UNIVERZITET U BEOGRADU
MATEMATIČKI FAKULTET

Jelena Ćosić

PRIMENA LEKSIČKIH RESURSA U SENTIMENT ANALIZI TEKSTA

master rad

Beograd, 2020.

Mentor:

dr Jelena Graovac, docent
Univerzitet u Beogradu, Matematički fakultet

Članovi komisije:

prof. dr Gordana Pavlović-Lažetić, redovni profesor
Univerzitet u Beogradu, Matematički fakultet

prof. dr Cvetana Krstev, redovni profesor
Univerzitet u Beogradu, Filološki fakultet

Datum odbrane: _____

Porodici, dečku i prijateljima za nesebičnu ljubav i podršku

Naslov master rada: Primena leksičkih resursa u sentiment analizi teksta

Rezime: Sentiment analiza je složen analitički proces koji se realizuje kroz primenu različitih metoda obrade prirodnog jezika, mašinskog učenja i računarske lingvistike u cilju sistematskog identifikovanja, dobijanja, kvantifikovanja i analize emocija i mišljenja iskazanih u tekstu. Primeri njene široke primene su analiza mišljenja o raznim proizvodima i uslugama dobijenih na različitim anketama, društvenim mrežama, forumima i blogovima, sa ciljem unapređenja donošenja odluka u trgovini, finansijskom sektoru, politici i u životu uopšte. Jedan od zadataka sentiment analize je klasifikacija dokumenata prema polaritetu iskazanog mišljenja u tekstu (pozitivno ili negativno). Rešavanju problema klasifikacije može se pristupiti primenom metoda zasnovanih na leksičkim resursima, primenom metoda mašinskog učenja ili hibridnim pristupom.

Cilj ovog rada je sentiment klasifikacija javno dostupnih korpusa filmskih recenzija na srpskom i engleskom jeziku, korišćenjem metoda zasnovanih na leksičkim resursima i metoda razvijenih hibridnim pristupom. Ideja je da kroz njihovu implementaciju i testiranje uočimo koje su sve mogućnosti i prednosti, ali i mane i nedostaci ovih pristupa. Oba pristupa će se zasnivati na primeni srpskog i engleskog Wordnet-a, leksičko-semantičke mreže za srpski i engleski jezik. Problem fleksije u srpskom jeziku biće rešen korišćenjem morfološkog rečnika za srpski jezik. Biće izvršena analiza i interpretacija dobijenih rezultata, njihovo poređenje i ispitivanje da li se uključivanjem morfoloških, sintaksičkih i semantičkih informacija sadržanih u leksičkim resursima može unaprediti proces sentiment analize na srpskom (kao jednom od morfološki bogatijih jezika) i engleskom (kao najrasprostranjenijem jeziku na svetu). Rezultati rada pokazuju da je sentiment analizu moguće uspešno izvesti nad nekim korpusima, dok je nad nekim korpusima to uslovljeno raznim jezičkim faktorima.

Ključne reči: sentiment analiza, nenadgledano učenje, nadgledano učenje, klasifikacija, lematizacija, leksički resursi, Wordnet, SVM

Sadržaj

1 Uvod **1**

2 Leksički resursi **4**

2.1 Wordnet 5

2.2 Elektronski rečnik za srpski jezik 9

3 Sentiment klasifikacija teksta **12**

3.1 Preprocesiranje teksta 13

3.2 Metode klasifikacije teksta 15

3.3 Evaluacija kvaliteta klasifikacije 19

4 Razvijene metode sentiment klasifikacije **23**

4.1 Metode zasnovane na leksičkim resursima 23

4.2 Hibridne metode 24

4.3 Korpusi filmskih recenzija 26

4.4 Python i korisni paketi 27

4.5 Implementacija 28

5 Rezultati **30**

5.1 Rezultati za korpus na engleskom jeziku 30

5.2 Rezultati za korpus na srpskom jeziku 32

6 Zaključak **38**

Bibliografija **40**

Glava 1

Uvod

U današnje vreme velikog i brzog razvoja modernih tehnologija, društvene mreže i razni portali igraju sve bitniju ulogu u svakodnevnom životu gotovo svakog pojedinca. Ogromna količina informacija je dostupna u svakom trenutku. Kako bi se te informacije mogle što efikasnije iskoristiti, potrebno ih je nekako urediti i klasifikovati. Glavni zadatak sentiment klasifikacije je upravo klasifikacija teksta prema polaritetu iskazanog mišljenja koje u sebi nosi. Ovom zadatku se može pristupiti na različite načine: primenom metoda zasnovanih na leksičkim resursima, primenom metoda mašinskog učenja ili hibridnim metodama. Cilj ovog rada je sentiment klasifikacija javno dostupnih korpusa filmskih recenzija na srpskom i engleskom jeziku primenom metoda zasnovanih na leksičkim resursima, kao i hibridnim metodama.

Sentiment analiza ili **analiza osećaja** odnosi se na primenu obrade prirodnog jezika, analize teksta, računarske lingvistike i biometrije za sistematsko identifikovanje, vađenje, kvantifikaciju i proučavanje afektivnih stanja i subjektivnih informacija [13]. Za razliku od činjenica, osećanja i mišljenja su veoma subjektivna. Zbog toga je potrebno da se analiziraju mišljenja većeg broja različitih ljudi, da bi se stekao pravi utisak o nekoj temi [4]. Tako se primenom različitih tehnika sentiment analize može iskazati sveukupno mišljenje javnosti o nekoj temi. Zbog toga sentiment analiza teksta nalazi široku primenu u obradi informacija dobijenih različitim anketama, društvenim mrežama, zatim u obradi filmskih recenzija, političkih mišljenja i tome slično. Sajtovi društvenih medija omogućavaju korisnicima slobodnu razmenu mišljenja i iskustava o procesu kupovine, kvalitetu proizvoda i usluga. Zbog svog značaja i široke rasprostranjenosti, onlajn recenzije postaju ključni tip korisnički generisanog sadržaja i zauzimaju centralno mesto u istraživanjima iz domena sentiment analize [4]. Za tu svrhu su nastale aplikacije koje koriste komentare preuzete

sa Twitter-a ili Facebook-a o nekoj određenoj temi i određuju raspoloženje korisnika na ovim društvenim mrežama prema prikupljenim podacima. Slično tome, tipične aplikacije integrisane u sajtove elektronske trgovine su upravo sistemi za davanje preporuka na osnovu korisničkih recenzija. Zahvaljujući tome, korisnici sajta ne pretražuju direktno recenzije, već dobijaju preporuku proizvoda koji najviše odgovara njihovim zahtevima. Takođe, poslovno informisanje je usko povezano sa oblastima primene sentiment analize jer se upotrebom podataka, tehnologija i analiza kompanijama pružaju uvid i saznanja neophodna za donošenje odluka i sprovođenje akcija sa ciljem ostvarenja ekonomske vrednosti. Primena sentiment analize je široko rasprostranjena i u domenu medicine. Naime, razumevanje iskustva pacijenta tokom zdravstvene zaštite predstavlja centralni deo u procesu pružanja nege i to je osnovni odraz kvaliteta zdravstvene zaštite. Dakle, zavisno od oblasti primene, istraživanja iz domena sentiment analize usmerena su ka unapređenju različitih socijalnih, ekonomskih, političkih i psiholoških aspekata svakodnevnog života [4].

Primena sentiment analize može da bude značajna i u okviru različitih zadataka iz oblasti obrade prirodnih jezika (engl. Natural Language Processing, NLP), kao sto su ekstrakcija informacija odvajanjem činjenica od mišljenja, identifikovanje višeznačnosti reči i drugo [4].

Osnovni zadatak u sentiment analizi teksta je klasifikacija teksta prema polaritetu iskazanog mišljenja na pozitivno ili negativno. Da bi se sentiment analiza teksta uspešno izvršila, neophodno je stvoriti adekvatne uslove za njenu realizaciju, što se može obezbediti odgovarajućim istraživanjima i tehnikama preprocesiranja.

Sentiment analiza teksta predstavlja proces klasifikacije koji se obavlja na jednom od tri moguća nivoa. **Sentiment klasifikacija na nivou dokumenta** ima za cilj da klasifikuje dokumente na one u kojima je iskazan pozitivan, odnosno negativan sentiment. Kod ovakvog pristupa se uzima u obzir celokupni dokument i polazi se od pretpostavke da se u dokumentu raspravlja o jednoj temi. Ovaj način klasifikacije često ne pruža dovoljno detalja o preovlađujućem mišljenju korisnika o različitim aspektima posmatranog entiteta, što je zahtev mnogih aplikacija. Za viši stepen detaljnosti, radi se analiza koja se bavi klasifikacijom sentimenta na nivou rečenice ili fraze prema aspektima analiziranog entiteta. **Sentiment analiza na nivou rečenice** zahteva da se prvobitno razdvoje subjektivne od objektivnih rečenica, potom se sentiment analizom utvrđuje da li je iskazan pozitivan ili negativan stav u subjektivnim rečenicama. Treći nivo sentiment analize podrazumeva **klasifikaciju reči ili fraza** prema polaritetu iskazanog sentimenta [4]. Jedna od osnovnih pode-

la svih metoda za klasifikaciju, koje se mogu koristiti za sentiment analizu teksta, je na **metode zasnovane na leksičkim resursima**, **metode mašinskog učenja** i **hibridne metode**. **Metode zasnovane na leksičkim resursima** koriste leksičke resurse za implementaciju algoritama za klasifikaciju teksta prema sentimentu. Za razvoj ovih algoritama nije neophodno postojanje labeliranih podataka (skupova tekstova koji su unapred već klasifikovani). Ukoliko su labelirani podaci ipak dostupni, oni se mogu iskoristiti u cilju testiranja razvijenih algoritama. **Metode mašinskog učenja** se mogu podeliti na metode nenadgledanog i metode nadgledanog učenja. Osnovna razlika između ova dva učenja je u tome što je kod metoda nadgledanog učenja neophodno imati na raspolaganju labelirane ulazne podatke, dok kod metoda nenadgledanog učenja takav skup nije potreban. **Hibridne metode** predstavljaju kombinaciju prethodno navedenih metoda.

U ovom radu akcenat će biti stavljen na analizu sentimenta tekstova filmskih recenzija na srpskom i engleskom jeziku, korišćenjem sentiment klasifikacije na nivou dokumenta u okviru metoda zasnovanih na leksičkim resursima i metoda dobijenih hibridnim pristupom. Nakon toga biće izvršeno upoređivanje dobijenih rezultata nad datim korpusima. U okviru metoda zasnovanih na leksičkim resursima biće izvršena klasifikacija korišćenjem leksičkih resursa za srpski i engleski jezik. Nakon toga, u okviru hibridnog pristupa, u cilju klasifikacije teksta, biće korišćen metod potpornih vektora (eng. Support Vector Machines (SVM)) kao jedan od najpopularnijih tradicionalnih algoritama mašinskog učenja. Ulazni podaci potrebni ovom algoritmu biće dobijeni prethodno razvijenom metodom zasnovanom na leksičkim resursima.

Ovaj rad je organizovan na sledeći način:

Nakon uvodnog dela o sentiment analizi, u poglavlju 2 je detaljno opisana struktura svih dostupnih leksičkih resursa koji su korišćeni (srpski i engleski Wordnet i elektronski rečnik za srpski jezik). Poglavlje 3 sadrži opis sentiment klasifikacije teksta, opis algoritama koji su korišćeni u ovom radu, kao i način na koji je izvršena evaluacija dobijenih rezultata. Na koji način su konkretne metode implementirane opisano je u poglavlju 4. Pored toga, u ovom poglavlju objašnjena je i sama struktura projekta, kao i koji su korpusi filmskih recenzija (na srpskom i engleskom jeziku) korišćeni kao ulazni podaci. U poglavlju 5 prikazani su svi dobijeni rezultati klasifikacije, kao i deo sa nekim problemima koji su se javljali prilikom klasifikacije. Poslednje poglavlje 6, sadrži osnovne zaključke o radu i mogućnosti daljeg unapređenja prikazanog pristupa rešavanju problema klasifikacije filmskih recenzija.

Glava 2

Leksički resursi

Jezički resursi predstavljaju skupove različitih jezičkih podataka, kao i opisa koji se nalaze u mašinski čitljivom obliku. Koriste se za obradu podataka koji su prikazani prirodnim jezikom i mogu se koristiti u unapređivanju ili evaluaciji raznih alata za obradu prirodnog jezika [5]. Ovi resursi sadrže sve neophodne i relevantne informacije za željenu obradu jezika. **Leksički resursi**, kao posebna vrsta jezičkih resursa, su resursi širokog spektra. Leksički resursi sadrže elektronske rečnike i semantičke mreže za određeni jezik. Takođe, leksički resursi, pored jednojezičnih mogu biti i višejezični. U tom slučaju mogu se napraviti međusobne relacije među rečima iz više jezika (ili iz istog jezika), koje mogu biti veoma korisne. Pored toga, primena leksičkih resursa je mnogostruka. Na primer, njihovom primenom moguće je poboljšati performanse veb pretraživača [9], dok je korišćenjem višejezičnih leksičkih resursa moguće sastaviti dvojezične terminološke liste [11].

Leksički resursi za srpski jezik predstavljaju veoma bitan segment za ovaj rad. To su leksički resursi, koji su od strane stručnjaka u toj oblasti posebno razvijeni za srpski jezik. Leksički resursi za srpski jezik sadrže leksičko-semantičku mrežu za srpski jezik (srpski Wordnet), kao i elektronski rečnik za srpski jezik, koji su od izuzetnog značaja za obradu koja će biti izvršena.

Pored leksičkih resursa za srpski jezik, u ovom radu se koriste i **leksički resursi za engleski jezik**. Većina resursa za engleski jezik (stop reči, elektronski rečnik, ...) dostupna je u okviru različitih paketa programskog jezika Python (`nltk.corpus`, `nltk.stem`, ...). S obzirom na to i dobijeni rezultati su relevantni za upoređivanje. Dakle, upoređivanjem ocena dobijenih nakon izvršene klasifikacije srpskog korpusa sa ocenama dobijenim nakon izvršene klasifikacije engleskog korpusa dobija se i informacija da li je korišćeni pristup pri klasifikaciji srpskog korpusa dovoljno dobar.

Ukoliko su dobijene ocene približno jednake, samim tim je i klasifikacija srpskog korpusa uspešno izvršena i korišćeni pristup je dobar.

2.1 Wordnet

Wordnet predstavlja jednu vrstu leksičko-semantičke mreže nekog jezika. Zbog svoje strukture ima široku primenu u obradi prirodnih jezika i računskoj lingvistici. Wordnet predstavlja skup koncepata povezanih semantičkim relacijama u semantičku mrežu. Sadrži imenice, glagole, prideve i priloge (ignoriše predloge, veznike i ostale vrste reči) koji su grupisani u skupove sinonima (sinsetove) od kojih svaki označava različit pojam. Sinsetovi su međusobno povezani leksičkim i semantičkim odnosima, na osnovu njihovog značenja. Tako dobijena mreža smisleno povezanih reči i pojmova može se lako pretraživati i koristiti. Ono što je veoma bitno za Wordnet je da se reči koje se nalaze u neposrednoj blizini u okrivu mreže ne razlikuju semantički [1][19]. Dakle, sinsetovi predstavljaju skup sinonima neke reči - različite reči sa istim ili približnim značenjem. Pored toga, svaki sinset, sadrži i kratku definiciju, kao i, u većini slučajeva, jednu ili više rečenica koje opisuju upotrebu članova sinseta. Reči koje imaju više različitih značenja mogu se pojaviti u više različitih sinsetova. Zbog toga je svaki par, koji označava reč i njeno značenje, jedinstven u Wordnet-u [1] [19].

Prvi razvijeni Wordnet je Wordnet za engleski jezik, poznatiji kao **Prinstonski Wordnet (PW)**. Nosi ovakav naziv jer su ga sačinili naučnici sa Prinstonskog univerziteta[1] pod vođstvom profesora psihologije Džordž Armitaž Milera[2]. Džordž Miler i Kristijan Felbaum[3] dobili su nagradu Antonio Zampolli[4] za svoj rad na Prinstonskom Wordnet-u 2006. godine. Prinstonski Wordnet sadrži 155 327 reči organizovanih u 175 979 sinsetova za ukupno 207 016 parova reč-značenje.

[1]**Princeton University** je privatni univerzitet koji se nalazi u Prinstonu, Nju Džerziju, Sjedinjenim Američkim Državama.

[2]**George Armitage Miller**(3. februar 1920 – 22. jul 2012) bio je američki psiholog, koji je bio jedan od osnivača kognitivne psihologije, i šireg polja kognitivne nauke. Takođe je doprineo formiranju polja psiholingvistike. Miler je napisao nekoliko knjiga i usmeravao razvoj Wordneta [20].

[3]**Christiane Fellbaum** je profesor na programu lingvistike i odeljenju za računarske nauke na Univerzitetu Prinston.

[4]Nagrada koja se dodeljuje u čast sećanja na profesora Antonija Zampolija, naučnika pionira i vizionara, koji je međunarodno priznat u oblasti računske lingvistike i tehnologija za obradu prirodnih jezika.

Wordnet je razvijen i za mnoge druge svetske jezike. Mnogi koncepti u Wordnet-u su specifični za određene jezike, a najtačnije mapiranje koje je obavljeno između jezika iznosi oko 94% [14].

Globalna Wordnet Asocijacija (GWA)[5] je javna i nekomercijalna organizacija koja sadrži platformu za diskusiju, deljenje i povezivanje wordneta za sve jezike u svetu. GWA takođe promoviše standardizaciju wordneta po jezicima kako bi se omogućila ujednačenost u nabrajanju sinsetova na svim jezicima, koliko je to moguće. GWA sadrži listu svih Wordnet-ova razvijenih širom sveta [21].

Na slici 2.1 se može videti jedan isečak iz Wordnet-a za engleski jezik, dok je na slici 2.2 prikazan deo hijerarhijske organizacije u Wordnet-u.

Dog (noun)

- **dog**, domestic dog, Canis familiaris - a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds; *"the dog barked all night"*

- frump, **dog** - a dull unattractive unpleasant girl or woman; *"she got a reputation as a frump"*; *"she's a real dog"*

- **dog** - informal term for a man; *"you lucky dog"*

- cad, bounder, blackguard, **dog**, hound, heel - someone who is morally reprehensible; *"you dirty dog"*

- frank, frankfurter, hotdog, hot dog, **dog**, wiener, wienerwurst, weenie - a smooth-textured sausage of minced beef or pork usually smoked; often served on a bread roll

- pawl, detent, click, **dog** - a hinged catch that fits into a notch of a ratchet to move a wheel forward or prevent it from moving backward

- andiron, firedog, **dog**, dog-iron - metal supports for logs in a fireplace; *"the andirons were too hot to touch"*

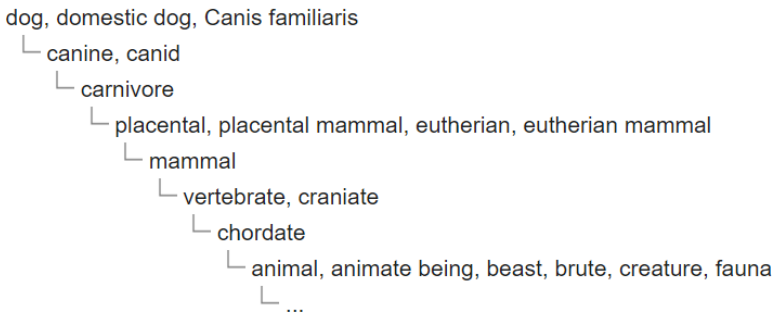
Dog (verb)

- chase, chase after, trail, tail, tag, give chase, **dog**, go after, track - go after with the intent to catch; *"The policeman chased the mugger down the alley"*; *"the dog chased the rabbit"*

Slika 2.1: Isečak iz Wordnet-a za reč „dog”

Na najvišem nivou, ove hijerarhije su organizovane u 25 početnih „stabala” za imenice i 15 za glagole, koje se kasnije granaju. Svi su povezani jedinstvenim početnim sinsetom. Hijerarhije za imenice daleko su dublje od hijerarhija za glagole. Pridevi nisu organizovani u hijerarhijska stabla.

[5]The Global WordNet Association (eng.)



Slika 2.2: Primer hijerarhije u Wordnet-u

Srpski Wordnet

Srpski Wordnet[6] je leksičko-semantička mreža napravljena za srpski jezik. Struktura je kao u i ostalim svetskim Wordnet-ima i sastavljen je takođe od sinsetova povezanih u jedinstvenu celinu. Postoji i veza između Prinstonskog Wordneta i srpskog Wordneta. Naime, ID koji predstavlja jedinstveni broj koji određuje svaki sinset, u sebi može sadržati ID sinseta u Prinstonskom Wordnet-u koji se odnosi na isti pojam, ukoliko taj pojam postoji u Prinstonskom Wordnet-u. Prateći ovu vezu, može se sinsetu iz srpskog Wordnet-a pridružiti odgovarajući sinset iz Wordnet-a na engleskom jeziku.

Sadržaj srpskog Wordnet-a je dostupan u *xml* formatu. Svaki sinset tag je sačinjen od nekoliko podtagova od kojih su najbitniji [10]:

- *<ID>* - jedinstveni broj koji određuje sinset; može sadržati ID sinseta iz engleskog Wordnet-a koji se odnosi na isti pojam, ukoliko pojam postoji u engleskom Wordnet-u; npr. *ENG30-15156001-n* je ID is srpskog Wordnet-a i sadrži deo *15156001* koji je iz Wordnet-a na engleskom jeziku
- *<SYNONYM>* - u okviru ovog taga nalaze se svi sinonimi za određeni pojam koji predstavlja taj čvor
  - *<LYTERAL>* - predstavlja podtag u okviru taga *<SYNONYM>* u okviru koga je navedena reč koja predstavlja sinonim; može ih biti nekoliko
- *<DEF>* - predstavlja objašnjenje pojma, tj. definiciju; uglavnom je predstavljena u par kratkih rečenica
- *<ILR>* - predstavlja ID sinseta koji je u nekoj relaciji sa trenutnim sinsetom

[6] http://dcl.bas.bg/bulnet/

  - <*TYPE*> - predstavlja podtag u okviru taga <*ILR*> u kome je navedena pomenuta relacija sa nekim drugim sinsetom (spisak nekih vrsta međusobnih relacija između sinsetova dat je u tabeli 2.1)

- <*SENTIMENT*> - predstavlja sentiment pojma; u okviru podtagova <*POSITIVE*> i <*NEGATIVE*> su prikazani numerički podaci koji oslikavaju da li dati pojam ima pozitivno ili negativno značenje u srpskom jeziku u zavisnosti koja vrednost je veća

Relacija	**Značenje**
hypernym	nadređenost
hyponym	podređenost
eng_derivative	izvedena reč
near_antonym	antonimija
verb_group	uzrokuje – uzrokovan
holo_part	deo – celina
holo_member	član – celina
be_in_state	biti u stanju – stanje nečega
similar_to	sličnost

Tabela 2.1: Spisak nekih relacija u srpskom Wordnet-u

Deo srpskog Wordnet-a koji je popunjen imenicama konstruisan je kao mreža hijerarhijskih čvorova između kojih su uspostavljene relacije *nadređenosti (eng. hypernym)* i *podređenosti (eng. hyponym)* pojmova. Jedan pojam je u relaciji *podređenosti* u odnosu na drugi pojam ukoliko sadrži sva svojstva kao i nadređeni pojam, ali može imati i neka svoja specifična svojstva, samo za njega karakteristična. Na primer, sinset koji sadrzi imenicu „ustanova" (u značenju zgrade u kojoj se obavlja neka delatnost) nadređen je sinsetu koji sadrži literale „kaznena institucija, kazneni zavod, pritvorna jedinica", dok je podređen sinsetu koji sadrži literal „objekat" (u značenju zgrade u kojoj se živi ili nešto obavlja). Takođe, imenice u srpskom Wordnet-u mogu biti povezane i drugim relacijama, kao sto su *deo – celina (eng. holo_part)* i *član – celina (eng. holo_member)*. Ovim relacijama se pokazuje da li je neka imenica ustvari deo neke veće celine ili ona predstavlja član neke celine sa kojom je u relaciji. Na primer, sinset koji sadrži imenicu „tramvaj" je u relaciji *deo – celina* sa sinsetom koji sadrži imenicu „tramvajska linija", dok sinset koji sadrži imenicu „suncokret" je u relaciji *član – celina* sa sinsetom koji sadrži literal „rod Helianthus". Još jedna zanimljiva relacija koja se uspostavlja između imenica suprotnog značenja je relacija *antonimije (eng. near_antonym)*. Takođe, ovu relaciju

moguće je uspostaviti i među drugim vrstama reči. Prateći tok povezanosti može se napraviti čitava mreža imenica koje su povezane ovim, ali i drugim relacijama [1].

Za povezivanje različitih vrsta reči u okviru srpskog Wordnet-a koristi se relacija *biti u stanju – stanje nečega (eng. be_in_state)*, koja povezuje imenice i prideve. Na primer, sinset koji sadrži pridev „važan" u relaciji *biti u stanju – stanje nečega* je sa sinsetom koji sadrži imenicu „važnost", a u relaciji *antonimije* sa sinsetom koji sadrži pridev „nevažan". Dakle, relacija *antonimije* uspostavlja se i između prideva u srpskom Wordnet-u [1].

Relacija koja se često uspostavlja među glagolima je relacija *uzrokuje– uzrokovan (eng. verb_group)*. Na primer, sinset koji sadrži glagol „prodati" u relaciji *uzrokuje– uzrokovan* je sa sinsetom koji sadrži glagol „prodavati se", a ovaj sa sinsetom koji sadrži glagol „prodavati" [1].

Na osnovu prethodno ilustrativnog opisa srpskog Wordnet-a može se zaključiti da on predstavlja jednu veoma gusto povezanu mrežu različitih pojmova razvrstanih u sinsetove. Isečak iz srpskog Wordnet-a prikazan je na slici 2.3[7].

Nažalost, uočeni su nedostaci u srpskom Wordnet-u, koji su kasnije opisani. Kako je jedan od ciljeva ovog rada i unapređenje srpskog Wordnet-a i ispitivanje efekta ovog unapređenja na rezultate sentiment analize teksta, uočeni nedostaci su prevaziđeni tako što su pronađene greške ispravljene.

2.2 Elektronski rečnik za srpski jezik

Elektronski rečnik predstavlja rečnik koji se nalazi u elektronskoj formi i koristi se u procesu obrade teksta [6]. Njegova glavna svrha je upotreba u procesu obrade prirodnog jezika, a može se koristiti i za preprocesiranje teksta [1]. Sadrži korisne informacije koje pomažu razrešavanju problema kod sintaksičke i semantičke obrade teksta [6]. Model **elektronskog rečnika za srpski jezik** i za druge slovenske jezike, razvijen je polazeći od metodologije koja je nastala u okviru mreže RELEX[8] [1]. Jedan primer obrade odrednice u rečniku je:

dvojica, N623+MG+Pl[9],

gde kôd N623 označava flektivnu klasu, MG prirodni muški rod, a Pl prirodnu množinu [6]. Ovaj elektronski rečnik za srpski jezik iskorišćen je za dobijanje odgovarajućeg izlaza za sve reči iz dostupnih korpusa na srpskom jeziku. Tom prilikom

[7]Slika preuzeta iz [1].

[8]http://infolingu.univ-mlv.fr/english/Relex/Relex.html

[9]Primer je iz rečnika tipa DELAS - rečnik prostih reči u osnovnom obliku (prostih lema)[1]

```
<SYNSET>
      <ID>ENG30-02083346-n</ID>
      <SYNONYM>
            <LITERAL>pas</LITERAL>
      </SYNONYM>
      <DEF>Bilo koji od raznovrsnih
           sisara koji obicyno imaju
           dugu nxusxku i kandye.
      </DEF>
      <POS>n</POS>
</SYNSET>

<SYNSET>
      <ID>ENG30-02084071-n</ID>
      <SYNONYM>
            <LITERAL>pas</LITERAL>
            <LITERAL>pseto</LITERAL>
            <LITERAL>domacxi pas</LITERAL>
      </SYNONYM>
      <DEF>Pripadnik Canis familiaris,
           srodan vuku, pripitomlxen od
           preistorijskog doba;
           postoje mnoge rase.
      </DEF>
      <POS>n</POS>
      <ILR>ENG30-02083346-n
            <TYPE>hypernym</TYPE>
      </ILR>
      <ILR>ENG30-07994941-n
            <TYPE>holo_member</TYPE>
      </ILR>
</SYNSET>
<SYNSET>
      <ID>ENG30-02158846-n</ID>
      <SYNONYM>
            <LITERAL>rep</LITERAL>
      </SYNONYM>
      <DEF>Upadlxivo oznacyen ili oblikovan
           zadnxi deo.</DEF>
      <POS>n</POS>
      <ILR>ENG30-02084071-n
            <TYPE>holo_part</TYPE>
      </ILR>
</SYNSET>

<SYNSET>
      <ID>ENG30-07994941-n</ID>
      <SYNONYM>
            <LITERAL>cyopor</LITERAL>
      </SYNONYM>
      <DEF>Grupa zxivotinxa koje love.</DEF>
      <POS>n</POS>
</SYNSET>
```

Slika 2.3: Isečak iz srpskog Wordnet-a (XML reprezentacija)

upotrebljen je i tager koji je razvijen i opisan u sklopu rada [8]. Dobijene informacije o rečima koriste se dalje za preprocesiranje tekstova iz korpusa. Dakle, svakoj reči je dodeljena odgovarajuća vrsta reči - **tag** kao i **lema**, koja predstavlja osnovni oblik reči. Vrste reči su pridružene rečima u zavisnosti od konteksta reči u rečenici. Na slici 2.5 može se videti deo jednog teksta iz korpusa sa posebno obeleženim vrstama reči. Sve vrste reči koje su korišćene u rečniku date su na slici 2.4. Takođe, posebnom vrstom reči su označeni specijalni interpunkcijski znakovi koji predstavljaju kraj rečenice - SENT, dok su ostali interpunkcijski znakovi označeni kao PUNCT.

Tag	Značenje
N	Imenica
A	Pridev
V	Glagol
ADV	Prilog
PREP	Predlog
PRO	Zamenica
CONJ	Veznik
INT	Član
PAR	Rečca
ABB	Skraćenica
NUM	Broj
RN	Rimski broj
PREF	Prefiks
PUNCT	Interpunkcijski znak
SENT	Kraj rečenice

Slika 2.4: Značenje tagova u rečniku

```
Ovo     PRO ovaj
je PRO ona
film    N:m film
koji    PRO koji
ne PAR ne
spada   V:m spada
u  PREP     u
taj     PRO taj
žanr    N:m žanr
,  PUNCT    ,
ali     CONJ    ali
po PREP     po
mom     PRO moj
skromnom    A:aen   skroman
mišljenju   N:n mišljenje
ovo     PRO ovaj
je PRO ona
njegov PRO njegov
najbolji    A:cem   dobar
film    N:m film
.  SENT     .
```

Slika 2.5: Isečak iz jednog teksta iz korpusa

Glava 3

Sentiment klasifikacija teksta

Klasifikacija teksta predstavlja svrstavanje teksta u jednu ili više predefinisanih kategorija. Na primer, novinski članci su najčešće organizovani po rubirkama kojima pripadaju, naučni radovi su klasifikovani po oblastima i podoblastima koje obrađuju, medicinski kartoni pacijenata su često razvrstani prema više kriterijuma – istorijat bolesti, brojevi osiguranja itd. U navedenim primerima svaka instanca (novinski članak, naučni rad, medicinski karton itd.) se može predstaviti nekim izabranim skupom njenih atributa. Svakoj instanci se može dodeliti oznaka klase – ciljna vrednost, kojoj instanca pripada. Upravo ovo su tzv. labelirani podaci. Problem klasifikacije se sastoji u određivanju ovih oznaka klasa na osnovu atributa instance. Matematički gledano, problem klasifikacije se može posmatrati kao aproksimacija funkcije koja svakoj instanci dodeljuje oznaku klase kojoj ta instanca pripada [3]. Po broju elemenata skupa labeliranih podataka, klasifikacija može biti **binarna** (dve klase) i **višeklasna** (više klasa). Ipak neki tekst može spadati u više klasa u isto vreme. Tada se radi o **višeznačnoj** klasifikaciji [15]. U ovom radu biće korišćena samo binarna klasifikacija i to prema sentimentu koji je iskazan u korpusima filmskih recenzija na srpskom i engleskom jeziku. U zavisnosti od toga da li je u tekstu iskazan pozitivan ili negativan sentiment, tako će i biti klasifikovan u jednu od ove dve klase. Naime, **klasifikacija prema sentimentu** je vrsta klasifikacije u kojoj se dokumenti razvrstavaju po određenim klasama u odnosu na sentiment koji je u njima iskazan. Ova klasifikacija specifična je po tome što je za njenu uspešnu realizaciju neophodno pre toga izvršiti odgovarajuća istraživanja i tehnike preprocesiranja, što je upravo čini i zahtevnijom u odnosu na druge vrste klasifikacije.

U procesu klasifikacije mogu se javiti određeni problemi. Jedan od njih jeste i **predimenzionisanost**, odnosno veliki broj atributa kojim se dokument predstavlja.

Postoje različiti pristupi za rešavanje ovog problema, a jedan od njih jeste **selekcija atributa** (eng. feature selection), kod koga se na osnovu odredenog znanja bira podskup početnog skupa atributa, tj. biraju se oni atributi koji su od najvećeg značaja [1]. Eliminacija stop reči je upravo jedna takva metoda i biće detaljno opisana u narednom poglavlju u okviru preprocesiranja teksta.

3.1 Preprocesiranje teksta

Kao što je u prethodnom poglavlju i rečeno, da bi se uspešno izvršila klasifikacija prema sentimentu neophodno je izvršiti kvalitetno preprocesiranje teksta, koje se obično sastoji iz različitih faza:

- ujednačiti pismo (prevesti tekst na latinično pismo - srpski korpus)
- rešiti problem lematizacije
- ukloniti stop reči
- ukloniti određene vrste reči kao na primer zamenice, brojeve, predloge ...

Svaka od faza tekstove obrađuje i prosleđuje narednoj fazi. Pomenute faze se mogu razlikovati u zavisnosti od jezika.

Od velike važnosti je da svi tekstovi i leksički resursi koji se koriste budu napisani na istom pismu. Kako su u srpskom jeziku podjednako zastupljena oba pisma, ćirilično i latinično, i kako su leksički resursi koji će biti korišćeni u ovom radu napisani na latinici, neophodno je da svaki **ćirilični tekst bude prebačen u latinično pismo**. Ovo u mnogome olakšava kasniju manipulaciju podacima jer će svi podaci biti zapisani u istom pismu - latiničnom. Dakle, ova faza nije primenljiva na korpus tekstova na engleskom jeziku.

Jedna od najbitnijih faza je svaku reč iz teksta pronaći u elektronskom rečniku koji je na raspolaganju i zameniti je njenom lemom koja se tamo nalazi. Ovaj proces se naziva proces **lematizacije**. Na slikama 3.1 i 3.2[1] nalaze se primeri kako se lematizacija primenjuje na različite vrste reči i šta se dobija kao izlaz. Lematizacija podiže preciznost rada modela jer se smanjuje kompleksnost pisanog sadržaja svođenjem reči na osnovni oblik, a samim tim postoji i manje varijacija u analiziranim sadržajima.

[1]Slika preuzeta sa http://kavita-ganesan.com/text-preprocessing-tutorial

- voleću, volite, voleli, voleo → voleti
- am, are, is → be
- car, cars, car's, cars' → car

Slika 3.1: Lematizacija

	original_word	lemmatized_word
0	trouble	trouble
1	troubling	trouble
2	troubled	trouble
3	troubles	trouble

	original_word	lemmatized_word
0	goose	goose
1	geese	goose

Slika 3.2: Lematizacija

Naredna faza se odnosi na **uklanjanje stop reči** iz teksta [12]. Stop reči predstavljaju reči koje su učestale u govornom i pisanom jeziku, a ne donose neko novo znanje, kao što su npr. veznici. Iako se stop reči obično odnose na najčešće reči u nekom jeziku, ne postoji jedinstveni univerzalni spisak stop reči koji koriste svi alati za obradu prirodnih jezika, a zapravo ni svi alati ne koriste takvu listu. Neki alati posebno izbegavaju uklanjanje ovih stop reči da bi podržali pretragu fraza. Osnovna svrha uklanjanja stop reči jeste da se poveća efikasnost algoritma smanjenjem broja reči na ulazu. Primenjuje se u procesu preprocesiranja teksta jer stop reči neće biti pronađene ni u engleskom, a ni u srpskom Wordnet-u. Takođe, neophodno je **zanemariti neke vrste reči** koje nisu od značaja za dalju obradu. Vrste reči koje je potrebno zadržati su npr: imenice, pridevi, prilozi. Ove vrste reči sa sobom nose određenu jačinu i sentiment i doprinose boljoj klasifikaciji teksta.

Nakon završenog prepocesiranja teksta formiran je finalni skup reči koji je od značaja za dalju obradu.

3.2 Metode klasifikacije teksta

Metode klasifikacije teksta se mogu podeliti na: **metode zasnovane na leksičkim resursima**, **metode mašinskog učenja** i **hibridne metode**. **Metode zasnovane na leksičkim resursima** pripadaju takozvanom algoritamskom pristupu koji ne podrazumeva postojanje labeliranih podataka. Za dati tekst na ulazu, razvija se algoritam koji korišćenjem domenskog znanja i različitih leksičkih resursa vrši preslikavanje tog teksta u jednu od unapred definisanih klasa (u ovom slučaju pozitivan ili negativan sentiment). Na primer, za dati tekst na ulazu, na osnovu nekog algoritma koji koristi leksičke resurse dobija se odgovarajući izlaz. **Hibridne metode** predstavljaju kombinaciju metoda zasnovanih na leksičkim resursima i metoda mašinskog učenja.

Metode mašinskog učenja

Sve **metode mašinskog učenja** se mogu svrstati u jednu od dve glavne kategorije – nenadgledano i nadgledano učenje. Razlika je jednostavna, ali veoma važna. Dok algoritmi nadgledanog učenja uz vrednosti za ulaz, imaju na raspolaganju i vrednosti za izlaz koji mu odgovara (na skupu za obučavanje), kod algoritama nenadgledanog učenja dostupne su samo vrednosti za ulaz, tj. nisu dostupni labelirani podaci. Klasifikaciju dokumenata pomoću mašinskog učenja moguće je jedino izvršiti korišćenjem metoda **nadgledanog učenja**. Osnovna karakteristika ovog učenja je da se podaci sastoje iz parova (ulaz, izlaz), gde ulaz predstavlja podatak iz kog se uči, a dostupan je i željeni izlaz, tj. šta je potrebno naučiti. Naziv je inspirisan načinom učenja gde učitelj učeniku zadaje zadatke, ali nakon njegovih odgovora mu daje i tačna rešenja, radi poređenja [2]. Prilikom upotrebe nadgledanog učenja pri obradi prirodnih jezika, uz tekstove iz korpusa koji se obrađuje, dobijaju se, na početku obrade, i labelirani podaci, tj. podaci o tome kojoj klasi svaki tekst pripada. Problem je pronaći odnos između podataka koji predstavljaju ulazne i izlazne podatke, jer se na osnovu tog odnosa, za neke naredne ulaze, vrši predviđanje izlaza. Ulazni i izlazni podaci se najčešće predstavljaju preko vektora x i y. Vektor x je vektor promenljivih koje se nazivaju *atributi (eng. features)*, a vektor y je uglavnom jedna promenljiva i naziva se *ciljna promenljiva (eng. target variable)*. Može se desiti da vektor x bude i višedimenzionalan, kao i slučaj da oba vektora uopšte ne sadrže numeričke podatke. Funkcije koje izražavaju vezu između vrednosti atributa i vrednosti ciljne promenljive nazivaju se *modeli mašinskog učenja*. Oni ne mogu u

potpunosti da opišu zavisnosti koje važe među promenljivim, ali se od njih očekuje da vrše dobru generalizaciju, tj. da ne prave velike greške. Takođe, ne očekuje se da neće uopšte biti grešaka, jer je to idealan slučaj koji nije realan [2].

Ulazni podaci, takozvani atributi u nadgledanom učenju mogu se posmatrati i kao uslovi pod kojima nastaje i od kojih direktno zavisi neki ishod nazvan ciljna promenljiva. Dodavanjem većeg broja promenljivih u atribute može se napraviti jača veza sa ciljnom promenljivom, ali ne može se očekivati da je moguće otkriti sve faktore koji utiču na neku pojavu. Zato se može desiti da se za iste vrednosti atributa dobiju različite vrednosti ciljne promenljive [2].

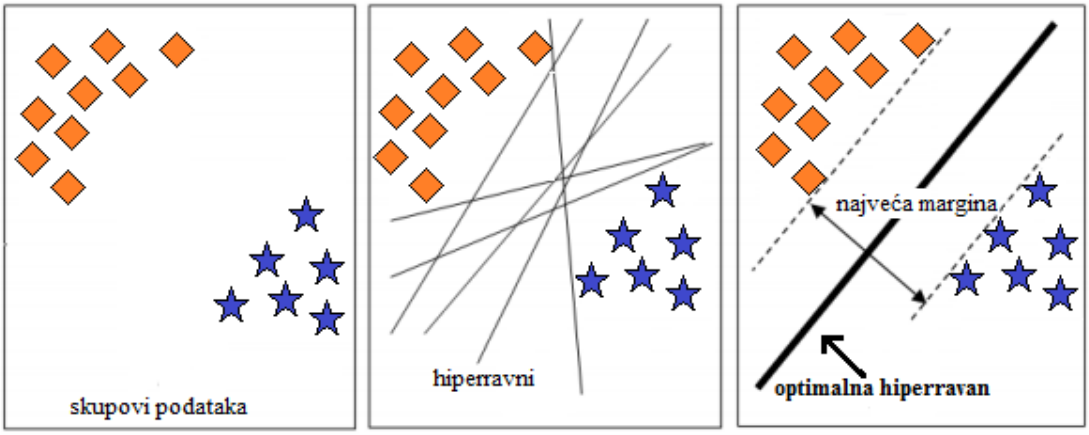
Pouzdanost predviđanja je veoma bitna kod algoritama mašinskog učenja. U nastavku ćemo razmotriti algoritam - **metod potpornih vektora** koji pripada modelu zasnovanom na širokoj margini (*eng. large margin*). Svrha ovog modela je proučavanje kada se predviđanje može smatrati pouzdanim. Koristi se koncept bezbednog odstojanja, tj. praznog prostora između objekata koji se posmatraju [2].

Metod potpornih vektora

Metod potpornih vektora (Support Vector Machine - SVM) je jedan od najvažnijih algoritama mašinskog učenja. On je zasnovan na ideji o geometrijski razdvojenim skupovima objekata. To se može opisati i predstaviti slikovito, radi lakšeg razumevanja. Dakle, može se pretpostaviti da postoje dva skupa tačaka tako raspoređenih u ravni da se može povući prava koja će ih razdvojiti tako da svi elementi jedne klase budu sa jedne strane prave, a elementi druge klase sa druge strane nacrtane prave. Ova prava se naziva hiperravan i može ih biti više od kojih su neke bolje od drugih. Takođe, postoji veliki broj jednostavnih algoritama kojima se mogu odrediti ove hiperravni, ali samo SVM algoritam se koristi ukoliko je potrebno pronaći optimalnu hiperravan. Optimalna hiperravan je ona prava koja ima najveću marginu, tj. ona koja je na najvećoj udaljenosti od oba skupa podataka. Tako se za svaku hiperravan može odrediti margina, a na osnovu nje videti da li je ta hiperravan optimalna ili ne. To je prikazano na slici 3.3 [2].

Ono što se može zaključiti sa slike je da ukoliko bi neka druga prava bila odabrana za optimalnu hiperravan, postojao bi veći rizik da se neki podaci nađu sa druge strane prave, a da tu ne pripadaju. Sada će i matematički biti opisan ovaj metod.

Neophodno je prvo definisati pretpostavke i oznake kako bi se dobile ispravne jednačine. Neka je dimenzija prostora d, a svaki podatak i prikazan u obliku vektora $x_i = (x^1, x^2, \ldots, x^d)$, gde svakom vektoru x_i odgovara tačno jedna vrednost za $y_i \in$



Slika 3.3: Određivanje optimalne hiperravni

$\{-1, 1\}$. Na osnovu principa linearne algebre sledi da je jednačina hiperravni, u oznaci R, jednaka $ax+b=0$, gde su a i b vektori u kanonskom obliku [2]. Rastojanje bilo kog podatka x od hiperravni R može se izračunati preko formule:

$$r_0(x, R) = \frac{ax + b}{\|a\|}$$

Ukoliko se širina margine, koja treba da bude što veća, označi sa ρ, onda je njena vrednost ustvari $\rho = 2 \min_x r_0$. Vektori a i b se kanonski određuju tako da rastojanje između paralelnih hiperravni od optimalne hiperravni bude 1 [2]. Odatle sledi da su jednačine tih paralelnih hiperravni jednake:

$$ax_i + b = 1$$

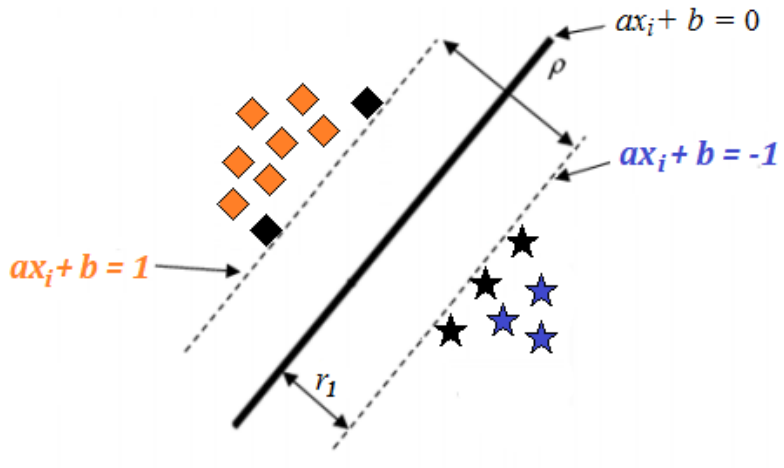
$$ax_i + b = -1$$

Potporne vektore predstavljaju podaci koji se nalaze na hiperravnima paralelnim optimalnoj hiperravni. Nazvani su na ovaj način jer deluju kao da pružaju potporu datom sistemu od tri hiperravni. Njima je inspirisan i naziv samog metoda [2]. Dakle, rastojanje potpornog vektora od optimalne hiperravni jednaka je $r_1(x, R) = \frac{1}{\|a\|}$, odakle sledi da je širina margine jednaka $\rho = 2r_1 = \frac{2}{\|a\|}$ [2].

Ako se svakom objektu iznad potpornog vektora $ax_i+b=1$ dodeli vrednost klase $y_i = 1$, a svakom objektu ispod potpornog vektora $ax_i + b = -1$ dodeli vrednost klase $y_i = -1$, onda se svi ti uslovi zajedno mogu predstaviti kao:

Za svaki par (x_i, y_i) važi:

$$ax_i + b \geq 1, \quad ukoliko \quad je \quad y_i = 1$$



Slika 3.4: Jednačine potpornih vektora

i

$$ax_i + b \leq -1, \quad ukoliko \quad je \quad y_i = -1$$

što predstavlja uslov da su svi objekti u klasama što više udaljeni od optimalne hiperravni nego što su potporni vektori, koji su na rastojanju 1 [2]. Dakle, da bi se rešio problem određivanja optimalne hiperravni i potpornih vektora, potrebno je rešiti sledeći optimizacioni problem:

$$\min_{a,b} \frac{\|a\|}{2},$$

$$y_i(ax + b) \geq 0, \quad y_i \in \{-1, 1\}, \quad i = 1, \ldots, d.$$

Metod potpornih vektora je jedan od najkorisnijih i najuspešnijih metoda mašinskog učenja i ima raznovrsnu primenu. Neke od primena su:

- u bioinformatici (za predviđanje ponašanja novih gena (gen/protein se prikazuje kao vektor dužine n, gde je n ili eksperimentalno određena ili vrednost koja određuje aktivnost gena), za klasifikacije gena/proteina na osnovu porekla ...)
- kod programa za prepoznavanje glasa i rukopisa
- za klasifikaciju dokumenata
- u medicini (obrađivanjem velikog broja izveštaja lekara, dobijaju se informacije potrebne za proučavanje bolesti i uspešno lečenje)

- za prepoznavanje objekata i lica ljudi (kod programiranja robota)
- u ekonomiji (za analizu cene proizvoda na tržištu)
- za razvrstavanje elektronske pošte
- u bankama (za procenu kreditne sposobnosti osobe)
- u trgovini i na internetu (za profilisanje kupca/korisnika usluga)

3.3 Evaluacija kvaliteta klasifikacije

Postoji više načina za određivanje da li je izvršena klasifikacija zadovoljavajuća. Naime, u zavisnosti od broja klasa, klasifikacija se može podeliti na **binarnu** (sa samo dve klase) i **višeklasnu** (sa više mogućih klasa) za klasifikovanje podataka [1].

U slučaju binarne klasifikacije, moguće je dobiti 4 vrste brojčanih podataka, uz pomoć kojih se daljim izračunavanjima mogu dobiti procene za izvršenu klasifikaciju. Ti podaci se prikazuju u tzv. **matrici konfuzije** (slika 3.5).

Predložena klasa / Stvarna klasa	POZITIVNA	NEGATIVNA
POZITIVNA	TP	FN
NEGATIVNA	FP	TN

Slika 3.5: Matrica konfuzije - binarna klasifikacija

Značenje svakog podatka prikazanog na prethodnoj slici je sledeće:

- **TP** (eng. True Positives) = **Stvarno Pozitivni (SP)**: predstavljaju broj dokumenata koji su na ulazu u klasifikaciju bili pozitivni i kao rezultat klasifikacije prepoznati kao pozitivni dokumenti

- **TN** (eng. True Negatives) = **Stvarno Negativni (SN)**: predstavljaju broj dokumenata koji su na ulazu u klasifikaciju bili negativni i kao rezultat klasifikacije prepoznati kao negativni dokumenti

- **FP** (eng. False Positives) = **Lažno Pozitivni (LP)**: predstavljaju broj dokumenata koji su na ulazu u klasifikaciju bili negativni ali kao rezultat klasifikacije prepoznati kao pozitivni dokumenti

- **FN** (eng. False Negatives) = **Lažno Negativni (LN)**: predstavljaju broj dokumenata koji su na ulazu u klasifikaciju bili pozitivni ali kao rezultat klasifikacije prepoznati kao negativni dokumenti

Matrica konfuzije se može napraviti i za višeklasnu klasifikaciju. Na slici 3.6 prikazan je slučaj za tri klase. U ovom slučaju, postoji mala razlika u načinu računanja TP, TN, FP i FN podataka. U zavisnosti za koju klasu se konstruiše matrica, TP podaci predstavljaju ukupan broj dokumenata koji na ulazu u klasifikaciju, ali i kao rezultat klasifikacije pripadaju klasi za koju se konstruiše matrica, dok TN podaci predstavljaju ukupan broj dokumenata koji na ulazu, ali i kao rezultat klasifikacije ne pripadaju klasi za koju se konstruiše matrica. FP podaci predstavljaju ukupan broj dokumenata koji na ulazu u klasifikaciju pripadaju ostalim klasama, ali kao rezultat klasifikacije pripadaju klasi za koju se konstruiše matrica, dok je za FN podatke obrnuto.

Veoma bitne ocene koje se mogu dobiti pomoću prethodno prikazanih podataka su:

- **Preciznost** (eng. Precision) koji predstavlja tačnost klasifikacije tj. koliki procenat podataka neke klase je ispravno klasifikovan. Računa se preko formule:

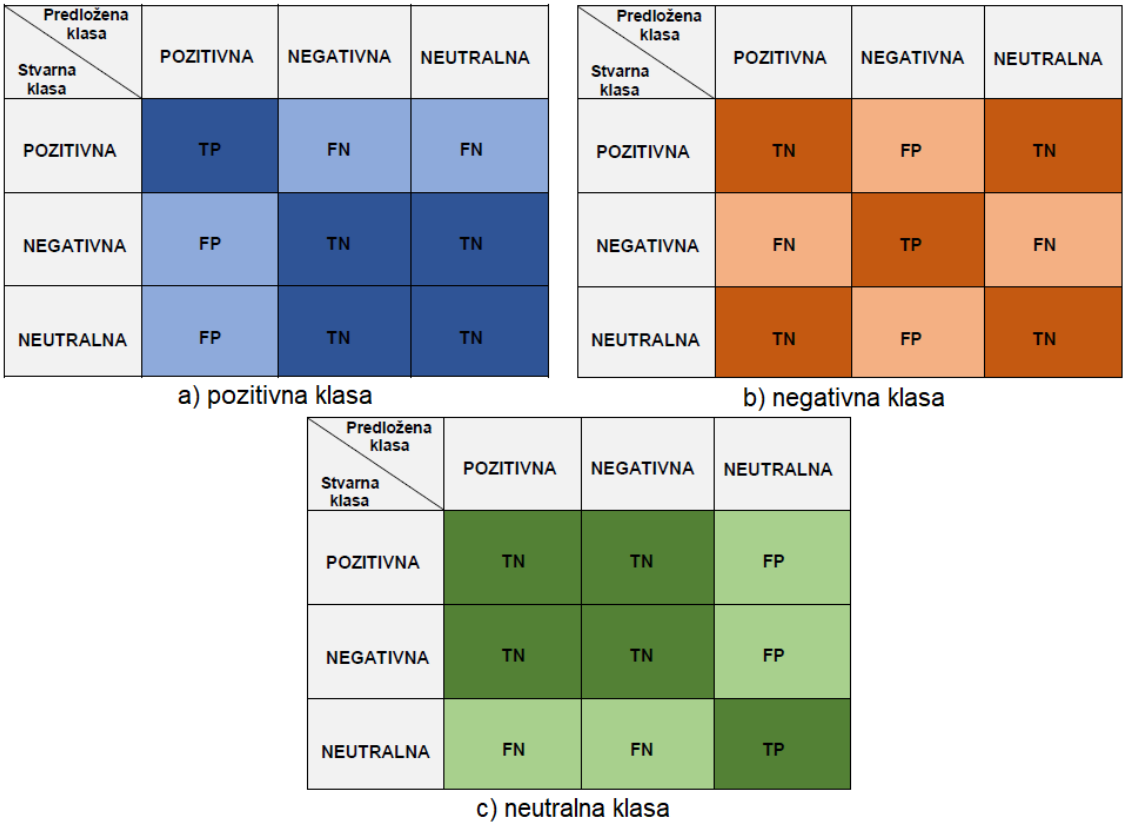
$$P = \frac{TP}{TP + FP}$$

- **Odziv** (eng. Recall) koji predstavlja koliko podataka neke klase klasifikator može da uoči. Računa se na sledeći način:

$$R = \frac{TP}{TP + FN}$$

- **F-mera** (eng. F-measure) predstavlja harmonijsku sredinu preciznosti i odziva.

$$F = \frac{2 * P * R}{P + R}$$



Slika 3.6: Matrica konfuzije - višeklasna klasifikacija (3 klase)

- **Tačnost** (eng. Accuracy) predstavlja procenat tačno klasifikovanih podataka od svih klasifikovanih podataka. Izražava se preko formule:

$$acc = \frac{TP + TN}{TP + TN + FP + FN}$$

U slučaju višeklasne klasifikacije, tada se može posmatrati kvalitet klasifikacije i na globalnom nivou. Naime, svakoj klasi se da podjednak značaj i izračunava se tzv. **makro-prosek** koji se računa tako što se izračunaju vrednosti TP, TN, FP i FN za svaku klasu, a pomoću njih zatim i preciznost, odziv, F-mera i tačnost za svaku klasu i na kraju izračuna njihova prosečna vrednost.

Primer izračunavanja makro-proseka za tri klase:

Prva klasa:

$$P_1 = \frac{TP_1}{TP_1 + FP_1}, R_1 = \frac{TP_1}{TP_1 + FN_1}, F_1 = \frac{2 * P_1 * R_1}{P_1 + R_1}, acc_1 = \frac{TP_1 + TN_1}{TP_1 + TN_1 + FP_1 + FN_1}$$

Druga klasa:

$$P_2 = \frac{TP_2}{TP_2 + FP_2}, R_2 = \frac{TP_2}{TP_2 + FN_2}, F_2 = \frac{2 * P_2 * R_2}{P_2 + R_2}, acc_2 = \frac{TP_2 + TN_2}{TP_2 + TN_2 + FP_2 + FN_2}$$

Treća klasa:

$$P_3 = \frac{TP_3}{TP_3 + FP_3}, R_3 = \frac{TP_3}{TP_3 + FN_3}, F_3 = \frac{2 * P_3 * R_3}{P_3 + R_3}, acc_3 = \frac{TP_3 + TN_3}{TP_3 + TN_3 + FP_3 + FN_3}$$

Makro-prosek:

$$P_{gl} = \frac{P_1 + P_2 + P_3}{3}, R_{gl} = \frac{R_1 + R_2 + R_3}{3}, F_{gl} = \frac{F_1 + F_2 + F_3}{3}, acc_{gl} = \frac{acc_1 + acc_2 + acc_3}{3}$$

Takodje, može se izračunati i **mikro-prosek** kod koga se veći značaj daje klasama koje sadrže veći broj dokumenata. Najpre se izračunavaju vrednosti TP, FN, FP i TN za svaku klasu pojedinačno. Zatim se izračunaju vrednosti TP', TN', FP' i FN' i to TP' kao suma svih TP, TN' kao suma svih TN itd. Na kraju se vrednosti mera izračunavaju za dobijene sumirane vrednosti TP', TN', FP' i FN'[2].

[2]Teorijski deo za ovo poglavlje preuzet je iz [1].

Glava 4

Razvijene metode sentiment klasifikacije

U ovom poglavlju će biti prikazene metode koje su razvijene u cilju rešavanja problema sentiment klasifikacije teksta. Pored toga, biće opisani i dostupni korpusi filmskih recenzija, kao i na koji način su same metode implementirane u programskom jeziku *Python*.

4.1 Metode zasnovane na leksičkim resursima

Za implementaciju metode zasnovane na leksičkim resursima na raspolaganju su leksički resursi kao što su srpski i engleski Wordnet i elektronski rečnik za srpski jezik, koji su detaljno opisani u poglavlju 2. Cilj ove metode je da se za svaki ulazni dokument na osnovu raspoloživih resursa odredi da li pripada pozitivnoj ili negativnoj klasi, tj. da li je u njemu iskazan pozitivan ili negativan sentiment

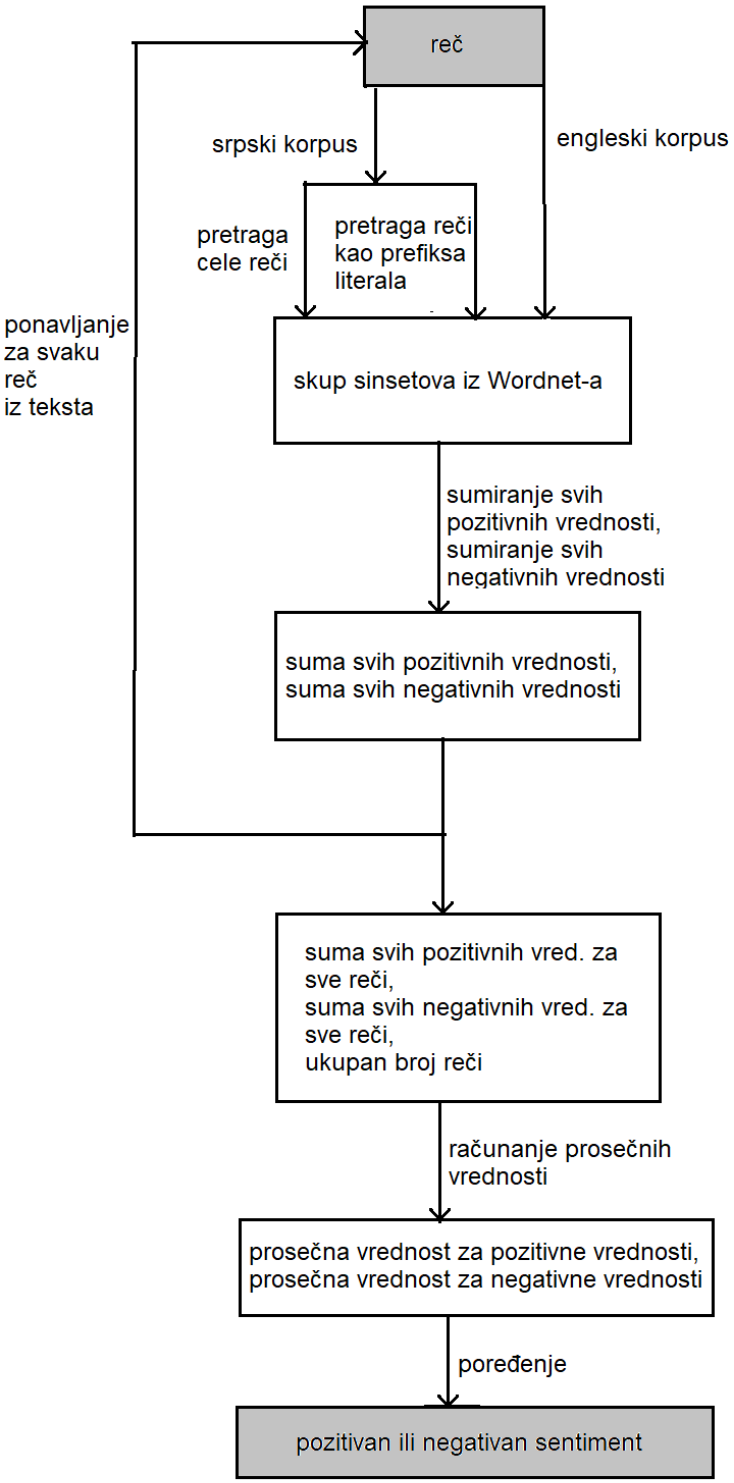
Preprocesiranjem su tekstovi spremni za glavni deo metode, a to je klasifikacija tekstova prema sentimentu. U narednom koraku potrebno je za svaku reč iz teksta pronaći sve sinsetove koji joj odgovaraju u Wordnet-u. Za tekstove i reči iz engleskog korpusa, neophodno je da se reč poklopi sa literalom iz sinseta i po obliku i po vrsti reči. Na taj način može se dobiti ceo skup sinsetova koji odgovaraju određenoj reči. Ovo je potrebno uraditi za svaku reč iz teksta koji je obrađen u fazi preprocesiranja. Nakon sumiranja pozitivnih i negativnih vrednosti koje u sebi sadrži svaki sinset, dobija se vrednost sentimenta za reč. Nakon obrade svih reči izračunava se i vrednost sentimenta za ceo tekst kao prosečna vrednost sentimenata svih reči iz tog teksta. Slično tome se računa i sentiment tekstova iz srpskog korpusa dokumenata,

s tim što mogu postojati varijacije, radi poboljšanja konačnih rezultata klasifikacije. Naime, umesto zahteva da se čitava reč poklapa sa literalom iz sinseta, može se zahtevati da literal počinje traženom rečju, tj. da reč bude prefiks literala iz sinseta. Ovo ima smisla jer npr. ukoliko je reč „užas", na ovaj način biće pronađeni i sinsetovi sa literalima: „užasavajući", „užasan", „užasnuti" koji imaju iste negativne sentimente, što može doprineti boljoj klasifikaciji teksta. Ovaj način pretraživanja Wordnet-a nije primenjen pri obradi engleskog korpusa. Ceo postupak prikazan je na grafiku 4.1. Kao sto se može videti, ne radi se razrešavanje problema višeznačnosti reči. Isto tako, za razvoj samog algoritma ne podrazumeva se postojanje labeliranih podataka. Kako su oni ipak dostupni, bice korišćeni u cilju testiranja razvijene metode. Prilikom analize rezultata dolazi se do zaključka da ima dosta pogrešno ili nelogično dodeljenih sentimenata u srpskom Wordnet-u, koji su uzrokovani pogrešno ili nelogično dodeljenim sentimentima u engleskom Wordnet-u. U cilju dobijanja boljih rezultata potrebno je ispraviti sentimente za sve literale za koje je to uočeno u okviru srpskog Wordnet-a. Značajno bi bilo videti kako se te ispravke odnose i na engleski Wordnet, ali to neće biti obrađeno u okviru ovog rada.

4.2 Hibridne metode

Metod nadgledanog učenja koji je korišćen u okviru hibridne metode, u ovom radu, jeste metod potpornih vektora (SVM). Za implementaciju metoda, potrebni su atributi kao i vrednosti ciljne promenljive. Neophodne podatke bilo je potrebno obraditi i smestiti u odgovarajuće strukture podataka koje su pogodne za dalju obradu. Deo podataka je potrebno odvojiti za treniranje, a jedan manji deo za testiranje ovog metoda. Pre toga, neophodno je kreirati odgovarajući model. Svaki tekst je klasifikovan i pripada pozitivnoj ili negativnoj klasi. Klase za svaki tekst iz korpusa poređane u listi predstavljaju vrednosti ciljne promenljive. Dakle, dužina liste jednaka je broju dokumenata u korpusu. Vrednosti za atribute potrebno je smestiti u retku matricu[1]. Prvo, za sve različite reči u celom korpusu kreira se rečnik, gde je za svaku reč sačuvan jedinstven broj i sentiment koji joj odgovara. Sentiment za svaku reč je izračunat korišćenjem prethodno opisanog metoda zasnovanog na leksičkim resursima. Nakon toga, za svaki tekst iz korpusa kreira se lista. Svaka lista sadrži parove iz prethodno kreiranog rečnika koji odgovaraju rečima u datom tekstu. U retkoj matrici svaka vrsta odgovara jednom tekstu. Broj kolona

[1]Retka matrica je matrica koja sadrži veliki broj nula.



Slika 4.1: Klasifikacija teksta prema sentimentu

je određen veličinom kreiranog rečnika. Prolazeći kroz listu parova koja odgovara nekom tekstu i upoređujući jedinstveni broj, kojim je reč određena u rečniku, i redni broj kolone popunjava se polje u matrici vrednošću sentimenta trenutne reči. Postupak se ponavlja sve dok se ne popuni cela matrica vrednostima iz kreiranih listi, na odgovarajućim mestima. Većina polja u matrici ostaće nepopunjena, tj. biće im vrednost 0. Na primer, retka matrica napravljena za korpus koji sadrži samo 2 recenzije prikazana je na slici 4.2. Na ovaj način dobijene su vrednosti za atribute. Nakon izračunatih atributa i vrednosti ciljne promenljive može se primeniti metod SVM. Kao rezultat izvršavanja ovog metoda dobijaju se ocene poput preciznosti, odziva, F-mere i tačnosti za dati korpus.

4.3 Korpusi filmskih recenzija

Razvijene metode za sentiment klasifikaciju biće testirane na korpusima filmskih recenzija na srpskom i engleskom jeziku.

Korpus filmskih recenzija na srpskom jeziku sadrži posebno obrađene komentare o filmovima koje su gledaoci postavljali na forumima kao sto su: *2kokice.com, filmskerecenzije.com, filmskihitovi.blogspot.com, happynovisad.com, kakavfilm.com, popboks.com, yc.rs i mislitemojomglavom.blogspot.com*. Postoje dve klase po kojima su klasifikovane date recenzije na osnovu ocena gledalaca (ocene su u rasponu od 1 do 10). Klasifikacija na klase (pozitivnu, negativnu) je izvršena tako da recenzije sa ocenama od 1 do 4 pripadaju negativnog klasi, a recenzije sa ocenama od 7 do 10 pozitivnoj klasi. Svaka od ovih klasa (pozitivna, negativna) sadrži po 841 recenziju. Recenzija sa najmanjim brojem reči ima samo 21 reč, dok recenzija sa najvećim brojem reči ima 1851 reč. Prosečna dužina recenzija u broju reči iznosi oko 480 reči. Korpus je dostupan u [17] i njegova veličina je 4.8MB [16].

Korpus filmskih recenzija na engleskom jeziku je direktorijum sastavljen od recenzija u obliku tekstualnih dokumenata. Ukupno 2000 filmskih recenzija je raspoređeno prema tome da li je u njima iskazan pozitivan ili negativan sentiment (po 1000 recenzija za svaku klasu - pozitivnu i negativnu). Recenzija sa najmanjim brojem reči ima samo 16 reči, a recenzija sa najvećim brojem reči ima 2366 reči. Prosečna dužina recenzija u broju reči iznosi oko 635 reči. Korpus je dostupan u [18] i njegova veličina je 7.4MB.

```
sparse_matrix =  [[0        0         0         0         0         0         0         0         0
                   0.5      0.3125    0         0         0         0         0.1875    0         0
                   0.25     0         0         0         0         -0.125    0         0         0
                   0.125    0         -0.125    0         0         0         0         0         0
                   0        0         0         0         0         0         0         0         0
                   0        0         -0.125    0         0         0         0         0         0
                   0        -0.00025  0         0.5       0.5       0.3125    -0.125    -0.25     0
                   0        0.25      0         0         0         0.75      0.125     0.00075   0
                   0        0         0         0         0         0         0         0         0
                   0        0         0         0         0         0         0         0         0
                   0        0         0         0         0         0         0         0.75      0
                   0        0         0         0         0         0         0         0         0
                   0.125    0         0         0.6667    0         0         0         0         0
                   0        0         0         0         0         0         0         0         0
                   0        0         0         0         0         -0.04167  0         0         0.0005
                   0        -0.625    0.3437    0.25      0         0         0         0         -0.3334
                   -0.25    0         0         0.00075   0         0         0         0         0
                   0        0         0         0         0         0         0         0         0
                   0        0         0         0         0         0         0         0         0
                   0        0         0         0         0         0         0         0         0
                   0        0         0         0         0         0         0         0         0
                   0        0         0         0         0         0         0         0         0
                   0        0         0         0         0         0         0         0         0
                   0        0         0         0         0         0         0         0         0
                   0        0         0         0         0         0         0         0         0
                   0        0         0         0         0         0         0         0         0
                   0        0         0         0         0         0         0         0         0
                   0        0         0         0         0         0         0         0         0
                   0        0         0         0         0         0         0         0         0
                   0        0         0         0         0         0         0         0         0
                   0        0         0         0         0         0         0         0         0
                   0        0         0         0         0         0         0         0         0
                   0        0         0         0         0         0         0         0         0
                   0        0         0         0         0         0         0         0         0
                   0        0         0         0         0         0         0         0         0
                   0        0         0         0         0         0         ]
                 [ 0        0         0         0         0         0         0         0         0
                   0        0         0         0         0         0         0         0         0
                   0        0         0         0         0         0         0         0         0
                   0        0         0         0         0         0         0         0         0
                   0        0         0         0         0         0         0         0         0
                   0        0         -0.125    0         0         0         0         0         0
                   0        0         0         0         0         0         -0.5      0         0
                   0        0         0         0         0         2.25      0         0.0015    0
                   0        0         0         0         0         0         0         0         0
                   0        0         0         0         0         0         0         0         0
                   0        0         0         0         0         0         0         0         0
                   0        0         0         0         0         0         0         0         0
                   0        0         0         0         0         0         0         0         0
                   0        0         0         0         0         0         0         0         0
                   0        0         0         0         0         0         0         0         0
                   0        0         0.34375   0         0         0         0         0         0
                   0        0         0         0         0         0         0         0         -0.1875
                   0        0         0         0         0         0         0.3125    0.3125    0
                   0        0         0         0         0         -1        0.5       0         0
                   0        0         -0.375    -0.125    0         0         0         0         0
                   0        0         0         0         0         0.375     0         0         0
                   0        0         0.5       0         0.25      0         0.125     0         0
                   0        0.41666   0         0         0         0.125     0         0         0
                   0        0         0         0         0         0         0         -0.25     0
                   0.25     0         0         0         0.125     0         0         0         0
                   0.125    0         0         0         0         0         0         0         0
                   0        -0.125    0         0         0         0.25      0         0         0
                   0        0         0         0         0         -0.5      0.125     0         0
                   0        0         0         0         0         0.125     0.125     -0.625    0
                   0        0         0         -0.25     0         0         0         0.375     0
                   0        0         0.125     0         0         0         0         0         0
                   0.00088  0         0         0         0         0         ]]
```

Slika 4.2: Retka matrica

4.4 Python i korisni paketi

Aplikacija za primenu leksičkih resursa u sentiment analizi teksta implementirana je u programskom jeziku *Python*. Kako Python u sebi sadrži puno ugrađenih paketa za rad u oblasti obrade prirodnih jezika, na raspolaganju je više opcija

za rešavanje problema sentiment analize teksta, bar za korpus na engleskom jeziku. U tu svrhu najvećim delom su korišćene funkcije iz paketa *nltk*, od kojih su najznačajnije funkcije za stop reči, lematizaciju. Wordnet koji se nalazi u okviru paketa *nltk* je Prinstonski Wordnet. Tako je iz paketa `nltk.corpus` korišćena funkcija za dobijanje skupa svih stop reči u engleskom jeziku po imenu `stopwords` preko poziva `set(stopwords.words('english'))`. Isto tako, iz ovog paketa, iskorišćena je i funkcija za tagove u Wordnet-u, pa su tako tagovi koji se pojavljuju: `ADJ, ADV, NOUN, VERB`. Iz paketa `nltk.stem` i njegove klase `WordNetLemmatizer` za lematizaciju je korišćena funkcuja sa pozivom `lemmatize(self, word, pos=NOUN)` i na taj način se za prosledjenu reč i tag koji joj odgovara dobija lema koja je potrebna. Za dobijanje tagovanog teksta korisna jeste kombinacija poziva funkcija `word_tokenize` i `pos_tag` iz paketa `nltk`. Prvo se pozivom funkcije `word_tokenize(text)` dobija tekst kao skup reči, a zatim se svakoj reči iz teksta dodeljuje tag vrste reči funkcijom `pos_tag`. Prethodno navedene funkcije i paketi korišćeni su, pre svega, za engleski korpus, dok su za srpski korpus korišćeni leksički resursi koji su opisani ranije u ovom radu. Još jedan veoma koristan paket za ovu aplikaciju jeste paket `sklearn`. Naime, u ovom paketu se nalaze razne funkcije za implementaciju metoda nadgledanog učenja. Kako je u ovoj aplikaciji korišćen i metod nadgledanog učenja - SVM, paket `sklearn` je poslužio za njegovu implementaciju preko poziva funkcija za pravljenje modela, primenu samog metoda i na kraju za prikazivanje rezultata klasifikacije preko funkcija `train_test_split`, zatim `SVC` i `classification_report`. Deo aplikacije koji se odnosi na pretraživanje Wordnet-a i izračunavanje sentimenata, implementiran je na isti način za oba korpusa, gde su veoma bitnu ulogu imale pomoćne metode koje se nalaze u okviru paketa: `os, csv, pandas, re, pathlib, string, transliterate, numpy`, kao što su npr. metode za učitavanje korpusa, parsiranje i slično. Pored toga, za neophodne testove korišćen je paket `unittest`.

4.5 Implementacija

Kao što je u prethodnom poglavlju i naglašeno, aplikacija za primenu leksičkih resursa u sentiment analizi teksta napisana je u *Python*-u. Javno je dostupna za preuzimanje na adresi ovde.

Svi potrebni podaci za aplikaciju (ulazni i izlazni) raspoređeni su u dva direktorijuma: `input_data` i `output_data` (nisu javno dostupni), dok je ceo kôd aplikacije smešten u direktorijumu `src`. Glavna skripta za pokretanje cele aplikacije - `main.py`

nalazi se unutar direktorijuma `src`. Pored toga, u njemu je smešten i direktorijum `util`, sa svim potrebnim funkcijama, direktorijum `tests` koji sadrži sve testove za implementirane funkcije i direktorijum `data` gde se nalaze engleski wordnet i dokument sa sinsetovima iz srpskog wordnet-a u kojima su pogrešno dodeljeni sentimenti. U direktorijumu `util` smeštene su sve skripte i klase koje su neophodne za izvršavanje aplikacije:

- `loader.py` - sadrži funkcije za učitavanje i parsiranje korpusa, pojedinačnih dokumenata, stop reči, ...
- `converter.py` - sadrži funkcije za različite konverzije: prevođenje sa ćiriličnog pisma na latinično, pretvaranje brojeva iz *float* u *string*, ...
- `constants.py` - sadrži konstante koje se često pojavljuju: *POSITIVE*, *NEGATIVE*, *NEUTRAL*
- `wordnet_helper.py` - sadrži sve potrebne funkcije za rad sa srpskim i engleskim Wordnet-om: smeštanje učitanih podataka iz Wordnet-ova u pogodne strukture podataka, preprocesiranje tekstova (ujednačavanje pisma, lematizacija, uklanjanje stop reči, ...), računanje sentimenta za svaku reč iz teksta, računanje sentimenta za svaki tekst u korpusu, računanje ocena za svaki korpus, ...
- `classifier_helper.py` - sadrži sve potrebne funkcije za implementaciju SVM klasifikatora: smeštanje podataka u odgovarajuće strukture podataka, kreiranje modela, klasifikacija SVM klasifikatorom

Glava 5

Rezultati

U ovom poglavlju će biti prikazani svi dobijeni rezultati nakon izvršene klasifikacije. Takođe, biće prikazane i razne varijacije koje su sprovođene radi poboljšanja samih rezultata. Na samom kraju biće izloženi i neki problemi koje su se našli kao smetnja za dobijanje zadovoljavajućih rezultata.

5.1 Rezultati za korpus na engleskom jeziku

Na slici 5.1 su prikazane ocene klasifikacije engleskog korpusa filmskih recenzija primenom metoda zasnovanih na leksičkim resursima.

PRECIZNOST	ODZIV	F1-MERA	TAČNOST
58.11%	85.20%	69.09%	61.90%

Slika 5.1: Ocena klasifikacije engleskog korpusa

Na slici 5.2 su prikazane ocene klasifikacije engleskog korpusa filmskih recenzija primenom hibridnih metoda, gde je udeo dela za testiranje 20%. Rezultati koji su prikazani dobijeni su pomoću klasifikacionog izveštaja nakon primene SVM klasifikatora[1].

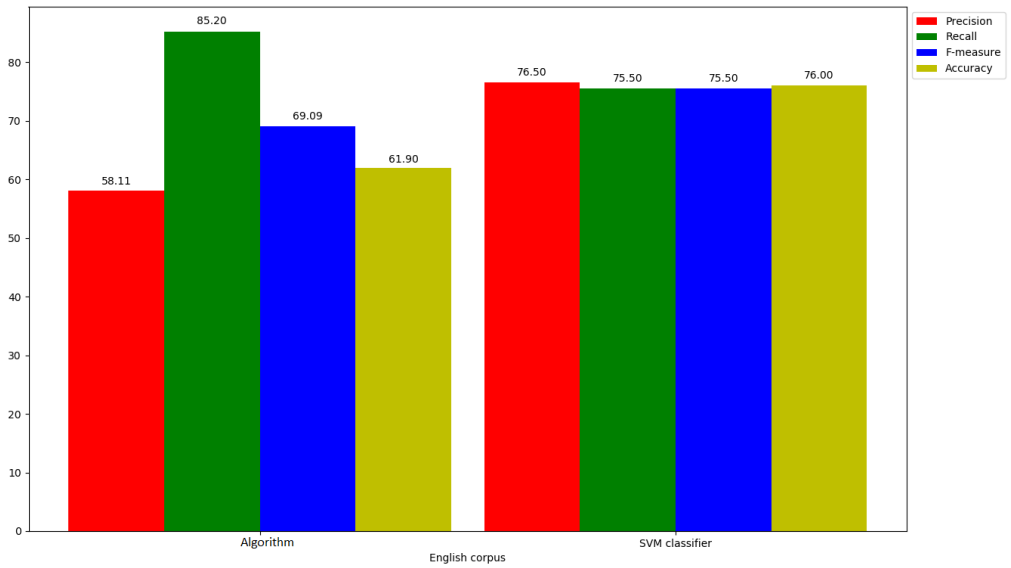
Na slici 5.3 mogu se videti grafički prikazani rezultati klasifikacije primenom oba pristupa (algoritamskog i hibridnog). Ono što se sa grafikona može videti i zaključiti

[1]Kolona *PODRŠKA* predstavlja koliki broj dokumenata iz dela za testiranje je nakon klasifikacije raspoređen u svakoj klasi.

	PRECIZNOST	ODZIV	F-MERA	PODRŠKA
POZITIVNA KLASA	74%	75%	75%	192
NEGATIVNA KLASA	77%	76%	76%	208
MIKRO PROSEK	76%	76%	76%	400
MAKRO PROSEK	75%	75%	75%	400
PROSEČNI UZORAK	76%	76%	76%	400

Slika 5.2: Ocena klasifikacije engleskog korpusa SVM klasifikatorom

jeste da su ocene koje su dobijene SVM klasifikatorom mnogo ujednačenije nego ocene dobijene primenom leksičkih resursa. Takodje, ove ocene su dosta bolje, što je i očekivano ponašanje, s tim što su i druge ocene dale zadovoljavajuće rezultate. Može se zaključiti da je klasifikacija primenom leksičkih resursa izvršena zadovoljavajuće za engleski korpus.



Slika 5.3: Engleski korpus - odnos ocena dobijenih različitim pristupima

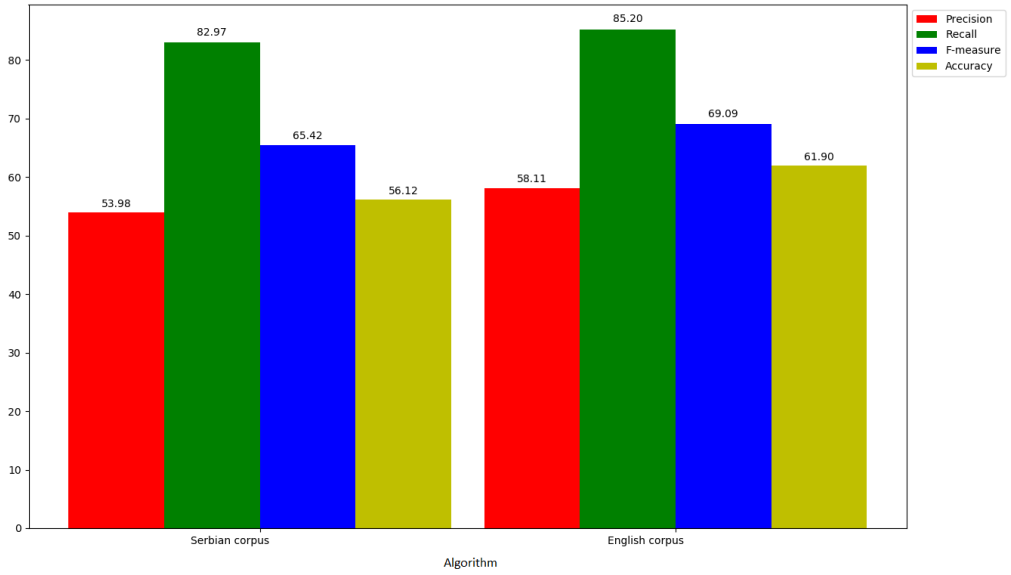
5.2 Rezultati za korpus na srpskom jeziku

Na slici 5.4 su prikazane ocene klasifikacije srpskog korpusa filmskih recenzija primenom metoda zasnovanih na leksičkim resursima.

PRECIZNOST	ODZIV	F1-MERA	TAČNOST
53.98%	82.97%	65.42%	56.12%

Slika 5.4: Ocena klasifikacije srpskog korpusa

Na slici 5.5 se može videti da su rezultati za srpski korpus za nijansu lošiji u odnosu na engleski korpus, ali ipak u granicama očekivanih (F-mera oko 65%).



Slika 5.5: Odnos dobijenih ocena algoritamskim pristupom nad srpskim i engleskim korpusom

Na slici 5.6 su prikazane ocene klasifikacije srpskog korpusa filmskih recenzija primenom hibridnih metoda, gde je deo za testiranje 20%. Rezultati koji su prikazani dobijeni su pomoću klasifikacionog izveštaja nakon primene SVM klasifikatora.

	PRECIZNOST	ODZIV	F-MERA	PODRŠKA
POZITIVNA KLASA	70%	66%	68%	180
NEGATIVNA KLASA	64%	68%	66%	157
MIKRO PROSEK	67%	67%	67%	337
MAKRO PROSEK	67%	67%	67%	337
PROSEČNI UZORAK	67%	67%	67%	337

Slika 5.6: Ocena klasifikacije srpskog korpusa SVM klasifikatorom

Na slici 5.7 može se videti grafikon sa grafički prikazanim rezultatima klasifikacije primenom oba pristupa (algoritamskog i hibridnog). I ovde se može videti da su rezultati dobijeni SVM klasifikatorom dosta ujedančeniji ali i lošiji u odnosu na rezultate za engleski korpus. To se delimično može objasniti time što je engleski Wordnet mnogo bogatiji od srpskog Wordnet-a i sadrži mnogo više sinsetova. U narednom poglavlju će više pažnje biti posvećeno upravo metodama kako da se prevaziđu pomenuti nedostaci u srpskom Wordnet-u.

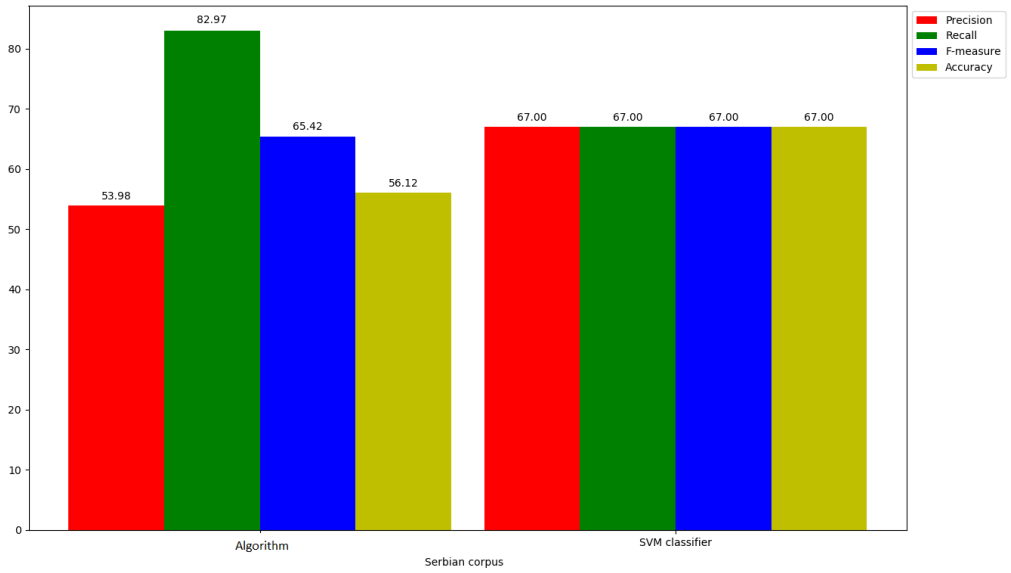
Problemi prilikom klasifikacije korpusa na srpskom jeziku

Neki od problema koji su se pojavili prilikom klasifikacije, najviše srpskog korpusa, a samim tim i uzroci loše dobijenih rezultata se mogu svrstati u nekoliko grupa:

- Nedostaci srpskog Wordnet-a
- Greške u srpskom korpusu dokumenata

Nedostaci srpskog Wordnet-a

Jedan od razloga zašto nisu dobijeni zadovoljavajući rezultati nakon klasifikacije jeste to što srpski Wordnet ima u sebi nedostataka, koji su svakako uzrokovani i nedostacima u engleskom Wordnet-u. Prvi nedostatak je to što ima **mnogo manje reči i sinsetova** nego engleski Wordnet, pa samim tim i neke reči koje se nalaze

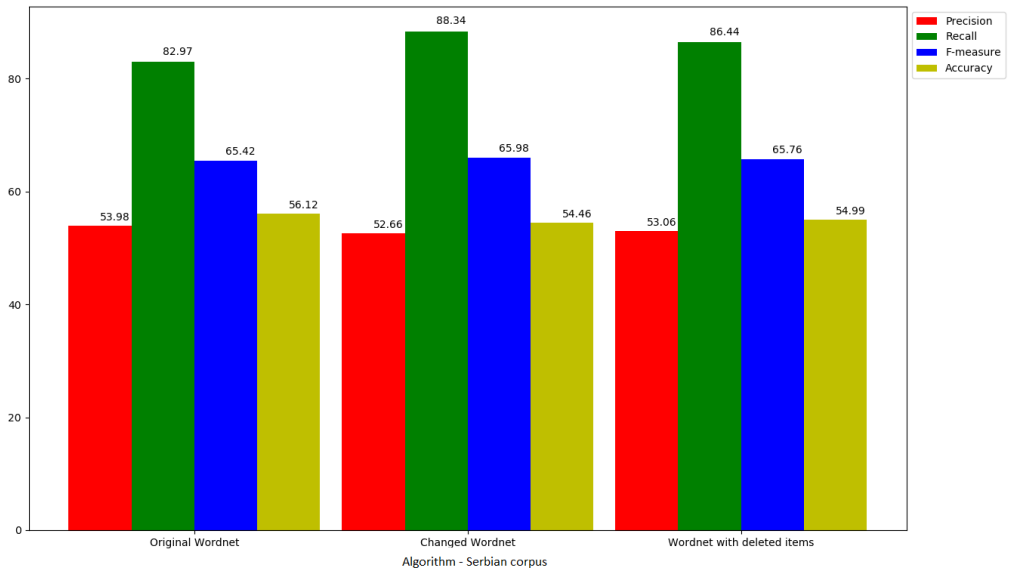


Slika 5.7: Srpski korpus - odnos ocena dobijenih različitim pristupima

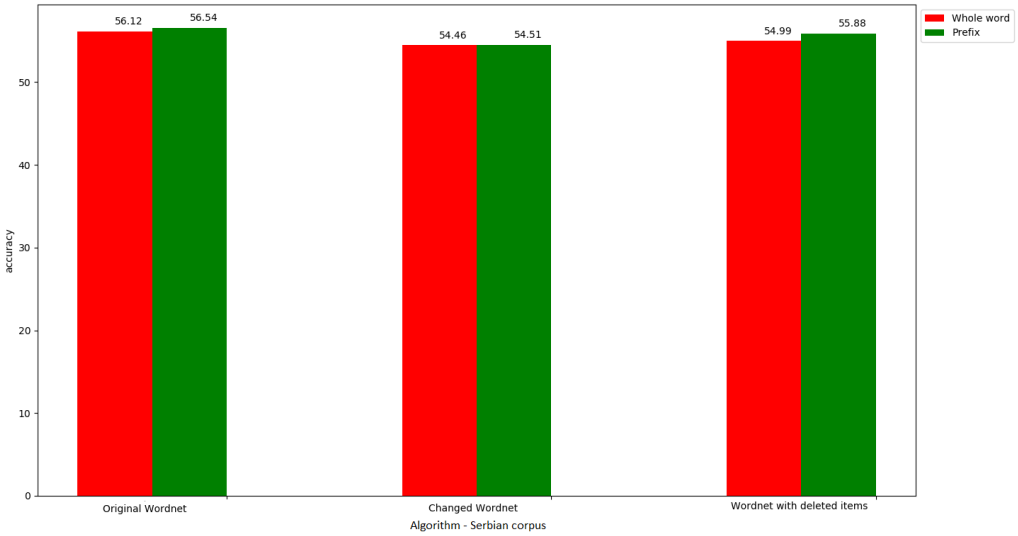
u dokumentima za klasifikaciju neće biti pronađene i klasifikovane. Naime, može se desiti da samo mali broj reči iz dokumenta uopšte bude pronađen i klasifikovan, a ostale reči se onda i ne uzimaju u razmatranje jer nemaju dodeljen sentiment (ni pozitivan ni negativan). Na taj način se može desiti da veoma bitne reči za klasifikaciju tog dokumenta uopšte ne budu razmatrane, pa onda i sam dokument na kraju bude loše klasifikovan, tj. bude mu dodeljena klasa kojoj inicijalno ne pripada. Jedino rešenje za ovaj nedostatak je dodavanje novih sinsetova u srpski Wordnet. Metoda koja je u ovom radu primenjena radi izbegavanja ovog problema je metoda u kojoj se menja pristup pretraživanja srpskog Wordnet-a. Naime, umesto kompletnog upoređivanja trenutne reči sa literalima iz Wordnet-a korišćen je pristup gde se posmatra da li literali počinju trenutnom rečju tj. da li je trenutna reč prefiks nekog literala u srpskom Wordnet-u.

Drugi nedostatak koji je primećen je da u srpskom Wordnet-u postoji dosta sinsetova u kojima je **sentiment loše dodeljen literalima**. Uzrok ovog nedostatka je i loše dodeljen sentiment literalima u engleskom Wordnet-u. Dešava se da se u sinsetu koji sadrži literale pozitivnog značenja nalazi veća numerička vrednost za negativan sentiment nego za pozitivan, kako bi trebalo da bude. Postoje slučajevi gde je obrnuto, tj. negativni literali imaju veću numeričku vrednost za pozitivan

sentiment nego za negativan, a trebalo bi da bude suprotno. Da bi se ovaj problem nekako prevazišao, ispravljene su ili uklonjene vrednosti za uočene literale sa greškama. U jednom slučaju su numeričke vrednosti koje su pogrešno dodeljene literalima zarotirane i postavljene na svoje mesto tj. na svim mestima gde je potrebno zamenjene su vrednosti za pozitivan i negativan sentiment u sinsetu. U drugom slučaju, svi literali koji su imali loše dodeljene sentimente su obrisani, zajedno sa sentimentom. Ove izmene u srpskom Wordnet-u takođe su korišćene u klasifikaciji dokumenata iz srpskog korpusa. Dokument sa pronađenim sinsetovima u kojima je sentiment loše dodeljen literalima iz sinseta dostupan je ovde. Dobijeni rezultati, zajedno sa rezultatima korišćenjem originalnog srpskog Wordnet-a, prikazani su na slici 5.8. Sa datog grafikona možemo videti da su se rezultati za nijansu poboljšali i da su najbolji rezultati dobijeni korišćenjem izmenjenog Wordnet-a, a najlošiji korišćenjem originalnog Wordnet-a. Na slici 5.9 su prikazane vrednosti za ocenu tačnosti za kombinaciju prethodno opisanih pristupa za prevazilaženje problema. Ono što se može zaključiti sa grafikona jeste da ova kombinacija ne daje znatno poboljšanje rezultata, ali poboljšanja ipak ima za nijansu. Takođe se može zaključiti da u ovom slučaju originalni Wordnet daje najbolje rezultate.



Slika 5.8: Srpski korpus - odnos ocena dobijenih korišćenjem različitih verzija Wordnet-a



Slika 5.9: Srpski korpus - odnos ocena tačnosti dobijenih korišćenjem kombinacija više pristupa

Ima prostora za dalje unapređenje i usavršavanje srpskog Wordnet-a. Kako je i predstavljeno u radu [7] moguće je, na primer, za validaciju i usavršavanje srpskog Wordnet-a koristiti razne tekstualne leksičke resurse kao što su odgovarajući korpusi i rečnici.

Greške u srpskom korpusu dokumenata

Još jedan problem koji se javio prilikom klasifikacije dokumenata iz srpskog korpusa je to što u srpskom korpusu postoji dosta sintaksnih grešaka. Kako srpski korpus predstavlja skup filmskih recenzija preuzetih sa različitih sajtova za ocenjivanje filmova, moguće je očekivati da one sadrže različite slovne i druge jezičke greške. Filmske recenzije su u neizmenjenom obliku korišćene za klasifikaciju pa svaka jezička greška koja se u njima nađe utiče na ishod klasifikacije. Na mnogim mestima je uočeno da su u rečima izostavljena slova, duplirana slova ili čak zamenjena nekim slovom koje tu ne pripada. Naravno, takve reči kasnije neće biti pronađene u Wordnet-u pa samim tim i neće uticati na konačan sentiment celog teksta, što značajno utiče na tačnost klasifikacije. Na nekim mestima, pak uočeno je da su susedne reči u tekstovima spojene u jednu reč. Kao takva reč, onda ona nema nikakvog smisla, a nema ni uticaja na klasifikaciju, kao u prethodno navedenoj

situaciji. Neki primeri tako spojenih reči koje se pojavljuju su: *„najboljianalitičar”, „izvršinapad”, „jeunapređen”, „dalekouspešniji”, „smrtiurađeno”, „izazivatismejanje”, „lošeodrađeno”.*

Prethodno navedeni problemi vezani za srpski korpus dokumenata su svakako nešto što se može očekivati i sa čim se treba pomiriti jer je u pitanju korpus koji nije prerađivan i dorađivan da bi bio u perfektnom stanju za klasifikaciju već je korišćen takav kakav jeste. Samim tim se i ne mogu očekivati rezultati na željenom nivou. Slični problemi postoje i u engleskom korpusu, ali oni nisu analizirani jer to nije tema ovog rada.

Glava 6

Zaključak

U današnje vreme, sve veći broj javnih diskusija odvija se na društvenim mrežama i drugim interaktivnim platformama. Takođe, veliki uticaj na donošenje odluka korisnika o kupovini nekog proizvoda, imaju komentari i ocene na forumima i portalima. Primer korisnički generisanog sadržaja u kojima korisnici iskazuju svoja mišljenja i iskustva su recenzije na raznim portalima. Ovakav način prikazivanja podataka predstavlja vredan izvor informacija, a samim tim dovodi do potrebe za analiziranjem i klasifikacijom istih, radi njihove dalje upotrebe. Upravo to je i zadatak sentiment analize, razvrstati tekstove (recenzije) po raznim kategorijama na osnovu mišljenja koje je u njima iskazano.

U ovom radu je prikazano na koji način je izvršena klasifikacija filmskih recenzija datih u korpusima na srpskom i engleskom jeziku na pozitivne i negativne, kao i na koji način leksički resursi utiču na klasifikaciju ovih recenzija.

Nakon uvodnog dela i upoznavanja sa temom ovog rada kao i njegovim ciljem, data su detaljna objašnjenja svih algoritama koji su korišćeni u sentiment analizi. Korišćena su dva različita pristupa u rešavanju istog problema, a to su algoritamski (metode zasnovane na leksičkim resursima), i hibridni pristup (kombinacija metoda zasnovanog na leksičkim resursima i metoda mašinskog učenja). Za oba pristupa potrebno je bilo koristiti elektronski rečnik za srpski jezik, kao i srpski i engleski Wordnet. Nakon izvršenog metoda zasnovanog na leksičkim resursima, obrađeni i u odgovarujućem formatu prosleđeni su podaci kao ulazni atributi metodi mašinskog učenja (u ovom slučaju SVM). Kako svaki jezik ima svoje specifičnosti, u razvoju sentiment analize u aplikacijama veoma je bitno obratiti pažnju i na specifičnosti samog jezika na kom su tekstovi za klasifikaciju napisani kao i na koji način se one odražavaju na preprocesiranje podataka. Kako je sprski jezik jedan od morfološki

bogatijih jezika, elektronski rečnik za srpski jezik upotrebljen je u te svrhe za dokumente na srpskom jeziku. Leksički resursi koji su korišćeni bili su od velike važnosti za samu klasifikaciju. Pored toga, imali su uticaja i na dobijene rezultate. Međutim, uočeni su neki njihovi nedostaci, gde se vidi mogućnost za njihovo poboljšanje, a samim tim i poboljšanje samog procesa klasifikacije. Doprinos ovog rada može se videti i u tome što su neki delovi iz srpskog Wordnet-a, u kojima su uočene greške i nelogičnosti, ispravljene ili obrisane, upravo radi poboljšanja klasifikacije dokumenata na srpskom jeziku.

Zaključak koji se može izvesti na kraju ovog rada jeste da kvalitet klasifikacije pomoću sentiment analize teksta umnogome zavisi od kvaliteta leksičkih resursa koji su na raspolaganju, kao i od samih korpusa koji se obrađuju. Prostor za napredak postoji u poboljšanju srpskog Wordnet-a, njegovom dopunjavanju, pored ispravljanja postojećih nedostataka, kao i sređivanju korpusa pre obrade. Takođe, rad se može nadograditi primenom više metoda mašinskog učenja.

Bibliografija

[1] Jelena Graovac. „Prilog metodama klasifikacije teksta: matematički modeli i primene". *Doktorska disertacija*. Matematički fakultet Univerziteta u Beogradu, Beograd, 2014.

[2] Mladen Nikolić i Anđelka Zečević. „Mašinsko učenje". *Skripta*. Matematički fakultet Univerziteta u Beogradu, Beograd, 2019.

[3] Predrag Janičić i Mladen Nikolić. „Veštačka inteligencija". *Skripta*. Matematički fakultet Univerziteta u Beogradu, Beograd, 2020.

[4] Olivera Grljević. „Sentiment u sadržajima sa društvenih mreža kao instrument unapređenja poslovanja visokoškolskih institucija". *Doktorska disertacija*. Univerzitet u Novom Sadu, Ekonomski fakultet, Subotica, 2016.

[5] Jelena D. Mitrović. „Elektronski jezički resursi i alati za obradu srpskog jezika i njihovo unapredjivanje putem modela grupne raspodele rada". *Doktorska disertacija*. Filološki fakultet Univerziteta u Beogradu, Beograd, 2018.

[6] Duško Vitas i Cvetana Krstev. „Srpski jezik i SNTPI". Matematički fakultet Univerziteta u Beogradu, Beograd, 2009.

[7] Cvetana Krstev, Gordana Pavlović-Lažetić, Duško Vitas, and Ivan Obradović. "Using Textual and Lexical Resources in Developing Serbian Wordnet". Romanian Journal of Information Science and Technology, Publishing House of the Romanian Academy, 2004.

[8] Ranka Stanković, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Škorić. "Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian". Marseille, France, 2020.

[9] Cvetana Krstev, Ranka Stanković, Duško Vitas, and Ivan Obradović. "The Usage of Various Lexical Resources and Tools to Improve the Performance of Web Search Engines". Proceedings of the Sixth Interantional Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, 2008.

[10] Miljana Mladenović, Jelena Mitrović, and Cvetana Krstev. "Developing and Maintaining a WordNet: Procedures and Tools". Proceedings of Seventh Global WordNet Conference 2014, eds. Heili Orav, Christiane Fellbaume, Piek Vossan, University of Tartu, Tartu, Estonia, pp. 55-62, 2014, ISBN 978–9949–32–492–7, January 25-29, 2014.

[11] Branislava Šandrih, Cvetana Krstev, and Ranka Stanković. "Two approaches to compilation of bilingual multi-word terminology lists from lexical resources". Natural Language Engineering, Cambridge University Press, 2020.

[12] Miljana Mladenović, Jelena Mitrović, Cvetana Krstev, and Duško Vitas. "Hybrid sentiment analysis framework for a morphologically rich language." Journal of Intelligent Information Systems 46.3 (2016): 599-620.

[13] Christos Troussas and Maria Virvou. "Advances in Social Networking-based Learning: Machine Learning-based User Modelling and Sentiment Analysis". Intelligent Systems Reference Library 181, 2020.

[14] Francis Bond and Ryan Foster. "Linking and Extending an Open Multilingual Wordnet". Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 2013.

[15] Yiming Yang and Thorsten Joachims. "Text categorization". Scholarpedia, 2008.

[16] Vuk Batanović, Boško Nikolić, and Milan Milosavljević. "Reliable Baselines for Sentiment Analysis in Resource-Limited Languages: The Serbian Movie Review Dataset". Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, 2016.

[17] *Korpus na srpskom jeziku.* `https://github.com/vukbatanovic/SerbMR`.

[18] *Korpus na engleskom jeziku.* `http://www.cs.cornell.edu/people/pabo/movie-review-data`.

[19] Princeton University. *About WordNet.* `https://wordnet.princeton.edu/`. 2010.

[20] *George Armitage Miller*. `https://en.wikipedia.org/wiki/George_Armitage_Miller`. 2020.

[21] *The Global WordNet Association (2010-02-04)*. `http://globalwordnet.org/`. 2014.

Biografija autora

Jelena Ćosić rođena je 25. februara 1994. godine u Aleksandrovcu. Osnovnu školu „Ivo Lola Ribar" završila je u Aleksandrovcu kao nosilac diplome „Vuk Karadžić". Gimnaziju pri srednjoj školi „Sveti Trifun" u Aleksandrovcu završila je sa odličnim uspehom 2013. godine kada je i upisala Matematički fakultet, smer „Računarstvo i informatika". Osnovne studije na Matematičkom fakultetu završila je 2018. godine, kad je svoje školovanje nastavila na master studijama Matematičkog fakulteta, takođe na istom smeru. Od početka 2018. godine zaposlena je u IT industriji kao programer. Osnovne oblasti njenog interesovanja su programski jezik Java i relacione baze podataka. Hobi joj je učenje španskog jezika i treniranje latinoameričkih plesova.