Sentiment analiza recenzija hotela na srpskom i engleskom jeziku i izdvajanje aspekata

Nataša Ivanović, Vera Kovačević
Univerzitet u Novom Sadu, Fakultet tehničkih nauka, Novi Sad, Srbija
{natasaivanovic, kovacevic.r214.2021}@uns.ac.rs

Apstrakt – Za hotelske kompanije, presudnu ulogu u poslovanju ima povratna informacija od strane posetilaca. Određivanjem sentimenta recenzija koje ostavljaju gosti i izdvajanjem aspekata usluge može se odrediti da li je većina posetilaca bila zadovoljna i koje aspekte bi trebalo unaprediti. Takođe, značajna stavka je i analiza recenzija na različitim jezicima, što bi omogućilo hotelskim kompanijama da uzmu u obzir i mišljenje turista sa drugih govornih područja. Ovaj rad se bavi skupom podataka koji se sastoji od recenzija na srpskom i engleskom jeziku. Za detekciju jezika predložen je klasifikator *Multinominal Naive Bayes* koji je postigao preciznost od 0.1. Klasifikatori *SVM* i *Random Forest* predloženi su za određivanje sentimenta. *SVM* je postigao veću preciznost – 0.955 nad srpskim *dataset*-om i 0.701 nad engleskim. Za problem izdvajanja aspekata (engl. *NER – Named-entity recognition*) predloženi su različiti modeli u zavisnosti od jezika – za engleske recenzije predložen je prilagođeni *SpaCy* model sa preciznošću od 0.402, a za srpske rekurentna neuronska mreža *BiLSTM-CRF* sa preciznošću od 0.828.

Ključne reči – hotelske recenzije, *NLP*, sentiment analiza, *MultinominalNB, SVM, Random Forest, NER, SpaCy, BiLSTM-CRF*

I. UVOD

Hoteli primenjuju različite tehnike procenjivanja stepena zadovoljstva posetilaca – to često uključuje knjige utisaka u hodnicima i na recepciji, slanje anketa, kao i usmeno traženje fidbeka. Iako osoblje ulaže napore da se gosti osećaju što prijatnije tokom boravka u hotelu, neretko se dešava da posetioci ne preuzimaju inicijativu za popunjavanje knjiga utisaka i nisu otvoreni da odmah i na licu mesta podele svoje iskustvo sa zaposlenima. Istraživanja su pokazala da ljudi imaju tendenciju da izbegavaju neprijatne situacije, pa se često dešava da ne prijave probleme koji su im pokvarili utisak tokom boravka u hotelu. Odsustvo ovakvih povratnih informacija doprinosi degradaciji ugleda hotela [1].

Intuitivne, *user-friendly* i lako dostupne *online* platforme za recenziranje turističkih destinacija i smeštaja značajno doprinose prevazilaženju prepreka u komunikaciji i olakšavaju posetiocima da međusobno dele svoja iskustva. Kako korisnički nalog na ovakvim platformama ne zahteva mnogo informacija (posetioci mogu da se predstave pod pseudonimom), ljudima je manje neprijatno da napišu iskreno mišljenje. Istraživanja pokazuju da je ovakav vid ostavljanja fidbeka danas veoma popularan među ljudima koji putuju – uočen je rapidan rast korišćenja *online* platformi od 98% koji iz godine u godinu ne jenjava [2]. Tipično, recenzije sadrže sveukupnu ocenu, (najčešće u formi *star-rating*-a) i tekst u slobodnoj formi koji opisuje usluge hotela. U kontekstu rekomendacionih sistema, ove recenzije služe kao sredstvo preporuke i pomažu posetiocima da lakše evaluiraju kvalitet usluge [3].

Mala je verovatnoća da osoba koja putuje nikada nije pisala recenzije tokom/nakon putovanja ili ih bar koristila kao značajan izvor informacija u fazi traženja destinacije i smeštaja. Većina ljudi neće rezervisati smeštaj u hotelu ukoliko o njemu ne mogu pročitati utiske drugih ljudi ili ukoliko su utisci negativni. Sistemi ocenjivanja danas predstavljaju neodvojiv segment turizma - razumevanje ključnih faktora koji utiču na stepen zadovoljstva posetilaca ima značajan uticaj na biznis. *Online* platforme za recenziranje turističkih destinacija i smeštaja značajno su olakšale posao menadžmentu u kontekstu dostupnosti relevantnih informacija – ali, s druge strane, takvih informacija je sve više i menadžment se neretko susreće sa *information overload*-om [4].

Mogućnost efikasne automatske obrade korisničkih recenzija predstavlja krucijalan element uspešnog savremenog hotelijerstva. Kako se na osnovu korisničkih komentara može zaključiti kojim aspektima usluge su posetioci bili zadovoljni, a kod kojih postoji prostor za napredak, automatizacija obrade ovakvih informacija doprinela bi bržem uočavanju problema i unapređenju usluge.

U radu je predstavljeno automatsko određivanje sentimenta recenzija i izdvajanje aspekata na koje je posetilac bio fokusiran prilikom ostavljanja recenzije. Cilj je postizanje lakšeg uočavanja elemenata koji predstavljaju izvor (ne)zadovoljstva posetilaca. Za postizanje ovog cilja odabrane su metode mašinskog učenja, jer su se pokazale kao efikasne u rešavanju veoma sličnih problema. Ove metode omogućavaju obradu velikog broja recenzija – posledično, menadžment će imati uvid u mišljenje većeg skupa ljudi, pa će samim tim dobijene informacije biti pouzdanije.

Dataset za potrebe ovog rada konstruisan je od recenzija na srpskom i engleskom jeziku, prikupljenih sa sajtova *booking.com* i *Kaggle*. Primećeno je da se recenzije koje nisu napisane na ciljnom jeziku u većini srodnih radova posmatraju kao *outlier*-i. Ideja je da se isproba drugačiji pristup i u obzir uzme i mišljenje stranaca sa drugih govornih područja, jer se ono može razlikovati od domaćih turista zbog kulturoloških razlika.

Za fazu detekcije jezika biće korišćen *MultinomialNB* (*Multinominal Naive Bayes*) model, a za sentiment analizu biće isprobani modeli *SVM* (*Support-vector machine*) i *RandomForest*. Za određivanje aspekata, odn. *NER* (*Named-entity recognition*) biće primenjeni različiti pristupi u zavisnosti od jezika na kojem su pisane recenzije. Za engleski jezik biće primenjen *SpaCy* model prilagođen odabranom skupu aspekata. Za srpski jezik biće primenjena rekurentna neuronska mreža – *BiLSTM* sa *CRF* (*Conditional random field*) slojem. Za evaluaciju modela biće primenjene mere *f1-score* i *accuracy*.

U narednom poglavlju biće dat pregled srodne literature koja je poslužila kao početna tačka istraživanja ovog problema. U poglavlju III. biće predstavljena primenjena metodologija. Nakon toga biće diskutovani postignuti rezultati i greške koje su modeli napravili.

II. Pregled relevantne literature

Pri izradi rešenja korišćena su saznanja iz naučnih radova u oblasti mašinskog učenja koji se bave detekcijom jezika, sentiment analizom i izdvajanjem aspekata. U nastavku će biti predstavljeni neki od relevantnih radova.

Rad [5] predlaže sistem za preporuku hotela baziran na sentiment analizi recenzija hotela. Skup podataka je preuzet sa sajta *Tripadvisor.com* i sadrži preko 50000 anotiranih recenzija hotela na engleskom jeziku. Recenzije su klasifikovane na pozitivne, negativne i neutralne, a zatim su kategorizivane na osnovu različitih aspekata na koje se odnose (npr. higijena, hrana, osoblje). Odrađeno je pretprocesiranje teksta, a zatim su primenjene tehnike vektorizacije *TF-IDF, Word2Vec* i *BERT*. Za klasifikaciju je korišćen ansambl dva *BERT* modela i *Random Forest* klasifikatora, a nakon toga recenzije su podeljene u kategorije na osnovu predefinisanih aspekata računajući *levenhstein* sličnost stringova sa odabranim rečima za svaku kategoriju. Model je postigao *Macro f1-score* od 84% i *accuracy* od 92%, što su solidni rezultati u odnosu na stanje u oblasti. Iz tog razloga, u ovom projektu će biti isprobane neke od navedenih metoda i tehnika evaluacije rešenja, ali će se pristup izdvajanju aspekata razlikovati.

Iz rada [6] preuzeta je ideja o treniranju *BiLSTM-CRF* mreže. Rad se bavi sentiment analizom recenzija hotela i izdvajanjem aspekata. Skup podataka je takođe prikupljen sa sajta *Tripadvisor.com* i sastoji se od skoro 76000 recenzija, ali je zbog nebalansiranosti podataka (82% recenzija pozitivno) za treniranje navedenog modela ručno kreiran i anotiran novi skup od 3500 rečenica. Aspekti su anotirani primenom *IOB* (engl. *Inside-Outside-Beginning*) formata za enkodiranje. Podaci su prosleđeni *BiLSTM-CRF* modelu koji ekstraktuje entitete zajedno sa pridruženim sentimentom. Na osnovu rezultata se može zaključiti da je model uspešno pronalazio aspekte kao što su *food, stuff, internal room facilities, view* i dr. Nedostatak ovog rada je što je rešenje fokusirano samo na engleski jezik.

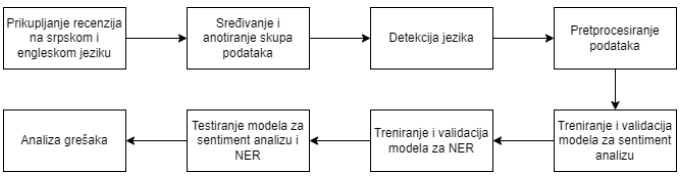
Bilingvalni pristup sentiment analizi obrađen je u radovima [7] i [8]. U radu [7] predstavljena je sentiment analiza komentara o filmovima na engleskom i kineskom jeziku, kako bi se stekla slika o stavovima ljudi iz različitih kultura. Skup podataka je formiran ručno prikupljanjem po 1000 pozitivnih i 1000 negativnih komentara na svakom jeziku. Engleski i kineski jezik nisu posmatrani odvojeno – svaki komentar je tretiran kao strim teksta koji može sadržati i engleske i kineske reči. Strim teksta je segmentisan i uklonjene su stop reči. Stem reči su konvertovane u *feature* vektore nad kojima su primenjena dva modela, *SVM* i *N-Gram*. Korišćene metrike (*precision, recall* i *f-score*) pokazale su da je *SVM* dao bolje rezultate. Takođe, sa engleskim komentarima postignuta je značajno veća vrednost *accuracy* (98%) u odnosu na kineske komentare (85%), što se može objasniti razlikom u prirodi jezika (npr. u engleskom jeziku postoji delimiter, što olakšava pretprocesiranje). Predstavljeno rešenje se razlikuje od ostalih jer nudi mogućnost simultane analize dva različita jezika. U ovom projektu će biti isproban *SVM* model koji se pokazao kao dobar izbor za posmatrani problem.

U radu [8] predstavljeno je rešenje koje se bavi identifikacijom jezika primenjujući *N-Gram* i *LIGA* (*Language Identification Graph-based Approach*) pristup. Skup podataka se sastoji od 9066 *tweet*-ova napisanih na 6 jezika – engleski, nemački, španski, francuski, italijanski i holandski. Iz ovog rada preuzeta je ideja da se ne primenjuje simultana analiza recenzija na dva jezika, nego da se prvo izvrši detekcija jezika, a zatim odradi odgovarajuće pretprocesiranje podataka u zavisnosti od jezika, kako bi rezultati bili što efikasniji.

U radu [9] dat je pregled različitih biblioteka i alata napisanih u *Python*-u, *Java*-i ili *Cython*-u koji sadrže pretrenirane *NER* modele koji se mogu modifikovati i dotrenirati kako bi se prilagodili određenim zahtevima. Neki od njih su *SpaCy, Apache OpenNLP* i *TensorFlow*. Pri poređenju rezultata zaključeno je da *SpaCy* model pruža najbolje performanse, odakle potiče i ideja da se on iskoristi i prilagodi u ovom projektu za izdvajanje aspekata na engleskom jeziku, pošto se navedene biblioteke i alati baziraju samo na engleskom jeziku.

III. Metodologija

Postupak sprovođenja sentiment analize i izdvajanje aspekata odvija se u nekoliko faza (Slika 1) – svaka od njih biće opisana u ovom poglavlju. Najpre je bilo neophodno formirati *dataset* prikupljanjem recenzija na engleskom i srpskom jeziku – zatim je vršeno anotiranje tog skupa podataka. Faza detekcije jezika primenjena je kako bi se preciziralo koji tip pretprocesiranja i koji model treba primeniti nad recenzijama. Nakon toga odrađeno je treniranje, validacija i testiranje modela za sentiment analizu i *NER*. Konačno, sprovedena je analiza recenzija nad kojima su modeli pravili greške.



Slika 1 Pregled koraka sprovedenih tokom razvijanja rešenja

A. Prikupljanje i anotiranje podataka

Skup podataka za sentiment analizu se sastoji od engleskih i srpskih recenzija. Recenzije su prikupljane skrejpovanjem veb-sajta za recenziranje turističkih destinacija i smeštaja, kao i korišćenjem javno dostupnog skupa podataka – što će biti detaljnije objašnjeno u nastavku ovog poglavlja.

a. Engleski jezik

Engleske recenzije su preuzete iz skupa podataka sa sajta *Kaggle* [10] koji se sastoji od 10000 hotelskih recenzija. Izdvojeno je i anotirano 1044 recenzija. Od 26 atributa koje sadrži skup podataka izdvojeni su *reviews.text* i *reviews.rating*, koji predstavljaju tekst recenzije i ocenu koju je posetilac dao

hotelu, u rasponu od 1 do 5. Anotiranje recenzija je sprovedeno na osnovu ocene:

- ocena 1 ili 2 – negativan sentiment;
- ocena 3 – neutralan sentiment;
- ocena 4 ili 5 – pozitivan sentiment.

Prilikom prikupljanja recenzija vodilo se računa o balansiranosti podataka – uočeno je da je neutralnih recenzija veoma malo u odnosu na negativne i pozitivne. U cilju postizanja boljih rezultata, odlučeno je da se neutralne recenzije pripoje negativnim i da se u rešenju stavi fokus na predikciju dve klase. Klase su predstavljene numeričkim vrednostima: -1 za negativan sentiment i 1 za pozitivan sentiment. Takođe, primećeno je da značajan broj recenzija pored komentara posetioca sadrži i odgovor menadžera – ovi delovi su uklanjani.

Za potrebe izdvajanja aspekata iz engleskih recenzija korišćen je podskup *dataset*-a prikupljenog za sentiment analizu. Pomoću *SpaCy* alata [11] anotirano je 500 engleskih recenzija. U svakoj od njih označene su reči koje predstavljaju jedan od sledećih tagova: *PERSONNEL, WEATHER, FOOD, LOCATION, HYGIENE, FURNITURE, PRICE*. Izbor ovih tagova nastao je kao posledica čitanja i analize prikupljenih recenzija – uočeno je da su ovo najčešći aspekti koji sa sobom nose sentiment.

a. Srpski jezik

Za prikupljanje dela skupa podataka na srpskom jeziku korišćen je alat za skrejpovanje *ParseHub* [12] pomoću kojeg je izvršeno parsiranje stranica sa sajta *booking.com* [13]. Prikupljeno je 1389 recenzija, pri čemu se vodilo računa o izbalansiranosti obe klase. Kako je veb-sajt organizovan tako da postoje predefinisana polja za pohvale i zamerke, to je iskorišćeno u svrhe skrejpovanja i razdvajanja pozitivnih i negativnih recenzija pomoću alata. Numeričke predstave klasa su iste kao u slučaju engleskih recenzija: -1 za negativan i 1 za pozitivan sentiment.

Uočen je zanemarljivo mali broj rečenica napisan na ćirilici. Ovakve recenzije su smatrane *outlier*-ima i uklanjane su iz skupa podataka. Recenzije koje su se po dužini značajno izdvajale od ostatka *dataset*-a razdvajane su na logičke celine i na taj način konvertovane u više recenzija.

Za potrebe izdvajanja aspekata iz srpskih recenzija ručno je anotirano 300 recenzija primenom *IOB2* (*Inside-Outside-Beginning*) [14] formata za enkodiranje. Ovaj format podrazumeva da se u frazi koja predstavlja neki entitet prva reč označi sa nazivom entiteta i slovom *B* (npr. *B-FOOD*), a svaka sledeća slovom *I* (npr. *I-FOOD*). Slovom *O* označavaju se fraze koje nisu od značaja za detekciju aspekata, odn. ne pripadaju ni jednom od prethodno navedenih tagova. Isti skup tagova je primenjen na recenzije pisane na oba jezika.

B. Eksplorativna analiza podataka

Posmatrani su skupovi recenzija po grupama pozitivno – negativno, kako bi se uočilo koje reči nose izražen pozitivan, odn. negativan sentiment i da li se neke reči ponavljaju u oba sentimenta. Za svaku klasu izvučeno je top 50 najčešće ponavljanih reči koje su predstavljene pomoću *cloud of words*.

Kod engleskih recenzija (Slika 2) uočeno je da se u obe grupe veoma često ponavljaju sledeće reči: *room, hotel, night, bed, staff, breakfast, time, bathroom, location, parking*. Ove reči su uključene u listu *stop words* tokom pretprocesiranja. Kako je korišćen američki *dataset* kao izvor podataka, primećeni su nazivi gradova *San Diego* i *San Francisko*. Oni su takođe uklonjeni u fazi pretprocesiranja.

Slika 2 Ponavljanje reči u engleskim negativnim recenzijama (levo) i pozitivnim recenzijama (desno)

Kod srpskih recenzija (Slika 3) uočeno je da se ponavljaju reči koje nemaju semantičku vrednost u kontekstu sentimenta – pomoćni glagoli, veznici i predlozi koji su uključeni u listu *stop* reči. Takođe, uočeno je da se često ponavljaju grupacije reči, odn. sintagme poput sledećih: *za svaku preporuku, ljubazno osoblje, odlična lokacija, pogled sa terase, loša zvučna izolacija, kao na slikama, buka iz kafića* i sl. Iz ovog razloga odlučeno je da se isprobaju *bigram*-i i *trigram*-i prilikom treniranja modela.

Slika 3 Ponavljanje reči u sprskim negativnim recenzijama (levo) i pozitivnim recenzijama (desno)

Izvršena je analiza ponavljajućih znakova interpunkcije. Provereno je da li je upotreba uzastopnih tačaka, uzvičnika i upitnika povezana sa sentimentom recenzija. Pošlo se od pretpostavke da bi ovakav stil pisanja mogao biti povezan sa frustracijom gosta i ukazivati na nezadovoljstvo.

Tabela 1 prikazuje ukupan broj recenzija na svakom jeziku kod kojih je uočeno ponavljanje znakova interpunkcije. Upitnici i uzvičnici se pojavljuju u zanemarljivo malom broju recenzija. Primećeno je da se tačke u negativnim recenzijama na engleskom jeziku ističu, međutim ostavljanje ovih znakova interpunkcije nije doprinelo boljim rezultatima prilikom treniranja. U finalnom rešenju odlučeno je da se svi znaci interpunkcije uklone u fazi pretprocesiranja.

	Engleski jezik		Srpski jezik	
	Negativno	Pozitivno	Negativno	Pozitivno
Tačka	221	81	41	54
Upitnik	0	0	1	0
Uzvičnik	8	9	13	10

Tabela 1 Pregled broja recenzija sa ponavljajućim znakovima interpunkcije

Sprovedena je analiza korišćenja *Caps Lock* slova u recenzijama jer u *online* komunikaciji ovo implicira glasno pričanje. Ideja je da se proveri da li možda ovakav stil pisanja može da se dovede u vezu sa pozitivnim/negativnim sentimentom. Rezultati su sledeći:

- engleske recenzije – 346 negativnih, 280 pozitivnih (izostavljeno je pisanje zamenice *I*);
- srpske recenzije – 102 negativne, 126 pozitivnih.

Zaključeno je da ljudi pišu na ovaj način i kada su nezadovoljni i kada su uzbuđeni. Tokom pretprocesiranja sve reči su svedene na mala slova.

C. Pretprocesiranje podataka

Saznanja iz faze eksplorativne analize podataka iskorišćena su tokom pretprocesiranja. U nastavku će biti predstavljeni koji su to koraci sprovedeni u ovoj fazi. Kontekst određenih koraka se ponavlja u oba jezika, ali je realizacija drugačija – stoga će pretprocesiranje biti predstavljeno odvojeno, za svaki jezik. Za konverziju string reprezentacije u numeričke vrednosti primenjen je *TfidfVectorizer* iz *sklearn* biblioteke.

Nad recenzijama pisanim na engleskom jeziku sprovedene su sledeće akcije:

- Uklanjanje *stop* reči pomoću biblioteke *nltk* i liste reči prikupljene tokom eksplorativne analize;
- Ispravljanje *typo* grešaka u recenzijama pomoću biblioteke *spellchecker*;
- Svođenje teksta na mala slova;
- Uklanjanje specijalnih karaktera;
- Uklanjanje *emoji*-ja pomoću rečnika gde ključevi predstavljaju reči *excellent, great, bad* i *ok*, a vrednosti predstavljaju liste *emoji*-ja koji odgovaraju opisu iz ključa;
- Uklanjanje *url*-ova pomoću *regex pattern*-a;
- *Lemmatization*, odn. morfološka analiza reči – različiti oblici reči mapirani su na osnovni oblik pomoću *WordNetLemmatizer*-a iz biblioteke *nltk*;
- *Stemming*, odn. svođenje reči na korenski oblik, pomoću *PorterStemmer*-a iz biblioteke *nltk*.

Nad recenzijama pisanim na srpskom jeziku sprovedene su sledeće akcije:

- Uklanjanje stop reči pomoću ručno pravljene liste koja uključuje zamenice, predloge, veznike (ne računajući *iako, ni, niti, a, ali* jer upućuje na negativan sentiment), priloge, skraćenice i često ponavljane netačne formulacije (*ustvari, odole, kolko, al*);
- Svođenje teksta na mala slova;
- Uklanjanje specijalnih karaktera;
- Uklanjanje emoji-ja pomoću rečnika gde ključevi predstavljaju reči *odlično, super, loše* i *okej*;
- *Šišanje* latinice, odn. uklanjanje dijakritika kako model ne bi iste reči tretirao kao različite;
- *Lemmatization* i *stemming* po uzoru na rešenje [15] koje je prilagođeno specifičnostima srpskog jezika.

D. Detekcija jezika

Engleski i sprski jezik su značajno različiti po određenim slovima i gramatici (npr. padeži u srpskom jeziku otežavaju svođenje reči na korenski jezik). Faza detekcije jezika je uvedena kako bi se mogla doneti odluka koji koraci pretprocesiranja treba da budu sprovedeni nad recenzijom i koji model treba da bude iskorišćen u svrhe predikcije. Posledično, za samu detekciju nije vršeno nikakvo pretprocesiranje, jer bi se na taj način uklonili elementi i skupovi karaktera koji nedvosmisleno ukazuju na određeni jezik (npr. dijakritike u srpskom jeziku, ponavljanje uzastopnih samoglasnika u engleskom jeziku i sl.). Za potrebe rešavanja ovog problema isprobani su modeli *SVM* i *Multinomial Naive Bayes*, jer su se u srodnim radovima postigli značajno dobre rezultate prilikom detekcije jezika.

E. Sentiment analiza

Za potrebe sentiment analize korišćeni su i poređeni modeli *SVM* i *RandomForest*. Ovi modeli su izabrani jer su preporučeni kao odgovarajući modeli u srodnoj literaturi.

SVM se smatra jednim od uspešnijih algoritama za problem binarne klasifikacije. Za potrebe ovog rada testirani su različiti kerneli – *linear, rbf, poly* i *sigmoid*. Isprobane su različite vrednosti za parametre *C* i *gamma*. Posmatrano je kako se model ponaša u zavisnosti od toga da li je treniran nad *unigram*-ima, *bigram*-ima ili *trigram*-ima.

RandomForest se smatra veoma dobrim modelom za rešavanje multiklasnih problema, ali postiže dobre rezultate i u binarnoj klasifikaciji. Smatra se da je prilikom korišćenja ovog modela opasnost od *overfitting*-a na niskom nivou, ali faza treniranja traje duže u odnosu na *SVM*. Za potrebe ovog rada isprobani su parametri *max_depth, max_features, min_samples_leaf, min_samples_split, n_estimators*. Takođe je posmatrano kako se model ponaša u zavisnosti od odabranog *n-gram*-a.

F. Izdvajanje aspekata iz recenzija – NER

NER predstavlja proces automatskog izdvajanja entiteta iz teksta. Kako je priroda engleskog i srpskog jezika veoma različita, izdvajanje aspekata iz recenzija je odrađeno drugačije u zavisnosti od toga na kom jeziku je pisana recenzija.

Za primenu *NER*-a nad engleskim recenzijama korišten je prilagođeni *SpaCy* model. Ovaj model zahteva da *dataset* bude u jasno definisanom formatu – sve reči u recenzijama su obeležene odgovarajućim tagovima (lista tagova je definisana u poglavlju o anotiranju podataka). Preciznost ovog modela zavisi od broja iteracija tokom faze treniranja i različite vrednosti ovog parametra su isprobane. Pre svake iteracije odrađeno je mešanje podataka, kako bi se izbegao scenario da model pravi generalizacije na osnovu redosleda recenzija.

Za primenu *NER*-a nad srpskim recenzijama korištena je rekurentna neuronska mreža *BiLSTM* sa *CRF* slojem. Bidirekcioni *LSTM* se sastoji od dva *LSTM*-a – jedan uzima ulaznu reč u *forward*, a drugi u *backward* direkciji. Korišćenjem izlaza ove dve mreže dobija se *context* koji sadrži informaciju o okruženju svakog tokena. *Context* zatim služi kao ulaz u *CRF* sloj koji generiše predikciju. Ovaj sloj je značajan jer ume da nauči korisne *constraints* koji poboljšavaju *accuracy* rezultat – npr. tag prve reči u rečenici treba da počne sa *B*- ili *O*, ali ne *I*. Prilikom treniranja isprobani su sledeći parametri koji se tiču *BiLSTM* mreže – *dropout* i *recurrent_dropout*.

IV. Rezultati i diskusija

Prikupljeni su skupovi podataka od 1044 engleske i 1389 srpskih recenzija. Za fazu testiranja odmah je izdvojeno i sklonjeno 10% podataka od svakog *dataset*-a. U slučaju engleskog, skup podataka za testiranje sadrži 104 recenzije, a 940 recenzija se nalazi u trening skupu. Za srpski – 135 i 1254 recenzija, respektivno. Prilikom podele *dataset*-ova vodilo se računa da budu izbalansirani – pa tako i trening i test skupovi podataka imaju 50:50 odnos pozitivnih i negativnih recenzija. Za potrebe treniranja i testiranja faze detekcije jezika korišćene su unije test (239 recenzija) i trening *dataset*-ova (2194 recenzija).

U fazi detekcije jezika ciljna varijabla može imati vrednosti -1 ili 1, gde je -1 indikator da je recenzija napisana na engleskom jeziku, a 1 je indikator da je u pitanju srpski jezik.

Slično važi i za kontekst sentiment analize: -1 indicira negativan sentiment, a 1 pozitivan. U slučaju *NER*-a očekivani izlaz je recenzija u kojoj je svaka reč označena odgovarajućim tagom.

A. *Detekcija jezika*

Na osnovu rezultata *MultinominalNB* modela za detekciju jezika ispostavilo se da problem razlikovanja srpskog i engleskog jezika nije težak. Prilikom faze treniranja korišćena je podela 80:20. Model je i nad validacionim i nad test skupom podataka postigao *accuracy* i *f-score* od 1.0. Nije bilo potrebe za *fine-tuning*-om parametara, jer se model iz prve odlično snašao u ovoj fazi. Pretpostavlja se da je razlog za to priroda engleskog i srpskog jezika – čoveku bi bilo intuitivno jasno da su u pitanju dva različita jezika i bez da poznaje jedan od njih, prosto zbog različitih slova i konstrukcije rečenice. Verovatno je na sličan način model uspeo da uoči različite šablone rasporeda karaktera u rečenicama. Problem dodatno olakšava činjenica da je u pitanju binarna klasifikacija dva veoma različita jezika.

B. *Sentiment analiza*

Na odvojenim trening skupovima srpskih i engleskih recenzija trenirana su dva modela – *SVM* i *RandomForest*. Korišćena je podela 80:20. Vršena je unakrsna validacija nad trening delom skupa (*k-fold*, gde je k = 5) pomoću *GridSearchCV*-ja, a zatim je testirano nad validacionim skupom šta je model naučio. Finalno testiranje je odrađeno nad test skupom, podacima koje model nije video tokom trening faze.

Za potrebe *fine-tuning*-a iskorišćene su mogućnosti koje nudi *GridSearchCV*, odn. pomoću njega je birana kombinacija parametara sa kojima su modeli postigli najbolje rezultate. U nastavku će biti predstavljeni posmatrani parametri za svaki model.

Za pretvaranje tokena u numeričke vrednosti korišćen *TfidfVectorizer* – isprobane su različite vrednosti njegovog parametra *ngram_range* – (1, 1), (1, 2) i (1, 3). Pretpostavka je bila da će korišćenje *bigram*-a ili *trigram*-a doprineti boljim rezultatima u srpskom jeziku, jer je uočeno ponavljanje određenih sintagmi.

Za potrebe treniranja *SVM* modela prosleđeni su parametri sa sledećim vrednostima:

- *kernel*: *rbf, poly, sigmoid*;
- *C*: 0.1, 1, 10, 100, 200, 1000;
- *gamma*: 1, 0.1, 0.01, 0.001, 0.0001.

Izdvojeni su najbolji *SVM* modeli za srpski i engleski jezik. Tabela 2 prikazuje rezultate.

	Srpski jezik	Engleski jezik
ngram_range	(1, 2)	(1, 1)
C	20	10
kernel	*sigmoid*	*rbf*
gamma	-	1
f-score (validation)	0.95618	0.85638
f-score (testing)	0.95588	0.70192

Tabela 2 Pregled najboljih SVM modela

S obzirom na to da je rešavan problem binarne klasifikacije, očekivano je bilo da će se linearni *SVM* pokazati bolje u odnosu na druge kernele. Desilo se suprotno – na osnovu rezultata zaključuje se da određivanje linearne granice odluke nije najbolji izbor za rešavanje ovog problema.

Za potrebe treniranja *RandomForest* modela prosleđeni su parametri sa sledećim vrednostima:

- *max_depth*: 10, 20, 30, 40;
- *max_features*: *auto, sqrt*;
- *min_samples_leaf*: 1, 2;
- *min_samples_split*: 2, 5;
- *n_estimators*: 200, 400, 600.

Izdvojeni su najbolji *RandomForest* modeli za srpski i engleski jezik. Tabela 3 prikazuje rezultate.

	Srpski jezik	Engleski jezik
ngram_range	(1, 3)	(1, 1)
max_depth	30	40
min_samples_leaf	2	2
min_samples_split	5	5
n_estimators	600	600
f-score (validation)	0.91633	0.83511
f-score (testing)	0.95555	0.70192

Tabela 3 Pregled najboljih RandomForest modela

Iz priloženih rezultata se može zaključiti da je *SVM* dao bolji rezultat od *RandomForest* modela. Pretpostavka da će *bigram*-i i *trigram*-i biti pogodni u slučaju srpskih recenzija je tačna, što se pokazalo u slučaju oba modela. Ono što takođe može primetiti je razlika između rezultata na srpskom i na engleskom jeziku – rezultati postignuti nad skupom srpskih recenzija su veoma dobri, dok su u slučaju engleskog znatno slabiji. Pretpostavlja se da je jedan od razloga taj što su engleske recenzije mnogo duže od srpskih. Detaljnija analiza grešaka biće data u sledećem poglavlju.

C. *Izdvajanje aspekata – NER*

Anotirano je ukupno 500 engleskih recenzija za potrebe NER-a. Odvojeno je 10% *dataset*-a za testiranje. Preostalih 450 recenzija korišćeno je u svrhe treniranja. Primenjena je podela 80:20 na trening i validacioni skup.

Model je treniran u 10 iteracija i postigao je sledeće rezultate nad test skupom podataka:

- *Precision*: 0.40163
- *Recall*: 0.43956
- *F-Score*: 0.41970

Anotirano je ukupno 300 srpskih recenzija. Isto kao i u slučaju engleskog jezika, odvojeno je 10% *dataset*-a za testiranje. Preostalih 270 recenzija iskorišteno je za treniranje. Takođe je primenjena podela 80:20 na trening i validacioni skup.

Istrenirani *BiLSTM-CRF* model je postigao *accuracy* od 0.82845 nad testing skupom podataka.

Iz navedenog se može uočiti da *BiLSTM-CRF* daje znatno bolji rezultat u odnosu na *SpaCy*. Ovo nije očekivano ponašanje jer je engleski *dataset* veći u odnosu na srpski. U narednom poglavlju biće analizirane greške koje su modeli napravili nad testnim skupom podataka.

V. ANALIZA GREŠAKA

Za potrebe analize grešaka rešenje je implementirano tako da se nakon testiranja modela u *csv* fajlu čuvaju recenzije nad kojim je vršeno testiranje, prediktovani sentiment i originalni sentiment. Za potrebe *NER*-a u posebnim fajlovima su čuvani rezultati u obliku rečenica gde je svaka reč označena prediktovanim tagom. Izlazni fajlovi i izvorni kod su dostupni na repozitorijum [16].

A. Sentiment analiza

Modeli su postigli odlične rezultate kada je u pitanju sentiment analiza recenzija pisanih na srpskom jeziku. Međutim, rezultati nisu savršeni i uočeni su određeni problemi. Tabela 4 prikazuje reprezentativne primere sa kojima se modeli nisu najbolje snašli.

	Original	Predikcija
Narušavaju vašu privatnost i krše zakon time što vam pri dolasku slikaju lične karte.	-1	1
Urednost i čistoća kako smeštaja, tako i SPA centra.	1	-1
Hladna voda u bazenu. Isto je tako bilo i pre 5 godina.	1	-1

Tabela 4 Pregled srpskih recenzija nad kojima su modeli grešili

U prvom primeru model greši jer rečenica sadrži neutralne reči – *zakon*, *dolazak*, *slikati*, *lična karta*. Reči koje bi mogle da ukažu na negativan sentiment (*kršiti*, *narušavati*) se nisu mnogo puta (ili uopšte) pojavljivale u trening skupu podataka, što je doprinelo da model generiše netačnu predikciju. Povećanje trening skupa podataka bi moglo da smanji verovatnoću grešaka ovog tipa.

U drugom i trećem primeru posetilac je recenziju uneo na veb-sajtu u polje predviđeno za pozitivne utiske – rečenice su napisane u obliku tvrdnje i ne sadrže tipične reči koje bi mogle da ukažu na pozitivan sentiment. Naprotiv, kako se reči i sintagme poput *čistoća smeštaja*, *urednost*, *hladna voda* uglavnom spominju u negativnoj konotaciji, model je pravio greške prilikom predikcije.

U slučaju engleskih recenzija primećeno je da su značajno duže od recenzija pisanih na srpskom jeziku. Uočeno je da se u tim dugačkim recenzijama neretko nalaze izmešani pozitivni i negativni utisci – pretpostavlja se da veb-sajt sa kojeg je vršeno skrejpovanje recenzija nije sadržao odvojena polja za ovu svrhu. U nastavku je primer jedne takve recenzije (crvenom bojom su označeni negativni delovi, a zelenom bojom pozitivni):

If you like King rooms, be prepared not to have a bath. None of their King rooms have baths, only showers. With regard to noise: the rooms at the freeway end of the hotel are very noisy. We were living out of our suitcases because of the ridiculous size of the drawers in the room. The pool is the size of a large hot-tub. Breakfast had many items but lacked quality: the only good things we found were yoghurt and fresh fruit. Hotel staff were helpful and did their best to resolve our issues.

Ovaj problem bi mogao da se umanji razbijanjem dugačkih engleskih recenzija na nekoliko kraćih rečenica, gde bi bilo jasnije određeno šta nosi pozitivan, a šta negativan sentiment.

Takođe, potencijalan problem predstavlja i anotiranje sentimenta na osnovu ocena koje su ostavljali korisnici. Ocene nisu konzistentno merilo, jer predstavljaju subjektivnu procenu posetioca.

B. Izdvajanje aspekata, NER

Pošto je prilagođeni *SpaCy* model dao neočekivano loš rezultat, i u ovom slučaju je izvršena analiza grešaka. Uočeno je da je sačuvani model u rečenicama test skupa izdvajao sledeće fraze kao najčešće indikatore za navedene tagove:

- *LOCATION: location, Westfield shopping centre on Market street, panoramic view, walking distance, few minutes walk, smartly located, city views, San Francisco;*
- *PERSONNEL: welcome, front desk, staff, housekeeping, friendly, efficient, amazing service;*
- *PRICE: expensive, cheap, pricey, budget, charged;*
- *HYGIENE: clean, well maintained;*
- *FURNITURE: bed, room, mini-suite, pool area, bathroom, TV;*
- *FOOD: restaurant, food, sushi bar, lunch order, breakfast.*

Deluje da je *accuracy* znatno manji nego što bi trebalo. Iščitavanjem postignutih rezultata stiče se utisak da se model relativno dobro snašao u predikcijama.

S obzirom na to da je anotiranje rađeno ručno, uočeno je da postoje propusti u samom procesu anotiranja skupa podataka. Na primer, nešto što predstavlja hranu nije označeno kao *FOOD*, a model je to predvideo – što je u biti ispravno. Dakle, model je radio tačne predikcije, ali prilikom računanja rezultata predikcija je bila poređena sa *OTHER* tagom i to je negativno uticalo na *accuracy*.

Pored toga, sam način merenja rezultata ne uzima u obzir činjenicu da u ovom problemu može postojati više smislenih rešenja. Na primer, neke veće fraze se mogu razdvojiti na više manjih pri čemu će i dalje pripadati istoj kategoriji (npr. *Westfield shopping centre on Market street*).

Nasuprot *SpaCy* modelu, *BiLSTM-CRF* model je postigao mnogo veću tačnost, nego što bi se moglo zaključiti na osnovu analize rezultata. Uočeno je da model za svoje predikcije koristi samo dva taga. Kako je ručno anotiranje vremenski zahtevan proces i podložan greškama, prikupljeni skup podataka je veoma mali. Posledično, model je naučio da je najefikasnije da predviđa samo dva taga, jer tako daje bolji rezultat nego kada pokušava da predviđa sve tagove. Pretpostavlja se da bi sa većim skupom podataka i rezultati bili bolji i smisleniji.

VI. ZAKLJUČAK

Rad se bavi temom sentiment analize hotelskih recenzija na srpskom i engleskom jeziku i izdvajanjem aspekata na koje se recenzija odnosi, u cilju automatizovanog određivanja uzroka (ne)zadovoljstva posetilaca. Kako sistemi ocenjivanja danas predstavljaju neodvojiv segment turizma, rešenje predstavljeno u radu pomoglo bi menadžmentu u bržem i efikasnijem razumevanju ključnih faktora koji utiči na stepen zadovoljstva posetilaca.

Za detekciju jezika predložen je model *MultinomialNB* koji je postigao preciznost od 1.0 nad testnim skupom podataka. Za potrebe sentiment analize predloženi su modeli *SVM* i *RandomForest*. *SVM* model je postigao za nijansu bolje

rezultate od *RandomForest*-a – u slučaju recenzija pisanih na srpskom jeziku postignut je *f-score* od 0.955; u slučaju engleskog ta vrednost iznosi 0.701. Za potrebe izdvajanja aspekata na srpskom jeziku korišten je *BiLSTM-CRF* sa preciznošću od 0.828. Za *NER* nad recenzijama na engleskom jeziku primenjen je *SpaCy* model i postignuta je preciznost od 0.402.

Uočeno je da su postignuti lošiji rezultati za sentiment analizu na engleskom jeziku u odnosu na srpski jezik – unapređenje bi se moglo postići razdvajanjem dugačkih engleskih recenzija na kraće recenzije za koje se nedvosmisleno može definisati sentiment. Unapređenje izdvajanja aspekata moglo bi se postići većim skupom podataka i preciznijim anotiranjem, kako model ne bi snosio posledice zbog ljudskih grešaka. Dakle, u kontekstu *NER*-a glavni nedostatak je ograničenje u pogledu skupa podataka.

Ono po čemu se ovaj rad izdvaja od većine srodnih radova je činjenica da je uključena detekcija jezika koja doprinosi prevazilaženju kulturoloških razlika kada su u pitanju utisci o turističkim destinacijama. U ovom rešenju kombinovana su saznanja koja se tiču detekcije jezika, sentiment analize i izdvajanja aspekata. Ovo predstavlja logičan sled akcija koje bi bile neophodne kada bi se ovakav sistem zaista koristio u turističkoj industriji.

Ono što predstavlja problem kada je u pitanju sentiment analiza (i sistemi za recenziranje generalno) je mogućnost postojanja *fake* recenzija, odn. utisaka postavljenih od strane botova u cilju unapređenja ugleda hotela. Mnogi veb-sajtovi nude mogućnost korisnicima da lajkuju one recenzije koje smatraju istinitim i relevantnim. Predloženo rešenje bi se moglo unaprediti korišćenjem informacije o tome koliki broj ljudi neku recenziju smatra korisnom, kao i implementacijom *opinion spam* detektora.

LITERATURA

[1] Berezina, Katerina, et al. "Understanding satisfied and dissatisfied hotel customers: text mining of online hotel reviews." Journal of Hospitality Marketing & Management 25.1 (2016): 1-24.

[2] O'connor, Peter. "User-generated content and travel: A case study on Tripadvisor. com." Enter. Vol. 2008. 2008.

[3] O'Mahony, Michael P., and Barry Smyth. "Learning to recommend helpful hotel reviews." Proceedings of the third ACM conference on Recommender systems. 2009.

[4] Mellinas, Juan Pedro, and Sofía Reino. "eWOM: the importance of reviews and ratings in tourism marketing." Strategic perspectives in destination marketing. IGI Global, 2019. 143-173.

[5] Biswarup Ray, Avishek Garain, Ram Sarkar (2020). An ensemble-based hotel recommender system using sentiment analysis and aspect categorization of hotel reviews

[6] Thang Tran Xuan, Van-Nam Huynh, Hung Ba (2019). Measuring Hotel Reviews Sentiment: An Aspect-Based Sentiment Analysis Approach

[7] Wu He, Chuanyi Tang (2014). A Billingual Approach for Conducting Chinese and English Social Media Sentiment Analysis

[8] Erik Tromp, Mykola Pechenizkiy. Graph-Based N-Gram Language Identification on Shot Texts

[9] Hemlata Shelar, Gagandeep Kaur, Neha Heda, Poorva Ahrawal (2020). Named Entity Recognition Approaches and Their Comparison for Custom NER Model. Science & Technology Libraries

[10] https://www.kaggle.com/meetnagadia/hotel-reviews

[11] http://agateteam.org/spacynerannotate

[12] ParseHub, https://www.parsehub.com/

[13] https://www.booking.com/reviews

[14] Krishnan, Vijay, and Vignesh Ganapathy. "Named entity recognition." Stanford Lecture CS229 (2005).

[15] https://github.com/nikolamilosevic86/SerbianStemmer

[16] https://github.com/natasa-ivanovic/siap