.

SVEUČILIŠTE U ZAGREBU
FAKULTET ORGANIZACIJE I INFORMATIKE
VARAŽDIN

Stjepan Marić, Ante Marić

PROCESIRANJE PRIRODNOG JEZIKA ZA RAZUMIJEVANJE POVRATNIH INFORMACIJA KLIJENATA U UGOSTITELJSKOJ INDUSTRIJI

PROJEKT

Varaždin, 2026.

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Stjepan Marić, Ante Marić

Matični broj: 0016166230, 0016168875

Studij: Informacijski i poslovni sustavi

PROCESIRANJE PRIRODNOG JEZIKA ZA RAZUMIJEVANJE POVRATNIH INFORMACIJA KLIJENATA U UGOSTITELJSKOJ INDUSTRIJI

PROJEKT

Mentor:

doc. dr. sc. Bogdan Okreša Đurić

Varaždin, siječanj 2026.

Stjepan Marić, Ante Marić

Izjava o izvornosti

Izjavljujem da je ovaj projekt izvorni rezultat mojeg rada te da se u izradi istoga nisam koristio drugim izvorima osim onima koji su u njemu navedeni. Za izradu rada su korištene etički prikladne i prihvatljive metode i tehnike rada.

Autorica potvrdila prihvaćanjem odredbi u sustavu FOI Radovi

Sažetak

Ovaj projekt obrađuje problem razumijevanja povratnih informacija klijenata u ugostiteljskoj industriji primjenom metoda procesiranja prirodnog jezika. Na skupu recenzija restorana provodi se obrada teksta, vektorizacija značajki te treniranje klasifikatora za predviđanje postavljenih sentimenta (pozitivno, neutralno i negativno). Prikazani su postupci učenja, evaluacija performansi modela i demonstracija rada aplikacije kroz interaktivni unos recenzija.

Ključne riječi: procesiranje prirodnog jezika; analiza sentimenta; klasifikacija teksta; TF-IDF; Naive Bayes; recenzije restorana

Sadržaj

1. Uvod

U moderniziranome digitalnome okruženju, korisnici sve češće izražavaju svoja mišljenja i iskustva putem online recenzija. Navedena situacija se posebno osjeti u ugostiteljskoj industriji, gdje su povratne informacije gostiju izuzetno važne, s obzirom na to da izravno utječu na reputaciju objekta, a samim time i na odluke budućih gostiju. Zbog velikog broja recenzija, ručna analiza za određene objekte, postala je nepraktična te vremenski zahtjevna.

Procesiranje prirodnog jezika omogućuje automatiziranu obradu te razumijevanje tekstualnih podataka. Na taj način, stavovi i emocije izražene u recenzijama mogu se učinkovito prepoznati i klasificirati. Analiza sentimenta jedna je od najčešće korištenih metoda u tom području te omogućuje klasifikaciju tekstova prema njihovom tonu, a u ovome projektu, dijeli se na pozitivne, neutralne i negativne.

Zanimanje za optimizaciju poslovanja ugostiteljskih objekata potaknulo je odabir navedene teme. Potreba je pomoći ugostiteljskim objektima u boljem razumijevanju povratnih informacija klijenata. Cilj je primjenom metoda procesiranja prirodnog jezika razviti model koji automatski analizira tekstualne recenzije te predviđa njihov sentiment.

2. Razrada teme

U ovom poglavlju predstavljeni su osnovni pojmovi i metode koje čine teorijsku podlogu projekta, nakon čega se objašnjava način njihove primjene u rješavanju problema analize sentimenta.

2.1. Analiza sentimenta

Analiza sentimenta je područje obrade prirodnog jezika koje se odnosi na automatsko prepoznavanje emocionalnog tona teksta. Cilj ove analize je odrediti korisnikov stav prema određenom subjektu. Taj stav može biti pozitivan, negativan, ili neutralan. Ova tehnika primjenjuje se u različitim industrijama na recenzije proizvoda i usluga, na komentare na društvenim mrežama te na različite oblike povratnih informacija korisnika u globalu. Analiza sentimenta predstavlja jedno od najčešćih područja primjene obrade prirodnog jezika, posebno u kontekstu klasifikacije korisničkih recenzija [1].

U kontekstu ugostiteljske industrije, analiza sentimenta pruža brzu i automatiziranu procjenu zadovoljstva na temelju pisanih recenzija. Tako olakšava posao i nema potrebe za ručnom analizom velikih količina podataka.

2.2. Obrada prirodnog jezika

Procesiranje ili obrada prirodnog jezika (engl. Natural Language Processing - NLP) područje je umjetne inteligencije koje je obrađivano na laboratorijskim vježbama. Navedeno područje bavi se analizom te obradom ljudskog jezika pomoću računalnih sustava.

Budući da je tekst nestrukturiran oblik podataka, prije primjene istoga u modelima strojnog učenja, potrebno je provesti odgovarajuću predobradu.

Za svrhu ovog projekta, primijenjene su osnovne tehnike obrade prirodnog jezika, kao što su uklanjanje interpunkcije, pretvaranje teksta u mala slova te uklanjanje čestih riječi bez značajnoga semantičkog značenja. Koristeći ove metode i postupke, uklanja se višak u podatcima te se poboljšava kvaliteta ulaznih značajki koje se koriste za treniranje modela.

2.3. Bag-of-Words model

Bag of Words (BoW) model jedan je od najjednostavnijih i najčešće korištenih pristupa za numeričku reprezentaciju teksta. U ovom modelu dokument se promatra kao skup riječi te se bilježi isključivo učestalost ponavljanja pojedinih riječi, a njihov redoslijed se zanemaruje. Bag-of-Words metoda predstavljaja temeljni pristup u problemima klasifikacije teksta [2].

Svaka riječ u skupu podataka predstavlja jednu značajku, a dokument se opisuje vektorom koji sadrži broj ponavljanja tih riječi. BoW model ne uzima u obzir semantičke odnose

između riječi, ali daje zadovoljavajuće rezultate kod klasifikacije teksta upravo zbog svoje jednostavnosti i učinkovitosti.

2.4. TF-IDF reprezentacija

TF-IDF (Term Frequency - Inverse Document Frequency) je metoda koja predstavlja nadogradnju BoW modela. Osnovna ideja ove metode je smanjiti utjecaj često ponavljanih riječi iz dokumenata te povećati važnost riječi karakterističnih za određene tekstove. TF-IDF metoda predstavlja temeljne pristupe u problemima klasifikacije teksta [2].

Komponenta TF predstavlja učestalost ponavljanja riječi, a IDF mjeri rijetkost riječi, kada se ove dvije mjere kombiniraju, dobiva se reprezentacija teksta koja precizno odražava informativnu vrijednost riječi što kasnije rezultira boljim performansama klasifikacijskih modela.

2.5. Nadzirano strojno učenje

Nadzirano strojno učenje predstavlja pristup u kojem se modeli uče na temelju unaprijed postavljenih podataka. Svaki se ulazni primjer povezuje s pripadajućom izlaznom oznakom što omogućuje modelu da nauči odnose između ulaza i izlaza.

U ovom projektu definirane su tri oznake sentimenta i mapirane u tri klase: pozitivan, negativan i neutralan. One su dobivene iz numeričkih ocjena recenzija. Na temelju tako označenih podataka, treniran je model koji pomaže predviđati sentiment novih, prethodno neviđenih recenzija.

2.6. Multinomial Naive Bayes klasifikator

Radi se o jednom od najčešće korištenih algoritama za klasifikaciju teksta. Temelji se na Bayesovom teoremu te pretpostavci da su ponavljanja pojedinih riječi međusobno nezavisna.

Zbog svoje jednostavnosti, brzine izvođenja i dobre prilagodbe značajkama kao što su broj pojavljivanja riječi, ovaj se algoritam koristi kao početni model u analizi sentimenata, pa tako i u ovom projektu.

2.7. Formalizam umjetne inteligencije

Primijenjen je formalizam nadziranog strojnog učenja. On predstavlja jedan od temeljnih područja umjetne inteligencije. Problem analize sentimenata formuliran je kao problem klasifikacije teksta, pri čemu je cilj svakom tekstualnom zapisu pridružiti jednu od definiranih klasa sentimenta.

Tekstualni podatci transformirani su u numerički oblik korištenjem Bag of Words modela i TF-IDF metode, čime je omogućena primjena statističkih algoritama strojnog učenja. Za kal-

sifikacijski model korišten je Mulltinominal Naive Bayes koji se temelji na Bayesovom teoremu i maloprije definiranoj pretpostavci uvjetne nezavisnosti značajki. Ovakav pristup predstavlja standardni formalizam umjetne inteligencije za rješavanje problema obrade prirodnog jezika i analize sentimenta. Nadzirano strojno učenje predstavlja jedno od temeljnih područja umjetne inteligencije [3].

3. Kritički osvrt

U početnoj fazi izrade projekta planirano je korištenje vlastitog skupa podataka prikupljenog na temelju recenzija jednog ugostiteljskog objekta u Varaždinu, pri čemu su recenzije bile napisane na hrvatskom jeziku. Cilj je bio razviti model koji bi analizirao povratne informacije korisnika na hrvatskom jeziku. Međutim, prikupljeni skup podataka bio je izrazito neuravnotežen, s velikim udjelom pozitivnih recenzija. To je uzrokovalo situaciju gdje je trenirani model u većini slučajeva predviđao isključivo pozitivan sentiment, što je ukazivalo na nepouzdanost dobivenih rezultata. Zbog navedenih ograničenja, donesena je odluka o promjeni skupa podataka te je korišten veći i raznovrsniji skup recenzija na platformi Kaggle, koji sadrži recenzije na engleskom jeziku. To je omogućilo da model bude istreniran na više različitih primjera i da daje pouzdane rezultate. Na novome skupu podataka model je pokazao zadovoljavajuću sposobnost prepoznavanja pozitivnog i negativnog sentimenta. Iz toga razloga, odlučili smo se za taj skup podataka i nastavili s tim modelom kao glavnim primjerom za ovaj projekt. Rezultati su bili puno bolji, ali je uočeno ako postoje teškoće u prepoznavanju neutalne klase sentimenta. Takvo ponašanje može se pripisati činjenici da neutralne recenzije često imaju obrasce koji su vrlo slični pozitivnim ili negativnim komentarima. Osim toga, nutralna klasa je također bila slabije zastupljena što je dodtano otežalo proces učenja pouzdanih granica. Navedena ograničenja proizlaze i iz odabira korištenih metoda. Bag of Words i TF-IDF ne uzimaju u obzir kontekst niti redoslijed riječi, a Multinomial Naive Bayes pretpostavlja nezavisnost značajki što pojednostavljuje stvarnu strukturu jezika, zato se jasno vidi prepoznatljivost pozitivnih i negativnih obrazaca, dok su suptilni i neutralni izrazi teže prepoznatljivi. Sve u svemu, razvijeno rješenje zapravo predstavlja funkcionalan i razumljiv sustav za osnovnu analizu sentimenta tekstualnih recenzija, a to potvrđuju dobiveni rezultati. Jasno je vidljiva primjenjivost, ali i smjerovi mogućih poboljšanja, kao što su to uravnoteženiji skupovi podataka, naprednije tehnike procesiranja teksta ili složenijih modela strojnog učenja.

4. Opis implementacije

Sustav je napravljen u programskom jeziku Python korištenjem biblioteka koje su nužne za obradu prirodnog jezika, strojnog učenja i vizualizaciju. Također, sustav se sastoji od nekoliko cjelina koje zajedno omogućuju učitavanje podataka, njihovu obradu, treniranje modela te analizu sentimenta novih recenzija.

4.1. Skup podataka

Na početku se učitava skup podataka koji sadrži tekstualne recenzije restorana i pripadajuće numeričke ocjene. Korišten je javno dostupan skup podataka s platforme Kaggle. Podaci se učitavaju iz CSV datoteke te se zadržavaju samo stupci koji su relevantni za analizu, odnosno tekst recenzije i ocjena korisnika. Zapisi s nedostajućim vrijednostima uklanjaju se kako bi se osigurala ispravnost daljnje obrade. Korišten je javno dostupan skup podataka s platforme Kaggle [4]. Numeričke ocjene recenzija mapiraju se u oznake sentimenta korištenjem jednostavne funkcije koja ocjene prevodi u tri klase: negativan, neutralan i pozitivan sentiment. Pri tome se ocjene 1 i 2 mapiraju u negativan sentiment, ocjena 3 je neutralan, dok se ocjene 4 i 5 mapiraju u pozitivan sentiment.

```
def mapiraj_sentiment(ocjena):
    if ocjena <= 2:
        return "negative"
    elif ocjena == 3:
        return "neutral"
    else:
        return "positive"
```

4.2. Obrada teksta

Nadalje, tekstualni podaci prolaze kroz postupak obrade kako bi se mogli koristiti u daljnjoj analizi. To je omogućeno implementacijom funkcije za obradu teksta. Funkcija uklanja interpunkcijske znakove, pretvara tekst u mala slova te uklanja česte riječi koje nemaju značajno značenje. Rezultat funkcije je lista tokena koja predstavlja obrađeni tekst recenzije.

```
def obradi_tekst(tekst):
    tekst = str(tekst)
    bez_interpunkcije = [z for z in tekst if z not in string.punctuation]
    bez_interpunkcije = "".join(bez_interpunkcije)
    tokeni = [
        word.lower()
        for word in bez_interpunkcije.split()
        if word.lower() not in stop_rijeci
```

```
    ]
    return tokeni
```

4.3. Izgradnja i treniranje modela

Za izgradnju klasifikacijskog modela korištena je klasa *Pipeline* iz biblioteke scikit-learn koja omogućuje povezivanje više koraka obrade u jedinstveni tok. Pipeline se sastoji od pretvorbe teksta u Bag-of-Words reprezentaciju, primjene TF-IDF transformacije te klasifikacije pomoću Multinomial Naive Bayes algoritma.

```
model = Pipeline([
    ("vreca_rijeci", CountVectorizer(analyzer=obradi_tekst)),
    ("tfidf", TfidfTransformer()),
    ("klasifikator", MultinomialNB(
        class_prior=[
            priori_klasa["negative"],
            priori_klasa["neutral"],
            priori_klasa["positive"]
        ]
    ))
])
```

Model se trenira na dijelu podataka predviđenom za učenje, dok se preostali dio koristi za testiranje. Na taj način omogućuje se objektivna procjena uspješnosti modela na prethodno neviđenim podacima.

4.4. Evaluacija modela

Nakon treniranja modela provodi se evaluacija rezultata na testnom skupu podataka. Izračunavaju se osnovne mjere uspješnosti klasifikacije, uključujući preciznost, odaziv i F1-mjeru za svaku klasu sentimenta. Rezultati evaluacije su prikazani pomoću matrice zabune koja daje uvid u točnost i pogreške klasifikacije.

4.5. Interaktivni rad aplikacije

Na samom kraju, razvijena je jednostavna konzolna aplikacija koja omogućuje korisniku unos novih tekstualnih recenzija. Uneseni tekst prolazi kroz isti proces obrade kao i trenirajući podaci, nakon čega istrenirani model predviđa pripadajući sentiment te ispisuje rezultat korisniku.

5. Prikaz rada aplikacije

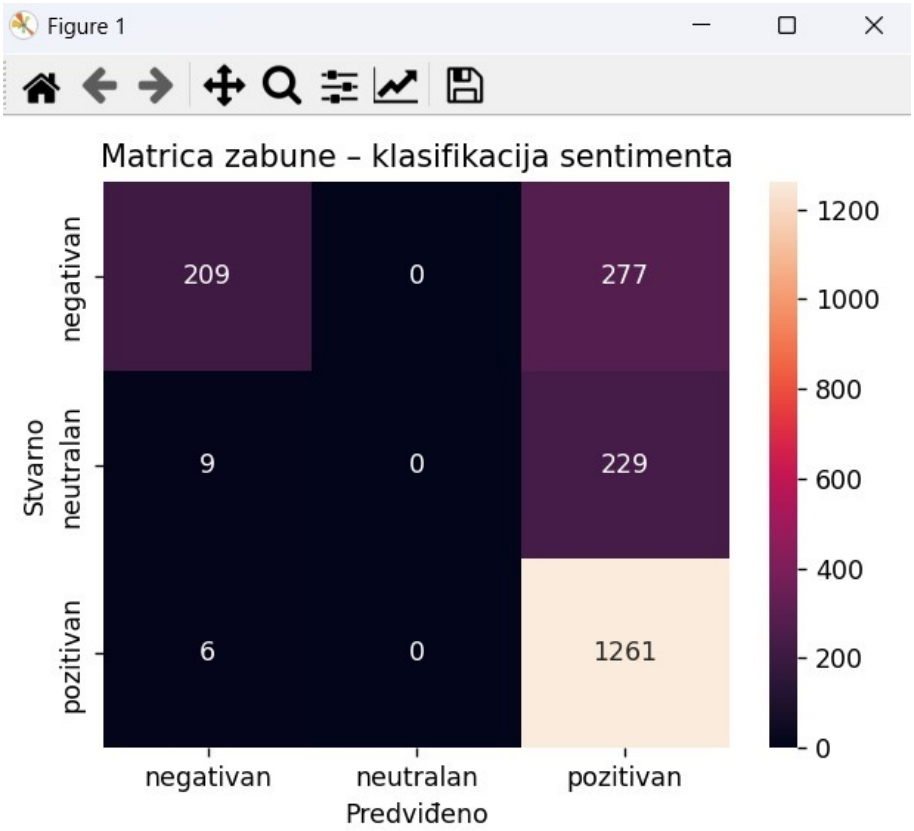
U ovom poglavlju prikazan je rad implementirane aplikacije kroz konkretne primjere izvođenja programa i dobivene rezultate. Aplikacija se pokreće izvođenjem Python skripte iz konzole, nakon čega se automatski provodi učitavanje podataka, treniranje modela i evaluacija rezultata.

Na slici 1 prikazan je ispis rezultata treniranja i testiranja modela u konzoli. Prikazan je ukupan broj recenzija, raspodjela klasa sentimenta u skupu podataka te klasifikacijski izvještaj koji uključuje mjere preciznosti, odziva i F1-mjere za svaku klasu sentimenta.

```
PS C:\Users\stjep\Desktop\FAKS\UUi\kod_csv_readme> python  a_s_maric.py
Broj recenzija: 9954

Raspodjela klasa:
pozitivne recenzije: 6334
negativne recenzije: 2428
neutralne recenzije: 1192

Treniranje modela...
Model je uspješno treniran.

Klasifikacijski izvještaj:

Klasa: negativan
  Preciznost: 0.93
  Odaziv: 0.43
  F1-mjera: 0.59
  Broj uzoraka: 486

Klasa: neutralan
  Preciznost: 0.00
  Odaziv: 0.00
  F1-mjera: 0.00
  Broj uzoraka: 238

Klasa: pozitivan
  Preciznost: 0.71
  Odaziv: 1.00
  F1-mjera: 0.83
  Broj uzoraka: 1267

Ukupna točnost: 0.74
Makro prosjek F1: 0.47
Ponderirani prosjek F1: 0.67
```

Slika 1: Ispis rezultata treniranja i evaluacije modela u konzoli

Na slici 2 prikazan je grafički prikaz matrice zabune, koji omogućuje vizualni uvid u točnost klasifikacije modela za sve tri klase sentimenta. Na temelju matrice moguće je uočiti broj ispravno i pogrešno klasificiranih recenzija za svaku klasu.



Slika 2: Matrica zabune za klasifikaciju sentimenta

Na slici 3 prikazan je interaktivni rad aplikacije. Korisnik unosi vlastitu tekstualnu recenziju putem konzole, nakon čega aplikacija, koristeći istrenirani model, predviđa sentiment unesene recenzije i ispisuje rezultat. Time je prikazana praktična primjena sustava na novim, prethodno neviđenim podacima.

```
Analiza sentimenta recenzija
(upiši 'exit' za izlaz)

Unesite recenziju: The food was cold and tasteless, and the staff was rude.
Ova recenzija je negativan.

Unesite recenziju: Excellent food, friendly staff and a great overall experience.
Ova recenzija je pozitivan.

Unesite recenziju: exit
Izlaz iz aplikacije.
```

Slika 3: Interaktivni unos recenzija i predikcija sentimenta

6. Zaključak

U ovom projektu razvijen je sustav za analizu sentimenta tekstualnih recenzija u ugostiteljskoj industriji primjenom metoda procesiranja, odnosno obrade prirodnog jezika te nadziranog strojnog učenja. Cilj projekta bio je automatizirati proces razumijevanja povratnih informacija klijenata na temelju recenzija te demonstrirati primjenu metoda umjetne inteligencije obrađenih na kolegiju.

U razradi teme prikazani su temeljni teorijski koncepti na kojima se temelji projekt, a to obuhvaća obradu prirodnog jezika, Bag-of-Words, TF-IDF te Multinomial Naive Bayes klasifikator. Na temelju navedenih metoda, izrađena je konzolna aplikacija u programskom jeziku Python koja omogućuje treniranje modela, evaluaciju njegove uspješnosti te predikciju sentimenta novih recenzija.

Rezultati evaluacije pokazali su da model uspješno prepoznaje pozitivne i negativne recenzije, dok su slabiji rezultati zabilježeni kod neutralne klase sentimenta. Takvo ponašanje posljedica je karakteristika korištenog skupa podataka te ograničenja korištenih metoda, što je detaljnije opisano u kritičkom osvrtu.

Projekt je pokazao da se relativno jednostavnim metodama strojnog učenja može izraditi funkcionalan i razumljiv sustav za analizu sentimenta korisničkih recenzija. Ostvareni rezultati potvrđuju praktičnu primjenjivost razvijenog rješenja te otvaraju prostor za buduća poboljšanja, kao što su korištenje uravnoteženijih skupova u budućnosti ili izrada naprednijih modela.

Popis literature

[1] B. Pang, L. Lee i S. Vaithyanathan, „Thumbs up? Sentiment Classification using Machine Learning Techniques," *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, 2002.

[2] T. Joachims, „Text categorization with Support Vector Machines: Learning with many relevant features," *Proceedings of the European Conference on Machine Learning*, 1998.

[3] S. Russell i P. Norvig, *Artificial Intelligence: A Modern Approach*. Pearson, 2021.

[4] J. Beach, *Restaurant Reviews Dataset*, `https://www.kaggle.com/datasets/joebeachcapital/restaurant-reviews`, 2023.

Popis slika