Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.85
DOI: https://doi.org/10.24867/28BE39Sarsanski

EKSPERIMENTI SA VEKTORSKIM REPREZENTACIJAMA TEKSTA ZA EKSTRAKTIVNU SUMARIZACIJU

EXPERIMENTS WITH TEXT EMBEDDING METHODS FOR EXTRACTIVE SUMMARIZATION

Aleksa Šaršanski, *Fakultet tehničkih nauka, Novi Sad*

Oblast – RAČUNARSTVO I AUTOMATIKA

Kratak sadržaj – *U ovom radu je implementirano rešenje problema ekstraktivne sumarizacije komentara za smeštaj u hotelima koje su napisali korisnici. Rešenje je implementirano koristeći FastText, BERT i GPT 3.5 modele dubokog učenja. U radu je detaljno opisan način funkcionisanja sva tri pomenuta modela, kao i načini njihove upotrebe prilikom rešavanja ovog problema. Skup podataka koji je korišćen u ovom rešenju sastoji se od javno dostupnih komentara korisnika, kreiran procesom scrape-ovanja. U radu je prilikom evaluacije rezultata upoređena efikasnost sva tri navedena pristupa rešavanja problema.*

Ključne reči: *Veštačka inteligencija, duboko učenje, ekstraktivna sumarizacija, NLP, FastText, BERT, GPT*

Abstract – *In this paper is implemented a solution to the problem of extractive summarization of user-written hotel accommodation reviews. The solution uses FastText, BERT, and GPT-3.5 deep learning models. The paper provides a detailed description of how each of these models works, as well as their application methods in addressing this problem. The dataset used in this solution consists of publicly available user comments obtained through the web scraping process. The effectiveness of all three mentioned approaches in solving the problem is compared during the evaluation of the results.*

Keywords: *Artificial Intelligence, Deep learning, Extractive summarization, NLP, FastText, BERT, GPT*

1. UVOD

Sumarizacija teksta predstavlja proces sažimanja teksta, čiji je rezultat kraći tekst koji sadrži sve bitne informacije i suštinu iz originalnog teksta. U oblasti mašinskog učenja, sumarizacija teksta spada u NLP (*Natural Language Processing*) probleme. Obzirom da rezultat sumarizacije teksta nije jednoznačan, svrstava se među teže probleme u oblasti NLP-a. Ekstraktivna sumarizacija predstavlja izdvajanje najvažnijih rečenica iz teksta i tako formira rezultujući tekst. U ovom pristupu se ne menja originalni tekst, već se samo izdvajaju njegovi najrelevantniji delovi. Njena primena se može naći u novinarstvu, sažimajući duge novinske članke, u apstrakciji dugih pravnih dokumenata, u brzom pregledu dugih medicinskih izveštaja, u analizama raznih poslovnih dokumenata... Ovaj rad se bavi problemom ekstraktivne sumarizacije komentara na online platformama za pretragu i rezervaciju smeštaja. Komentari putnika znaju da budu veoma dugački, da sadrže mnogo informacija koje korisnicima nisu presudne pri donošenju odluke o rezervaciji i obično zahtevaju mnogo izdvojenog vremena za analizu i izvlačenje bitnih informacija.

Motivacija za rešavanje ovog problema jeste da se putnicima olakša odabir smeštaja, tako će se iz svakog komentara izvući najrelevantniji delovi i samim tim značajno uštedeti vreme koje je potrebno korisniku da donese odluku koji će smeštaj rezervisati. Iako je veoma korisna, ova funkcionalnost je retko zastupljena na spomenutim platformama. Implementacija ekstraktivne sumarizacije teksta u ovom radu obuhvata 3 različita pristupa, odnosno upotrebu 3 različita modela dubokog učenja – FastText, BERT i GPT 3.5 modela. Prva dva pristupa (FastText i BERT) koriste TextRank algoritam za rangiranje rečenica. Implementacija pomoću GPT 3.5 modela obuhvata upotrebnu svih njegovih slojeva neuronske mreže, bez dodatnih modifikacija i klasifikacija izlaza.

2. PREGLED STANJA U OBLASTI

2.1. Radovi koji se bave TextRank algoritmom

U svom radu [1] 2004. godine, TextRank model je predstavljen od strane Rada Mihalcea i Paul Tarau. On se zasniva na grafovima i ima dve primene u rešavanju problema iz oblasti obrade prirodnog jezika. Prva njegova primena je ekstrahovanje ključnih reči iz teksta. Za ovu primenu je upotrebljen skup podataka koji se sastoji od kolekcije od 500 sažetaka iz Inspec baze podataka i odgovarajućih manuelno dodeljenih ključnih reči. Evaluacija rezultata je odrađena korišćenjem *precision*, *recall* i *F-measure* metrika, pri čemu su za evaluaciju rezultata korišćeni sistemi za *state-of-the-art* ekstrakciju ključnih reči iz Hulth 2023 eksperimenta.

Fokus druge primene TextRank algoritma jeste na ekstrakciji rečenica iz teksta. Efikasnost ovog eksperimenta proverena je pomoću *single-document* sumarizaciju nad skupom od 567 novinskih članaka iz Document Understanding Evaluations 2002.

2.2. Radovi koji koriste LLM-ove (Large Language Modele)

U maju 2023. godine, Abhishek Kumar je u svom radu [2] uradio istraživanje na temu poređenja efikasnosti nekoliko modela dubokog učenja za rešavanje problema ekstrak-

---

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bio dr Aleksandar Kovačević, red. prof.

tivne sumarizacije teksta. Modeli koji su obuhvaćeni radom su: BERT, GPT-2, KL-summarizer, Luhn, LEX i Word Rank. Evaluacija rezultata izlaza modela je rađena pomoću tri tehnike, Rouge score, BERT score i Mover score, gde su rezultati upoređeni sa ljudskim, ručno sumarizovanim komentarima iz dva skupa podataka, BBC New Datsets i DailyMail/CNN datasets. U ovom istraživanju su se kao najefikasniji modeli za ekstraktivnu sumarizaciju teksta pokazali GPT-2 i Luhn, uzimajući u obzir rezultate sve tri navedene tehnike evaluacije.

3. TEORIJSKI POJMOVI I DEFINICIJE

3.1 Vestačka inteligencija

Vestačka inteligencija (eng. *Artificial Intelligence*) – na osnovu pronalaska šablona unutar velike količine podataka omogućava računarima da sami donose odluke ili izvršavaju određene aktivnosti kako bi ostvarili konkretne ciljeve [3].

3.2 Mašinsko učenje

Mašinsko učenje (eng. *Machine Learning*) je podoblast veštačke inteligencije koja se bavi razvojem algoritama i računarskih sistema koji imaju sposobnost adaptacije na nove, do sada neviđene situacije. To im omogućava velika količina podataka u kojoj nalaze šablone koji im omogućavaju da donose odluke uz minimalne ljudske intervencije.

3.3 Duboko učenje

Duboko učenje (eng. *Deep Learning*) – predstavlja podoblast mašinskog učenja koja je zasnovana na veštačkim neuronskim mrežama, algoritmima koji su inspirisani ljudskim mozgom. Veštačka neuronska mreža je matematički model sastavljen od veštačkih neurona – perceptrona gde svaki veštački neuron predstavlja pojednostavljen model biološkog neurona [3].

3.4 NLP – *Natural Language Processing*

NLP (*Natural Language Processing*) je grana veštačke inteligencije koja se bavi obradom prirodnog, ljudskog jezika, bilo to obrada i analiza teksta ili govora. Ovaj metod kombinuje računarsko-lingvisitčko modelovanje ljudskog jezika sa statističkim modelima mašinskog učenja kako bi omogućila računarima da prepoznaju, razumeju i generišu tekst i govor [4]. Neke od najčešćih primena NLP-a su:

- Prepoznavanje govora (eng. *Speech recognition*),
- Pretvaranje teksta u govor (eng. *Text-to-Speech*),
- Segmentacija reči (tokenizacija),
- Analiza sentimenta (eng. *Sentiment Analysis*),
- Sumarizacije teksta (eng. *Text summarization*),
- Generisanje prirodnog jezika (eng. *Natural language generation*)...

3.5 Vektori reči (*Word embeddings*)

S obzirom da modeli mašinskog učenja rade samo sa brojevima (oni su matematički modeli), njima reči koje su sastavljanje od slova ne znače ništa. Iz tog razloga, reči je potrebno nekako prikazati pomoću brojeva, pa se zbog toga uvodi pojam vektora reči. Vektor reči predstavlja numeričku reprezentaciju reči u višedimenzionalnom matematičkom prostoru. Svaka reč je predstavljena pomoću jedinstvenog vektora koja joj određuje njenu poziciju u vektorskom prostoru. Vektorska reprezentacija reči omogućuje modelima mašinskog učenja da razumeju reči, na osnovu njihovog semantičkog i sintaksnog značenja. Slične reči se nalaze blizu u vektorskom prostoru, dok se različite reči nalaze daleko jedna od druge. Tako se na primer reč "jabuka" nalazi blizu reči "voće" u vektorkom prostoru, ali se nalazi daleko od reči "automobil". Najreprezentativniji primer kako vektori reči funckionišu su reči "*king*", "*queen*", "*man*" i "*woman*", gde za njihove vektorske reprezentacije važi "*king*" + "*woman*" – "*man*" ≈ "*queen*", odnosno "*king*" – "*man*" ≈ "*queen*" – "*woman*"

3.6 FastText algoritam

FastText je razvijen od strane Facebook-a 2016. godine. Koristi za dobijanje vektorskih reprezentacija reči koje potom koristi za rešavanje raznih NLP problema kao sto su klasifikacija teksta i dokumenata, modelovanje naslova, analiza sentimenta teksta... Ovaj algoritam čini neuronska mreža jednostavne arhitekture (bez mnogo slojeva) koja se odlično pokazala za kreiranje vektora reči, pri čemu se čuvaju semantičke i sintaksne osobine reči i zbog toga se često koristi. Ovaj algoritam umesto da kreira direktno vektor cele reči, deli reči na delove (eng. *n-gram*) i kreira vektor za svaki od *n-gram*-a, gde se rezultujući vektor reči dobija kao suma vektora svih *n-gram*-a te reči. Na ovaj način FastText čuva morfološku strukturu reči sto ga čini veoma moćnim i za kreiranje vektora reči složenijih jezika. Ovakav pristup mu omogućava da se dobro pokaže pri radu sa rečima koje se retko pojavljuju ili do sada nisu viđene sa njegove strane. Ovaj model je dostupan za 294 različita jezika.

3.7 Veliki jezički modeli (LLMs – *Large Language Models*)

LLM-ovi su veliki i složeni modeli dubokog učenja koji mogu da prepoznaju, sumarizuju, prevode, predviđaju i generišu tekst koristeći veoma velike skupove podataka. Koriste se za rešavanje raznih vrsta NLP zadataka, od najjednostavnijih do najsloženijih. Neki od najčešćih zadataka LLM-ova su: klasifikacija teksta (analiza sentimenta i emocija), sumarizacija teksta, prevođenje teksta sa jednog jezika na drugi, generisanje teksta, generisanje programskog koda...

Sastoje se od velikog broja slojeva neuronske mreže, obučavani su nad ogromnim skupom podataka (do nekoliko terabajta tekstualnih podataka) i sadrže veliki broj obučavajućih parametara (do nekoliko stotina milijardi parametara). Oni su obučavani na principu nenadgledanog učenja, što zahteva da se obučavaju na velikom broj nelabeliranih podataka.

3.8 BERT model

BERT (*Bidirectional Encoder Representations from Transformers*) je veliki jezički model koji radi na principu transformera. BERT predstavlja specifičan slučaj transformer modela jer se sastoji samo iz enkoder komponenti (ne sadrži dekoder komponente). Enkoder komponente su poređane jedna na drugu i vektori reči dobijeni iz jedne enkoder komponente se prosleđuju dalje na kodiranje narednoj enkoder komponenti. Ulaz u prvi enkoder je sekvenca reči.

BERT je pre-trenirani model, obučavan nad velikim skupom podataka nakon čega ga je u svrhe boljeg rešavanja određenog problema moguće do-trenirati novim skupom podataka koji pripadaju domenu kome pripada i sam rešavani problem.

3.9 GPT model

GPT 3 (*Generative Pre-Trained Trasformer*) je treća generacija velikog jezičkog modela dubokog učenja predstavljena 2020. godine od strane kompanije OpenAI u ciji rad i razvoj ulaze kompanije Microsoft. Ovaj pre-trenirani model generiše tekst na osnovu predikcije naredne reči u rečenici primenom sistema "učenja bez učitelja" (eng. *unsupervised learning*).
Za njegovo obučavanje korišćeno je 175 milijardi obučavajućih parametara i oko 570 gigabajta podataka koji su sakupljeni sa raznih internet stranica, foruma, knjiga, članaka...
GPT 3 je transformer model koji je za razliku od BERT modela kojeg čine enkoder blokovi, izgrađen od dekoder blokova. Izgrađen je od 96 *attention* blokova gde se svaki od njih sastoji od 96 *attention head* – ova.

3.10 TextRank algoritam

TextRank algoritam pripada grafovskim algoritmima i oslanja se na Google-ov PageRank algoritam. Koristi se rangiranje tekstova, rečenica i reči tako sto meri vezi između dve ili vise reci. Ima veliku primenu pri rešavanju raznih NLP problema kao sto su sumarizacija teksta, izdvajanje ključnih reci iz teksta, analiza citata...

4. METODOLOGIJA

4.1 Skup podataka

Skup podataka koji je korišćen u ovom radu za dotreniravanje i evaluaciju rezultata modela dubokog učenja za rešavanje problema ekstraktivne sumarizacije teksta je kreiran tako što je preuzet sadržaj komentara za hotele sa poznate internet stranice koja se bavi rezervacijom smeštaja. Svi komentari se odnose na utiske putnika koji su boravili u raznim hotelima na tropskoj destinaciji – Tajlandu.
Ovaj skup podatak sadrži 3983 komentara putnika koji se odnose na razne hotele na Tajlandu. Prilikom preuzimanja komentara, filtrirali su se samo oni komentari koji su na engleskom jeziku, jer su rešenja korišćena za rešavanja problema ekstraktivne sumarizacije teksta u ovom radu fokusirana samo na engleski jezik.

4.2 Pristup rešavanja problema pomoću FastText modela

Rešavanje problema ekstraktivne sumarizacije teksta pomoću FastText modela dubokog učenja se sastoji od dotreniravanja FastText modela pomoću domenskog skupa podataka opisanom u poglavlju 4.1 kako bi model dodelio vektorsku reprezentaciju rečima iz skupa podataka.
Na osnovu dobijenih vektorskih reprezentacija reči, traži se njihov prosek po rečenici pri čemu se dobija vektorska reprezentacija cele rečenice.
Tako dobijene vektorske reprezentacije rečenica se prosleđuju TextRank algoritmu koji ih rangira po važnosti i vraća željeni broj najbitnijih rečenica za komentar.

4.3 Pristup rešavanja problema pomoću BERT modela

Rešavanje problema ekstraktivne sumarizacije teksta pomoću BERT velikog jezičkog modela je u velikoj meri slično kao i rešavanje pomoću FastText modela koje je objašnjeno u prethodnom poglavlju. U tu svrhu je korišćena posebna vrsta BERT modela – *Sentence BERT* (SBERT). SBERT se od standardnog BERT modela razlikuje po tome što radi na nivou rečenica umesto na nivou reči kao standardan BERT model. Konkretno, umesto da kreira vektorske reprezentacije reči, on kreira vektorske reprezentacije rečenica. Nije potrebno računati prosečnu vrednost reči u rečenici da bi se dobila vektorska reprezentacija rečenice.
Kako bi odredio vektorsku reprezentaciju čitave rečenice, SBERT koristi Sijamsku arhitekturu (*Siamese architecture*). Sijamska arhitektura se sastoji od dve ili vise identičnih podmreža koje paralelno generišu vektore (u ovom slučaju vektore rečenica) i porede ih kako bi im odredile poziciju u vektorskom prostoru.
Ova arhitektura neuronskih mreža se obično koristi prilikom rešavanja problema pronalaska duplikata u tekstovima, pronalaska anomalija, semantičku pretragu, klasterovanje dokumenata...
U ovom slučaju Sijamske arhitekture, SBERT koristi paralelno dva standardna BERT modela, gde svakom prosleđuje po jedna rečenica i na izlaznim slojevima se određuje slično između njih i na osnovu toga im se dodeljuje vektorska reprezentacija. Dobijene vektorske reprezentacije rečenica se prosleđuju takođe TextRank algoritmu koji ih rangira po važnosti i vraća željeni broj najbitnijih rečenica za komentar.

4.4 Pristup rešavanja problema pomoću GPT 3.5 modela

Rešavanje problema ekstraktivne sumarizacije teksta. Pristup rešavanja problema ekstraktivne sumarizacije teksta pomoću GPT 3.5 velikog jezičkog modela se razlikuje od prethodnih rešenja u tome sto se ne radi dotreniravanje GPT 3.5 modela kao sto je to bio slučaj sa FastText i BERT modelima. U ovom pristupu se koriste već postojeće vektorske reprezentacije reči koje su kreirane prilikom obučavanja GPT modela. GPT zna kakav izlaz treba da generiše na osnovu *prompta* koji mu je prosleđen. *Prompt* predstavlja tekstualni fajl koji sadrži naredbe, odnosno instrukcije koje simuliraju komunikaciju čoveka i mašine i govore modelu kako bi trebao da se ponaša i kakav izlaz da generiše.

5. EKSPERIMENTI

5.1 Eksperiment 1

Eksperiment 1 obuhvata evaluaciju rešenja nad komentarima koji su već viđeni od strane modela prilikom dotreniravanja i nalaze se u obučavajućem skupu podataka. Kao sto je prethodno opisano, taj skup je dobijen na tri načina: bez primene lematizacije i steminga nad komentarima, sa primenom lematizacije nad komentarima i sa primenom steminga nad komentarima.
Komentari koji se u ovom slučaju koriste kao evaluacioni skup podataka, a zapravo su podskup obučavajućeg skupa podata su dužine između osam i dvanaest rečenica.

Rezultati svakog eksperimenta sadrže sumarizaciju komentara na tri i na pet rečenica kako bi se i proverila zavisnost očuvanja semantičkih informacija iz originalnog komentara od dužine sumarizovanog komentara. Evaluacioni skup broji 40 različitih originalnih komentara.

5.2 Eksperiment 2

Eksperiment 2 obuhvata evaluaciju rešenja nad komentarima koji do sada nisu viđeni od strane modela prilikom dotreniravanja i ne nalaze se u obučavajućem skupu podataka. Razlika između ova dva eksperimenta nije od velike važnosti za rešenje problema pomoću GPT 3.5 modela, s obzirom da nije rađeno njegovo dotreniravanje.
Obučavajući skup korišćen u ovom eksperimentu dobijen je na tri načina: bez primene lematizacije i steminga nad komentarima, sa primenom lematizacije nad komentarima i sa primenom steminga nad komentarima.
Komentari koji se u ovom slučaju koriste kao evaluacioni skup podataka su dužine između osam i dvanaest rečenica. Rezultati svakog eksperimenta sadrže sumarizaciju komentara na tri i na pet rečenica kako bi se i proverila zavisnost očuvanja semantičkih informacija iz originalnog komentara od dužine sumarizovanog komentara. Evaluacioni skup broji 40 različitih originalnih komentara.

5.3 Evaluacija

Za procenu efikasnosti FastText, BERT i GPT 3.5 modela dubokog učenja za problema ekstraktivne sumarizacije komentara upotrebljene su ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) metrike. Upotrebljene ROUGE metrike su ROUGE-1, ROUGE-2 (koje pripadaju ROUGE-N metrikama) i ROUGE-L metrike. Pored navedenih metrika, procena efikasnosti modela je rađena i po subjektivnom osećaju autora sumarizovanih komentara uključenih u evaluaciju. ROUGE-N metrike (ROUGE-1 i ROUGE-2) porede broj odgovarajućih "*n-gram*a" u tekstu, odnosno poređenje n broja reči 2 sumarizovana komentara. ROUGE-1 se koristi za poređenje svake pojedinačne reči, dok se ROUGE-2 metrika koristi sa pronalaženje poklapanja parova reči, to mogu biti na primer neke fraze koje označavaju da li je sačuvana semantika komentara prilikom sumarizacije. ROUGE-L metrika meri najdužu sekvencu reči koja se poklapa između generisanog i referentnog komentara.
Prilikom evaluacije eksperimenata, prikazane su vrednosti *F1 Score* metrike koja predstavlja prosečnu vrednost između ostalih *precision* i *recall* metrika. S obzirom da je za svaki eksperiment korišćeno 40 različitih komentara, prikazana vrednost *F1 Score* metrike za svaki testni slučaj predstavlja prosečnu vrednost F1 Score metrike za svih 40 komentara za taj testni slučaj.

6. REZULTATI I DISKUSIJA

U tabelama 3 i 4 možemo da vidimo modele koji su dali najbolje rezultate na oba eksperimenta za ekstraktivnu sumarizaciju komentara na 3 i na 5 rečenica. Prilikom ekstraktivne sumarizacije na 3 rečenice, po svim ROUGE metrikama najbolje se pokazao BERT model i to u Eksperimentu 2 gde je evaluacija rađena nad podacima koji do sada nisu viđeni od strane modela. BERT je za sva 3 načina pre-procesiranja obučavajućeg skupa podataka pokazao identične, najbolje rezultate. Situacija je ista što se tiče i ekstraktivne sumarizacije na 5 rečenica. BERT se takođe po svim metrikama pokazao kao najbolje rešenje nezavisno od načina pre-procesiranja podataka i to ponovo u Eksperimentu 2.

Na osnovu prethodnih zaključaka možemo da kažemo da se BERT model dubokog učenja pokazao najuspešnije u ovom radu prilikom rešavanja problema ekstraktivne sumarizacije komentara. Najbolje rezultate u pogledu na oba eksperimenta pokazao je prilikom ekstraktivne sumarizacije teksta na 5 rečenica korišćenjem podataka za evaluaciju koje do sada nije video (ne nalaze se u obučavajućem skupu podataka). Njegove performanse ne zavise od toga da li je nad podacima u obučavajućem skupu primenjen algoritam steminga i lematizacije ili nije.

7. ZAKLJUČAK

U ovom radu su implementirana tri pristupa rešavanja problema ekstraktivne sumarizacije komentara korisnika hotela sa poznate internet platforme za rezervaciju smeštaja na Tajlandu. Svaki od tri pristupa koristi model veštačke inteligencije za rešavanje ovog problema – FastText, BERT i GPT 3.5. Iako se u eksperimentima pokazalo da se za rešavanje ovog problema najbolje pokazao BERT model, FastText i GPT 3.5 su se pokazali dovoljno dobro u cilju istog. Zanimljivo proširenje zadatka koje bi skoro sasvim sigurno povećalo uspešnost njegovog rešavanja bi bilo da se ekstraktivna sumarizacija komentara pomoću određenih *prompt*-ova implementira koristeći novije velike jezičke modele kao što su: *GPT 4*, *Antropic Claud*, *Amazon Titan*, *Google Bard*, *Mistral*, *Mixtral* i drugi koji su javno dostupni pomoću svojih provajdera.

8. LITERATURA

[1] Rada Mihalcea and Paul Tarau, 2004. *TextRank: Bringing Order into Texts* https://aclanthology.org/W04-3252.pdf
[2] Abhishek Kumar, 2023. EXTRACTIVE TEXT SUMMARIZATION http://14.139.251.106:8080/jspui/bitstream/repository/20079/1/Abhishek%20Kumar%20Mtech.pdf
[3] Complete Guide to TensorFlow for Deep Learning with Python by Jose Portilla – Udemy course - https://www.udemy.com/course/complete-guide-to-tensorflow-for-deep-learning-with-python/
[4] What is natural language processing (NLP)? - https://www.ibm.com/topics/natural-language-processing

Kratka biografija:

Aleksa Šaršanski rođen je u Novom Sadu 23.09.1998. godine. Osnovne akademske studije završio je 2022. godine, na Fakultetu Tehničkih Nauka u Novom Sadu, smer Primenjeno softversko inženjerstvo. Godine 2022. upisuje master akademske studije na studijskom programu Primenjene računarske nauke, odsek Inteligentni sistemi kontakt: aleksa.berdih@gmail.com