Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.9
DOI: https://doi.org/10.24867/12BE48Matovic

ANALIZA SENTIMENTA TEKSTA NA SRPSKOM JEZIKU KORIŠĆENJEM DUBOKOG UČENJA

SENTIMENT ANALYSIS OF TEXT IN SERBIAN LANGUAGE USING DEEP LEARNING

Stevan Matović, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIKA I RAČUNARSTVO

Kratak sadržaj – *Analiza sentimenta je naučno polje koje se bavi analizom mišljenja, stavova i emocija ljudi koji su napisali određeni tekst. Ovakva analiza može se koristiti u svrhe praćenja brendova, poboljšanja korisničke podrške, analize proizvoda, istraživanja tržišta ili u svrhe kreiranja sistema za preporuke. U ovom radu obučeno je i upoređeno više modela mašinskog učenja za zadatku analize sentimenta. Recenzije korišćene kao skup podataka prikupljene su sa jednog od sajtova za dostavu hrane u Srbiji. Recenzije sadrže skup ocena koje će biti korišćene kao pokazatelj emocionalne polarnosti. Trenirana su i upoređena tri modela mašinskog učenja sa različitim vrstama vektorizacije i rezultati su upoređeni sa pristupom transfera učenja. Transfer učenja je metoda dubokog učenja gde se model obučen da reši jedan problem koristi kao polazna tačka pri rešavanju drugog problema. Za transfer učenja korišćen je model dubokog učenja zasnovan na transformerima.*

Ključne reči: *Analiza sentimenta, mašinsko učenje, duboko učenje, transfer učenja*

Abstract – *Sentiment analysis is a scientific field that deals with the analysis of opinions, attitudes and emotions of people who have written a certain text. Such analysis can be used for purposes like brand monitoring, customer service improvement, product analysis, market research, creating recommendation systems etc. In this paper, several machine learning models are trained and compared for the task of sentiment analysis. Reviews used as a dataset were collected from one of the food delivery websites in Serbia. Reviews contain a set of ratings that will be used as an indicator of emotional polarity. Three models of machine learning with different types of vectorization were trained and compared with results of transfer learning approach. Transfer learning is a method of deep learning where a model trained to solve one problem is used as a starting point in solving another problem. Deep learning model based on transformers is used for transfer learning.*

Keywords: *Sentiment analysis, Machine Learning, Deep Learning, Transfer Learning*

1. UVOD

Šta drugi ljudi misle o proizvodu ili temi oduvek je bio važan faktor za ljude pri donošenju odluka. Pre nego što se pojavio internet, mnogi su pitali prijatelje i poznanike za preporuku pri odabiru restorana, mesta koja vredi posetiti, određenog proizvoda itd.

Internet je je omogućio ljudima da saznaju mišljenja i iskustva velikog broja ljudi koji nisu njihovi prijatelji ili poznanici, ali to je takođe dovelo do povećanja broja ljudi koji dele svoje mišljenje sa nepoznatim ljudima preko interneta. Sve ovo je rezultiralo ogromnim količinama nestruktuiranih tekstualnih podataka.

Analiza sentimenta se bavi određivanjem emocionalne polarnosti takvog teksta. Pristupi za analizu sentimenta se razlikuju po metodama pretprocesiranja, kao i po samim algoritmima i tipovima algoritama. Zadatak ovog rada će biti poređenje različitih pristupa nadgledane analize sentimenta za recenzije na srpskom jeziku. Srpski jezik spada u grupu visoko flektivnih i morfološki bogatih jezika, koji koriste puno različitih sufiksa kako bi izrazili različite gramatičke, sintaktičke ili semantičke karakteristike.

Uporedićemo pristup transfera učenja (eng. *transfer learning*) sa tradicionalnim pristupima mašinskog učenja. Transfer učenja je metoda mašinskog učenja gde se model treniran na velikom skupu podataka da reši jedan problem koristi kao polazna tačka pri rešavanju drugog problema. Ovo je popularan pristup u dubokom učenju gde se modeli pretreniraju da reše problem uz korišćenje značajnih komputacionih i vremenskih resursa nad velikim količinama nelabeliranih ili labeliranih podataka da bi se potom prilagodili za druge zadatke (moguće i različite od zadatka za pretreniranje, ali u istom domenu) i na taj način iskoristili prethodno stečeno znanje.

Transfer učenje je imalo mnogo uspeha u domenu mašinske vizije, gde se danas modeli retko treniraju „od nule", već se prilagođavaju već pretrenirani modeli. Transfer učenje se pokazalo kao veoma korisno i u domenu obrade prirodnog jezika.

U ovom radu uporedićemo tradicionalne metode nadgledane analize sentimenta (Naive Bayes, Support Vector Machine, logistička regresija) sa metodom transfera učenja koristeći duboki model reprezentacije jezika BERT [1].

Pod pojmom dubokog učenja (eng. Deep Learning) smatra se primena veštačkih neuronskih mreža (skraćeno neuronskih mreža) koje se sastoje od više slojeva. Duboko učenje koristi sposobnost neuronskih mreža da

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

naprave reprezentacije različitih nivoa kompleksnosti u zavisnosti od dubine sloja. Što je dublji sloj to su reprezentacije kompleksnije, otuda naziv duboko učenje.

Zahvaljujući brzom razvoju u poslednjoj deceniji, neuronske mreže danas predstavljaju jedno od najznačajnijih i najprimenljivijih metoda mašinskog učenja.Bez obzira na veliki broj primena koje neuronske mreže imaju, ipak je važno istaći tipove problema u čijem rešavanju su se one pokazale najbolje - a to su problemi sa velikom količinom podataka koji se nalaze u svojoj sirovoj reprezentaciji. Ovo svakako znači da neuronske mreže nisu rešenje za svaki problem. Male količine podataka veoma lako dovode do overfit-ovanja, dok podaci koji nisu u svojoj sirovoj reprezentaciji već u vektorskom formatu sa već određenim atributima od interesa ne koriste osnovnu prednost neuronskih mreža, a to je svakako sposobnost da same konstruišu kompleksne reprezentacije nad sirovim podacima.

2. METODOLOGIJA I ALATI

U ovom poglavlju predstavljena je metodologija analize sentimenta teksta recenzija na srpkom jeziku koje nose pozitivan ili negativan sentiment. Detaljno će biti predstavljen metod prikupljanja podataka, tehnike preprocesiranja i samo obučavanje i evaluacija različitih modela mašinskog učenja.

2.1. Prikupljanje podataka

Podaci korišćeni u ovom radu prikupljeni su sa jednog od sajtova za dostavu hrane u Srbiji. Namena ovog sajta je povezivanje ljudi sa velikim brojem restorana koji nude uslugu dostave hrane. Nakon naručivanja hrane putem ovog servisa, korisnik ima opciju da unese recenziju za dati restoran. Recenzija može da sadrži naslov, tekst, sliku i set ocena. Korisnik ima opciju da dodeli 4 različite ocene i to za sledeće kategorije: kvalitet hrane, izbor hrane, cenu i uslugu.

Za svaku kategoriju korisnik bira da li će da je oceni sa ocenom od 1 do 5, što znači da svaka recenzija ima minimalno 0 ocena a maksimalno 4. Za prikupljanje podataka korišćen je python programski jezik, biblioteka Beautiful Soup [2] kao i Selenium WebDriver [3].

Ukupno je prikupljeno 75857 recenzija, od kojih je 60807 za restorane u Beogradu a 15050 za restorane u Novom Sadu. Skup podataka neće biti javno dostupan i korišćen je isključivo za eksperimentalne svrhe.

2.2. Pretprocesiranje podataka

Pretprocesiranje podataka je posebno važan korak za zadatke obrade teksta. Tekst se pretvara u savladiv oblik kako bi algoritmi mašinskog učenja mogli bolje da ga obrade. U ovom radu nekoliko različitih metoda pretprocesiranja su primenje a to su:

1. uklanjanja dijakritika
2. izbacivanje stop reči
3. normalizacija emotikona i znakova
4. prebacivanje u mala slova
5. stemovanje reči
6. upweighting
7. vektorizacija

Srpski jezik sadrži nekoliko slova sa dijakritičkim znakovima(č,ć,đ,š,ž). Pri korišćenju srpskog jezika na računaru korisnici često ignorišu dijakritičke znakove, zbog ovog razloga je odlučeno da se karakteri sa dijakritikom svedu na svoju osnovnu formu. Ovo se radi kako ne bi došlo do različite interpretacije tokena koji imaju isto značenje (npr. cevapi i ćevapi).

Tokenizacija je proces u kome se tekst (rečenice, paragrafi, dokumenti) predstavlja kao lista tokena. Token može biti na nivou reči, delova reči ili karaktera. U ovom radu rečenice su podeljene na tokene na nivou reči, tako da će rečenica „Hrana je odlična“ biti predstavljena kao sledeća lista tokena [„Hrana“, „je“, „odlična“].

Stop reči su skup reči koje se često koriste u jeziku. Primeri stop reči na srpskom su „a“, „ali“, „i“, „“ itd. Intuicija iza izbacivanja zaustavnih reči je da se uklanjanjem reči sa niskim informacijama iz teksta algoritam može fokusirati na važne reči.

Emotikoni su dobar pokazatelj emocionalnog polariteta. U [4], tweet-ovi koji se završavaju pozitivnim emotikonima poput „:)“ I „:-)“označeni su kao pozitivni a tweet-ovi koji se završavaju negativnim emotikonima poput „:(“ ili „:-(“ su označeni kao negativni.

U ovom radu emotikoni su zamenjeni rečima sličnog polariteta sentimenta. Na primer „:-))“ je zamenjen rečju „odlican“.

Prebacivanje u mala slova je jedna od najprostijih transformacija a ujedno i veoma efektivna. Sva velika slova se prebacuju u mala kako bi se izbegla različita interpretacija tokena sa istim značenjem (npr. Dostava i dostava).

Upweighting je tehnika gde se za neke reči računa kao da su se pojavile više puta nego što zapravo jesu kako bi se povećao njihov uticaj. U ovom radu ova tehnika se koristila za reči iz naslova recenzije (računate su kao da su se pojavile dva puta umesto jednom).

Kako bi modeli mašinskog učenja bili u stanju da razumeju tekstualne podatke neophodno je tekst predstaviti numeričkim vrednostima (vektorima) - ovaj proces naziva se vektorizacija. U ovom radu korišćeno je više tehnika vektorizacije kako bi se uporedili rezultati.

Prva tehnika je predstavljanje teksta kao set n-grama. N-grami su sekvence susednih tokena dužine N iz datog uzorka teksta. U ovom radu eksperimentisano je sa unigramima, bigramima i trigramima.

Drugi način vektorizacije je tf-idf vektorizacija. Tf-idf vektorizator konvertuje tekst u vektor tf-idf vrednosti množenjem broja ponavljanja tokena u recenziji (eng. term frequency ili tf) sa invertovatnim brojem dokumenata u kojima se token pojavljuje (eng. inverse document frequency ili idf).

Ovo znači da će reč imati veći uticaj ukoliko se pojavljuje više puta u recenziji, ali i ako je reč retka(ne pojavljuje se u puno drugih recenzija).

2.3. Treniranje modela

Nakon što su podaci pretprocesirani spremni su za algoritme mašinkog učenja. Analizom prethodnih radova na temu analize sentimenta i klasifikacije teksta generalno, utvrđeno je koji su algoritmi pokazali dobre performanse za ovaj problem i oni su analizirani i u ovom radu. To su algoritmi:

- Logistička regresija
- Naive Bayes
- SVM (Support Vector Machines)

Na performanse svakog od ovih algoritama utiču tehnike pretprocesiranja, kao i način formiranja finalnog skupa atributa. Upravo zato su svi algoritmi primenjeni na skupu, koji prolazi kroz isti podsistem za pretprocesiranje i formiranje skupa atributa. Optimizacija vrednosti hiperparametara svakog algoritma rađena je postupkom ugnježdene unakrsne validacije.

Osim ovih standardnih pristupa isproban je i model dubokog učenja BERT (Bidirectional Encoder Representations from Transformers).

Ovaj model je dizajniran tako da nauči duboke bidirekcione reprezentacije teksta tokom pretreniranja na velikim količinama nelabeliranih podataka. Rezultat ovog procesa, pretrenirani BERT se potom može prilagoditi (eng. fine-tuning) sa samo jednim dodatnim slojem kako bi rešio širok spektar NLP problema.

Pretreniranje BERT-a je izuzetno skupo, konkretno obuka za BERT Base izvedena je na 16 TPU čipova dok je obuka za BERT Large izvedena na 64 TPU čipa. Svako pretreniranje trajalo je 4 dana. Srećom Google je objavio više BERT-a pretreniranih modela. U ovom radu korišćen je BERT Base Multilingual Cased model pretreniran na na 104 jezika (uključujući srpski). Cased označava da se koristi originalna veličina slova i originalni markeri padeža i akcenta.

Prilagođavanje BERT-a za klasifikaciju sekvenci postiže se dodavanje dodatnog klasifikacionog sloja postojećem pretreniranom modelu. Prvi znak u svakom BERT ulazu je poseban token [CLS]. Ulaz u dodati klasifikacioni sloj je konačno skriveno stanje (izlaz transformera) za ovaj token. Ulazni tekst nije unapred pretprocesiran kao za ostale algoritme u ovom radu.

Model je prilagođen u Google Colab[5] okruženju pomoću Tensor procesne jedinice (TPU) sa sledećim hiperparametrima:

- TRAIN_BATCH_SIZE = 32
- EVAL_BATCH_SIZE = 8
- PREDICT_BATCH_SIZE = 8
- LEARNING_RATE = 2e-5
- NUM_TRAIN_EPOCHS = 3.0
- MAX_SEQ_LENGTH = 512

Svi modeli mašinskog učenja su trenirani nad istim trening skupom i testirani nad istim test skupom podataka. Trening i test skup su dobijeni tako što se originalni skup podelio u odnosu 80/20.

U ovom radu kao metrika za upoređivanje korišćena je mikro F1 mera.

3. REZULTATI I DISKUSIJA

U ovom radu istražena je upotreba unigrama, bigrama, trigrama i tf-idf vektorizacije na tri algoritma: Naive Baies, SVM i Logistička regresija. Ovi rezultati su upoređeni sa rezultatima dobijenim korišćenjem pretreniranog BERT modela. Kao metrika korišćena je mikro F1 mera. Rezultati su prikazani u tabeli 1.

Tabela 1. Rezultati eksperimenata

	No upweighting	Upweighting
Naive Bayes + unigrami	0.93	0.93
Naive Bayes + bigrami	0.91	0.92
Naive Bayes + trigrami	0.86	0.88
Naive Bayes + tf-idf	0.88	0.88
Logistička regresija + unigrami	0.93	0.93
Logistička regresija + bigrami	0.90	0.91
Logistička regresija + trigrami	0.85	0.86
Logistička regresija + tf-idf	0.93	0.93
SVM + unigrami	0.93	0.93
SVM + bigrami	0.9	0.91
SVM + trigrami	0.85	0.86
SVM + tf-idf	0.93	0.93
Bert	0.94	

Iz rezultata se može zaključiti sledeće:

- Bigrami rade lošije od unigrama. Ovo je zbog činjenice da su reprezentacije sa bigramima vrlo retko popunjene i ukupna ocena opada za sve algoritme. Bolje pristup bi mogao biti upotreba unigrama i bigrama zajedno.
- Trigrami su dali vrlo loše rezultate. To je zato što su reprezentacije trigramima još ređe popunjene nego od onih sa bigrama (434.333 ukupnih obeležja u poređenju sa 251.583 obeležja za bigrame).
- Za Naive Baies, tf-idf se loše pokazao, ovo je zbog toga što se koristio Multinomial Naive Baies. Ovaj algoritam je više pogodan za klasifikaciju za diskretne podatke.
- Za SVM i logističku regresiju tf-idf radi podjednako dobro kao unigrami.
- Za sve algoritme upweighting je poboljšao f1 metriku za bigrame i trigrame, to je možda zato

što upweighting dodaje nova obeležja i podaci postaju manje retko popunjeni.
- Upweighting međutim nije uticao na unigrame i tf-idf.
- BERT je nadmašio sve pomenute algoritme, bez prethodne obrade ulaznog teksta. To pokazuje da se pretreniranjem na velikim količinama tekstualnih podataka, duboki modeli obučavaju da shvate jezik na neki način, što im omogućava da prenesu to razumevanja na konkretne probleme. Prilagođavanje ovakvih modela nadmašuje modele koji su trenirani od nule.
- Naivni Baies, SVM i Logistička regresija i dalje daju jako dobre rezultate, za relativno malo vreme obučavanja.

Po današnjim standardima dubokog učenja, količina podataka prikupljena u ovom radu je i dalje veoma mala. Može se pretpostaviti da bi sa većom količinom podataka BERT pokazao još bolje performanse u odnosu na standardne modele mašinskog učenja.

4. ZAKLJUČAK

U ovom radu rešavan je problem analize sentimenta tekstualnih podataka na srpkom jeziku. Analiza sentimenta je naučno polje koje se bavi analizom mišljenja, stavova i emocija ljudi koji su napisali tekst. Tekstualni podaci koji su korišćeni su recenzije koje ostavljaju korisnici na sajtu za dostavu hrane. Detaljno su opisane su metode za prikupljanje podataka kao i za preprocesiranje istih.

Podaci su pretprocesirani nekim od standardnih tehnika pretprocesiranja u polju obrade teksta. Nad pretprocesiranim podacima, obučena su tri modela mašinska učenja - Naive Baies, SVM i logistička regresija. Svaki od ovih algoritama testiran je sa 4 različita tipa vektorizacije: unigrami, bigrami, trigrami i tf-idf. Pored ovoga je testirana i upweighting tehnika (brojanje nekih reči kao da su se pojavile dva puta).

Takođe je prilagođen pretrenirani model dubokog učenja BERT kako bi rešio ovaj zadatak analize sentimenta i upoređeni su rezultati sa prethodno navedenim pristupima. Pokazano je da je prilagođen pretrenirani model dubokog učenja nadmašio rezultate algoritama obučenih od nule čak i nad ovim relativno malim skupom podataka. Ovo pokazuje veliki potencijal tehnike transfera učenja.

Ovaj rad bi se mogao proširiti detaljnijim analizama nad većim skupom podataka, kao i poređenjem sa novijim modelima dubokih reprezentacija jezika.

5. LITERATURA

[1] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
[2]Leonard Richardson 2014, BeautifulSoup4 https://www.crummy. com/software/BeautifulSoup
[3] Jason Huggins, et al, 2004. Selenium, https://www.seleniumhq.org
[4] Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." CS224N project report, Stanford 1.12 (2009): 2009.
[5] https://colab.research.google.com/

Kratka biografija:

Stevan Matović rođen je u Beogradu 1995. god. Master rad na Fakultetu tehničkih nauka iz oblasti Računarstva i automatike – Sistemi za istraživanje i analizu podataka odbranio je 2020.god.

kontakt: stevan.matovic95@gmail.com