tvrtka svojim radom je započela u prosincu 2015. godine, a među grupom poduzetnika i istraživača bili su i Elon Musk i Sam Altman[9]. Pored svog primarnog alata razvili su i Dall-E, program pogonjen umjetnom inteligencijom koji služi za generiranje umjetnina iz zadanog teksta. Ovaj alat može razumjeti jezik u govoru i pismu što mu omogućuje da obradi ulazne te generira odgovarajuće izlazne podatke u tekstualnom, slikovnom ili zvučnom obliku. ChatGPT također može: odgovarati na pitanja neovisno o temi, rješavati matematičke zadatke, prevoditi tekst s jednog jezika na drugi, otkloniti pogreške u priloženom programskom kodu, napisati sastavak, priču, pjesmu i druge standardizirane oblike teksta te još mnogo toga[8]. Problem na kojeg se može naići tijekom korištenja ovog alata jest mogućnost dobivanja netočnih odgovora. Razlog tome je taj što ChatGPT nije direktno spojen s Internetom pa tako u određenim slučajevima može pružiti pogrešan odgovor iz razloga što podaci nisu ažurirani i prilagođeni trenutnima. Također ima ograničeno znanje o svijetu i događajima nakon rujna 2021. godine te povremeno može generirati štetne upute ili pristrani sadržaj[10]. Iako je učenje svih trenutno dostupnih GPT modela završilo 2021. godine, kontinuirano se razvijaju novi modeli(npr. GPT-4) koji unaprjeđuju razumijevanje korisničkih upita i glatkoću generiranja odgovora. Pored toga, iako u radu ne koriste vezu s Internetom, njime se mogu koristiti pri specifičnim upitima koji su vezani za informacije o svijetu nakon vremena učenja modela (rujan 2021. godine).

3.1.1 Kako funkcionira ChatGPT?

ChatGPT funkcionira korištenjem naučenog Transformer modela koji je opisan u poglavlju 2. Jedna od ključnih značajki ovog alata je visoka razina učenosti pozadinskih modela koji omogućuju generiranje odgovora korisniku u takoreći realnom vremenu. Ovi se modeli uče upravo na interakciji s čovjekom kroz ojačano učenje putem povratnih informacija koje dobivaju od korisnika te na temelju modela nagrađivanja koji rangiraju najbolje odgovore. Pored toga što korisnici pasivno, sa-

mim korištenjem doprinose poboljšanju modela, oni mogu ocijeniti odgovor te se izjasniti je li isti bio koristan ili ne. Također, ako se razgovor nastavlja i korisnik riječima opisuje s kojim segmentima odgovora je zadovoljan, a s kojim nije, to dodatno poboljšava rad modela. Na Slici 3.1 koju objavljuje sam kreator OpenAI, nastoji se sažeto prikazati način rada mehanizma nagrađivanja implementiranog u rad aplikacije. U prvom koraku se uzorkuje zahtjev u obliku pitanja na kojeg čovjek daje željeno izlazno ponašanje (eng. output behavior). Ti se podaci koriste za precizno prilagođavanje modela nadziranim učenjem (eng. supervised learning). U sljedećem koraku uzorkuju se i zahtjev i izlazi (odgovori) koje čovjek ocjenjuje od najboljeg prema najgorem, a takvi podaci služe za učenje modela nagrađivanja. U posljednjem koraku zadaje se novi upit, gdje model generira odgovor koristeći dosad stečene podatke, a izračunata nagrada za taj odgovor koristi se za ažuriranje politike učenja modela.

ChatGPT je izvorno koristio veliki jezični model GPT-3, model strojnog učenja neuronskim mrežama i treću generaciju generativnog prethodno istreniranog Transformera (eng. Generative Pre-trained Transformer)[9]. Aplikacija ChatGPT trenutno u besplatnoj verziji daje pristup GPTu 3.5 koji je treniran na 175 milijardi parametara[11] i ograničen pristup GPTu 4o i 4o mini jezičnom modelu. Noviji GPT 4 može izvršiti kompleksnije zadatke u usporedbi s verzijom 3.5 kao što je opisivanje slikovnih podataka i generiranje detaljnijih odgovora do 25000 riječi. Taj model se trenutno naplaćuje i predstavlja najrazvijeniji alat tvrtke OpenAI[9].



Slika 3.1 Mehanizam nagrađivanja kojeg koristi ChatGPT (preuzeto iz [8])

3.1.2 Prednosti i nedostaci korištenja alata

Pojava umjetne inteligencije u formi jedinstvenog alata namijenjenog širokoj uporabi dovela je do brojnih podjela u mišljenjima ljudi te se počelo govoriti o raznim prednostima, ali i nedostacima i mogućim prijetnjama koje donosi ovakav razvitak umjetne inteligencije.
Neke od prednosti dostupnosti ovakvog alata su:

- Učinkovitost - umjetna inteligencija se može nositi s rutinskim zadacima obrade podataka za koje je inače potrebna ljudska inteligencija i napor, samim time se oslobađaju vrijeme, ljudski i tehnološki resursi za odrađivanje složenijih poslova,
- Ušteda financijskih sredstava - ako postoji aplikacija koja može jednako dobro ili bolje odrađivati određeni zadatak ili posao onda nema potrebe za zapošljavanjem dodatnih kadrova za to polje djelovanja,
- Kvaliteta sadržaja - s obzirom na to da ChatGPT raspolaže gomilom tekstualnih podataka na brojnim jezicima te poznaje jezične norme, korisnik može od

njega zatražiti da obogati priloženi tekst, smjesti ga u kontekst ili da pak samo ispravi gramatičke pogreške,

- Vrijeme odziva i dostupnost - AI modeli su dostupni 24 sata dnevno te mogu praktički bez prestanka odrađivati zadane zadatke,
- Pristupačnost za osobe s invaliditetom - unos naredbi i pitanja govorom prema ChatGPT-u ili srodnom alatu pomaže osobi da na lakši način i gotovo trenutno dođe do željenog odgovora putem sučelja kojim se je moguće lakše služiti nego nekim drugim alatima koji su namijenjeni isključivo osobama s invaliditetom.

Može se zaključiti kako novi alati pogonjeni umjetnom inteligencijom pružaju razne mogućnosti svim kategorijama korisnika te mogu znatno olakšati posao čovjeku. Međutim, kako ni čovjek ni alati koje je stvorio nisu savršeni u svom djelovanju tako se i glede ove teme mogu izdvojiti neki nedostaci:

- Složenost ljudskog jezika - ChatGPT je prilagođen da generira tekst na temelju unosa korisnika, u tom slučaju zbog kompleksnosti čovjekova jezičnog izražavanja odgovori se ponekad mogu činiti površnima i s nedostatkom željenog konteksta ili s druge strane potpuno neprirodnima iz očitog razloga što odgovore ne formira čovjek, već stroj,
- Neprovjerljivost pruženih podataka - odgovori jesu relevantni i sadrže većinom ispravne podatke, no u slučaju sumnje na istinitost istih, ne znamo otkuda je alat pribavio podatke te ne možemo temeljito provjeriti konkretan izvor,
- Naglasak na krivi pojam - alat ponekad može krivo shvatiti što korisnik traži od njega te se u tom slučaju zadržava na objašnjenju dijela koji nije primarni interes korisnika.

Naposljetku, pored opisanih razloga za i protiv služenja ovim alatom s tehnološkog i aspekta svakodnevnog korištenja pojavili su se i neki moralni problemi. Jedan od ključnih problema koje valja navesti u samom početku je pojava plagijata i zlouporaba. Navedeno se najviše može očitovati u obrazovanju i poslovnom svijetu gdje čovjek koristan alat može zlouporabiti kako bi si olakšao posao ili čak u potpunosti riješio neki zadatak nesamostalno. U svrhu rješavanja spomenutog problema razvijeni su alati kojima je moguće provjeriti je li sporni tekst generirala umjetna inteligencija

ili je on proizvod ljudske inteligencije. Testiranjem je ustanovljeno kako takvi alati nisu u potpunosti točni, no svakako su korisni za dodatnu provjeru ako se sumnja na korištenje umjetne inteligencije u situacijama gdje to nije dopušteno. Razlog tome su sličnosti ljudskog jezika i jezika naprednih jezičnih modela, ali i kontinuirana prilagodba modela koja rezultira generiranjem teksta koji sve više nalikuje ljudskom načinu izražavanja. Pored toga, uvijek se postavlja pitanje privatnosti i koliko su podaci koje dijelimo s alatom zaštićeni od trećih strana. Poznato je da aplikacija bilježi i uči iz razgovora s korisnicima i da se ti podaci čuvaju te postoji mogućnost da ChatGPT otkrije osjetljive i osobne informacije korisnika. Tako je zbog neusklađenosti rada aplikacije s europskim propisima o zaštiti privatnosti, 2023. godine OpenAI napravio izmjene sukladne spomenutim propisima[9]. Pored spomenutog, uvijek se nameće možda i glavna rasprava o tome koliko je razvoj umjetne inteligencije benevolentan prema čovjeku. Uvijek je prisutan strah kako bi umjetna inteligencija i alati koji su njom pogonjeni mogli trajno zamijeniti čovjeka u obavljanju određenih poslova. Tako bi mogli nestati ljudski oblici pružanja korisničke podrške, obrada i unos podataka, prevođenje i drugi slični poslovi. Od velike je važnosti da se očuva ljudsko djelovanje i kontrola nad uporabom umjetne inteligencije te da se ona koristi u savjetodavne svrhe i kao izvor novih ideja koje će naposljetku, uz temeljito prosuđivanje, u zbilju sprovesti čovjek.

3.2 BERT

BERT (Bidirectional Encoder Representations from Transformers) jezični model je program otvorenog koda koji služi za obradu prirodnog jezika. Predstavljen je 2018. godine od strane Googlea. Ovaj model je dizajniran da pomogne računalu razumjeti značenje dvosmislenog teksta stavljajući ga u ispravan kontekst uz pomoć okolnih riječi. BERT je prvobitno treniran podacima s Wikipedije, a precizno prilagođavanje (eng. fine tuning) je odrađeno na temelju skupova podataka pitanja i odgovora (eng. Q&A)[12]. Ključna razlika između ChatGPTa i BERTa je što prvi služi generiranju teksta i jezičnoj interakciji čovjeka i računala, dok drugi se većinom koristi za analizu postojećeg teksta i klasifikaciju podataka te nije dizajniran da kreira tekst.

3.2.1 Kako funkcionira BERT?

Kao i ChatGPT, BERT je utemeljen na Transformerima čiji je princip rada opisan u prethodnom poglavlju. Ono što je povijesno gledajući razlikovalo BERT od ostalih velikih jezičnim modela je to što je dizajniran da čita u oba smjera odjednom. Ovo se postiglo uvođenjem tehnologije Transformera u način rada BERTa, čime je tvrtka Google koristeću vlastitu inovativnu tehnologiju kreirala možda i najnapredniji sustav obrade prirodnog jezika u to vrijeme. Naime do tada je modelima bilo moguće pratiti tekst samo u jednome smjeru. Uvođenje modela Transformatora omogućilo je dvosmjernost, čime su dvije metode obrade prirodnog jezika postale povezane. Radi se o modeliranju maskiranog jezika (eng. Masked Language Models) i predviđanju sljedeće rečenice (eng. Next Sentence Prediction)[9]. Svrha MLMa je da jezičnom modelu omogući dublje razumijevanje riječi na temelju konteksta kojeg daju susjedne riječi. Tako su određeni dijelovi rečenice skriveni, a model se zatim trenira da predvidi te skrivene elemente uz pomoć ostalih riječi u okolini. Na taj način model shvaća i bilježi kontekst i konkretno značenje riječi u obrađenom primjeru. S druge strane, NSP ima za cilj predvidjeti jesu li dvije rečenice značenjski povezane ili je njihova blizina nasumična. Nadalje, BERT počiva i na mehanizmima pažnje uz pomoć kojih model stječe bolje razumijevanje odnosa riječi u rečenici i njihovo konkretno značenje u različitim primjenama. Spomenuto je od značajne važnosti jer promatrana riječ može mijenjati značenje i kontekst tijekom razvijanja rečenice. Me-

hanizmom pažnje i dvosmjernim čitanjem mogućnost pogreške se smanjuje jer model eliminira učinak *s lijeva na desno* koji neutemeljeno usmjerava riječi k određenom značenju kako rečenica napreduje.

3.2.2 Korištenje i tipovi modela

Spomenuto je kako je BERT pogodan za analizu teksta i klasifikaciju. Tako je Google namijenio BERT za optimizaciju tumačenja upita korisnika pri pretraživanju. Pored toga, može se koristiti za zadatke razumijevanja teksta poput prevođenja, za pitanja i odgovore (eng. Q&A), analizu značenja, itd[17]. Uzevši u obzir da je BERT program otvorenog koda, svatko se može njime služiti i prilagoditi ga svojim potrebama, tzv. precizno prilagođavanje (eng. fine tuning). Neke od inačica koje su prilagodili korisnici su[9]:

- DocBERT je model za zadatke vezane uz klasifikaciju dokumenata.
- BioBERT služi za *rudarenje* tekstova vezanog uz biomedicinu.
- VideoBERT učenje temelji na videozapisima s YouTubea.
- SciBERT se koristi za analizu i pročišćavanje znanstvenog teksta.
- BERTić je prethodno treniran na 8 milijardi tokena s web stranica na hrvatskom, bosanskom, srpskom i crnogorskom jeziku. Prilagođen je za zadatke vezane uz analizu i klasifikaciju teksta na spomenutim jezicima[18][19].

Poglavlje 4

Hardverski ubrzivači u obradi prirodnog jezika

Iz prethodnog poglavlja može se zaključiti kako je obrada prirodnog jezika u velikom uzletu i da je široko primjenjiva u brojnim segmentima interakcije čovjeka i računala. Od chatbota koji uspješno vrše komunikaciju s najširim spektrom korisnika pa sve do strogo namjenskih sustava koji analiziraju i po nekoliko milijardi tokena (npr. čitavi repozitoriji radova sveučilišta, podaci s Wikipedije, baze znanstvenih radova na određenu temu, itd.). Sve navedeno iziskuje namjenske komponente visokih performansi koje će optimalno i u željenom vremenu vršiti obradu zadanih podataka. Tradicionalni računalni procesori (eng. Central Processing Unit), iako široke namjene, relativno su loš odabir za obradu prirodnog jezika te njihova uporaba u ovom specifičnom polju može dovesti do uskih grla (eng. bottleneck) i usporenog rada. Za ovu namjenu najprikladniji tipovi hardvera koji optimalno odrađuju spomenute zadatke su grafički procesori (eng. Graphics Processing Unit) i jedinice za obradu tenzora (eng. Tensor Processing Unit). Oba se mogu pohvaliti jedinstvenom arhitekturom koja, ovisno o željenom načinu uporabe, pomiče granice mogućeg u polju obrade prirodnog jezika. U nastavku će biti obrađene navedene komponente, njihove mogućnosti i područja primjene.

4.1 GPU ubrzivači

Grafički procesori prvotno su bili namijenjeni obradi grafičkih elemenata i slika, no zbog svoje arhitekture i prednosti paralelne obrade podataka postale su temeljna komponenta u obradi prirodnog jezika za zadatke kao što su npr. množenje i konvolucija matrica, koji su ključni za treniranje modela dubokog učenja[20]. Za razliku od procesora (CPU) koji imaju jednu ili manji broj jezgara (u širokoj potrošnji do 24 jezgre), grafičke kartice trenutno raspolažu tisućama manjih jezgara koje su namijenjene za paralelno rješavanje zadataka. Važno je nadodati kako procesori, ukoliko se koriste za obradu prirodnog jezika, a za to nisu podobni radi svoje arhitekture i performansi, pored problema s popunjenosti memorije, mogu koristiti znatno više energije u usporedbi sa za to specijaliziranim komponentama. Poznato je i kako su grafički procesori u treniranju modela dubokog učenja i do 10 puta učinkovitiji od sustava baziranim samo na CPU[22].
S ovim mogućnostima, grafički procesori postaju pravi izbor za modele poput konvolucijskih neuronskih mreža (eng. Convolutional Neural Networks) i povratnih neuronskih mreža (eng. Recurrent Neural Networks). Pored toga, i u polju obrade prirodnog jezika koriste se i za svoju primarnu namjenu, a to su zadaci obrade slika i videa, pogonjenje aplikacija računalnog vida i video analitika[21]. Grafički procesori su zbog svojih svojstava primjenu našli i u znanstvenim sektorima kod složenih simulacija i u istraživanjima u polju medicine, biotehnologije, itd[21]. Pored toga što se napredne grafičke kartice koriste i za poboljšanje vizualnih detalja računalnih igara, kod onih koje su namijenjene platformama za virtualnu stvarnost (eng. Virtual Reality), koriste se i za značajke vođene umjetnom inteligencijom koje obogaćuju iskustvo igranja[21]. U ovu svrhu kreirane su i knjižnice specijalizirane za ovaj tip obrade podataka kako bi se performanse hardvera maksimalno iskoristile. Neke od tih knjižnica su TensorFlow (Google), PyTorch (MetaAI), CUDA (NVIDIA), itd.

4.1.1 Ograničenja grafičkih procesora

Kao i kod svih komponenti, tako i grafički procesori imaju određena ograničenja i nedostatke. Možda i glavni problem koji se nameće je ograničena količina memorije. Grafičke kartice su dizajnirane da imaju točno određenu količinu memorije koja se

ne može nadograditi kao što je to slučaj kod radne memorije u računalima. Isto tako, korištenje GPUa za ovu specifičnu svrhu iziskuje poznavanje određenih specijaliziranih okvira kao što je to za grafičke kartice NVIDIA-e *framework* CUDA. Kao daljnji nedostaci mogu se izdvojiti cijena i potrošnja električne energije. Grafičke kartice snažnih specifikacija dolaze i s visokom cijenom što potencijalno predstavlja izazov manjim tvrtkama u usponu ili pojedinačnim korisnicima. Kada se govori o povećanoj potrošnji energije, to za posljedicu ima povećanje pogonskih troškova te zahtjeva ugradnju odgovarajućih rashladnih sustava kako bi komponente mogle optimalno odrađivati zadatke.
Može se zaključiti kako su grafički procesori prikladni i visoko specijalizirani za obradu prirodnog jezika, no ipak iziskuju dobro poznavanje njihovih mogućnosti i ograničenja te predviđanje mogućih poteškoća i strateško planiranje pri višekorisničkom korištenju.

4.2 TPU akceleratori

Jedinice za obradu tenzora spadaju u integrirane sklopove za specifične namjene (eng. Application specific integrated circuit) kreirane od strane tvrtke Google s posebnom namjenom ubrzanja sustava strojnog učenja. Ovu vrstu hardvera nije moguće nabaviti za privatnu ili komercijalnu uporabu, već se usluga korištenja resursa TPUa pruža preko servisa u oblaku (eng. cloud service) od strane Googlea. Svojom posebnom arhitekturom TPU su posebno učinkoviti za rad s konvolucijskim neuronskim mrežama i modelima Transformera. Primjenjivi su u brojnim oblicima obrade prirodnog jezika od prijevoda, analize, obrade pa sve do potpuno funkcionalnog chat servisa, a primjena se širi i na računalni vid gdje ovaj tip hardvera s lakoćom odrađuje zadatke poput klasifikacije slika i prepoznavanja lica[21]. Nadalje, TPU se ističu i kod aplikacija koje zahtijevaju predviđanje i reagiranje u realnom vremenu poput sustava autonomne vožnje, sigurnosnih sustava unutar velikih postrojenja i sl. Pokazali su se iznimno učinkovitim u situacijama gdje je potrebno upravljanje velikim brojem podataka, istovremeno obavljanje više zadaća, ali uz zadržavanje optimalnog vremenskog odziva. Ova komponenta danas nerijetko pronalazi primjenu u zdravstvu i uvelike pomaže čovjeku u razumijevanju prirodnih procesa i predviđanju istih

prilikom otkrivanja lijekova i savijanja proteina[23].Također se koristi u znanstvenim područjima bioinformatike i genetike gdje se uz pomoć specijaliziranog hardvera mogu otkriti obrasci i predviđanja vezana uz varijacije ljudskih gena i bolesti povezanih s istima[21]. Što se tiče potrošnje električne energije, u usporedbi s grafičkim procesorima, korištenjem jedinica za obradu tenzora mogu se osjetno smanjiti pogonski troškovi. Tako na primjer jedna grafička kartica NVIDIA A100 ima snagu od 400 W, dok Google Cloud TPU verzije 4 ima maksimalnu snagu od 200-250 W[24]. Za veća postrojenja i opsežnije projekte s kontinuiranom potrebom za obradom podataka, korištenjem TPU hardverskog rješenja može doći do značajnih ušteda radi ekonomičnijeg rada ovog tipa sklopovlja.

4.2.1 Ograničenja jedinica za obradu tenzora

Iako se radi o komponenti koja je do sada najprikladnija za rješavanje zadataka vezanih uz duboko učenje i obradu prirodnog jezika, ona dolazi s određenim ograničenjima i stavkama na koje korisnik treba obratiti pozornost pri odabiru hardvera za specifične zahtjeve. Prvo ograničenje koje se nameće je softverska limitiranost. S obzirom na to da je čitav sustav *cloud TPU* servisa dizajniran od strane jedne tvrtke, hardverski dio sustava je optimiziran isključivo za vlastitu softversku podršku TensorFlow. Prethodno spomenuta knjižnica (eng. framework) je usko specijalizirana i prikladna za treniranje jezičnih modela, no može predstavljati izazov za programere koji se služe drugim okvirima. Pored toga, procesori (CPU) i grafičke procesori (GPU) puno su dostupniji i cjenovno prihvatljiviji te bi rad isključivo s jedinicama za obradu tenzora mogao dugoročno predstavljati financijski izazov pojedinca ili tvrtke koja se njima služi.

Poglavlje 5

Zaključak

U ovom radu objašnjeni su različiti aspekti i komponente koji omogućuju obradu prirodnog jezika. Ovo područje umjetne inteligencije trenutno doživljava rapidan uspon i širenje svoje primjene, kako u znanosti i poslovnom svijetu, tako i u privatne svrhe za asistenciju, savjete i zabavu. Ključna spoznaja ovog rada je da su veliki jezični modeli, zahvaljujući svojoj učinkovitosti pri obradi velikih količina podataka i složenih jezičnih struktura, postali nezamjenjivi alati u računalnoj obradi prirodnog jezika. Kroz ovaj rad opisani su temelji na kojima počiva čitav proces obrade prirodnog jezika, a Transformeri su prepoznati kao esencijalan faktor suvremenih jezičnih modela, što predstavlja značajan preokret u odnosu na prethodne modele poput povratnih neuronskih mreža (RNN). Upravo su Transformeri zaslužni za razvoj i efikasnost trenutno dostupnih velikih jezičnih modela. Danas postoje modeli koji na visokoj jezičnoj razini komuniciraju s korisnicima i raspolažu znanjem iz svih polja ljudskog djelovanja. Ti su modeli također prilagodljivi s obzirom na željenu funkciju i područje rada koje se od njih očekuje. Konkretno, u ovom radu su obrađena dva modela čije su glavne značajke, arhitektura i način rada također istaknuti. Dokazano je kako se neovisno o namjeni i uporabi radi o dva vrhunska primjera moderne tehnologije koju pogoni umjetna inteligencija. Spomenuti modeli, ChatGPT i BERT spadaju u najbolje jezične modele zbog razine svoje istreniranosti na velikim skupovima podataka. Ta stavka im je zajednička, iako ChatGPT (u verziji GPT-3.5) znatno prednjači s više od 175 milijardi parametara. Neovisno o svrsi korištenja spomenutih modela, ovdje se radi o vrhunskim tehnološkim izumima koji mogu do-

prinijeti brojnim poljima ljudskog djelovanja, ako se koriste na pravilan, savjestan i moralan način.

S razvitkom velikih jezičnih modela nastala je potreba za prilagodbom sklopovlja procesima učenja modela i paralelnoj obradi velikih količina podataka. Kreirane su komponente koje donedavno nisu postojale, a razlog tome bila je težnja za optimizacijom i ubrzanjem procesa strojnog učenja kao temelja velikih jezičnih modela i obrade prirodnog jezika. U prethodnom poglavlju obrađene su značajke grafičkih procesora (GPU) i jedinica za obradu tenzora (TPU). Može se zaključiti kako su grafički procesori u formi grafičke kartice kao komponente široko dostupne, s relativno pristupačnu cijenom s obzirom na dugotrajnost i obujam posla kojeg mogu odrađivati. No ipak treba imati u vidu kako se ovdje radi o hardveru s konačnom memorijom, stoga treba pripaziti na količinu podataka i naposljetku na željeno vrijeme izvršavanja zadataka. S druge strane, jedinice za obradu tenzora (TPU) predstavljaju inovativan i usko specijaliziran hardver namijenjen za treniranje većih i kompleksnijih modela dubokog učenja koji sadrže puno matričnih izračuna, kao što su i sami veliki jezični modeli.

Promatrajući tehnološki napredak današnjice, može se primijetiti kako njen razvoj teče brže nego ikada prije te da donosi brojne prednosti i olakšanje čovjeku u zdravstvenom, poslovnom i privatnom okruženju. Rad također ukazuje na to da će nadolazeći razvoj u području obrade prirodnog jezika ovisiti o relaciji između naprednih algoritama i specijaliziranog hardvera. Iako su dosadašnji modeli pokazali svoje dalekosežne mogućnosti, u području optimizacije resursa i povećanja energetske učinkovitosti u budućim istraživanjima uvijek će biti prostora za napredak i usavršavanje. Važno je napomenuti da svaka inovacija, koja na prvi pogled donosi isključivo prednosti, mora dolaziti s dozom opreza te da, unatoč današnjoj moći umjetne inteligencije, njezina uporaba mora biti zakonski regulirana kako ne bi došlo do zlouporabe alata i kreiranja plagijata u obrazovnim i poslovnim sustavima. Ključno je i očuvanje moralnih vrijednosti pri korištenju umjetne inteligencije i njezinih proizvoda te prisutnost ljudskog rasuđivanja kod donošenja odluka koje mogu utjecati na razvoj čovječanstva. Suvremena tehnologija može i treba biti usmjerena na razvoj novih ideja i poboljšanja kvalitete života, ali nikako ne smije u potpunosti zamijeniti ljudski faktor i čovjekovu završnu riječ.

Bibliografija

[1] What is a Large Language Model (LLM), s Interneta, https://www.geeksforgeeks.org/large-language-model-llm/, 20. svibnja 2024.

[2] Language Models Explained, s Interneta, https://www.altexsoft.com/blog/language-models-gpt/, 22. srpnja 2024.

[3] Tom B. Brown i drugi: Language Models are Few-Shot Learners, s Interneta, https://arxiv.org/abs/2005.14165, 5. rujna 2024.

[4] Josep Ferrer: How Transformers Work: A Detailed Exploration of Transformer Architecture, s Interneta, https://www.datacamp.com/tutorial/how-transformers-work?dc_referrer=https%3A%2F%2Fwww.google.com%2F, 22. srpnja 2024.

[5] Abid Ali Awan: Recurrent Neural Networks(RNN), s Interneta, https://www.datacamp.com/tutorial/tutorial-for-recurrent-neural-network, 23. srpnja 2024.

[6] Ashish Vaswani i drugi: Attention Is All You Need, s Interneta: https://arxiv.org/abs/1706.03762

[7] Transformer Architecture explained, s Interneta, https://medium.com/@amanatulla1606/transformer-architecture-explained-2c49e2257b4c, 1. kolovoza 2024.

[8] Chat GPT: What is it?, s Interneta, https://uca.edu/cetal/chat-gpt/, 14. kolovoza 2024.

[9] Amanda Hetler: What is ChatGPT?, s Interneta, https://www.techtarget.com/whatis/definition/ChatGPT, 14. kolovoza 2024.

[10] What is ChatGPT?, s Interneta, https://help.openai.com/en/articles/6783457-what-is-chatgpt, 14. kolovoza 2024.

[11] Alex Hughes: ChatGPT: Everything you need to know about OpenAI's GPT-4 tool, s Interneta, https://www.sciencefocus.com/future-technology/gpt-3, 23. kolovoza. 2024.

[12] Cameron Hashemi-Pour, Ben Lutkevich: BERT language model, s Interneta, https://www.techtarget.com/searchenterpriseai/definition/BERT-language-model, 26. kolovoza 2024.

[13] Kinza Yasar: What are masked language models(MLMs)?, s Interneta, https://www.techtarget.com/searchenterpriseai/definition/masked-language-models-MLMs, 26. kolovoza 2024.

[14] Rani Horev: BERT Explained: State of the art language model for NLP, s Interneta, https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270, 26. kolovoza 2024.

[15] BERT Fine Tuning with Cloud TPU: Sentence and Sentence-Pair Classification Tasks (TF 2.x), s Interneta, https://cloud.google.com/tpu/docs/tutorials/bert-2.x, 27. kolovoza 2024.

[16] Shar Narasimhan: NVIDIA Clocks World's Fastest BERT Training Time and Largest Transformer Based Model, Paving Path For Advanced Conversational AI, s Interneta, https://developer.nvidia.com/blog/training-bert-with-gpus/¸, 27. kolovoza 2024.

[17] BERT, s Interneta, https://www.nvidia.com/en-us/glossary/bert/, 6. rujna 2024.

[18] Nikola Ljubešić, Davor Lauc: BERTić - The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian, s Interneta, https://arxiv.org/abs/2104.09243, 5. rujna 2024.

[19] BERTIĆ - model na hrvatskom jeziku, s Interneta, https://ekoninfochecker.efri.uniri.hr/?p=735, 6. rujna 2024.

[20] NLP on GPUs and TPUs: Utilizing specialized hardware for efficient NLP processing, s Interneta, https://medium.com/@ghatifernado.inc/nlp-on-gpus-and-tpus-utilizing-specialized-hardware-for-efficient-nlp-proces 29. kolovoza 2024.

[21] Kelsie Anderson: What hardware should you use for ML inference?, s Interneta, https://telnyx.com/resources/hardware-machine-learning, 29. kolovoza 2024.

[22] Natural language processing, s Interneta, https://www.nvidia.com/en-us/glossary/natural-language-processing/, 29. kolovoza 2024.

[23] Accelerate AI development with Google Cloud TPUs, s Interneta, https://cloud.google.com/tpu?hl=en, 29. kolovoza 2024.

[24] Kurtis Pykes: Understanding TPUs vs GPUs in AI: A Comprehensive Guide, s Interneta, https://www.datacamp.com/blog/tpu-vs-gpu-ai?dc_referrer=https%3A%2F%2Fwww.google.com%2F, 7. rujna 2024.

Sažetak

Ovaj rad ima za cilj proučiti ključne komponente prisutne u procesu obrade prirodnog jezika. Rad počinje od temelja - modela transformera, gdje su razmotreni njihova arhitektura, način rada i uloga u razvoju suvremenih alata za obradu jezika. Transformeri su predstavljeni kao baza jezičnih modela radi učinkovitijeg načina obrađivanja složenih jezičnih struktura u usporedbi s prethodnim modelima. Nadalje, u radu su prikazani postojeći modeli ChatGPT i BERT koji dokazuju da se veliki jezični modeli mogu prilagoditi specifičnim namjenama, od generiranja teksta do analize i klasifikacije istog. Također su istražena rješenja za povećanje učinkovitosti jezičnih modela, kao što su grafički procesori (GPU) i jedinice za obradu tenzora (TPU), pri čemu je utvrđeno kako te tehnologije igraju ključnu ulogu u ubrzavanju i optimizaciji treniranja modela. Kroz rad se naglašava potreba za kontinuiranim razvojem i istraživanjem u području obrade prirodnog jezika, uz dozu opreza i potrebu za čovjekovim nadgledanjem svih procesa i rada umjetne inteligencije.

Ključne riječi — **modeli transformera, ChatGPT, BERT, hardverski akceleratori**

Abstract

This paper aims to study the key components present in the natural language processing. The paper starts from the foundation - the transformer model, where its architecture, operation and role in the development of modern language processing tools are discussed. Transformers are presented as a base of language models resulting in a more efficient way of processing complex language structures in comparison to previous models. Furthermore, the paper presents the existing ChatGPT and BERT models, which prove that large language models can be adapted to specific purposes, from text generation to analysis and classification. Solutions to increase the efficiency of language models, such as graphic processors (GPUs) and tensor processing units (TPUs) were also explained, finding that these technologies play a key role in speeding up and optimizing model training. The paper emphasizes the need for continuous development and research in the field of natural language processing, with a dose of caution and the need for human supervision of all processes and work of artificial intelligence.

Keywords **— transformer models, ChatGPT, BERT, hardware accelerators**