Primjena strojnog učenja za klasifikaciju lažnih recenzija

Šokota, Dora

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 742

PRIMJENA STROJNOG UČENJA ZA KLASIFIKACIJU LAŽNIH RECENZIJA

Dora Šokota

Zagreb, veljača 2025.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 742

PRIMJENA STROJNOG UČENJA ZA KLASIFIKACIJU LAŽNIH RECENZIJA

Dora Šokota

Zagreb, veljača 2025.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Zagreb, 30. rujna 2024.

DIPLOMSKI ZADATAK br. 742

Pristupnica:	**Dora Šokota (0036521795)**
Studij:	Računarstvo
Profil:	Računarska znanost
Mentor:	izv. prof. dr. sc. Mario Brčić

Zadatak: **Primjena strojnog učenja za klasifikaciju lažnih recenzija**

Opis zadatka:

U okviru završnog rada istražit će se metode strojnog učenja za detekciju lažnih recenzija na internetskim platformama. Lažne recenzije postaju sve veći problem u online poslovanju jer mogu značajno utjecati na percepciju korisnika i prodaju proizvoda. Cilj rada je implementirati i usporediti nekoliko algoritama strojnog učenja (npr. logistička regresija, stablo odluke, slučajne šume, SVM , neuralne mreže) kako bi se prepoznale karakteristike lažnih recenzija. Korištenjem javno dostupnih skupova podataka, potrebno je izraditi sustav za automatsku klasifikaciju recenzija kao lažnih ili stvarnih. Rad bi trebao uključivati analizu točnosti, preciznosti, odziva i F1 mjere različitih modela te zaključiti koji algoritam najbolje rješava ovaj problem.

Rok za predaju rada: 14. veljače 2025.

Sadržaj

1. Uvod

Digitalno doba značajno je promijenilo način na koji ljudi izražavaju mišljenja i razmjenjuju iskustva, posebice putem mrežnih recenzija. Mrežne recenzije, koje predstavljaju korisnička iskustva i stavove o proizvodima ili uslugama, neprestano dobivaju na značaju i utjecaju [1]. S obzirom na sve veću važnost recenzija, razumijevanje njihove vjerodostojnosti postaje ključno za donošenje informiranih odluka. Recenzije su od velike vrijednosti i za potrošače i za poduzeća. Potrošači kroz njih dobivaju korisne informacije o proizvodima i uslugama, dok poduzeća dobivaju povratne informacije [2]. Potencijalni kupci često se oslanjaju na recenzije prilikom donošenja odluka o kupnji. Istodobno, poduzeća ih koriste kao izvor informacija za poboljšanje svojih proizvoda, usluga ili marketinških strategija [3]. Stoga su recenzije vrijedan izvor za donošenje odluka.
Danas, postoji velik broj mrežnih stranica na kojima korisnici mogu objavljivati recenzije i dijeliti svoja iskustva o proizvodima ili uslugama. Među najpopularnijima su TripAdvisor, Amazon, Yelp i mnoge druge. U ovom radu koriste se recenzije sa stranice Yelp.
Yelp[4] je jedna od najpopularnijih platformi za objavljivanje recenzija restorana, trgovina i raznih drugih poslovnih subjekata. Platforma omogućuje korisnicima ocjenjivanje poslovanja putem zvjezdica, kao i ostavljanje detaljnih tekstualnih recenzija, što pridonosi njenoj pouzdanosti i popularnosti.

Pozitivne recenzije mogu značajno doprinijeti financijskom uspjehu i ugledu poduzeća, dok negativne recenzije mogu narušiti njihovu reputaciju i prodaju [5]. Kako recenzije postaju sve utjecajnije u donošenju odluka potrošača, raste i interes za njihovu manipulaciju u marketinške svrhe. Neka poduzeća angažiraju stručnjake kako bi napisali lažne recenzije s ciljem promoviranja svojih proizvoda ili narušavanja ugleda konkurencije [6]. Takve lažne recenzije mogu biti pozitivne, radi poticanja kupnje, ili negativne, radi odvraćanja kupaca od proizvoda konkurentskih tvrtki [6].

Izvori lažnih recenzija uključuju zaposlenike poduzeća, konkurenciju ili čak poznanike [7]. Štoviše, napredak umjetne inteligencije omogućio je razvoj automatiziranih sustava za generiranje lažnih recenzija. Ti sustavi oponašaju jezik i stil stvarnih korisnika, prilagođavajući sadržaj određenim proizvodima te u kratkom vremenu stvaraju velike količine podataka [7]. Ovi automatizirani alati predstavljaju značajan izazov u razlikovanju lažnih od stvarnih recenzija [8].

Posljednjih godina otkrivanje lažnih recenzija privuklo je značajnu pozornost. Lažne recenzije postale su ozbiljan problem jer mogu navesti potrošače na pogrešne odluke, što rezultira gubitkom povjerenja ne samo u proizvode, već i u same platforme za recenzije. Iako mnoge web stranice za recenzije integriraju klasifikatore kako bi filtrirale stvarne i lažne recenzije, problem ostaje složen, što zahtijeva razvoj naprednih metoda za prepoznavanje lažnih sadržaja.

Jedna od učinkovitih strategija za otkrivanje lažnih recenzija su metode strojnog učenja. Primjena strojnog učenja omogućuje analizu podataka recenzija i prepoznavanje obrazaca koji pomažu u razlikovanju stvarnih od lažnih recenzija. Stoga, cilj ovog rada je istražiti i implementirati različite algoritme nadziranog strojnog učenja za automatsko otkrivanje lažnih recenzija na platformi Yelp. U okviru rada, analizirat će se i usporediti performanse algoritama kao što su logistička regresija, potporni vektorski strojevi, algoritmi bazirani na stablima i neuronske mreže, s ciljem pronalaženja optimalnog modela za ovu svrhu. Također će se istražiti i različite značajke recenzija, kako bi se našle optimalne kombinacije značajki za postizanje najbolje klasifikacije. Konačni cilj je osvijestiti ovaj problem i doprinijeti razvoju učinkovitijih metoda za njegovo rješavanje.

U prvom poglavlju rada obrađena je definicija i problematika lažnih recenzija. Drugo poglavlje uključuje analizu skupa podataka i relevantne značajke lažnih recenzija. U trećem poglavlju opisane su korištene metode strojnog učenja, dok u posljednjem poglavlju se analiziraju i uspoređuju dobiveni rezultati.

2. Problem lažnih recenzija

Recenzije koje su objavili ljudi koji se nisu osobno susreli s predmetima ili uslugama koje recenziraju smatraju se lažnim recenzijama [9]. Sukladno tome, korisnik koji objavljuje lažne recenzije smatra se lažnim recenzentom (engl. *spamer*) [9]. Recenzija koja nije lažna naziva se autentičnom, istinitom ili pravom recenzijom te se recenzenti koji objavljuju takve recenzije smatraju pravim, autentičnim recenzentima [10].
Lažne recenzije imaju cilj zvučati što autentičnije tj. realističnije kako bi bile uvjerljive i utjecajnije. Također, često su napisane na način koji pojačava njihov učinak, bilo pozitivan ili negativan, na proizvod ili uslugu [11]. Lažne recenzije mogu se podijeliti u tri glavne kategorije [12]:

- **Neistinite recenzije** - Lažne recenzije koje pišu korisnici kako bi naštetiti ugledu poduzeća ili pozitivne recenzije kako bi se poduzeće promoviralo
- **Recenzije robnih marki** - Recenzije koje samo komentiraju marku proizvoda i ne daju recenziju proizvoda
- **Nerecenzije** - Recenzije koje su irelevantne i ne nude istinsko mišljenje ili su jednostavno reklame

Posljednje dvije vrste nazivaju se ometajućim neželjenim mišljenjima (engl. *disruptive spam opinions*) i uzrokuju malu prijetnju te ih svatko može lako prepoznati prilikom čitanja [12]. Međutim, prvu kategoriju, neistinite recenzije, puno je teže razlikovati od pravih recenzija. To potvrđuje i istraživanje koje je ručnim označavanjem postiglo samo oko 60% točnosti [13]. To ukazuje na ozbiljnost problema i potrebu za automatiziranim metodama. Primarni izazov za ručno označavanje je nedostatak bilo kakvih razlikovnih riječi koje bi jasno ukazivale na lažnu recenziju [14].

Proteklih godina predloženo je mnogo različitih pristupa kako bi se riješio problem vjerodostojnosti informacija na društvenim medijima [15]. Procjena vjerodostojnosti informacija bavi se analizom sadržaja koji su generirali korisnici, karakteristikama autora te društvenim odnosima koji povezuju korisnike platforme [15]. Što se tiče otkrivanja lažnih recenzija, predložene su tri osnovne tehnike za njihovo prepoznavanje [16]:

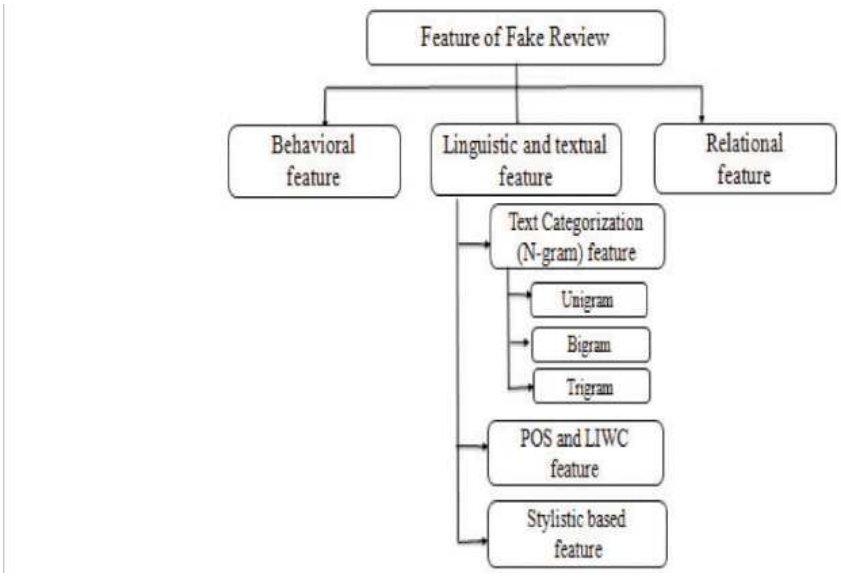
- **Pristup usmjeren na recenziju** - Analizira sadržaj recenzije kako bi utvrdio njezinu autentičnost. Koriste se značajke poput sličnosti sadržaja recenzije, prepoznavanje obrazaca jezika ili analize stila pisanja.
- **Pristup usmjeren za recenzente** - Istražuje ponašanje korisnika uzimajući u obzir sve podatke o korisnicima i sve recenzije koje su napisali. Koriste se značajke poput IP adrese, starosti računa i broja napisanih recenzija.
- **Pristup usmjeren na proizvod** - Analizira podatke vezane uz proizvod koristeći značajke poput cijene proizvoda i prodajnog ranga.

Većina istraživanja o izazovu otkrivanja lažnih recenzija oslanja se na tehnike strojnog učenja [17]. Prvi pristupi bili su usmjereni na recenzije, točnije na jednostavne tekstualne značajke izvučene iz teksta recenzije. Istraživanja su bila fokusirana na jezične značajke i na sličnost sadržaja recenzija, smatrajući dvije recenzije koje su jako slične lažnima. No, mnogo lažnih recenzija nastaje kopiranjem rečenica iz istinitih recenzija te pristupi usmjereni samo na recenziju ne uspijevaju identificirati takve recenzije kao lažne [10]. Iz tog je razloga predložen pristup usmjeren na recenzente koji se pokazao učinkovitiji od jezičnih značajki [15]. Pristupi usmjereni na recenzente izdvajaju karakteristična ponašanja lažnih recenzenata, no ne mogu poslužiti kao konačan dokaz lažnih recenzija [10]. Pristup usmjeren na proizvode nije toliko razmatran zbog nedostatka dostupnih informacija specifičnih za neki proizvod ili uslugu, te značajke koje se temelje na proizvodu ne pomažu izravno u prepoznavanju lažnih recenzija [10].
Najbolje rezultate u otkrivanju lažnih recenzija dali su pristupi koji se temelje na nadziranim ili polunadziranim tehnikama strojnog učenja koje uzimaju u obzir i pristup usmjeren na recenzije i pristup usmjeren na recenzente [15].

Međutim, još uvijek postoji mnogo izazova poput poteškoća u izradi skupova podataka i slabe sposobnosti prilagodbe domene algoritma [2]. Najveći izazov u otkrivanju lažnih recenzija je nedostatak temeljnih istinitih podataka s kojima bi se mogli trenirati i testirati modeli [17]. Kako bi se mogli razviti modeli koji mogu identificirati lažne recenzije, prvo se mora imati skup postojećih lažnih i pravih recenzija s kojima se model može trenirati. Neka istraživanja su pokušala prevladavati ovaj izazov stvarajući generirane lažne recenzije [18]. Takvim recenzijama nedostaje autentičnosti te ni na koji način ne predstavljaju ponašanje stvarnog lažnog recenzenta pa su kritizirane kao zamjena za lažne recenzije [19]. Drugi pristup, koji je ujedno i korišten u ovom radu, je korištenje recenzija koje su označene i filtrirane s pomoću algoritma na platformama.

2.1. Značajke lažnih recenzija

Za učinkovito otkrivanje lažnih recenzija, ključno je razumjeti različite značajke koje se mogu koristiti u analizi. Najčešće korištene značajke su one usmjerene na recenzente ili na samu recenziju. Također se značajke dijele na ponašajne, jezične i relacijske značajke. Detaljna podjela je prikazana na slici 2.1. Razumijevanje i pravi odabir ovih značajki pomaže u razvoju modela koji mogu razlikovati lažne od pravih recenzija.



Slika 2.1. Prikaz značajki recenzija. Slika preuzeta s [16]

2.1.1. Značajke usmjerene na recenziju

Značajke usmjerene na recenziju su značajke koje se mogu izdvojiti iz teksta recenzije, kao i metapodaci povezani s njom. Korištenje ovih značajki je najčešći oblik otkrivanja lažnih recenzija [1]. Ove značajke su se pokazale korisnima u otkrivanju lažnih recenzija [1].

Jezične značajke i svojstva teksta recenzije jedne su od važnijih značajki za otkrivanje lažnih recenzija [16]. One uključuju složenost teksta, prosječan broj riječi po rečenici, stilske značajke teksta i druge elemente [20].

N-gram je kontinuirani niz od n riječi iz teksta. Niz od jedne riječi naziva se unigram, dok su nizovi od dvije riječi bigrami, a od tri riječi trigrami. Ova značajka koristi se u procesu vektorizacije teksta s pomoću učestalosti termina, što omogućuje brojanje učestalosti pojavljivanja pojmova u svakoj recenziji.

Također, jedna od jezičnih značajki je **označavanje vrsta riječi** (engl. *part-of-speech-tagging*, POS). To je postupak označavanja u kojem se svakoj riječi u tekstu dodjeljuje oznaka koja odgovara njezinoj gramatičkoj kategoriji. Dodjeljivanje POS oznaka riječi pruža dodatne informacije o samoj riječi što može pridonijeti određivanju konteksta u rečenici. U prethodnim radovima je utvrđeno da prave recenzije koriste više imenica, prijedloga i pridjeva, dok lažne recenzije koriste više glagola i priloga [20].

Još jedna bitna značajka je **sentiment** teksta recenzije. Istraživanja pokazuju da lažne recenzije sadrže veći postotak riječi koje izražavaju pozitivne emocije nego prave pozitivne recenzije; slično tome, lažne negativne recenzije koriste više negativnih izraza nego prave negativne [20].

Jedna od glavnih značajki za prepoznavanje lažnih recenzija je **duljina teksta recenzija**. Oko 80% lažnih korisnika nema recenziju dulju od 135 riječi, dok više od 92% pouzdanih korisnika piše recenzije duže od 200 riječi [1]. Također, gledajući omjer riječi i stop-riječi, prosjek pravih riječi iznosi 1.22, dok omjer kod lažnih recenzija iznosi 0.96 [21]. Ovaj omjer ukazuje na smislenost recenzije. Veći dio riječi trebao bi biti povezan s informativnim sadržajem, s obzirom na to da stop-riječi ne prenose nikakvo semantičko značenje. S ovim omjerom mogu se filtrirati i duge lažne recenzije koje imaju manju informativnu vrijednost.

Očekuje se da pravi korisnici ocjene proizvod ili uslugu na temelju prosječne ocjene tog proizvoda [22]. Lažni korisnici namjerno koriste ili vrlo visoku ili vrlo nisku ocjenu kako bi promovirali ili degradirali uslugu i proizvod. Zbog toga njihove ocjene često odstupaju od prosječne ocjene proizvoda ili usluge [1]. Pokazano je da prave recenzije odstupaju svojom ocjenom do 0.6 na ljestvici od 5 zvjezdica, dok većina lažnih recenzija ima odstupanje veće od 2 [5].

Značajke metapodataka recenzije su bilo koje informacije o recenziji koje opisuju njen sadržaj, ali nisu stvarni sadržaj. One uključuju identifikaciju recenzenta, IP adresu računala s kojeg je recenzija napisana, ocjenu i slično. S pomoću ovih značajki mogu se otkriti anomalije ponašanja korisnika, ali ove značajke nisu dostupne u mnogim izvorima podataka što ograničava njihovu korisnost [2].

Iako se većina istraživanja usredotočuje na značajke usmjerene na recenziju, sadržaj lažnih značajki postaje sve sličniji onom u autentičnim značajkama te korištenje samo značajki usmjerenih na recenzije postaje izazovno [19]. Zbog toga treba uzeti u obzir i ponašanje korisnika koji pišu lažne recenzije [2].

2.1.2. Značajke usmjerene na recenzente

Značajke usmjerene na recenzente obuhvaćaju ponašanje korisnika na platformi i pružaju ključne informacije za prepoznavanje lažnih recenzija. Ove značajke uključuju ukupni broj recenzija korisnika, omjer negativnih i pozitivnih recenzija, broj prijatelja na platformi i slično. Prepoznavanje lažnih korisnika može biti korisno budući da takvi korisnici dijele karakteristike profila i obrasce aktivnosti [1]. Također, ako se recenzent identificira kao netko tko piše lažne recenzije, sve njegove recenzije lako se mogu kategorizirati kao lažne [22].

Lažni recenzenti ponašaju se drugačije od pravih korisnika, često pišu veliki broj recenzija u kratkom vremenskom razdoblju i koriste ekstremne ocjene [16]. Uočeno je da oko 75% lažnih korisnika napiše više od 5 recenzija u jednom danu, dok 90% legitimnih korisnika nikada ne napiše više od jedne recenzije dnevno [1].
Osim toga, lažni recenzenti pišu recenzije samo u tim kratkim intervalima i obično nisu dugogodišnji aktivni članovi [5]. Pokazano je da je 80% lažnih korisnika aktivno samo 2 mjeseca, dok su pravi korisnici aktivni više od 10 mjeseci [5]. Posljedično, lažni korisnici

obično imaju i manji broj prijatelja na platformi ili ih uopće nemaju.
Lažni korisnici također imaju manji ukupni broj recenzija jer nisu stvarno zainteresirani za proizvode ili usluge. Prema istraživanjima, 80% lažnih korisnika ima ukupno 11 recenzija [5].
Također, lažne recenzije mogu biti korištene za promociju ili degradaciju ciljanih poduzeća. To rezultira time da lažni korisnici često imaju puno veći postotak pozitivnih recenzija ili negativnih recenzija, dok pravi korisnici imaju ravnomjerno raspoređen raspon ocjena [5].
Još jedan pokazatelj lažnih korisnika je maksimalna sličnost njihovih recenzija. Lažni korisnici često pišu slične ili identične recenzije za različite proizvode [2]. Ova sličnost se najčešće mjeri s pomoću kosinusne sličnosti između bilo koje dvije recenzije korisnika. Istraživanja pokazuju da oko 70% korisnika koji pišu prave recenzije imaju vrlo malu sličnost, dok većina lažnih korisnika pokazuje visoku sličnost u svojim recenzijama [5]. Uzimajući u obzir sve navedene značajke, prepoznavanje lažnih korisnika može značajno doprinijeti učinkovitijem otkivanju lažnih recenzija.

Analiza različitih pristupa i značajki koji se koriste u detekciji lažnih recenzija jasno pokazuje složenost problema lažnih recenzija. Stoga je jedan od ciljeva ovog rada istražiti i identificirati značajke koje najviše doprinose uspješnoj klasifikaciji lažnih recenzija. Poseban naglasak stavljen je na usporedbu učinkovitosti korištenja jezičnih značajki dobivenih iz samih recenzija, s pristupima koji se oslanjaju na numeričke značajke koje opisuju ponašanje recenzenata. Dodatno, analiziran je i učinak kombiniranja obje vrste značajki, s ciljem utvrđivanja optimalnog skupa značajki za klasifikaciju lažnih recenzija. Ovim pristupom, pruža se dublji uvid u važnost različitih vrsta značajki u kontekstu problema lažnih recenzija. U sljedećim poglavljima detaljno su opisane korištene značajke i modeli.

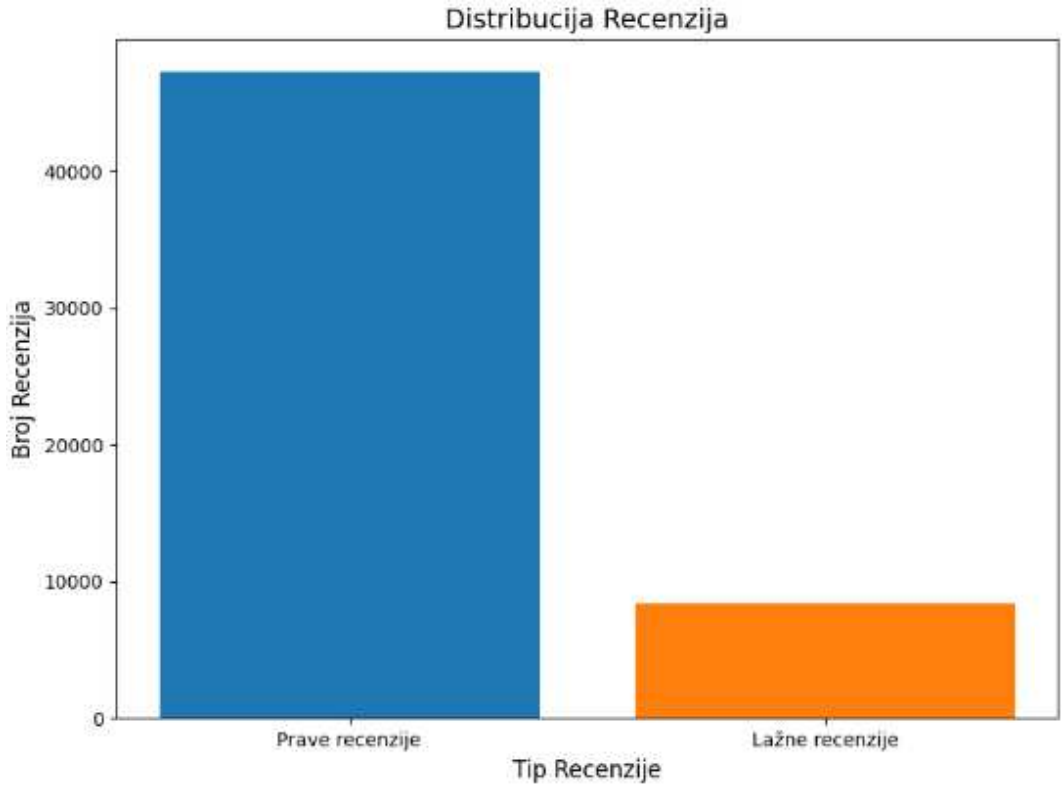
3. Skup podataka

U ovom diplomskom radu korištena je stranica Yelp.com, jedna od najpoznatijih platformi za recenzije, koja omogućuje korisnicima da ostavljaju svoje recenzije o različitim poslovnim subjektima. Yelp ujedno filtrira sumnjive ili lažne recenzije i čini ih javno dostupnima. Iako je algoritam za filtriranje recenzija tajan, istraživanja su pokazala njegovu pouzdanost [23]. Istraživanje je potvrdilo da Yelp ne koristi samo jezične značajke, već i različite značajke ponašanja korisnika, uz dodatne parametre koji nisu javno dostupni, poput IP adresa i geo-lokacije. Na platformi Yelp, recenzije su podijeljene u dvije kategorije: preporučene i filtrirane recenzije. Preporučene recenzije su vidljive na glavnoj stranici objekata, a Yelpov algoritam ih smatra vjerodostojnima. S druge strane, filtrirane recenzije nisu prikazane na glavnoj stranici objekta, već su dostupne putem poveznice na dnu stranice. Yelp priznaje da njihov algoritam za filtriranje recenzija čini pogreške, te je do sada ukupno uklonio 9% svojih recenzija, dok je 16% filtrirao [4].

Za potrebe ovog diplomskog rada korištena je kombinacija dvaju skupova podataka, Yelp i YelpZip. Skup podataka Yelp predstavlja podskup podataka o korisnicima, poduzećima i recenzijama s platforme Yelp te je dostupan u obliku JSON datoteka. Sadrži ukupno 6,990,280 recenzija. Platforma Yelp koristi svoj algoritam za filtriranje koji razlučuje lažne recenzije i odvaja ih u filtrirani popis dostupan na dnu stranice putem poveznice. Skup podataka YelpZip [24] je prikupljen sa stranice Yelp te sadrži oko 600000 recenzija restorana. Kako platforma Yelp prikazuje preporučene recenzije, one se u ovom skupu podataka smatraju stvarnim recenzijama, dok se filtrirane recenzije tretiraju kao lažne recenzije. Recenzije su napisane za 5044 restorana i napisalo ih je 32832 korisnika.

Konačni skup podataka koji je korišten u ovom radu dobiven je kombiniranjem ovih dvaju skupova podataka. Iz skupa podataka YelpZip preuzete su označene recenzije restorana, dok su iz skupa Yelp dohvaćene značajke korisnika. Podaci su povezani na teme-

lju teksta recenzije i imena restorana. Ovim spajanjem podataka dobivene su informacije o korisnicama uz oznake lažnih recenzija.
Konačni skup podataka sadrži 55609 recenzija napisane o 778 restorana od strane 32793 korisnika. Skup sadrži 47259 pravih recenzija i 8350 lažnih recenzija. Ova raspodjela je vidljiva na slici 3.1. Yelp filtrira oko 16% recenzija kao lažnih, što ukazuje na iskrivljenu klasnu distribuciju koja će biti razmatrana u daljnjim poglavljima.



Slika 3.1. Prikaz raspodjele lažnih i pravih recenzija u skupu podataka

Konačni skup podataka sadrži sljedeće informacije:

- Ime restorana
- Datum objavljivanja recenzije
- Prosječna ocjena restorana
- Tekst recenzije
- Ocjena recenzije
- Oznaka jeli recenzija lažna ili prava

- Broj primljenih korisnih glasova za recenziju
- Datum kada se korisnik pridružio Yelpu
- Broj recenzija koje je korisnik napisao
- Broj korisnih glasova koje je korisnik dobio na svim svojim recenzijama
- Broj korisnikovih prijatelja

3.1. Analiza skupa podataka

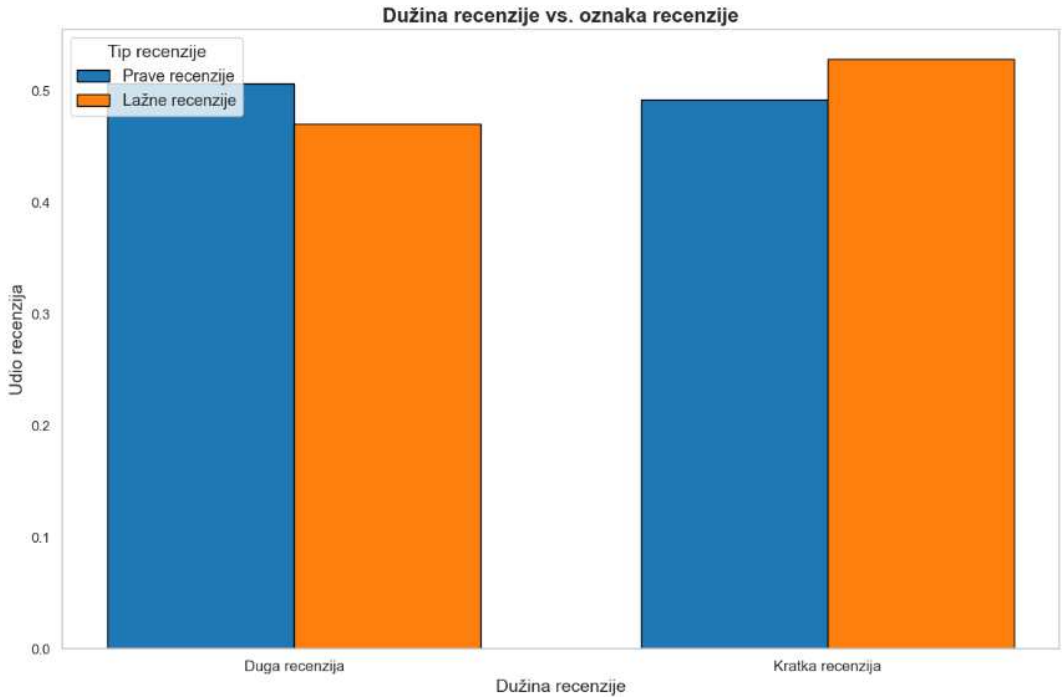
Analiza značajki skupa podataka ključna je za razumijevanje razlika između lažnih i pravih recenzija te omogućuje dobivanje uvida koji mogu pomoći u izgradnji klasifikacijskog modela. U ovom odjeljku analiziraju se različite karakteristike recenzija i korisnika kako bi se identificirali obrasci ponašanja i tekstualne značajke koje mogu ukazivati na lažne recenzije. Analizirane značajke mogu se podijeliti u dvije kategorije: značajke usmjerene na recenziju i značajke usmjerene na recenzente. Značajke usmjerene na recenziju su značajke koje su konstruirane pomoću informacija sadržanih u recenziji. Nasuprot tome, značajke usmjerene na recenzente gledaju sve recenzije i informacije koje je napisao određeni korisnik.
Cilj ove analize je zaključiti koje značajke pružaju najviše informacija za razlikovanje lažnih od pravih recenzija, te su stoga najprikladnije za korištenje u modelu klasifikacije.

3.1.1. Analiza značajki recenzija

Analiza značajki recenzija izvedene su iz samog teksta recenzije, a mogu se podijeliti u dvije kategorije, jezične značajke i značajke vezane uz sadržaj recenzije. U nastavku se detaljnije analiziraju značajke povezane sa sadržajem recenzije.

Duljina teksta recenzija često se koristi kao značajka za razlikovanje lažne od prave recenzije. Pretpostavka je da će lažne recenzije biti kraće jer autori nemaju stvarno iskustvo s proizvodom ili uslugom, imaju ograničeno znanje o njemu te ne žele uložiti puno vremena u pisanje detaljnog opisa.
Za analizu, izračunat je broj riječi u svakoj recenziji. Skup podataka je podijeljen u dvije

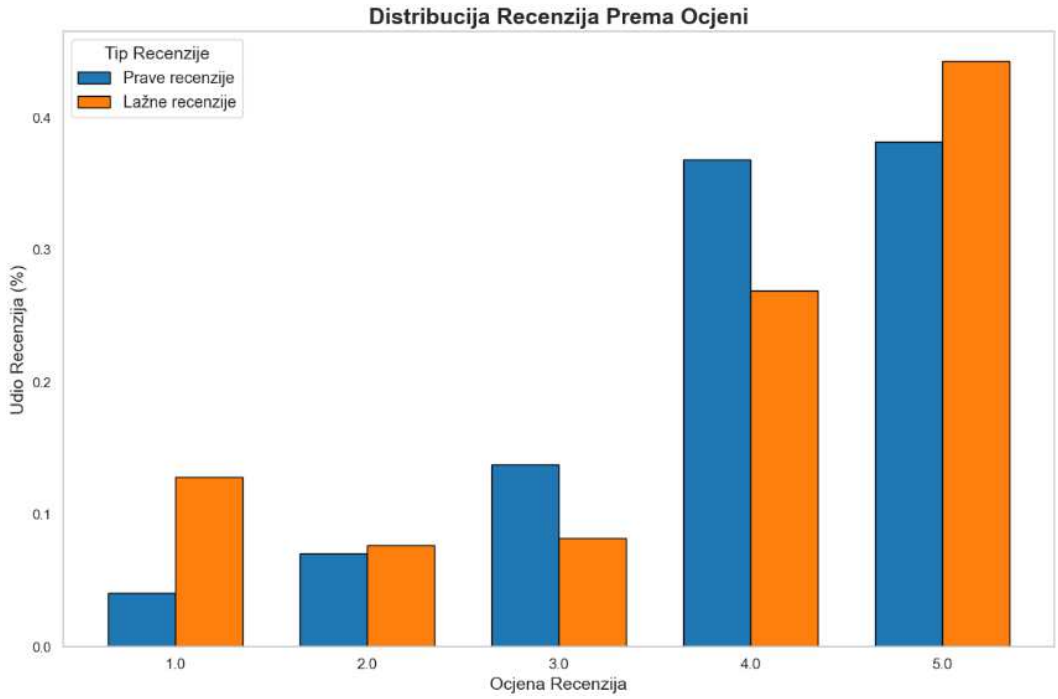
skupine: "kratke recenzije" koje imaju manje od 70 riječi i "duge recenzije" koje imaju više od 70 riječi. Prag od 70 riječi odabran je na temelju medijana duljine svih recenzija u skupu podataka. Slika 3.2. prikazuje distribuciju recenzija prema duljini teksta.
Rezultati pokazuju da lažne recenzije imaju nešto veći udio kratkih recenzija (53%) u usporedbi s pravim recenzijama (49%). Iako razlika nije značajna, duljina teksta može biti korisna značajka u kombinaciji s drugim značajkama za detekciju lažnih recenzija.



Slika 3.2. Prikaz distribucije recenzija prema duljini recenzije

Ocjena koju je korisnik dao uz recenziju je pokazatelj korisnikovog zadovoljstva proizvodom ili uslugom. Pretpostavka je da će lažne recenzije često koristi ekstremne ocjene (1 ili 5 zvjezdica) kako bi manipulirale prosudbom drugih korisnika.
Za analizu, izračunati su i postotci pozitivnih (5 zvjezdica) i negativnih recenzija (1 zvjezdica) za lažne i prave recenzije. Slika 3.3. prikazuje udio lažnih i pravih recenzija za svaku ocjenu.
Rezultati pokazuju da lažne recenzije imaju najveći udio u ocjeni 5, koja čini čak 44% svih lažnih recenzija. Također, značajan udio lažnih recenzija nalazi se i u ocjeni 1 s 13% udjela, što potvrđuje sklonost korištenju ekstremnih ocjena za stvaranje lažnog dojma. S druge strane, prave recenzije imaju uravnoteženiju raspodjelu ocjena. Najveći udio pravih recenzija ima u ocjeni 5, s 38%, ali značajan broj korisnika daje i srednje ocjene (3 i 4 zvjezdice).

Ova analiza pokazuje da ocjena recenzije korisna značajka za model klasifikacije. Ekstremne ocjene mogu povećati vjerojatnost da je recenzija lažna, dok srednje ocjene mogu smanjiti tu vjerojatnost.

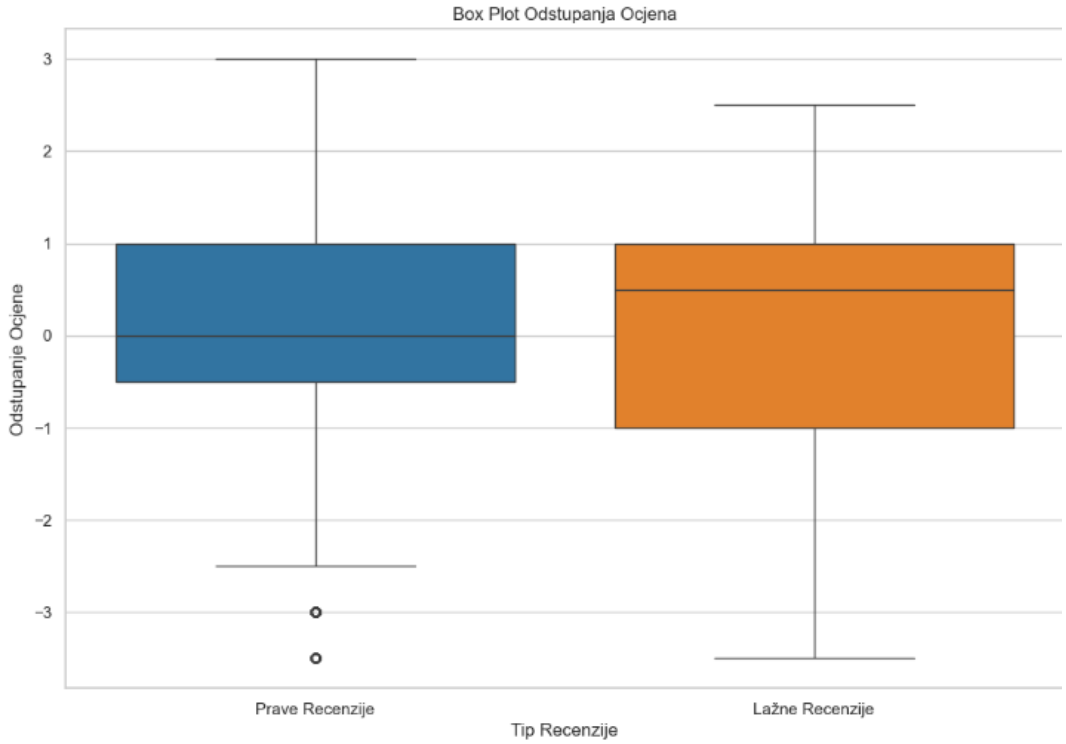


Slika 3.3. Prikaz distribucije recenzija prema ocjeni recenzije

Odstupanje ocjene recenzije od prosječne ocjene restorana mjeri razliku između ocjene koju je korisnik dao i prosječne ocjene tog restorana. Pretpostavka je da će lažni recenzenti, s ciljem manipulacije, davati ocjene koje značajno odstupaju od prosjeka.

Slika 3.4. prikazuje raspodjelu odstupanja za prave i lažne recenzije. Analiza pokazuje da prave recenzije imaju prosječno odstupanje od 0.06, dok lažne recenzije imaju prosječno odstupanje od -0.09. Iako postoji razlika u prosječnim vrijednostima, raspon odstupanja je sličan za obje skupine, s većinom vrijednosti između -1 i 1.

S obzirom na malu razliku u raspodjeli i preklapanje raspona odstupanja između lažnih i pravih recenzija, ova značajka se ne koristi u modelu klasifikacije.

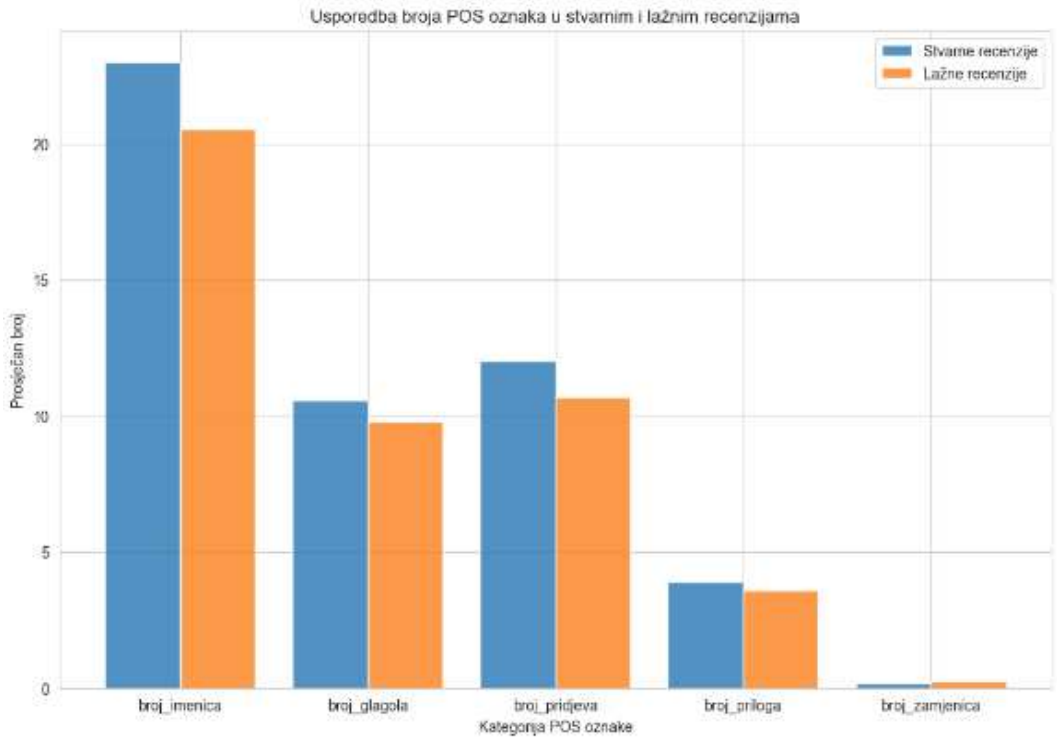


Slika 3.4. Prikaz raspodjele odstupanja ocjene recenzije

Analiza gramatičkih kategorija riječi (**POS oznaka**) može otkriti razlike u stilu pisanja između lažnih i pravih recenzija. U radu su korištene sljedeće POS oznake: imenice, glagoli, pridjevi, prilozi i zamjenice. Za automatsko označavanje gramatičkih kategorija korištena je biblioteka nltk [25].

Slika 3.5. prikazuje prosječan broj svake POS oznake u lažnim i pravim recenzijama. Iako razlike nisu velike, uočeno je da prave recenzije koriste više imenica i pridjeva, što ih čini informativnijima i konkretnijima.

Ova analiza sugerira da se POS oznake mogu koristiti kao značajke u modelu klasifikacije. Na primjer, veći udio imenica i pridjeva može povećati vjerojatnost da se recenzija označi kao autentična.

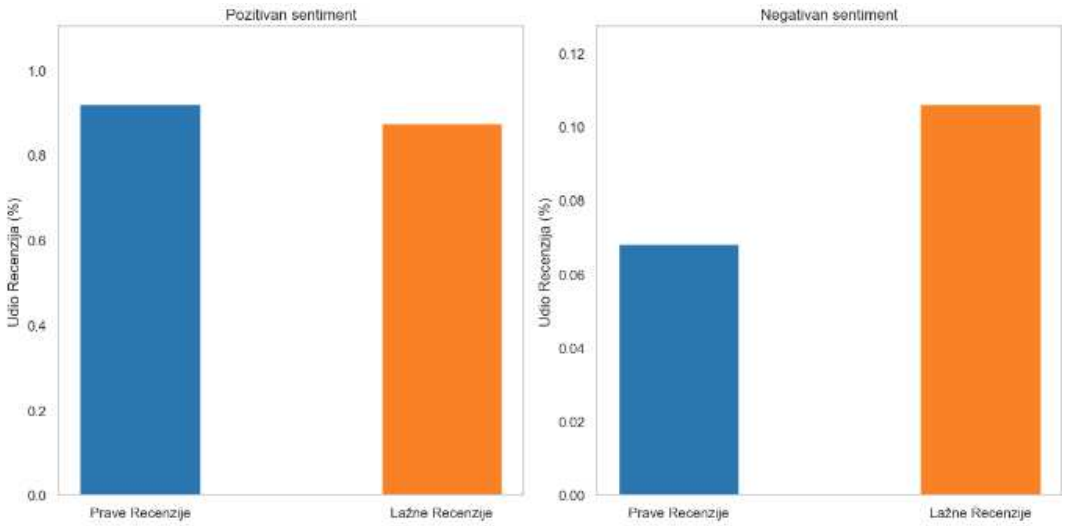


Slika 3.5. Prikaz raspodjele odstupanja ocjene recenzije

Sentiment recenzije određuje je li tekst pozitivan, negativan ili neutralan. Za analizu sentimenta korištena je biblioteka VADER koja je dizajnirana za analizu sentimenta u kratkim tekstovima, poput recenzija ili komentara [26]. Za izračunavanje sentimenta u obzir uzima velika slova i interpunkciju, tako da riječi napisane velikim slovima ili popraćene interpunkcijskim znakovima mogu imati jači emocionalni intenzitet.

VADER koristi unaprijed definiran rječnik riječi u kojem svaka riječ ima pridruženu ocjenu koja izražava njen sentiment. Ukupna ocjena sentimenta recenzije kreće se u rasponu od -1 (veoma negativno) do 1 (veoma pozitivno). Recenzije s ocjenom iznad 0.05 smatraju se pozitivnima, one s ocjenom ispod -0.05 negativnima, dok se sve ostale smatraju neutralnima. Distribucija lažnih i pravih recenzija prema sentimentu, prikazana na slici 3.6.

Analiza pokazuje da i lažne i prave recenzije pretežno pokazuju pozitivan sentiment. Međutim, udio pozitivnog sentimenta je nešto veći kod pravih recenzija, dok je udio negativnog sentimenta veći kod lažnih recenzija. S obzirom na to da postoje blage razlike u udjelima koje mogu ukazivati na obrasce izražavanja mišljenja između pravih i lažnih recenzija, sentiment recenzije može biti koristan za model klasifikacije.

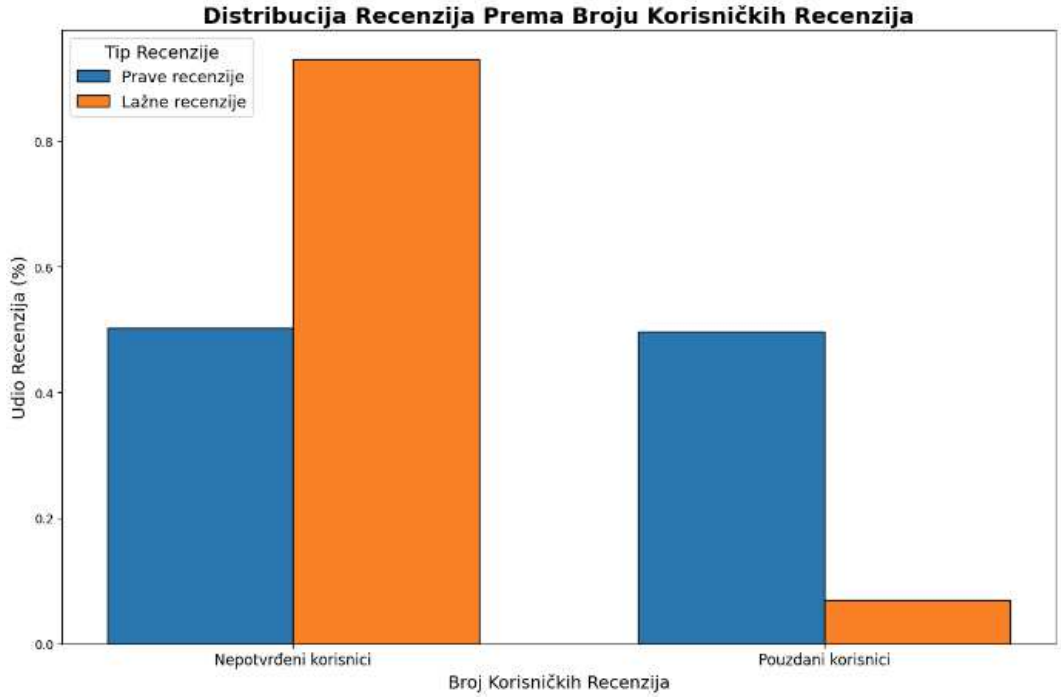


Slika 3.6. Prikaz distribucije recenzija prema sentimentu

3.1.2. Analiza značajki recenzenta

Prepoznavanje lažnih recenzenata ključno je za poboljšanje točnosti detekcije lažnih recenzija, budući da lažni recenzenti često dijele specifične karakteristike profila i obrasce aktivnosti. U nastavku su analizirane značajke povezane s korisnicima koji su pisali recenzije, s ciljem identificiranja tih karakteristika.

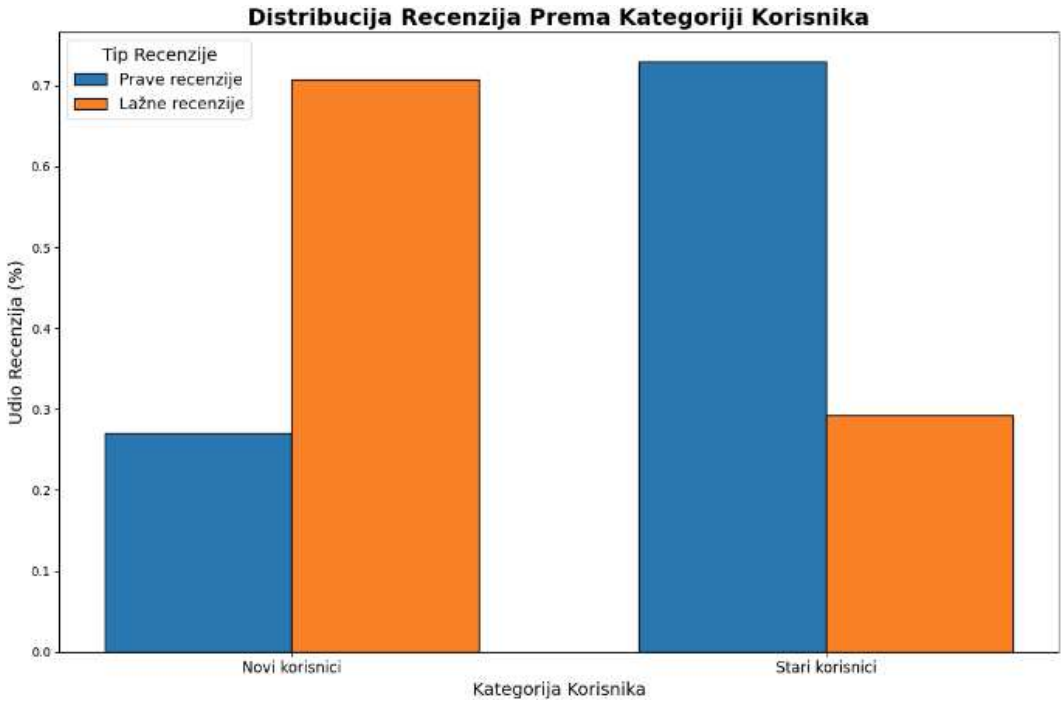
Broj ukupnih recenzija recenzenta je broj recenzija koje je korisnik napisao te može biti indikator njegove autentičnosti. Pretpostavka je da lažni korisnici nisu dugogodišnji i aktivni članovi platforme, te stoga imaju manji ukupan broj recenzija.
Za analizu, korisnici su podijeljeni u dvije skupine. Ako korisnik ima više od 20 recenzija, smatra se "Pouzdanim korisnikom", u suprotnom korisnik se smatra "Nepotvrđenim korisnikom". Granica od 20 recenzija određena je na temelju distribucije ovog atributa i eksperimentiranja s različitim granicama. Slika 3.7. prikazuje postotak lažnih i pravih recenzija u odnosu na to je li korisnik Pouzdan ili Nepotvrđen.
Rezultati pokazuju da je vjerojatnost da je lažnu recenziju napisao Pouzdani korisnik samo 6%, dok je vjerojatnost da ju je napisao Nepotvrđeni korisnik čak 93%. To pokazuje da je ova značajka vrlo korisna za model klasifikacije. U modelu, korisnici s manjim brojem recenzija će imati veću vjerojatnost da budu označeni kao lažni.



Slika 3.7. Prikaz distribucije recenzija prema ukupnom broju korisničkih recenzija

Trajanje članstva korisnika na platformi može bit pokazatelj njegove autentičnosti. Pretpostavka je da će lažni korisnici imati kraće članstvo jer ne koriste platformu. Za analizu, izračunata je razlika između datuma recenzije i datuma kada se korisnik pridružio platformi. S obzirom na to da neki korisnici imaju više recenzija u skupu podataka, za svakog je odabrana maksimalna vrijednost članstva. Korisnici su zatim podijeljeni u dvije skupine: korisnici čije članstvo traje više od 60 dana i korisnici čije članstvo traje manje od 60 dana. Slika 3.8. prikazuje postotak lažnih i pravih recenzija u odnosu na trajanje članstva.

Rezultati pokazuju da su korisnici koji su članovi platforme manje od 60 dana napisali 71% lažnih recenzija, dok je 73% pravih recenzija napisano od korisnika s dugotrajnim članstvom. Ovo pokazuje da je trajanje članstva korisna značajka za model klasifikacije. U modelu, kraće trajanje članstva može povećati vjerojatnost da se recenzija označi kao lažna.

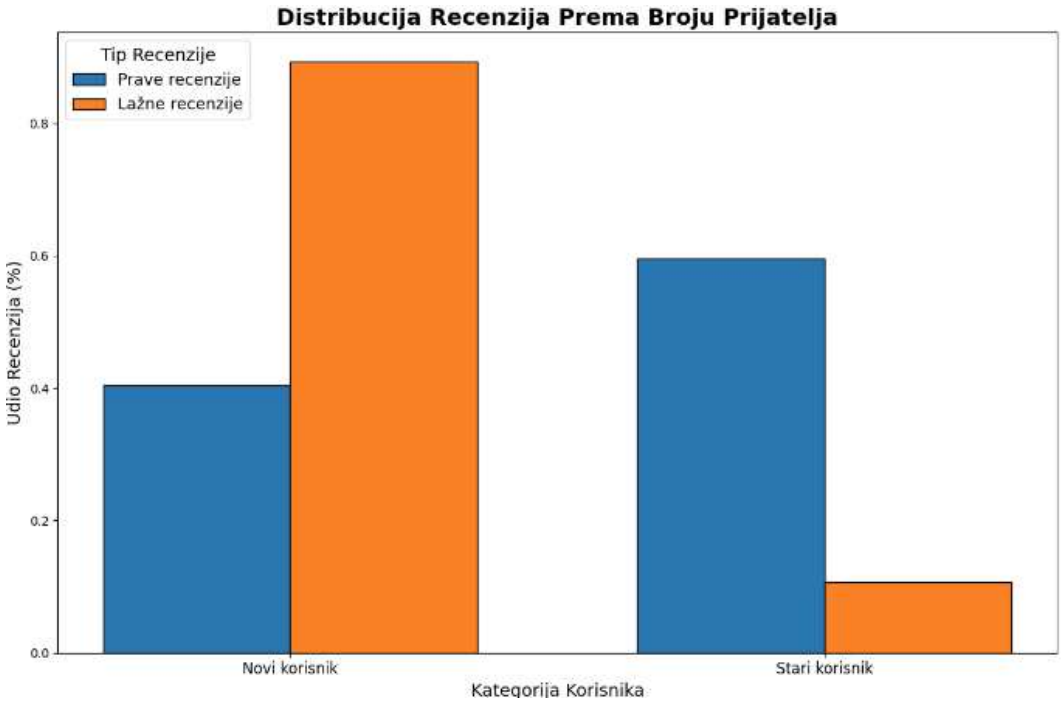


Slika 3.8. Prikaz distribucije recenzija prema trajanju članstva korisnika

Broj prijatelja koje korisnik ima na platformi može biti pokazatelj njegove autentičnosti i reputacije. Pretpostavka je da korisnici s većim brojem prijatelja ne bi riskirali svoj ugled pisanjem lažnih recenzija.

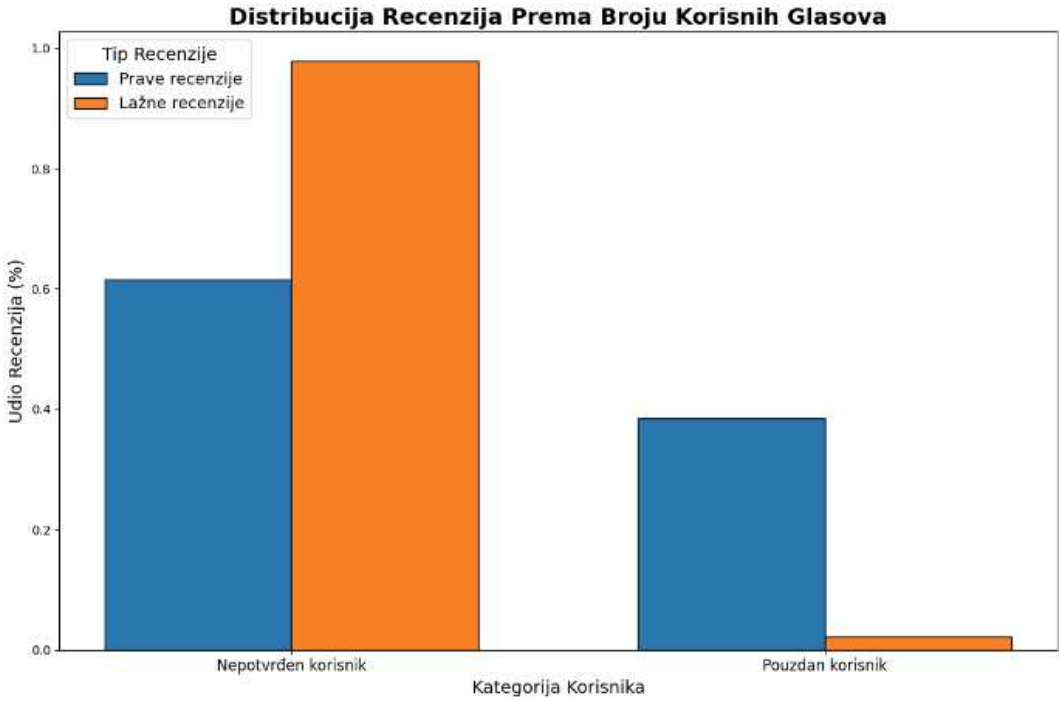
Za analizu, korisnici su podijeljeni u dvije skupine: oni s više od 15 prijatelja i oni s 15 ili manje prijatelja. Slika 3.9. prikazuje postotak lažnih i pravih recenzija u odnosu na broj prijatelja.

Rezultati pokazuju da su korisnici koji imaju manje od 15 prijatelja napisali skoro 90% lažnih recenzija, što pokazuje da je ova značajna korisna za izgradnju modela klasifikacije. U modelu, manji broj prijatelja značajno povećava vjerojatnost da se recenzija označi kao lažna.



Slika 3.9. Prikaz distribucije recenzija prema broju prijatelja korisnika

Ukupni broj korisnih glasova na recenzijama korisnika može pomoći u prepoznavanju lažnih korisnika. Pretpostavka je da će korisnici čije recenzije drugi smatraju korisnima i relevantnima imati veći broj korisnih glasova, te da samim time nisu lažne. Za analizu, korisnici su podijeljeni u skupine: "Pouzdani korisnici" koji imaju više od 30 korisnih glasova i "Nepotvrđeni korisnici" s manje od 30 korisnih glasova. Slika 3.10. prikazuje postotak lažnih i pravih recenzija u odnosu na broj korisnih glasova.
Rezultati pokazuju da su Nepotvrđeni korisnici napisali 98% lažnih recenzija, što pokazuje da je ova značajka korisna za izgradnju modela klasifikatora. U modelu klasifikacije, nizak broj korisnih glasova može značajno povećati vjerojatnost da se recenzija označi kao lažna.



Slika 3.10. Prikaz distribucije recenzija prema broju korisnih glasova na recenzijama

4. Modeli strojnog učenja

Strojno učenje jedan je od pristupa za prepoznavanje lažnih recenzija. Algoritmi strojnog učenja mogu se podijeliti u četiri glavne kategorije: nadzirano, nenadzirano, polunadzirano i učenje s pojačanjem. Ovaj rad koristi nadzirano strojno učenje, gdje model uči iz označenih podataka i predviđa pripadnost recenzija određenim kategorijama.
Klasifikacija je vrsta nadziranog učenja u kojoj model pokušava predvidjeti ispravnu oznaku za zadane ulazne podatke. Može biti binarna, višeklasna, klasifikacija s više oznaka i neuravnotežena klasifikacija. U ovom radu provodi se binarna klasifikacija, gdje se recenzije klasificiraju kao lažne ili autentične.
Proces klasifikacije lažnih recenzija uključuje nekoliko ključnih koraka:

- **Prikupljanje podataka** - Prikupljanje podataka o recenzijama
- **Predobrada podataka** - Primjena koraka pretprocesiranja podataka poput uklanjanja interpunkcijskih znakova, brojeva i zaustavnih riječi (engl. *stopwords*)
- **Ekstrakcija i odabir značajki** - Izdvajanje različitih vrsta značajki iz prethodno obrađenih podataka
- **Konstrukcija i testiranje modela klasifikatora** - Odabir modela za klasifikaciju te evaluacija s pomoću testnih podataka.

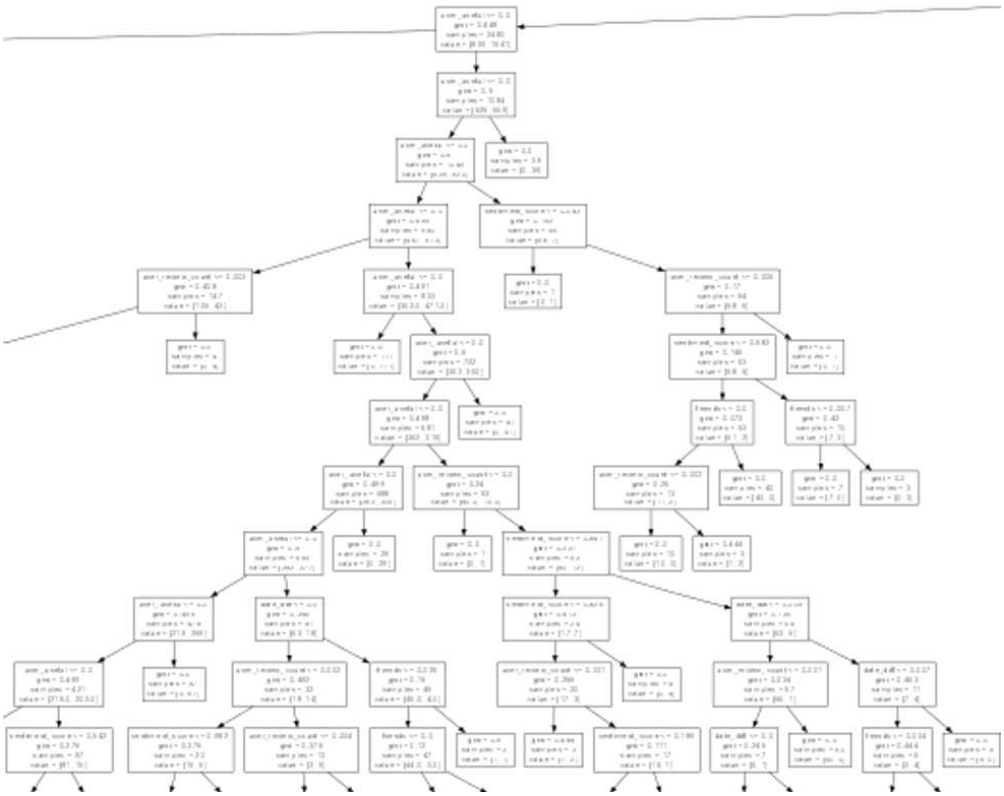
Svi koraci su detaljnije objašnjeni u nastavku i u sljedećim poglavljima.

4.1. Algoritmi

Za rješavanje problema klasifikacije lažnih recenzija korišteni su različiti algoritmi strojnog učenja koji su objašnjeni u nastavku.

Logistička regresija (engl. *Logistic regression*, LR) je jednostavan i interpretativan klasifikacijski model koji predviđa vjerojatnost pripadnosti primjerka jednoj od dviju klasa. Temelji se na sigmoidnoj funkciji koja ograničava vrijednosti izlaza između 0 i 1. Odabir granične vrijednosti za klasifikaciju može značajno utjecati na rezultate evaluacije modela. Iako se granična vrijednost obično postavlja na 0.5, prilagodba ove granice može poboljšati performanse, posebno u slučajevima kada su klase nebalansirane.
Kako bi se riješio problem prenaučenosti, često se primjenjuju tehnike regularizacije, pri čemu je L2 regularizacija jedna od najefikasnijih. Složenost modela logističke regresije kontrolira hiperparametar C. Manje vrijednosti parametra povećavaju regularizaciju i smanjuju prenaučenost, dok veće vrijednosti omogućuju bolju prilagodbu modela podacima.

Stablo odlučivanja (engl. *Decision tree*, DT) je klasifikacijski algoritam koji koristi strukturu u obliku stabla za donošenje odluka. Sastoji se od čvorova koji predstavljaju uvjete za podjelu podataka, te grana koje vode do krajnjih čvorova, odnosno listova, koji definiraju konačne klase. Proces klasifikacije započinje od korijenskog čvora, gdje se podaci dijele na temelju atributa, a svaka podjela se nastavlja dok se ne postigne kriterij zaustavljanja.
Algoritmi za izgradnju stabla koriste različite kriterije za odabir najboljeg atributa prema kojem će se izvršiti podjela. Jedan od kriterija je Gini indeks koji mjeri homogenost podataka u čvorovima te se koristi za odabir podjele koja rezultira što homogenijim čvorovima. Stablo odlučivanja ima nekoliko hiperparametara koji se mogu optimizirati poput maksimalne dubine stabla te minimalnog broja instanci u terminalnom čvoru. Ovi parametri pomažu u sprječavanju prenaučenosti i ograničavaju složenost modela.
Na slici 4.1. prikazan je dio stabla odlučivanja generiranog u ovom radu. Vizualizacija prikazuje neke čvorove s pripadajućim uvjetima i podjelama u procesu klasifikacije recenzija.



Slika 4.1. Dio generiranog stabla odlučivanja koje prikazuje proces klasifikacije recenzija

Nasumična šuma (engl. *Random forest*, RF) je ansambl metoda koja koristi više stabala odluke za poboljšanje točnosti i robusnosti predikcija. Algoritam gradi skup stabala na temelju nasumično odabranih podskupova podataka. Svako stablo u šumi donosi svoju odluku, a konačna klasifikacija se određuje glasanjem među svim stablima. Proces počinje odabirom broja stabala koja će se generirati, kao i broja atributa koji će se koristiti za svako stablo. Za svako stablo koristi se samo dio odabranih atributa. S ovim se omogućuje da svako stablo bude različito, čime se povećava raznolikost modela i smanjuje prenaučenost. Mogu se optimizirati isti hiperparametri kao i kod stabla odluka uz dodatak broja stabala koji određuje koliko će stabala biti uključeno u šumu.

XGBoost (engl. *Extreme Gradient Boosting*) se temelji na metodi gradijentnog pojačanja. Ovaj algoritam kombinira više stabala odlučivanja u jedan model, pri čemu svako novo stablo pokušava ispraviti pogreške prethodnih stabala. Proces počinje s inicijalnom predikcijom, nakon čega se izračunavaju reziduali, odnosno pogreške između stvarnih i predviđenih vrijednosti, te se svako sljedeće stablo trenira na rezidualima, čime se kontinuirano poboljšava ukupna točnost.

Jedna od ključnih prednosti XGBoosta je njegova učinkovitost i skalabilnost na velikim skupovima podataka. Važni hiperparametri su broj stabala, koji određuje koliko će stabala biti uključeno u model, te brzina učenja koja određuje koliko će doprinos svakog novog stabla biti umanjen prilikom ažuriranja predikcije. Manja brzina učenja omogućava dobivanje boljih rezultata, ali zahtijeva više stabala za postizanje istog učinka.

Stroj potpornih vektora (engl. *Support Vector Machine*, SVM) je algoritam strojnog učenja koji se koristi za klasifikaciju i regresiju. SVM radi na principu pronalaženja optimalne hiperravnine koja razdvaja različite klase podataka u prostoru značajki. Ova hiperravnina se bira tako da maksimizira marginu, odnosno udaljenost između hiperravnine i najbližih uzoraka svake klase. Pronalazak hiperravnine je jednostavan ako su podaci linearno odvojivi. Međutim, u slučajevima kada nisu, koristi se jezgrena funkcija (engl. *kernel*) koja omogućava transformaciju podataka u višu dimenziju. Jezgrena funkcija je jedan od ključnih hiperparametara koji se mogu optimizirati. Popularne jezgrene funkcije uključuju linearnu, polinomijalnu i Gaussovu radijalnu (RBF) jezgru. Drugi bitni hiperparametar bitan za optimizaciju je hiperparametar C, koji kontrolira ravnotežu između maksimizacije margine i minimizacije pogrešaka klasifikacije. Manja vrijednost znači da će model dozvoljavati više pogrešaka i obrnuto, veća vrijednost znači da će model više kažnjavati pogreške i bit će složeniji. Dodatno, hiperparametar γ koji se koristi s RBF jezgrenom funkcijom utječe na složenost oblika granice odluka.

Neuronska mreža (engl. *Neural network*, NN) je složeni model strojnog učenja koji se značajno razlikuje od tradicionalnih statističkih modela opisanih ranije [2]. Dok se tradicionalni modeli oslanjaju na ručno definirane značajke i pravila za klasifikaciju, neuronske mreže imaju sposobnost automatskog učenja i prilagodbe složenih obrazaca iz podataka.

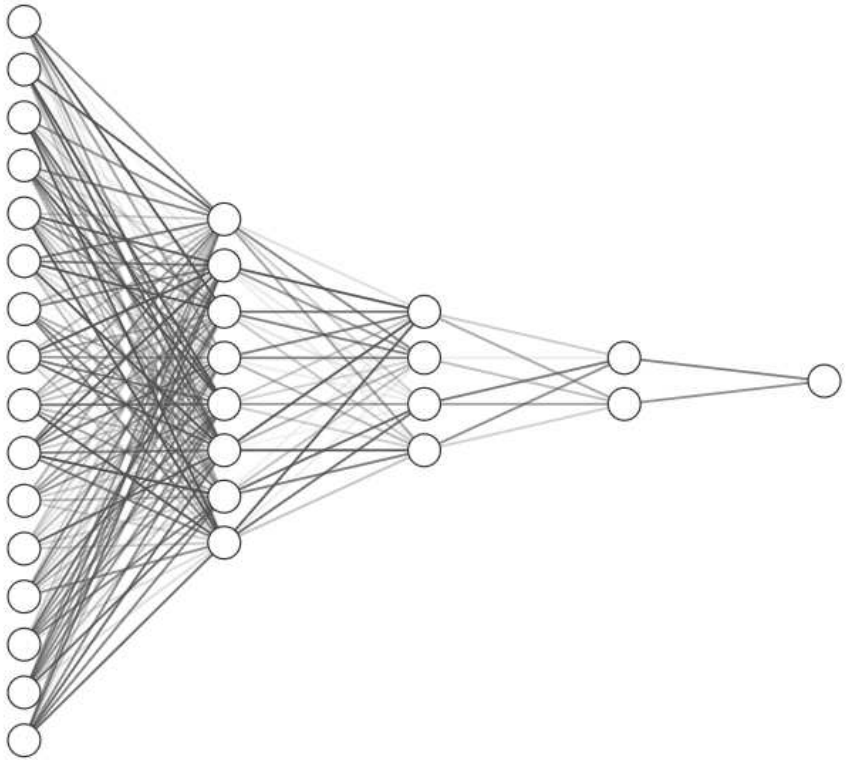
Neuronska mreža sastoji se od tri vrste slojeva: ulaznog sloja, jednog ili više skrivenih slojeva i izlaznog sloja. Ulazni sloj prima podatke, dok skriveni slojevi obrađuju informacije kroz niz transformacija, a izlazni sloj generira konačne predikcije. Svaki sloj se sastoji od međusobno povezanih neurona. Svaki neuron prima ulazne vrijednosti iz prethodnog sloja, množi ih s pridruženim težinama, te rezultat koji prolazi kroz aktivacijsku funkciju prosljeđuje sljedećem sloju. Ovaj proces naziva se propagacija unaprijed (engl. *forward propagation*). Aktivacijske funkcije, poput zglobnica, sigmoida ili hiperboličnog

tangensa, omogućuju modeliranje nelinearnih odnosa.

Cilj neuronske mreže je pronaći optimalne težine i pristranosti koje minimiziraju funkciju gubitka, koja mjeri razliku između stvarnih i predviđenih izlaza. Tijekom treniranja mreže koristi se algoritam povratnog širenja gubitka (engl. *backpropagation*) koji prilagođava težine na temelju gradijenta funkcije gubitka. Ključni hiperparametri koji se mogu optimizirati uključuju broj skrivenih slojeva i broj neurona po svakom sloju.

U ovom radu koristi se jednostavna arhitektura neuronske mreže. Model se sastoji od pet potpuno povezanih slojeva i optimiziran je korištenjem Adam optimizatora. U početnim slojevima se obrađuje veći broj značajki, dok dublji slojevi sažimaju informacije. U skrivenim slojevima koristi se nelinearna aktivacijska funkcija koja omogućuje modeliranje složenijih obrazaca u podacima, dok izlazni sloj koristi sigmoidnu aktivacijsku funkciju za binarnu klasifikaciju.

Kako bi se poboljšala stabilnost treniranja i ubrzala konvergencija, nakon pojedinih slojeva primjenjuje se normalizacija podataka. Također, kako bi se smanjio rizik od prenaučenosti, u mrežu su uključeni slojevi koji nasumično isključuju dio neurona tijekom učenja. Na slici 4.2. prikazana je arhitektura neuronske mreže korištene u ovom radu.



Slika 4.2. Okvirni prikaz arhitekture korištene neuronske mreže u ovom radu. Prikaz je napravljen pomoću [27].

U radu su za optimizaciju hiperparametara korišteni pristupi mrežnog pretraživanja (engl. *Grid Search*) i nasumičnog pretraživanja (engl. *Random Search*). Mrežno pretraživanje je metoda koja pretražuje unaprijed definiranu mrežu kombinacija hiperparametara. Omogućuje sistematsko ispitivanje svih mogućih kombinacija, ali može biti računski skupa ako se radi s velikim skupovima podataka. S druge strane, nasumično pretraživanje nasumično odabire kombinacije hiperparametara unutar zadanih vrijednosti. Može pronaći dobre kombinacije hiperparametara brže od mrežnog pretraživanja. Ovi pristupi za optimizaciju hiperparametara korišteni su za sve modele, čime je osigurano da svaki model bude optimiziran za maksimalno prepoznavanje lažnih recenzija.

4.2. Balansiranje podatka

Neuravnoteženi podaci predstavljaju jedan od glavnih izazova kod klasifikacije lažnih recenzija, jer značajno mogu utjecati na učinkovitost klasifikatora. Modeli imaju tendenciju favoriziranja većinske klase, što može rezultirati pristranim predviđanjima. Ako se u fazi treniranja modela ne uzme u obzir neuravnoteženost podataka, postoji rizik da klasifikator nauči prepoznavati isključivo obrasce većinske klase, dok zanemaruje manjinsku klasu, odnosno lažne recenzije. U ovom slučaju, klasifikator bi učio obrasce pravih recenzija, dok bi zanemarivao lažne recenzije [15].

Za rješavanje ovog problema koriste se tehnike balansiranja podataka, preuzorkovanja (engl. *oversampling* i poduzorkovanja (engl. *undersampling*). Preuzorkovanje uključuje povećanje broja uzoraka manjinske klase, dok poduzorkovanje podrazumijeva smanjenje broja uzoraka većinske klase. Cilj obje tehnike je osigurati da klasifikator jednako dobro nauči prepoznavati obje klase te smanjiti pristranost modela prema većinskoj klasi.

Jedan od najpoznatijih pristupa preuzorkovanja je algoritam SMOTE (engl. *Synthetic Minority Oversampling technique*) [28]. Ovaj algoritam funkcionira tako da nasumično odabire instancu manjinske klase i pronalazi njezinih k najbližih susjeda. Zatim se sintetički uzorak generira interpolacijom između odabrane instance i jednog od njezinih najbližih susjeda. Ovaj postupak se ponavlja dok se ne postigne željeni broj sintetičkih instanci manjinske klase [28]. Primjena SMOTE algoritma pomaže u poboljšanju ravnoteže između klasa i povećava učinkovitost klasifikacije. Ova metoda ima prednost nad nasumičnim poduzorkovanjem jer smanjuje rizik gubitka informacija većinske klase.

4.3. Evaluacija modela

Za evaluaciju svakog modela korištene su sljedeće metrike: točnost, preciznost, odziv i F1 rezultat. Izračun ovih mjera temelji se na sljedećim kategorijama:

- Stvarno negativni (engl. *True Negative*, TN) – recenzije koje su ispravno klasificirane kao prave.
- Lažno pozitivni (engl. *False Positive Negative*, FP) – recenzije koje su pogrešno klasificirane kao lažne, iako su zapravo prave.
- Lažno negativni (engl. *False Negative*, FN) – recenzije koje su pogrešno klasificirane kao prave, iako su zapravo lažne.
- Stvarno pozitivni (engl. *True Positive*, TP) – recenzije koje su ispravno klasificirane kao lažne.

Ove vrijednosti čine **matricu zabune**, koja daje uvid u izvedbu klasifikacijskog modela. Matrica sadrži dva retka i dva stupca, pri čemu redci predstavljaju stvarne klase, a stupci predviđene klase. Važno je napomenuti da u slučaju klasifikacije lažnih recenzija, lažne recenzije su zapravo pozitivni primjeri, dok su prave recenzije negativni primjeri. **Točnost** (engl. *Accuracy*) predstavlja omjer točno predviđenih recenzija (lažnih i pravih) u odnosu na ukupan broj recenzija. **Preciznost** (engl. *Precision*) mjeri omjer točno predviđenih pozitivnih slučajeva u odnosu na ukupan broj predviđenih pozitivnih slučajeva, odnosno koliko su predviđanja modela točna kada označi recenziju kao pravu. **Odziv** (engl. *Recall*) pokazuje koliko je stvarno pozitivnih recenzija model ispravno identificirao, odnosno koliki je udio stvarnih pravih recenzija koje su točno klasificirane.
F1 mjera (engl. *F1 score*) je harmonijska sredina preciznosti i odziva te osigurava balans između njih.
Osim navedenih metrika, za evaluaciju modela korištena je i površina ispod **ROC krivulje** (engl. *Receiver Operating Characteristics*, ROC), koja prikazuje odnos između stvarno pozitivnih i lažno pozitivnih primjera. Vrijednost površine ispod krivulje (engl. *Area under the curve*, AUC) je vrijednost između 0 i 1, gdje veća vrijednost označava bolju izvedbu modela. Ako je AUC-ROC jednak 0.5, to znači da model klasificira slučajno, dok vrijednost 1 označava savršenu klasifikaciju.

U slučaju neuravnoteženih podataka, točnost nije prikladna mjera jer može doseći visoke vrijednosti ako klasifikator sve primjere svrstava u većinsku klasu. Budući da je korišteni skup podataka neuravnotežen, fokus evaluacija stavljen je na F1 mjeru, odziv i preciznost. Te metrike, uz ROC-AUC vrijednost, daju sveobuhvatan uvid u sposobnost modela da prepozna pozitivnu klasu, odnosno lažne recenzije, što je i primarni cilj.

5. Analiza rezultata i rasprava

Metodama iz biblioteke Pandas učitani su prethodno opisani podaci o recenzijama. Nakon učitavanja, provedeni su standardni koraci predobrade kako bi se osigurala kvaliteta podataka i pripremili za učenje modela strojnog učenja.
Prvo je izvršena obrada teksta, koja uključuje sljedeće korake:

1. Čišćenje teksta - U ovom koraku su iz teksta recenzija uklonjeni svi simboli, brojevi i ostali šumovi

2. Tokenizacija - U ovom koraci je tekst recenzije podijeljen na pojedinačne riječi ili izraze koji se nazivaju tokeni

3. Uklanjanje zaustavnih riječi - U ovom koraku su uklonjene riječi koje dodaju malo ili nimalo vrijednosti značenju teksta. Za to je korišten zaustavnih riječi na engleskom jeziku biblioteke nltk [25].

4. Označavanje dijela govora - U ovom koraku se dodjeljuju gramatičke oznake svakom tokenu u tekstu na temelju njegove uloge u rečenici

5. Određivanje korijena riječi - U ovom koraku koristi se lematizacija (engl. *lemmatization*) koja ima cilj reducirati riječ na njen osnovni oblik, uzimajući u obzir kontekst i značenje riječi. Korišten je nltk WordNet Lemamtizer uz gramatičke oznake.

Nadalje, nakon što je tekstualni sadržaj pročišćen, provedena je vektorizacija teksta koja pretvara prirodni jezik u numerički format. Za vektorizaciju teksta korištena je frekvencija termina - inverzna frekvencija termina (engl. *Term Frequency - Inverse Document frequency*, TF-IDF) koja utvrđuje koliko je neka riječ važna za neki dokument tj. u ovom slučaju u recenzijama. Frekvencija termina (TF) predstavlja težinu određene riječi u danom dokumentu, tj. računa učestalost riječi u dokumentu u odnosu na ukupni broj

riječi u dokumentu. Inverzna frekvencija termina (IDF) mjeri logaritam omjera ukupnog broja dokumenata prema broju dokumenata koji uključuju riječ. Množenjem ovih dviju vrijednosti dobiva se konačna TF-IDF vrijednost.
Za izračun TF-IDF matrice korišten je TfidfVectorizer iz sklearn biblioteke [29]. Broj najčešćih riječi koje će se koristiti za vektorizaciju ograničen je na 5000. Također se ignoriraju sve riječi koje se pojavljuju u manje od 10 recenzija, te se ignoriraju sve riječi koje se pojavljuju u više od 80% recenzija. Kako bi se uhvatile neke fraze ili kombinacije riječi, koriste se N-grami točnije, unigrami, bigrami i trigrami.
Također, korištene su POS oznake koje su spremljene kao rječnici s učestalošću svake kategorije te su transformirane u vektorski format pomoću DictVectorizer kako bi se mogle koristiti kao značajka u modelima strojnog učenja. Ove dvije značajke korištene su zajedno kao jezične značajke dobivene iz sadržaja recenzije.
Osim obrade teksta, provedena je i predobrada numeričkih podataka. Svi numerički podaci normalizirani su u rasponu [0,1] kako bi se osigurala konzistentnost i olakšalo treniranje modela. Numeričke značajke koje su normalizirane uključuju: broj korisnih glasova koje je korisnik dobio, ukupan broj recenzija korisnika, broj prijatelja na platformi, duljina članstva na platformi, duljina recenzije te ocjena sentimenta recenzije.

Skup podataka podijeljen je u 2 podskupa, skup za treniranje s 80% podataka, te testni skup s 20% podataka te je sačuvana izvorna distribucija klasa u oba skupa. U testnom skupu podataka je 1736 lažnih recenzija i 9386 pravih recenzija. Za rješavanje problema neuravnoteženosti skupa podataka korišten je algoritam SMOTE, koji je prethodno opisan. Implementacija algoritma učitana je pomoću biblioteke ImbalancedLearn [30]. SMOTE je korišten samo na podacima za treniranje, pošto su se modeli testirali samo na originalnim podacima kako bi dobili nepristrane rezultate. Korištenje SMOTE algoritma na cijelom skupu podataka može dovesti do curenja podataka te samim time testni skup više nije reprezentativan za distribuciju podataka u stvarnom svijetu, što može dovesti do preoptimističnih rezultata.

Za klasifikaciju korišteni su prethodno opisani modeli strojnog učenja, logistička regresija, stablo odluke, slučajna šuma, XGBoost, SVM i neuronska mreža. Svaki model je treniran i evaluiran u tri scenarija:

1. Samo na numeričkim značajkama
2. Samo na jezičnim značajkama (TF-IDF i POS oznake)
3. Na kombinaciji svih značajki (numeričkih i jezičnih)

Cilj ovakvog pristupa je utvrditi doprinos različitih skupina značajki. Svaki model je dodatno optimiziran podešavanjem hiperparametara različitih arhitektura.
U nastavku je prikazana detaljna analiza rezultata za svaki model i svaki scenarij, s fokusom na preciznost, odziv i F1 mjeru.

5.1. Rezultati s numeričkim značajkama

Tablica 5.1. prikazuje rezultate evaluacije različitih modela strojnog učenja koji su trenirani isključivo na normaliziranim numeričkim značajkama. Evaluacija je provedena pomoću standardnih metrika: točnost, preciznost, odziv, F1 mjera i AUC-ROC vrijednost.

Modeli	**Točnost**	**Preciznost**	**Odziv**	**F1 mjera**	**AUC-ROC**
Logistička regresija	0.7437	0.3686	0.9009	0.5232	0.8964
Stablo odluke	0.8731	0.5760	0.7091	0.6357	0.8546
Slučajna šuma	0.8984	0.6591	0.7229	0.6896	0.9314
XGBoost	0.8996	0.6845	0.6601	0.6723	0.9299
SVM	0.7527	0.3765	0.8462	0.5211	0.8859
Neuronska mreža	0.7816	0.4086	0.8917	0.5604	0.9095

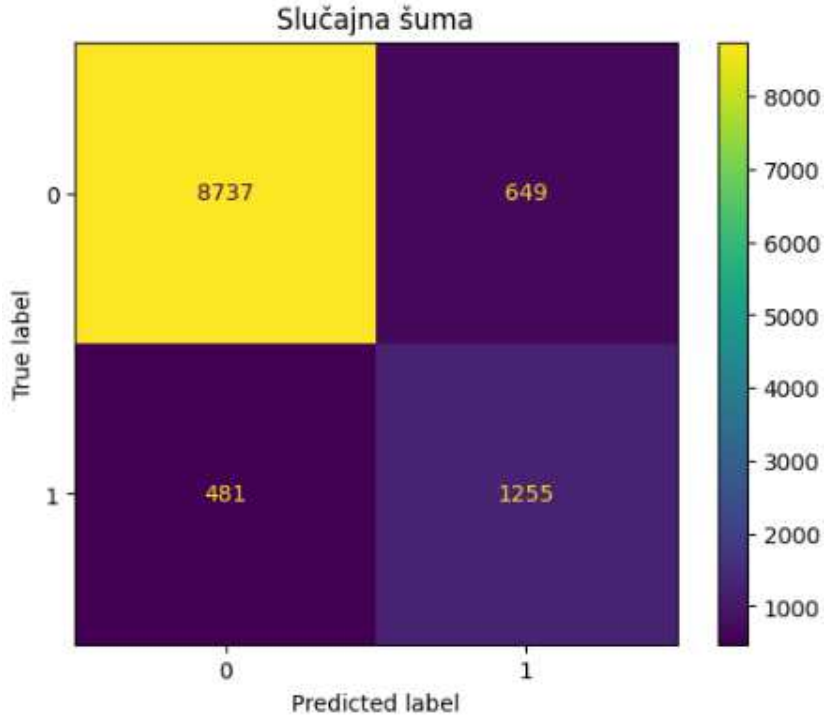
Tablica 5.1. Rezultati evaluacije modela s numeričkim značajkama

S obzirom na to da je skup podataka neuravnotežen, najvažnije metrike za interpretaciju rezultata su odziv, preciznost i F1 mjera, koje daju bolji uvid u performanse modela na manjinskoj klasi.
Modeli temeljeni na stablima (slučajna šuma i XGBoost) općenito pokazuju bolje perfor-

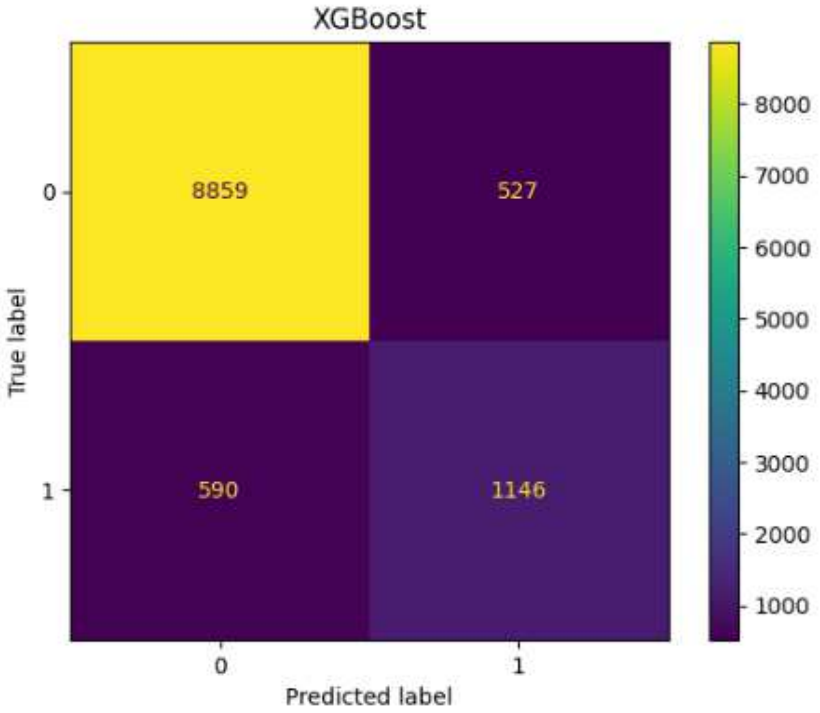
manse u smislu omjera preciznosti i odziva, što se vidi u visokim F1 mjerama, u usporedbi s ostalim modelima.

Logistička regresija i SVM pokazuju nisku preciznost, iako imaju vrlo visok odziv. To ukazuje da ovi modeli identificiraju većinu lažnih recenzija, ali također, jako puno pravih recenzija klasificiraju kao lažne. Neuronska mreža ima sličan problem, iako u manjoj mjeri. Ima veću preciznost, ali i dalje klasificira značajan broj pravih recenzija kao lažne. Stablo odluke pokazuje dobre performanse, no slučajna šuma i XGBoost pružaju još bolje rezultate.

Slučajna šuma i XGBoost postižu slične rezultate, pri čemu XGBoost ima nešto veću točnost i preciznost, dok slučajna šuma ima nešto veću F1 mjeru i odziv. Na slikama 5.1. i 5.2. prikazane su njihove matrice konfuzije koje prikazuju broj ispravno i pogrešno klasificiranih primjera za svaku klasu. Analiza matrica konfuzije pokazuje da model slučajne šume bolje identificira lažne recenzije, iako ima nešto veći broj pogrešno klasificiranih pravih recenzija. Budući da je cilj identificirati što više lažnih recenzija, makar to značilo da neke prave recenzije budu označene kao lažne, model slučajne šume pokazuje bolje rezultate od XGBoost-a.

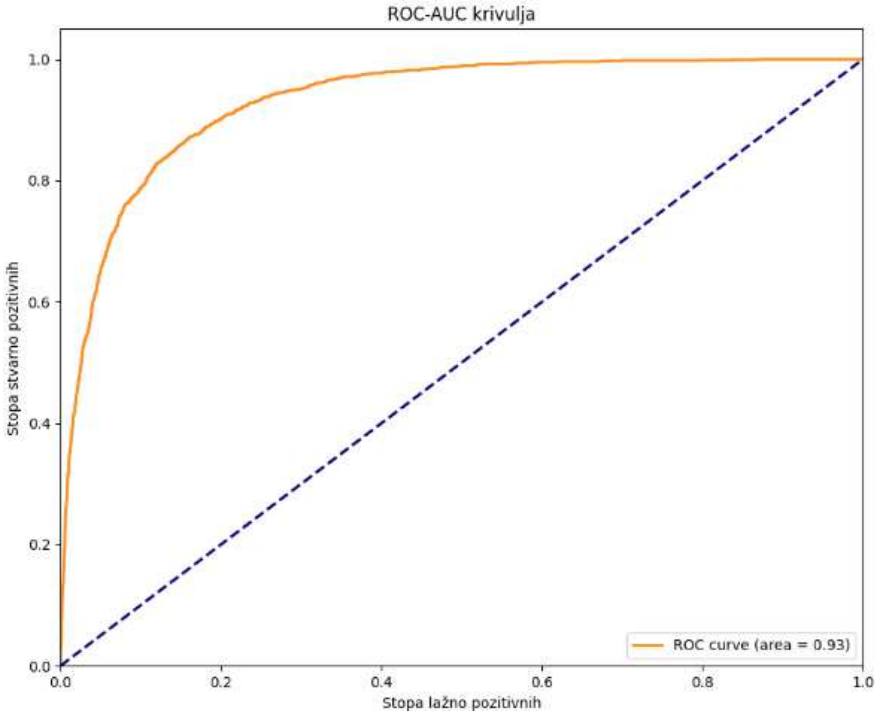


Slika 5.1. Prikaz matrice konfuzije za model slučajne šume



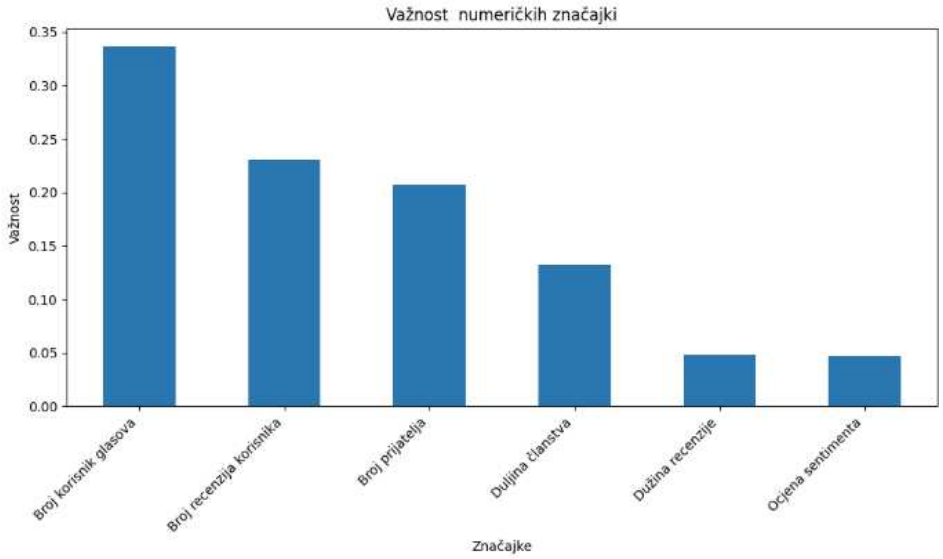
Slika 5.2. Prikaz matrice konfuzije za model XGBoost

Na slici 5.3. prikazana je ROC-AUC krivulja modela slučajne šume s vrijednošću 0.9314. ROC-AUC krivulja prikazuje stvarni pozitivni omjer u odnosu na lažni pozitivni omjer te daje mjeru sposobnosti modela da razlikuje prave i lažne recenzije.



Slika 5.3. Prikaz ROC-AUC krivulje za model slučajne šume

Također je provedena analiza relative važnosti svake od numeričkih značajki, koja je prikazana na slici 5.4. U klasifikaciji lažnih recenzija je najvažnija značajka broja korisnih glasova na recenzijama korisnika. Ova značajka je pokazala i najveće razlike između lažnih i pravih recenzija u prethodnoj analizi značajki, što dodatno potvrđuje njenu važnost.



Slika 5.4. Prikaz analize važnosti numeričkih značajki

5.2. Rezultati s jezičnim značajkama

Tablica 5.2. prikazuje rezultate evaluacije modela strojnog učenja koji su trenirani koristeći isključivo jezične značajke.

Modeli	**Točnost**	**Preciznost**	**Odziv**	**F1 mjera**	**AUC-ROC**
Logistička regresija	0.6888	0.2518	0.5040	0.3358	0.6616
Stablo odluke	0.7913	0.2095	0.1215	0.1538	0.5359
Slučajna šuma	0.8255	0.3186	0.1037	0.1564	0.6522
XGBoost	0.8417	0.4570	0.0766	0.1313	0.5921
SVM	0.6777	0.2621	0.5858	0.3622	0.7031
Neuronska mreža	0.6582	0.2307	0.5098	0.3177	0.6312

Tablica 5.2. Rezultati evaluacije modela s jezičnim značajkama

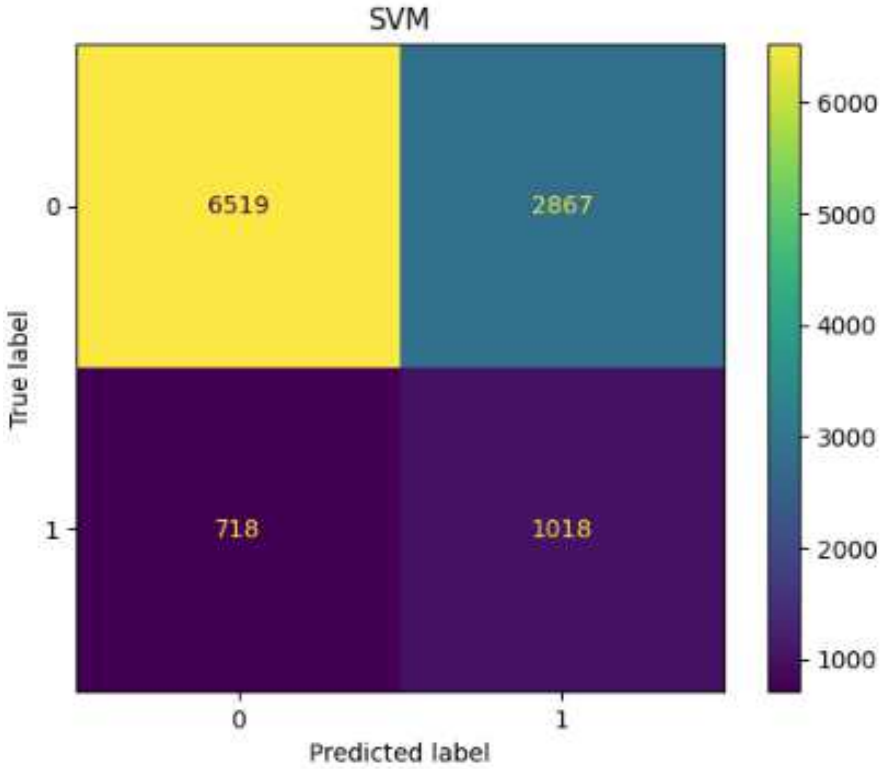
Općenito, performanse modela treniranih samo s jezičnim značajkama su značajno lošije u usporedbi s onima dobivenim korištenjem numeričkih značajki. AUC-ROC vrijednosti su niske za sve modele, što ukazuje na ograničenu sposobnost modela da razlikuju prave i lažne recenzije. Ovo pokazuje da je klasifikacija lažnih recenzija složenija kada se oslanja samo na tekst recenzije.

Modeli temeljeni na stablima (Stablo odluke, slučajna šuma i XGBoost) pokazuju izrazito loše rezultate. Imaju visoku točnost, ali njihova preciznost i odziv su niski. To znači da ovi modeli rijetko točno identificiraju lažne recenzije, a istovremeno klasificiraju prave recenzije kao lažne. Specifično, XGBoost ima najveću točnost, ali i najmanji odziv, što znači da klasificira sve recenzije kao prave.

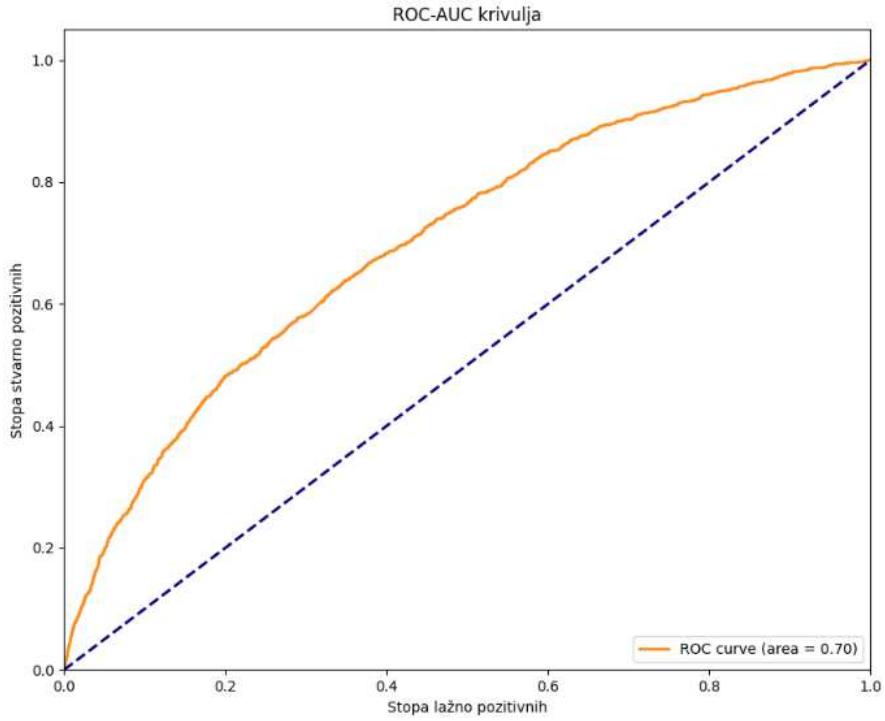
Neuronska mreža također postiže loše rezultate, s niskom točnošću, preciznošću i F1 mjerom. Jedan od razloga može biti korištenje TF-IDF vektorizacije, koja ne uzima u obzir redoslijed riječi u rečenici, semantičke odnose i kontekst. Složenije tehnike vektorizacije, poput ugrađivanja riječi (engl. *word embeddings*) ili transformera, vjerojatno bi dovele do boljih rezultata.

Logistička regresija, SVM imaju bolji odziv u usporedbi s ostalim modelima. Oni ipak uspijevaju identificirati dio lažnih recenzija. SVM se ističe s najvećim odzivom, što znači da najbolje identificira lažne recenzije. Na slici 5.5. prikazana je matrica konfuzije koja prikazuje koliko je primjera SVM ispravno i pogrešno klasificirano za svaku klasu. Iz matrice je vidljivo da SVM točno klasificira 1018 lažnih recenzija, ali pogrešno klasificira 2867 pravih recenzija. Također, na slici 5.6. je prikazana ROC-AUC krivulja za SVM s vrijednošću 0.7030, koja pokazuje umjerenu sposobnost modela da razlikuje prave i lažne recenzije.

Ovi rezultati pokazuju da jezične značajke same po sebi nisu dovoljne za pouzdano klasificiranje lažnih recenzija. Svi modeli pokazuju lošije rezultate što ukazuje na to da je potrebno koristiti dodatne značajke kako bi se poboljšala točnost klasifikacije.



Slika 5.5. Prikaz matrice konfuzije za model SVM



Slika 5.6. Prikaz ROC-AUC krivulje za model SVM

5.3. Rezultati sa svim značajkama

Konačno, tablica 5.3. prikazuje performanse modela strojnog učenja treniranih s kombinacijom numeričkih i jezičnih značajki.

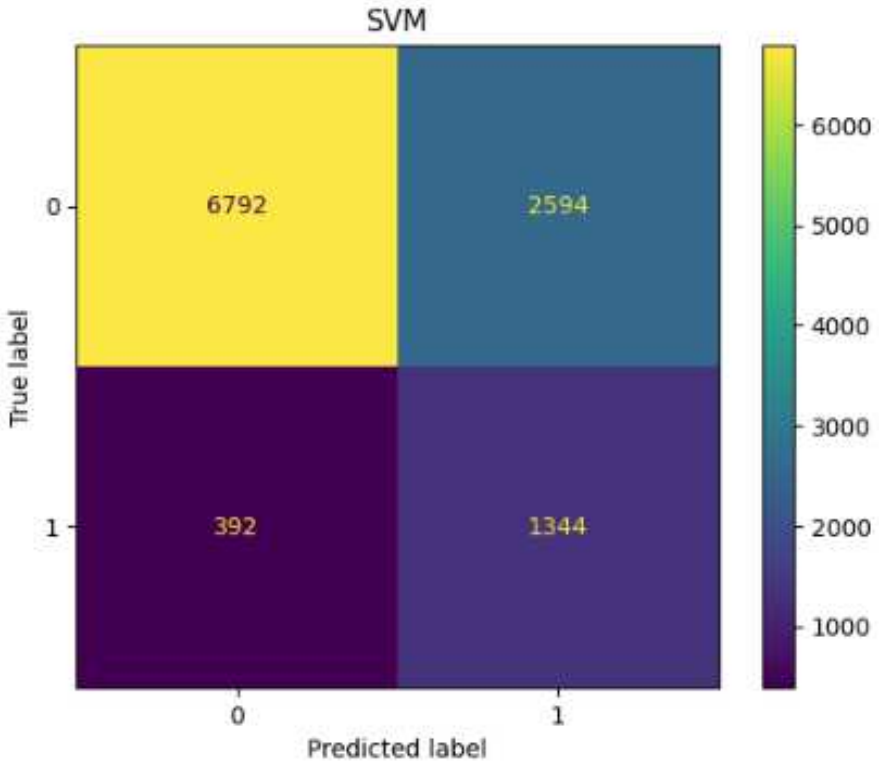
Modeli	**Točnost**	**Preciznost**	**Odziv**	**F1 mjera**	**AUC-ROC**
Logistička regresija	0.7839	0.3849	0.6441	0.4816	0.8149
Stablo odluke	0.8518	0.5189	0.6935	0.5937	0.8173
Slučajna šuma	0.8699	0.5650	0.7235	0.6345	0.9099
XGBoost	0.8979	0.7011	0.6025	0.6481	0.9229
SVM	0.7315	0.3413	0.7742	0.4738	0.8220
Neuronska mreža	0.6582	0.2307	0.5098	0.3177	0.6312

Tablica 5.3. Rezultati evaluacije modela

Treniranje modela s kombinacijom značajki nije rezultiralo značajnim poboljšanjem performansi modela u odnosu na one koji su trenirani isključivo na numeričkim značajkama.

Neuronska mreža se pokazala najmanje učinkovitom, s niskom preciznošću i odzivom. Logistička regresija i SVM pokazuju nešto bolje rezultate, no i dalje imaju nisku preciznost, što znači da generiraju veliki broj lažno pozitivnih klasifikacija. Kombinacija numeričkih i jezičnih značajki poboljšava njihove performanse u odnosu na samo jezične značajke.

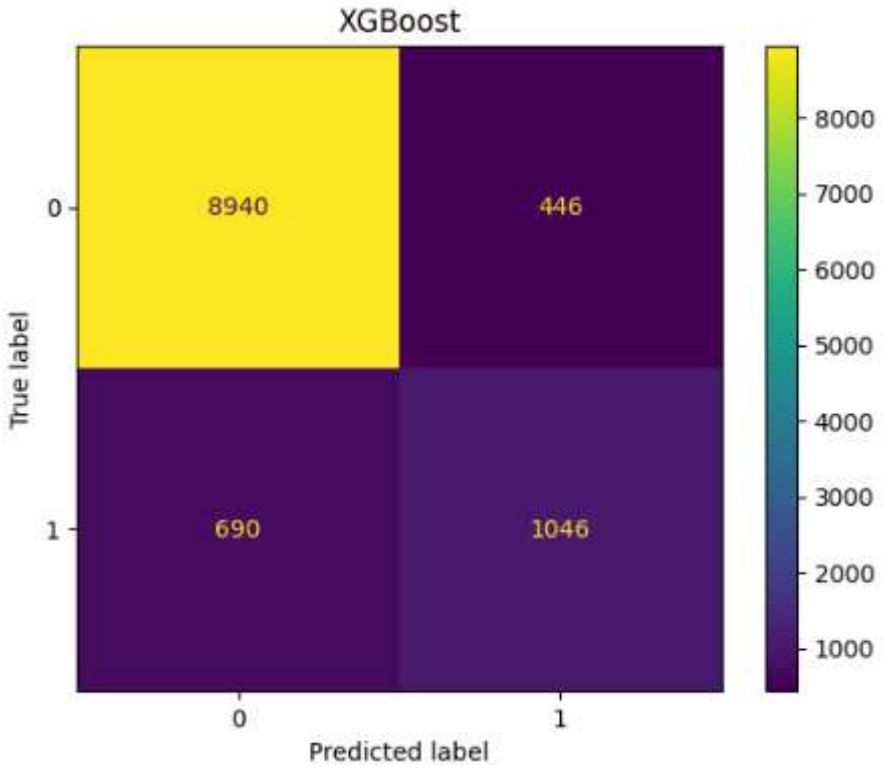
SVM model, premda ima nisku točnost i preciznost, postiže najveći odziv (77%) u usporedbi s ostalim modelima. SVM najbolje identificira lažne recenzije, iako uz cijenu puno pogrešnih klasifikacija. Matrica konfuzije za SVM model prikazana je na slici 5.7. Ona pokazuje da SVM ispravno klasificira 1344 od 1736 lažnih recenzija, ali također pogrešno klasificira 2594 prave recenzije kao lažne. Stoga, SVM model bi bio prikladan kada je puno važnije identificirati što više lažnih recenzija, čak i uz cijenu većeg broja pogrešnih klasifikacija.



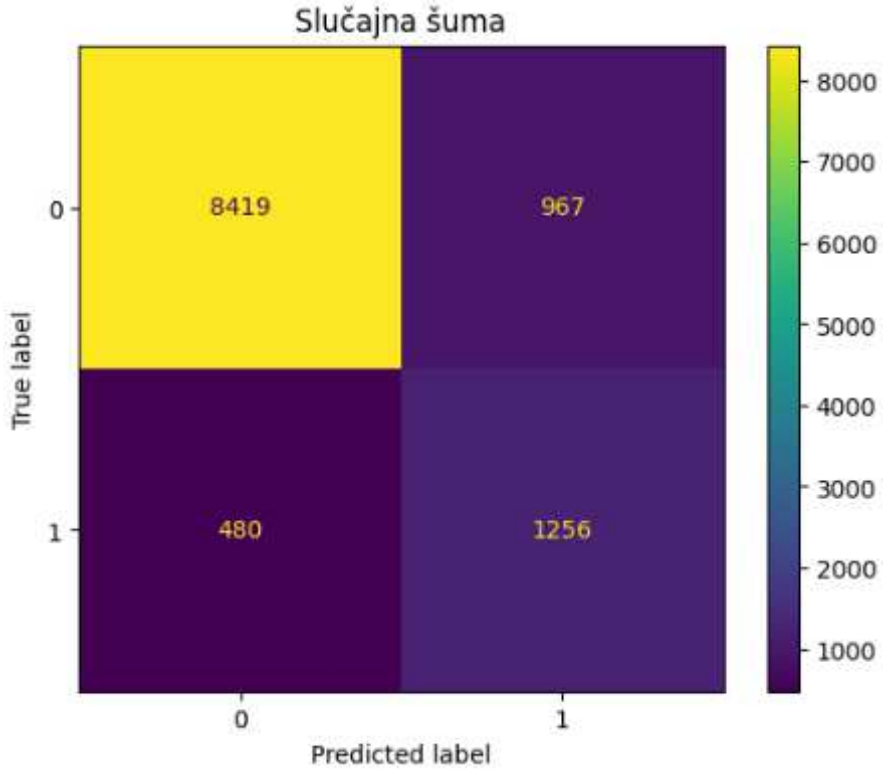
Slika 5.7. Prikaz matrice konfuzije za model SVM

Modeli temeljeni na stablima, konkretno slučajna šuma i XGBoost, pokazuju najbolje rezultate. Oba modela imaju visok odziv i F1 mjeru. XGBoost ima značajno veću preciznost, dok slučajna šuma ima bolji odziv. Njihove matrice konfuzije prikazane su na slikama 5.8. i 5.9. Na njima je vidljivo da XGBoost točno klasificira 1046 lažnih recenzija, dok slučajna šuma točno klasificira 1256 lažnih recenzija. S druge strane, XGBoost krivo klasificira samo 446 pravih recenzija, dok slučajna šuma krivo klasificira 967 pravih recenzija.

Iako XGBoost ima veću preciznost, što je poželjno ako je cilj minimizirati broj pravih recenzija koje se pogrešno označe kao lažne, slučajna šuma ima veći odziv i identificira veći broj lažnih recenzija. Kod detekcije lažnih recenzija važnije je identificirati što više lažnih recenzija čak i uz rizik pogrešne klasifikacije nekih pravih recenzija pa je slučajna šuma prikladniji izbor.

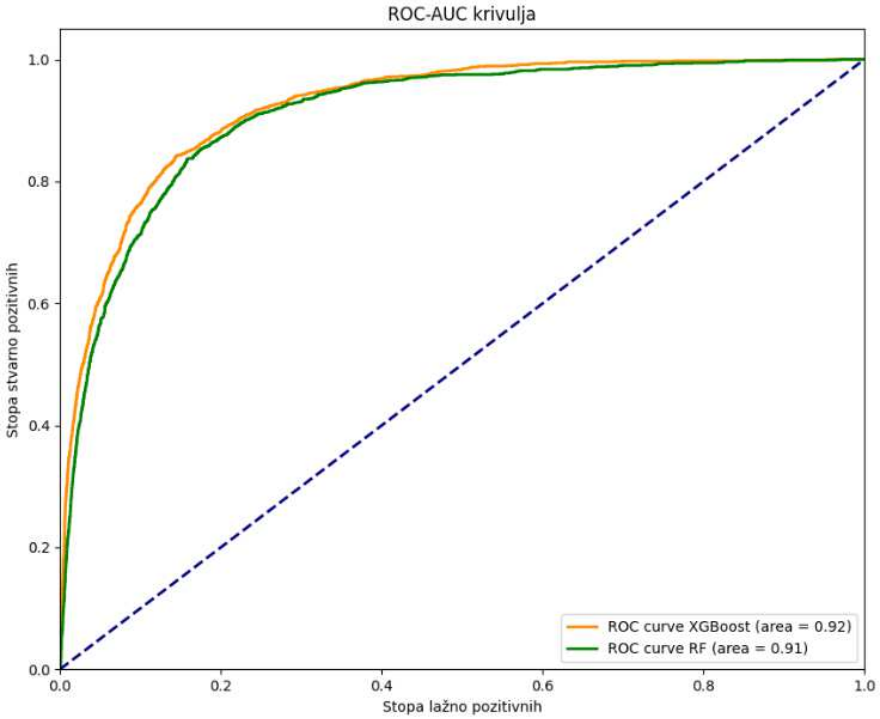


Slika 5.8. Prikaz matrice konfuzije za model XGBoost



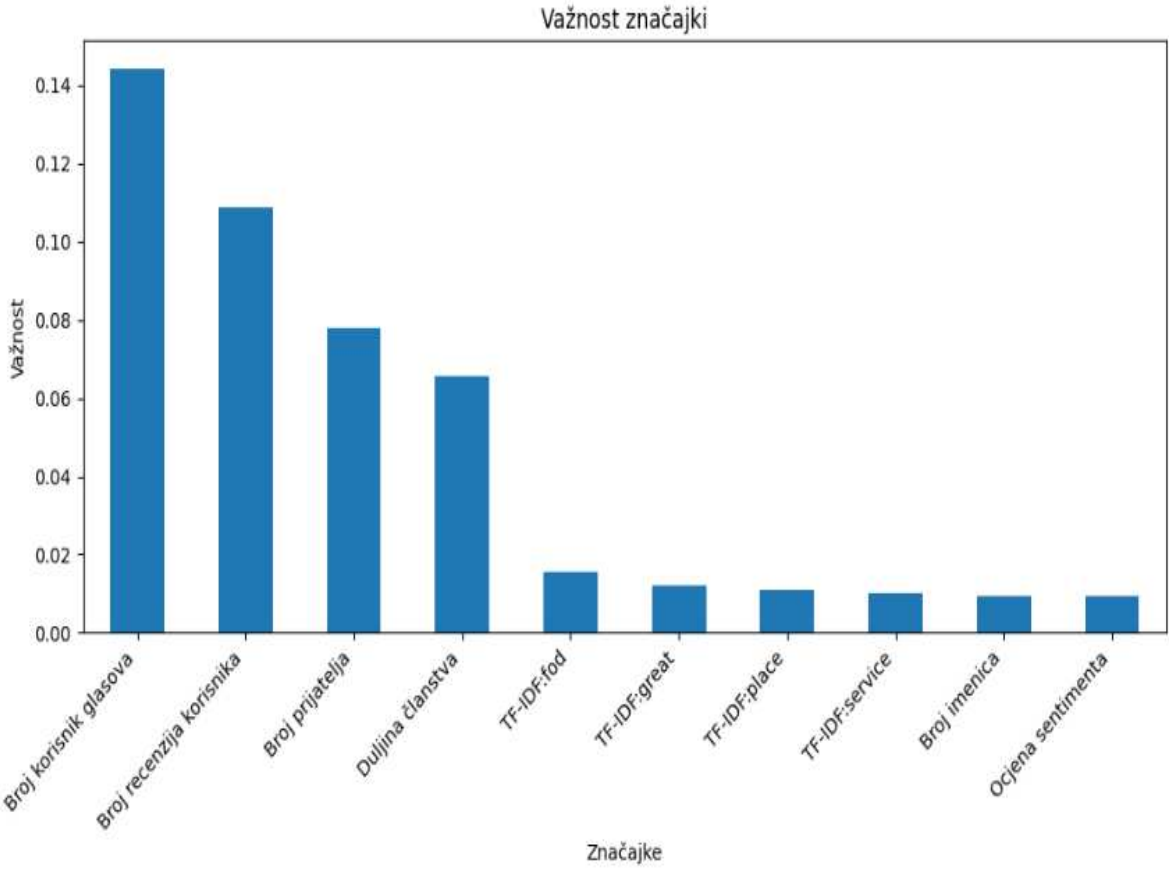
Slika 5.9. Prikaz matrice konfuzije za model slučajne šume

Na slici 5.10. prikazane su ROC-AUC krivulje za modele XGBoost i slučajne šume. Vidljivo je da oba modela imaju visoke ROC-AUC vrijednosti, što potvrđuje njihovu sposobnost da dobro razlikuju prave i lažne recenzije.



Slika 5.10. Prikaz ROC-AUC krivulje za modele XGBoost i slučajne šume

Također je provedena analiza relative važnosti svake od značajki, koja je prikazana na slici 5.11. Iz analize je vidljivo da su najvažnije značajke one numeričke koje opisuju recenzenta, kao što je broj korisnik glasova i broj recenzija. Jezične značajke imaju manju važnost, ali su isto prisutne u prvih 10 najvažnijih značajki što znači da ih ne možemo zanemariti.



Slika 5.11. Prikaz analize važnosti značajki

Analiza rezultata pokazuje da treniranje modela s kombinacijom jezičnih i numeričkih značajki nije dovelo do značajnog poboljšanja performansi u odnosu na one koji su trenirani isključivo na numeričkim značajkama. Štoviše, kombinacija numeričkih i jezičnih značajki nije dovela do poboljšanja performansi svih modela. Pojedini modeli su ostvarili bolje rezultate koristeći samo numeričke značajke, dok su drugi postigli bolje rezultate s kombinacijom. Ovo naglašava važnost odabira značajki i modela za preciznu detekciju lažnih recenzija. Nadalje, bolji rezultati numeričkih značajki pokazuju da jezične značajke u ovom skupu podataka nisu imale dovoljan utjecaj za poboljšanje točnosti klasifikacije lažnih recenzija.

U ovom radu, numeričke značajke uglavnom opisuju karakteristike recenzenata i njihovo ponašanje na platformi, dok su se jezične značajke fokusirale isključivo na analizu teksta same recenzije. Analiza najvažnijih varijabli za klasifikaciju u najboljem modelu koji koristi sve značajke pokazuje da su značajke usmjerene na recenzente učinkovitije od jezičnih značajki, te da bi se veća pažnja trebala posvetiti analizi ponašanja korisnika na platformi. Ipak, ova analiza također pokazuje da se nijedan skup varijabli ne može

zanemariti, budući da su i numeričke i jezične varijable zastupljene među 10 najvažnijih. Stoga, se ne treba odbaciti potencijal jezičnih značajki, već bi se trebale koristiti uz naprednije tehnike obrade teksta.

6. Zaključak

Mrežne recenzije igraju ključnu ulogu u donošenju odluka potrošača i poslovnom uspjehu. Međutim, manipulacije recenzijama u marketinške svrhe predstavljaju problem, jer lažne recenzije mogu navesti potrošače na pogrešne odluke, narušavajući povjerenje u platforme za recenzije. Otkrivanje lažnih recenzija zahtijeva detaljnu analizu sadržaja recenzija i ponašanja lažnih recenzenata. U ovom radu implementirani su i uspoređeni različiti modeli binarne klasifikacije s ciljem prepoznavanja karakteristika lažnih recenzija. Trenirani su modeli logističke regresije, modeli temeljeni na stablu, SVM i neuronske mreže. Korišten je skup podataka recenzija s platforme Yelp, te su analizirane različite značajke, uključujući numeričke značajke vezane uz recenzente te jezične značajke dobivene iz samog teksta recenzije. Svaki model je treniran na tri načina: isključivo na numeričkim značajkama, isključivo na tekstualnim značajkama, te na kombinaciji obje vrste značajki. Kombinacija značajki nije značajno poboljšala performanse modela u odnosu na korištenje samo numeričkih značajki. Najbolje rezultate postigli su modeli temeljeni na stablima, specifično slučajna šuma s F1 mjerom od 91%. Također, model XGBoost pokazuje dobre performanse, iako uz nešto niži odziv. Rezultati pokazuju da jezične značajke nisu same po sebi dovoljne za pouzdanu klasifikaciju, što ukazuje na potrebu za kombiniranjem različitih značajki za poboljšanje točnosti klasifikacije.

Unatoč dobrim rezultatima modela slučajne šume, važno je napomenuti da je ovaj model treniran samo na recenzijama restorana s jedne platforme, što može ograničiti njegovu primjenu na šire područje. Modeli bi se trebali dodatno testirati na većem skupu podataka kako bi se procijenila njegova učinkovitost u različitim domenama. Također, jedno od potencijalnih poboljšanja je izdvajanje više značajki iz teksta recenzija, kao što su stil pisanja i semantički odnosi riječi. Takve značajke mogle bi dodatno poboljšati sposobnost modela da prepozna obrasce u lažnim recenzijama. Dodatno, korištenje ši-

reg skupa značajki o ponašanju recenzenta, kao što su vremenski obrasci objavljivanja recenzija te analiza mreža recenzenata mogu pomoći u otkrivanju lažnih recenzija.

Literatura

[1] M. Crawford, T. M. Khoshgoftaar, J. D. Prusa, A. N. Richter, i H. Al Najada, "Survey of review spam detection using machine learning techniques", sv. 2, br. 1, str. 23. https://doi.org/10.1186/s40537-015-0029-9

[2] R. Yafeng. Learning to detect deceptive opinion spam: A survey | IEEE journals & magazine | IEEE xplore. [Mrežno]. Adresa: https://ieeexplore.ieee.org/abstract/document/8678638

[3] "DETECTION OF FAKE REVIEWS ON SOCIAL MEDIA USING MACHINE LEARNING ALGORITHMS". https://doi.org/10.48009/1_iis_2020_185-194

[4] Yelp. [Mrežno]. Adresa: https://www.yelp.ie/

[5] A. Mukherjee, V. Venkataraman, B. Liu, i N. Glance, "Fake review detection: Classification and analysis of real and pseudo reviews".

[6] N. S. Chowdhary i A. A. Pandit, "Fake review detection using classification", sv. 180.

[7] Fake review detection techniques, issues, and future research directions: a literature review | knowledge and information systems. [Mrežno]. Adresa: https://link.springer.com/article/10.1007/s10115-024-02118-2

[8] A. Palmer. People are using a.i. chatbots to write amazon reviews. Section: Technology. [Mrežno]. Adresa: https://www.cnbc.com/2023/04/25/amazon-reviews-are-being-written-by-ai-chatbots.html

[9] A. Heydari, M. a. Tavakoli, N. Salim, i Z. Heydari, "Detection of review spam: A survey", sv. 42, br. 7, str. 3634–3642. https://doi.org/10.1016/j.eswa.2014.12.029

[10] The effect of fake reviews on e-commerce during and after covid-19 pandemic: SKL-based fake reviews detection | IEEE journals & magazine | IEEE xplore. [Mrežno]. Adresa: https://ieeexplore.ieee.org/abstract/document/9716933

[11] R. Agarwal i D. K. Sharma, "Detecting fake reviews using machine learning techniques: a survey", u *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, str. 1750–1756. https://doi.org/10.1109/ICACITE53722.2022.9823633

[12] N. Jindal i B. Liu, "Opinion spam and analysis", u *Proceedings of the 2008 International Conference on Web Search and Data Mining*, ser. WSDM '08. Association for Computing Machinery, str. 219–230. https://doi.org/10.1145/1341531.1341560

[13] M. Ott, Y. Choi, C. Cardie, i J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination". https://doi.org/10.48550/arXiv.1107.4557

[14] A. Abbasi, Z. Zhang, D. Zimbra, H. Chen, i J. F. Nunamaker, "Detecting fake websites: The contribution of statistical learning theory", sv. 34, br. 3, str. 435–461, publisher: Management Information Systems Research Center, University of Minnesota. https://doi.org/10.2307/25750686

[15] J. Fontanarava, G. Pasi, i M. Viviani, "Feature analysis for fake review detection through supervised classification", u *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, str. 658–666. https://doi.org/10.1109/DSAA.2017.51

[16] N. A. Patel i R. Patel, "A survey on fake review detection using machine learning techniques", u *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, str. 1–6, ISSN: 2642-7354. https://doi.org/10.1109/CCAA.2018.8777594

[17] S. He, B. Hollenbeck, G. Overgoor, D. Proserpio, i A. Tosyali, "Detecting fake-review buyers using network structure: Direct evidence from amazon", sv. 119, br. 47, str. e2211932119. https://doi.org/10.1073/pnas.2211932119

[18] H. Sun, A. Morales, i X. Yan, “Synthetic review spamming and defense”.

[19] U. Aslam, M. Jayabalan, H. Ilyas, i A. Suhail, “A survey on opinion spam detection methods”, sv. 8, br. 9.

[20] J. K. Rout, S. Singh, S. K. Jena, i S. Bakshi, “Deceptive review detection using labeled and unlabeled data”, sv. 76, br. 3, str. 3187–3211. https://doi.org/10.1007/s11042-016-3819-y

[21] R. Kumar, S. Mukherjee, i N. P. Rana, “Exploring latent characteristics of fake reviews and their intermediary role in persuading buying decisions”, sv. 26, br. 3, str. 1091–1108. https://doi.org/10.1007/s10796-023-10401-w

[22] R. Mohawesh, S. Xu, S. N. Tran, R. Ollington, M. Springer, Y. Jararweh, i S. Maqsood, “Fake reviews detection: A survey”, sv. 9, str. 65 771–65 802, conference Name: IEEE Access. https://doi.org/10.1109/ACCESS.2021.3075573

[23] A. Mukherjee, V. Venkataraman, B. Liu, i N. Glance, “What yelp fake review filter might be doing?” sv. 7, br. 1, str. 409–418, number: 1. https://doi.org/10.1609/icwsm.v7i1.14389

[24] S. Rayana i L. Akoglu, “Collective opinion spam detection: Bridging review networks and metadata”, u *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, str. 985–994. https://doi.org/10.1145/2783258.2783370

[25] Nltk. [Mrežno]. Adresa: https://www.nltk.org/

[26] “vader”. [Mrežno]. Adresa: https://github.com/cjhutto/vaderSentiment

[27] Nnsvg. [Mrežno]. Adresa: https://alexlenail.me/NN-SVG/index.html

[28] N. V. Chawla, K. W. Bowyer, L. O. Hall, i W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique”, sv. 16, str. 321–357. https://doi.org/10.1613/jair.953

[29] “scikit”. [Mrežno]. Adresa: https://scikit-learn.org/stable/

[30] “imbalaned-learn”. [Mrežno]. Adresa: https://imbalanced-learn.org/stable/

Sažetak

Primjena strojnog učenja za klasifikaciju lažnih recenzija

Dora Šokota

Ovaj rad istražuje problem klasifikacije lažnih recenzija na internetskim platformama. U radu su implementirani i evaluirani modeli logističke regresije, SVM, modeli temeljeni na stablu i neuronske mreže, koristeći skup podataka s platforme Yelp. Analizirane su različite značajke, uključujući numeričke značajke usmjerene na recenzente i jezične značajke vezane uz tekst recenzije. Modeli su testirani na numeričkim značajkama, jezičnim značajkama, te na kombinaciji obje vrste značajki. Analizom rezultata pokazano je da modeli temeljeni na stablu najbolje klasificiraju lažne recenzije. U radu se opisuje arhitektura svakog korištenoj modela strojnog učenja.

Ključne riječi: strojno učenje; lažne recenzije; binarna klasifikacija

Abstract

Application of machine learning for fake reviews classification

Dora Šokota

This paper investigates the problem of classifying fake reviews on online platforms. In this paper, logistic regression models, SVM, tree-based models, and neural networks are implemented and evaluated using the Yelp data set. Various features are analyzed, including numerical features focused on reviewers and linguistic features related to the review text. The models are tested on numerical features, linguistic features, and a combination of both types of features. The analysis of the results shows that tree-based models are the best at classifying fake reviews. The paper describes the architecture of each machine learning model used.

Keywords: machine learning; fake reviews; binary classification