

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

Predlog projekta iz SIAP-a

---

## Definicija problema

Cilj projekta je klasifikacija recenzija hotela na pozitivne, negativne i neutralne, kako bi se odredilo da li su gosti zadovoljni uslugom, kao i izdvajanje aspekata na koje su gosti bili fokusirani prilikom ostavljanja recenzija. U projektu će biti uzete u obzir recenzije različitih hotela pisane na srpskom i engleskom jeziku.

## Motivacija

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 efikasno 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, koje bi zbog kulturoloških razlika u nekim slučajevima moglo da se razlikuje od mišljenja domaćih turista.

## Relevantna literatura

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

### Zadatak rada

Rad predlaže sistem za preporuku hotela baziran na sentiment analizi recenzija hotela. Recenzije su klasifikovane na pozitivne, negativne i neutralne, a zatim su kategorizovane na osnovu različitih aspekata na koje se odnose (npr. higijena, hrana, osoblje, lokacija).

### Metodologija

Odrađeno je pretprocesiranje teksta, a zatim su primenjene tehnike vektorizacije TF-IDF, Word2Vec i BERT. Za klasifikaciju je korišten 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 svaki kategoriju.

### Skup podataka

Skup podataka je preuzet sa sajta Tripadvisor.com i sadrži preko 50000 anotiranih recenzija hotela na engleskom jeziku.

### Evaluacija rešenja

Rešenje je evaluirano Macro F1-score i accuracy merama. Skup podataka je podeljen na trening, validacioni i test skup u razmeri 70:20:10.

### Rezultati

Model je postigao Macro F1-score od 84% i accuracy od 92.36%, što su solidni rezultati u odnosu na stanje u oblasti, te bismo primenile korištene metode i u našem projektu.

### Zaključci

U projektu bismo se fokusirale na sličnu temu i koristile bismo neke od navedenih metoda, kao i navedene tehnike evaluacije rešenja. Naš projekat bi se razlikovao po tome što bismo radile sa dva jezika i koristile druge metode za kategorizaciju.

## [2] A bilingual approach for conducting Chinese and English social media sentiment analysis

### Zadatak rada

Predstavljeno je rešenje koje se bavi analizom komentara o filmovima objavljenih na društvenim mrežama – ideja je da se stekne slika o stavovima ljudi iz različitih kultura sa različitim stepenom obrazovanja. Primenjena je sentiment analiza komentara pisanih na engleskom i kineskom jeziku (dva najrasprostranjenija jezika na svetu). Različiti jezici otežavaju da sentiment analiza održi prihvatljiv nivo preciznosti i konzistentnosti – cilj ovog rada je prevazilaženje ove barijere primenom bilingvalne sentiment analize.

### Metodologija

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 vectors nad kojima su primenjena dva modela – SVM i N-Gram. Analizirani su entiteti u pozitivnim i negativnim komentarima pomoću alata za vizuelizaciju kako bi se uočilo da li neke reči „vuku“ određeni sentiment.

### Skup podataka

Podaci su ručno prikupljeni – engleski komentari su prikupljeni sa Facebook-a i Tweeter-a (1000 pozitivnih i 1000 negativnih), a kineski komentari su prikupljeni sa Tianya i Weixin sajtova (1000 pozitivnih i 1000 negativnih).

### Evaluacija rešenja

Korištene su tri metrike – precision, recall i F-score. Dataset je podeljen na trening i test set u odnosu 80:20. Posmatrano je kako se metrike ponašaju za različite odnose pozitivnih i negativnih komentara u testnom skupu, kao i uticaj ponavljajućih sponova. Analizirano je koliko je proces time-consuming i kako analiza jednog i drugog jezika utiče na performanse.

### Rezultati

SVM se pokazao kao robustniji model u odnosu na N-Gram – postigao je bolje rezultate u sve tri metrike. Analiza kineskog teksta je time-consuming proces i značajno utiče na performanse za razliku od analize engleskog teksta (269s naspram 13s). Accuracy kineskih komentara je značajno niži od engleskih (85% naspram 98.5%).

### Zaključci

Rezultati pokazuju da je accuracy SVM značajno bolji od N-Gram-a. Klasifikacija engleskih reči je efikasnija od klasifikacije kineskih (što je logična posledica postojanja delimitera u engleskom). Lošiji rezultati analize kineskih komentara objašnjeni su činjenicom da i dalje postoji veoma malo alata za vizuelizaciju rezultata sentiment analize. Predstavljeno rešenje se izdvaja od ostalih jer nudi mogućnost simultane analize dva različita jezika (odakle smo i dobile inspiraciju za projekat). U našem projektu isprobale bismo SVM jer se pokazao kao dobar klasifikator. U odnosu na ovaj rad, drugačije bismo uradile sledeće – postojala bi faza detekcije jezika, pa bismo na osnovu toga primenjivale odgovarajuće pretprocesiranje teksta.

## [3] Measuring Hotel Review Sentiment: An Aspect-Based Sentiment Analysis Approach

### Zadatak rada

Rad se bavi analizom recenzija hotela i izdvajanjem aspekata na koje su korisnici usluga fokusirani prilikom ostavljanja komentara. Cilj ove analize je uočavanje onih elemenata koji mogu biti relevantni za unapređenje poslovanja hotela, na sveopšte zadovoljstvo posetilaca hotela i menadžera.

### Metodologija

Sprovedeni su aspect mining i sentiment analysis recenzija hotela. Pretprocesiranje teksta uključuje detekciju jezika, uklanjanje šuma i lematizaciju. Podaci su prosleđeni BiLSTM-CRF modelu koji ekstrahuje entitete zajedno sa pridruženim sentimentom. Implementiran je topic model korišćenjem LDA u cilju određivanja tema i ključnih reči koje pripadaju toj temi.

### Skup podataka

Podaci su prikupljeni skrejpovanjem sajta TripAdvisor – sve recenzije su pisane na engleskom jeziku. Dataset sadrži ukupno 75933 recenzija za 410 hotela u Ho Chi Minh-u (Vijetnam). Recenzije u rasponu od 3 do 5 zvezdica čine 82% datasea. Za treniranje BiLSTM-CRF modela ručno je kreiran i anotiran novi dataset koji sadrži 3500 rečenica. Aspekti su anotirani primenom IOB encoding formata.

## Evaluacija rešenja

Evaluacija ATE (aspect term extraction) i APC (aspect polarity classification) taskova je odrađena odvojeno. Primenjena je 10-fold kros validacija na trening setu. Za evaluaciju ATE korištene su metrike precision, recall i F1 score. Za evaluaciju APC je korišten accuracy.

## Rezultati

Topici poput Food, Stuff, Internal Room Facilities, View prisutni su u 69.6% recenzija. Topike u vezi sa hranom najčešće prati pozitivan sentiment (čak 86698 recenzija). Negativan sentiment pratio je topike koji se tiču Room i View (28170 recenzija). Posetioци koji su ostavili 5-star rating bili su najčešće zadovoljni lokacijom (96.20%) i restoranom (90%).

## Zaključci

Predstavljeno je rešenje koje kombinuje ATE-PC task sa LDA modelom u cilju analize hotelskih recenzija. Izdvojeni aspekti klasifikovani su u 11 topika izvedenih iz LDA modela. Ograničenje je što je rešenje fokusirano samo na engleski jezik. Iz ovog rada bismo preuzele ideju za treniranje BiLSTM-CRF modela, kao i analizu aspekata.

## [4] Graph-Based N-gram Language Identification on Short Texts

### Zadatak rada

Predstavljeno je rešenje koje se bavi identifikacijom jezika (LI). Autori ističu da se LI najčešće sprovodi nad dužim i dobro strukturiranim tekstovima. Zadatak ovog rada je da primeni LI na kratkim tweetovima, primenjujući N-gram pristup.

### Metodologija

Autori su u obzir uzeli ne samo prisustvo i ponavljanje reči, nego i njihov poredak. Da bi predstavili redosled reči kreirali su graf labeliranih podataka. Labela čvora predstavlja prisustvo neke reči u grafu. Težine čvorova predstavljaju stepen ponavljanja reči u određenom jeziku. Težine veze između čvorova oslikavaju gramatiku jezika – u ovom radu pod gramatikom se smatra redosled reči. Za rešavanje problema koriste se N-Gram i LIGA (Language Identification Graph-based approach). Pretprocesiranje teksta uključuje izbacivanje svih tweetova koji sadrže reči iz različitih jezika, uklanjanje linkova, specijalnih znakova i emojiја.

### Skup podataka

Dataset sadrži 9066 anotiranih podataka, odn. tweetova napisanih na 6 jezika - German, English, Spanish, French, Italian i Dutch.

## Evaluacija rešenja

Primenjena je 10-fold cross validacija 50 puta. Poređene su srednje vrednosti accuracy N-gram modela i LIGA modela.

## Rezultati

LIGA postiže bolje rezultate od N-grama (naročito u scenarijima kada je trening skup bio manji). Korišćenjem više od 50% dataseta u trening svrhe došlo se do zaključka da ne postoje značajne razlike u postignutim preciznostima oba modela.

## Zaključci

Autori ističu da LIGA postiže bolje rezultate od N-gram-a u detekciji jezika – naročito kada su u pitanju rečenice koje sadrže žargonizme. U našem projektu pokušale bismo da implementiramo ovaj pristup u cilju detekcije jezika pre određivanja sentimenta recenzije. Istražile bismo javno dostupne implementacije (npr. <https://github.com/llaisdy/liga>) i proverile da li nam detekcija jezika pre određivanja sentimenta donosi benefite.

## Skup podataka

U projektu bismo koristile skup podataka sa sajta Kaggle (<https://www.kaggle.com/meetnagadia/hotel-reviews>) koji sadrži 10000 recenzija hotela u SAD-u. Od 26 atributa izdvojile bismo reviews.title, reviews.test i reviews.rating, a podatke bismo anotirale ručno.

Za deo podataka na srpskom jeziku parsirale bismo html stranice sajta <https://www.booking.com/reviews> koji sadrži recenzije američkih hotela na srpskom jeziku. Ove podatke bismo takođe anotirale ručno, i povele bismo računa o tome da skup podataka bude balansiran u odnosu na jezik [2].

## Metodologija

Prvi korak bio bi pretprocesiranje podataka koje uključuje uklanjanje stop words, lematizaciju, uklanjanje znakova interpunkcije i sl. Zatim bismo iskoristile tehnike vektorizacije navedene u radu [1] – TF-IDF, Word2Vec, BERT. Za samu klasifikaciju isprobale bismo Random Forest klasifikator [1] i SVM [2]. Što se tiče prepoznavanja imenovanih entiteta (NER) primenile bismo metodu BiLSTM-CRF [3], a u zavisnosti od rezultata istražile bismo i druge metode, kao što je npr. SpaCy alat.

U slučaju da rezultati sentiment analize ne postignu zadovoljavajuće rezultate, pokušale bismo da kao prvi korak sprovedemo detekciju jezika. Na osnovu rezultata ove faze primenile bismo odgovarajuće pretprocesiranje teksta za odgovarajući jezik - engleski i srpski su dosta različiti po određenim slovima i gramatici (npr. padeži u srpskom jeziku otežavaju svođenje reči na korenski oblik) [4].

Koraci koje bismo isprobale za pretprocesiranje teksta i posmatrale kako utiču na performanse modela:

- Uklanjanje stop words
  - Engleske recenzije – upotrebom NLTK biblioteke.
  - Srpske recenzije – uklanjanje zamenica, predloga, veznika, rečca i skraćenica. Uočavanje potencijalnih izuzetaka poput odričnih zamenica niko/ništa i veznika bez jer mogu nositi negativan sentiment.
- Uklanjanje znakova interpunkcije (oba jezika na isti način)
  - Analiza da li je povećano ponavljanje znakova interpunkcija povezano sa raspoloženjem korisnika (npr. korišćenje više uzastopnih tačaka ili uzvičnika).
- Svođenje teksta na mala slova (oba jezika na isti način)
  - Analiza da li reči pisane caps lock-om nose sentiment.
- Prebrojavanje reči i uklanjanje onih koje se javljaju u svim sentimentima (oba jezika na isti način)
- Uklanjanje dijaktirika
  - Srpske recenzije – šišanje latinice.
- Stemming
  - Engleske recenzije – upotrebom NLTK biblioteke.
  - Srpske recenzije – listu sufiksa za stemming preuzele bismo sa sledećeg [izvora](#). Potencijalno bismo isprobale i [BERTić](#) – BERT prilagođen za srpski jezik.

## Metod evaluacije

Rezultate bismo evaluirale accuracy i F1-score merama, i podelile bismo skup podataka na trening, validacioni i test skup u razmeri 70:20:10 [1].

## Softver

Za implementaciju navedenih algoritama koristile bismo Python i odgovarajuće biblioteke. Za parsiranje html stranica koristile bismo Selenium driver i alat za parsiranje ParseHub. Takođe, po potrebi bismo isprobale i SpaCy alat za prepoznavanje imenovanih entiteta.

## Plan

- Prikupljanje podataka
- Eksplorativna analiza podataka
- Pretprocesiranje teksta
- Obučavanje modela
- Evaluacija modela
- Analiza grešaka

## Članovi tima

- Vera Kovačević, R214/2021
- Nataša Ivanović, R212/2021

## Prikupljanje podataka

- [illegible]



## Pretprocesiranje teksta

- English preprocessing
  - Uklanjanje stop words (nltk)
  - Spell checking
  - Lowercase teksta
  - Uklanjanje specijalnih karaktera
  - Uklanjanje emoji-ja
  - Uklanjanje url-ova
  - Lemming
  - Stemming
- Serbian preprocessing
  - Uklanjanje stop words (ručno pravljenje liste)
  - Lowercase teksta
  - Uklanjanje specijalnih karaktera
  - Uklanjanje emoji-ja
  - Šišanje latinice
  - Lemming i stemming (nikola98 SerbianStemmer)
  -
- **TODO:**
  - ~~Uklanjanje reči koje se često ponavljaju (neutralne — ima ih u svim sentimentima)~~
  - ~~Analiza da li znakovi interpunkcije utiču na sentiment~~
  - ~~Analiza da li caps lock utiče na sentiment~~
  - ~~Analiza da li neke reči nose određeni sentiment (npr. užasno, sjajno itd.)~~
  - Bert

## Metodologija

- Detekcija jezika
  - Trivijalno pretprocesiranje teksta (uklanjanje specijalnih karaktera i brojeva)
  - Tfidf vektorizacija teksta
  - MultinomialNB model za predikciju

```
>> Training phase for language detection model...
      precision    recall  f1-score   support

     0         1.00      1.00      1.00        18
     1         1.00      1.00      1.00        23

 accuracy          1.00          1.00          1.00         41
 macro avg          1.00          1.00          1.00         41
weighted avg          1.00          1.00          1.00         41
```

- Mali broj podataka je doveo do ovakvog rezultata za accuracy – prilikom testiranja uočeno je da model pravi greške ukoliko rečenica sadrži reči koje mogu pripadati u oba jezika (npr. no, a, to...)
  - **TODO:**
    - Dopuniti dataset
    - Isprobati nove modele/odraditi fine tuning MultinomialNB
- Sentiment analiza
  - Tfidf vektorizacija teksta
  - SVM model

```
>> Training phase for english reviews...
      precision    recall  f1-score   support

    -1         0.00      0.00      0.00         6
     0         0.00      0.00      0.00         2
     1         0.74      1.00      0.85        23

 accuracy          0.74          0.74          0.74        31
 macro avg          0.25          0.33          0.28        31
weighted avg          0.55          0.74          0.63        31

>> Training phase for serbian reviews...
      precision    recall  f1-score   support

    -1         0.00      0.00      0.00        10
     0         0.00      0.00      0.00         3
     1         0.58      1.00      0.73        18

 accuracy          0.58          0.58          0.58        31
 macro avg          0.19          0.33          0.24        31
weighted avg          0.34          0.58          0.43        31
```

- RandomForest model

```
>> Training phase for english reviews...
      precision    recall  f1-score   support

    -1         0.00      0.00      0.00         3
     0         0.00      0.00      0.00         2
     1         0.76      1.00      0.86        16

 accuracy          0.76          0.76          0.76        21
 macro avg          0.25          0.33          0.29        21
weighted avg          0.58          0.76          0.66        21

>> Training phase for serbian reviews...
      precision    recall  f1-score   support

    -1         0.00      0.00      0.00         7
     1         0.67      1.00      0.80        14

 accuracy          0.67          0.67          0.67        21
 macro avg          0.33          0.50          0.40        21
weighted avg          0.44          0.67          0.53        21
```

- **TODO:** Fine tuning modela i improvement dataseta

- **NER – Izdvajanje aspekata**

- BiLSTMCRF model za srpski jezik – dodata labela FOOD

```
Accuracy: 0.6111
Makaroni      : 0
su            : 0
italijanska   : 0
hrana         : 0
.             : 0
```

- Prilagođeni Spacy model za engleski jezik – dodata labela FOOD

```
Entities [('I', 'FOOD'), ('Maggi', 'FOOD')]
I ate Sushi yesterday. Maggi is a common fast food
PRECISION: 0.5
RECALL: 0.5
FSCORE: 0.5
```

- **TODO:**

- Dodati labele: 'PERSONNEL', 'WEATHER', 'FOOD', 'HYGIENE', 'FURNITURE', 'LOCATION'
- Anotirati podatke na engleskom pomoću <http://agateteam.org/spacynerannotate/>
- Anotirati podatke na srpskom ručno

- **Trenutni flow:**

- Radi predekiciju jezika
- Na osnovu jezika sprovodi se odgovarajuće pretprocesiranje teksta
- Vršiti se predikcija modelom za odgovarajući jezik
- **TODO:**
  - NER ekstrakcija aspekata – izlaz treba da prikaže i aspekte na koje se odnosi recenzija

```
Ovo je užasno do bola!! Negativno iskustvo! Sobe smrde, ništa nije kao na slici!!! Odvratno je i gadi mi se. >> Serbian
Prediction: [1]
```

## Analiza grešaka

Uočeno je da modeli prediktuju pozitivne recenzije čak i kada one to nisu – to je posledica nebalansiranog dataseta. Pretpostavlja se da će ovaj problem biti ublažen (prevaziđen) proširivanjem skupa podataka.

## Druga kontrolna tačka – Presek stanja

### Prikupljanje podataka

- Određivanje sentimenta
  - 1044 anotiranih recenzija na engleskom
    - Trening set: 940
      - Pozitivne recenzije: 470
      - Negativne recenzije: 470
    - Test set: 104
    - Recenzije preuzete iz dataseta navedenog u predlogu projekta
    - Uočeno je da značajan broj recenzija u sebi sadrži i odgovor menadžera – ovakvi delovi su uklanjani
    - Anotiranje vršeno na osnovu ocena dostupnih u datasetu (1-5):
      - 1 – 2 negativne recenzije
      - 4 – 5 pozitivne recenzije
      - 3 – prvobitno neutralan sentiment, ali pošto je ovakvih recenzija bilo veoma malo pripojene su negativnim recenzijama radi balansiranosti dataseta
  - 1389 anotiranih recenzija na srpskom
    - Trening set: 1254
      - Pozitivne recenzije: 627
      - Negativne recenzije: 627
    - Test set: 135
    - Prikupljanje podataka odrađeno korišćenjem ParseHub alata za skrejpovanje
    - Uočeno je da ljudi veoma često unose pozitivne recenzije u polja predviđena za negativne – ispravljanje ovih slučajeva rađeno ručno
    - Uočen je mali broj recenzija napisan na ćirilici – ovakve recenzije smatrane su outlierima i uklanjane su iz dataseta
    - Anotiranje rađeno ručno
- Izvlačenje aspekata (NER)
  - 500 recenzija na engleskom
    - Anotiranje vršeno pomoću Spacy alata
  - 300 recenzija na srpskom
    - Anotiranje vršeno ručno
  - 7 tagova: *PERSONELL, WEATHER, FOOD, HYGIENE, FURNITURE, LOCATION, PRICE*



## Najčešće ponavljane reči – engleske recenzije

## Najčesče ponavljane reči – engleske recenzije



## Najčešće ponavljane reči – srpske recenzije



Figure 8 Negativne recenzije



Figure 7 Pozitivne recenzije

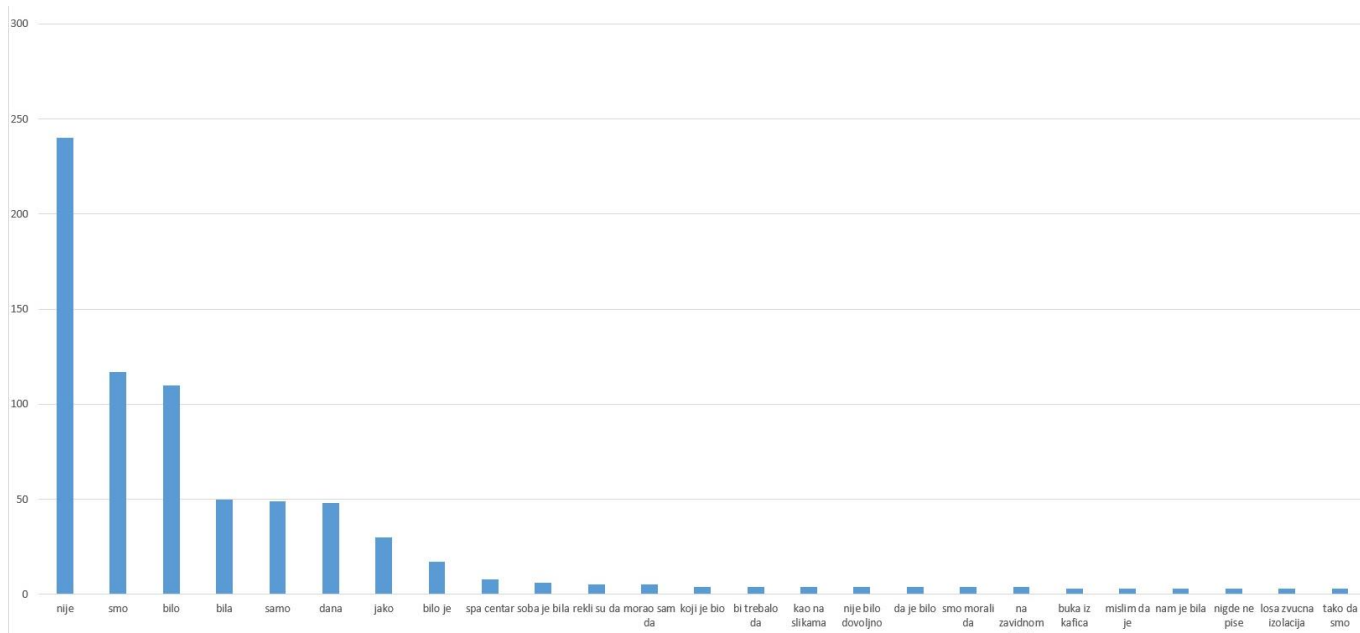


Figure 6 Top 25 najčešće ponavljanih reči u negativnim recenzijama

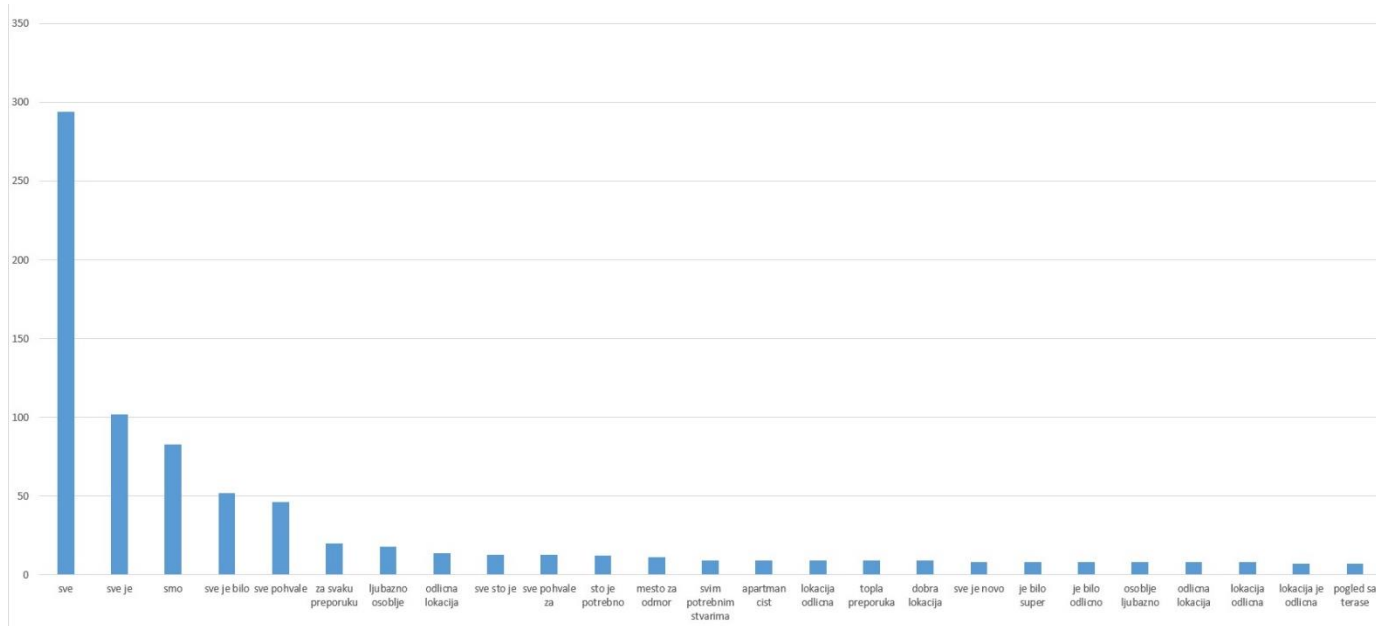


Figure 9 Figure 6 Top 25 najčešće ponavljanih reči u pozitivnim recenzijama

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.

Kod engleskih recenzija 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 reči tokom pretprocesiranja. Nazivi gradova *San Diego* i *San Francisco* su takođe izbačeni.

Kod srpskih recenzija uočeno je da se ponavljaju reči koje nemaju semantičku vrednost – te reči su izbačene tokom pretprocesiranja. Takođe, primećeno je da se često ponavljaju grupacije reči, odn. sintagme. Iz ovog razloga odlučeno je da se isprobaju *bigram*-i i *trigram*-i prilikom treniranja modela.

### 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.

Tačke:

- Srpski: 41 negativne, 54 pozitivne
- **Engleski: 221 negativne, 81 pozitivne** – nije pokazalo značajne razlike tokom treniranja

Uzvičnici:

- Srpski: 13 negativne, 10 pozitivne
- Engleski: 8 negativne, 9 pozitivne

Upitnici:

- Srpski: 1 negativne, 0 pozitivne
- Engleski: 0 negativne, 0 pozitivne

Zaključeno je da je za treniranje ipak bolje izbaciti sve znake interpunkcije.

### Analiza CAPS lock recenzija (as in “gost se dere”)

Pretpostavka je da su ljudi nezadovoljni kada se deru.

Engleske recenzije:

- 346 negativne,
- 280 pozitivne (izostavljeno je pisanje zamenice I)

Srpske recenzije:

- 102 negativne,
- 126 pozitivne

Zaključeno je da ljudi pišu caps lock-om i kada su nezadovoljni i kada su uzbuđeni i zadovoljni – tokom pretprocesiranja sve reči su svedene na mala slova.

## Metodologija

### Detekcija jezika

Training results				
	precision	recall	f1-score	support
0	1.00	1.00	1.00	258
1	1.00	1.00	1.00	181
accuracy			1.00	439
macro avg	1.00	1.00	1.00	439
weighted avg	1.00	1.00	1.00	439
Testing results				
	precision	recall	f1-score	support
0	1.00	1.00	1.00	136
1	1.00	1.00	1.00	104
accuracy			1.00	240
macro avg	1.00	1.00	1.00	240
weighted avg	1.00	1.00	1.00	240

Za detekciju jezika korišćen je MultinomialNB model – ispostavilo se da problem razlikovanja srpskog i engleskog jezika nije težak. Model se odlično snašao bez potrebe za fine tuning parametara.

Model je treniran na uniji skupa podataka za treniranje za sentiment analizu.

Model je testiran na uniji skupa podataka za testiranje za sentiment analizu.

### Sentiment analiza – SVM model

Skup podataka je podeljen na trening i validacioni skup u odnosu 80:20. Korišćenjem GridSearchCV vršena je unakrsna validacija nad trening delom skupa (k-fold, gde je k=5). Nakon toga je testirano ono što je SVM model naučio pomoću validacionog skupa podataka.

Za pretvaranje tokena (odn. reči iz ocena) u numeričke vrednosti korišten je TfidfVectorizer, a isprobane su različite vrednosti n\_gram\_range parametra – (1, 1), (1, 2), (1, 2).

Isprobana su četiri kernela: linear, polynomial, rbf i sigmoid.

#### SVM – linear – (1, 1) – srpski:

- Najbolji model:  
SVC(C=1, kernel='linear')  
TfidfVectorizer(lowercase=False, sublinear\_tf=True, use\_idf=False)
- Rezultat nad validacionim skupom: 0.9761
- Rezultat nad testnim skupom: 0.94814

#### SVM – linear – (1, 2) – srpski:

- Najbolji model:  
SVC(C=10, kernel='linear')  
TfidfVectorizer(lowercase=False, ngram\_range=(1, 2), sublinear\_tf=True, use\_idf=False)
- Rezultat nad validacionim skupom: 0.96813
- Rezultat nad testnim skupom: 0.94814

#### SVM – linear – (1, 3) – srpski:

- Najbolji model:  
SVC(C=10, kernel='linear')  
TfidfVectorizer(lowercase=False, ngram\_range=(1, 3), sublinear\_tf=True, use\_idf=False)
- Rezultat nad validacionim skupom: 0.96414
- Rezultat nad testnim skupom: 0.94814

#### SVM – linear – (1, 1) – engleski:

- Najbolji model:  
SVC(C=1, kernel='linear')  
TfidfVectorizer(lowercase=False, sublinear\_tf=True, use\_idf=False)
- Rezultat nad validacionim skupom: 0.86170

- Rezultat nad testnim skupom: 0.67307

#### **SVM – linear – (1, 2) – engleski:**

- Najbolji model:  
SVC(C=1, kernel='linear')  
TfidfVectorizer(lowercase=False, ngram\_range=(1, 2), sublinear\_tf=True, use\_idf=False)
- Rezultat nad validacionim skupom: 0.86792
- Rezultat nad testnim skupom: 0.70192

#### **SVM – linear – (1, 3) – engleski:**

- Najbolji model:  
SVC(C=1, kernel='linear')  
TfidfVectorizer(lowercase=False, ngram\_range=(1, 3), sublinear\_tf=True, use\_idf=False)
- Rezultat nad validacionim skupom: 0.84043
- Rezultat nad testnim skupom: 0.70192

#### **U grid searchu su isprobane i sledeće kombinacije SVM parametara:**

```
tuned_parameters = {
    'svm__kernel': ['rbf', 'poly', 'sigmoid'], 'svm__C': [0.1, 1, 10, 100, 200, 1000],
    'svm__gamma': [1, 0.1, 0.01, 0.001, 0.0001], 'vec__ngram_range': [(1, 1), (1, 2), (1, 3)]
}
```

#### **Srpski jezik:**

- Najbolji model:  
SVC(C=200, gamma=0.01, kernel='sigmoid')  
TfidfVectorizer(lowercase=False, ngram\_range=(1, 2), sublinear\_tf=True, use\_idf=False)
- Rezultat nad validacionim skupom: 0.95618
- Rezultat nad testnim skupom: 0.95588

#### **Engleski jezik:**

- Najbolji model:  
SVC(C=10, kernel='rbf', gamma=1)  
TfidfVectorizer(lowercase=False, ngram\_range=(1, 1), sublinear\_tf=True, use\_idf=False)
- Rezultat nad validacionim skupom: 0.85638
- Rezultat nad testnim skupom: 0.70192

### **Sentiment analiza – RandomForest model**

#### **U grid searchu su isprobane i sledeće kombinacije SVM parametara:**

```
tuned_parameters = {
    'rfc__max_depth': [10, 20, 30, 40], 'rfc__max_features': ['auto', 'sqrt'], 'rfc__min_samples_leaf': [1, 2],
    'rfc__min_samples_split': [2, 5], 'rfc__n_estimators': [200, 400, 600]
}
```

#### **RFM – (1, 1) – engleski:**

- Najbolji model:  
RandomForestClassifier(max\_depth=40, min\_samples\_leaf=2, min\_samples\_split=5, n\_estimators=600)  
TfidfVectorizer(lowercase=False, sublinear\_tf=True, use\_idf=False)

- **Rezultat nad validacionim skupom: 0.83511**
- **Rezultat nad testnim skupom: 0.70192**

#### **RFM – (1, 2) – engleski:**

- Najbolji model:  
RandomForestClassifier(max\_depth=20, max\_features='sqrt', min\_samples\_leaf=2, min\_samples\_split=5, n\_estimators=400)  
TfidfVectorizer(lowercase=False, ngram\_range=(1, 2), sublinear\_tf=True, use\_idf=False)
- Rezultat nad validacionim skupom: 0.8617
- Rezultat nad testnim skupom: 0.63461

#### **RFM – (1, 3) – engleski:**

- Najbolji model:  
RandomForestClassifier(max\_depth=30, min\_samples\_leaf=2, n\_estimators=600)  
TfidfVectorizer(lowercase=False, ngram\_range=(1, 3), sublinear\_tf=True, use\_idf=False)
- Rezultat nad validacionim skupom: 0.80319
- Rezultat nad testnim skupom: 0.64423

#### **RFM – (1, 1) – srpski:**

- Najbolji model:  
RandomForestClassifier(max\_depth=40, max\_features='sqrt', min\_samples\_leaf=2, min\_samples\_split=5, n\_estimators=200)  
TfidfVectorizer(lowercase=False, sublinear\_tf=True, use\_idf=False)
- Rezultat nad validacionim skupom: 0.92829
- Rezultat nad testnim skupom: 0.94074

#### **RFM – (1, 2) – srpski:**

- Najbolji model:  
RandomForestClassifier(max\_depth=30, min\_samples\_leaf=2, n\_estimators=600)  
TfidfVectorizer(lowercase=False, ngram\_range=(1, 2), sublinear\_tf=True, use\_idf=False)  
Rezultat nad validacionim skupom: 0.89641
- Rezultat nad testnim skupom: 0.94814

#### **RFM – (1, 3) – srpski:**

- **Najbolji model:**  
RandomForestClassifier(max\_depth=30, min\_samples\_leaf=2, min\_samples\_split=5, n\_estimators=600)  
TfidfVectorizer(lowercase=False, ngram\_range=(1, 3), sublinear\_tf=True, use\_idf=False)
- **Rezultat nad validacionim skupom: 0.91633**
- **Rezultat nad testnim skupom: 0.95555**

#### **Analiza grešaka sentiment analize**

U cilju analize grešaka prilikom testiranja modela čuvani su fajlovi koji sadrže originalan sentiment, prediktovani sentiment i originalnu recenziju. Na taj način su izdvojene recenzije na kojima model greši.

Primeri za srpski jezik:

- 1(p)|-1(o)|Narušavaju vasu privatnost i krse zakon time sto vam pri dolasku slikaju licne karte. (neutralne reci)
- -1(p)|1(o)|Urednost i čistoća kako smeštaja, tako i SPA centra. (field za positive)
- -1(p)|1(o)|Hladna voda u bazenu. Isto je tako bilo i pre 5 godina. (field za positive)

U datasetu za engleski jezik je primećeno da su recenzije veoma dugačke. Neretko se dešava da gosti u okviru jedne recenzije napišu mnogo negativnih stvari, ali i nekoliko pozitivnih (i obrnuto), što je donekle zbunjujuće za model.

- 1(p)|-1(o)| If you like King rooms, be prepared not to have a bath. None of their King rooms have baths, only showers. I could not properly wash my hair with a tiny showerhead. Many King rooms have communicating doors to the next room, which are good noise-carriers. With regard to noise: the rooms at the freeway end of the hotel are very noisy, much more so than other hotels we've stayed in on Hotel Circle South. We were living out of our suitcases because of the ridiculous size of the drawers in the room (basically the size of nightstand drawers - see picture). 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, cooked food was poor. Hotel staff were helpful and did their best to resolve our issues in a not-so-great hotel. Next San Diego visit: back to the Doubletree.

Potencijalan problem predstavlja i anotiranje sentimenta na osnovu ocena koje su ostavljali korisnici jer ocene nisu u 100% slučajeva konzistentne.

#### Izdvajanje aspekata – NER – engleski jezik + ANALIZA GREŠAKA

Prilagođeni Spacy model je postigao sledeće rezultate na test skupu:

- PRECISION: 0.40163
- RECALL: 0.43946
- FSCORE: 0.41970

Neki od uočenih rezultata po tagovima:

- 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 značajno manji nego što bi trebalo – iščitavanjem postignutih rezultata deluje da se model relativno dobro snašao. S obzirom da se ručno radi anotiranje, uočeno je da postoje propusti u anotiranju testnog seta – npr. nešto što je FOOD nije označeno tagom (previd), a model je to predvideo kao FOOD i to je negativno uticalo na accuracy.

#### Izdvajanje aspekata – NER – srpski jezik + ANALIZA GREŠAKA

Ručno anotiranje zahteva mnogo vremena – stoga je prikupljeni dataset na srpskom jeziku mali. To je uticalo na treniranje i testiranje – model je shvatio da mu je lakše da pogađa dva taga, jer će tako sigurno pogoditi nešto, nego da pokušava da pogađa sve tagove.

Postignut accuracy na test setu: 0.82845 (što deluje super, ali kada se analiziraju rezultati primećuju se samo dva taga u predikcijama).