UNIVERZITET U NIŠU
ELEKTRONSKI FAKULTET

Adela B. Ljajić

Obrada negacije u kratkim neformalnim tekstovima u cilju poboljšanja klasifikacije sentimenta

DOKTORSKA DISERTACIJA

Niš, 2019.

UNIVERSITY OF NIŠ
FACULTY OF ELECTRONIC ENGINEERING

Adela B. Ljajić

Processing Negation in Short Informal Texts for Improving the Sentiment Classification

DOCTORAL DISSERTATION

Niš, 2019.

Podaci o doktorskoj disertaciji

Mentor:	Prof. dr Suzana Stojković Univerzitet u Nišu, Elektronski fakultet
Naslov:	Obrada negacije u kratkim neformalnim tekstovima u cilju poboljšanja klasifikacije sentimenta
Rezime:	U ovoj disertaciji je dat predlog unapređenja metode za klasifikaciju kratkih neformalnih tekstova po sentimentu. Poboljšanje se pre svega zasniva na obradi pravila sintaksičke negacije u srpskom jeziku. Složenost gramatike srpskog jezika nameće potrebu da se fenomenu negacije u tekstovima na srpskom jeziku pristupi sistematski i da se pri njenoj obradi koriste i jezički resursi koji učestvuju u kreiranju pravila za obradu negacije. Korišćeni resursi su signali negacije, odrečni kvantifikatori, pojačivači negacije i neutralizatori negacije. Pored jezičkih resursa za primenu pravila negacije, pri klasifikaciji sentimenta korišćen je i opšti rečnik sentimenata. Evaluacija korišćene metode je rađena nad skupom tvitova na srpskom jeziku. Za evaluaciju su korišćene dve vrste metoda: metoda koje se zasniva na rečniku sentimenata kao i metoda bazirana na nadgledanom mašinskom učenju. Prezentovana metoda je u oba slučaja poređena sa dve osnovne (poredbene): prva koja ne obrađuje negaciju i druga koja obrađuje negaciju ali bez prezentovanih pravila za obradu sintaksičke negacije. U slučaju kada se primeni metoda zasnovana na rečniku sentimenata, tačnost klasifikacije je znatno veća u odnosu na dve poredbene metode a relativna poboljšanja ove metode u odnosu na prvu poredbenu metodu (koja ne obrađuje negaciju) su sledeća: za ceo skup tvitova 10.62%, za skup tvitova koji sadrže negaciju 26.63% i za skup tvitova koji sadrži negacije koje su obrađene korišćenim pravilima čak 31.16%. Kada se koristi metoda mašinskog učenja dobija se veća tačnost klasifikacije nego u slučaju metode zasnovane na rečniku sentimenata: za tri klase do 69.76% i za dve klase do 91.15%. Međutim, metodom mašinskog učenja se dobijaju manje vrednosti poboljšanja: za tri klase do 2.65% i za dve klase do 1.65%. Rezultati

	pokazuju statistički značajno poboljšanje ako se detektovana pravila negacije uključe u metodu klasifikacije kratkih neformalnih tekstova po sentimentu.
Naučna oblast:	Elektrotehničko i računarsko inženjerstvo (Računarstvo i informatika)
Naučna disciplina:	Procesiranje prirodnih jezika; Obrada teksta
Ključne reči:	analiza sentimenta, analiza teksta, detekcija negacije, pravila negacije, srpski jezik, Tviter, mašinsko učenje
UDK:	(004.738.5+004.774):(81'322+811.163.41)
CERIF klasifikacija:	P176: Veštačka inteligencija
Tip licence Kreativni zajednice:	**CC BY-NC-ND**

Data on Doctoral Dissertation

Doctoral Supervisor:	PhD Suzana Stojković, associate professor University of Niš, Faculty of Electronic Engineering
Title:	Processing Negation in Short Informal Texts for Improving the Sentiment Classification
Abstract:	In this dissertation, the method for classifying short informal texts by sentiment was proposed. The improvement was achieved by processing the rule of syntactic negation in the Serbian language. The complexity of the grammar of the Serbian language imposes the need to systematically approach the phenomena of negation and to use the linguistic resources involved in the creation of rules for the negation treatment in its processing. The resources used are negation signals, negative quantifiers, negation intensifiers, and negation neutralizers. In addition to language resources for the application of the rules of negation, the general sentiment lexicon of positive and negative terms was used in the classification by sentiment. The evaluation of the used method was performed over a set of tweets in Serbian. Lexicon based method, as well as the supervised method of machine learning, were used for evaluation. The method presented in both cases is compared with two baseline methods: the first one that does not process the negation and the other that processes the negation, but without the rules for processing a syntactic negation. In the case where a method based on sentiment lexicon was used, the accuracy of the classification is considerably higher in relation to the two baseline methods, and the relative improvements of this method with respect to the first baseline method are the following: for the entire dataset - up to 10.62%, for a set of tweets containing negation - up to 26.63% and for a set of tweets containing negations that were processed using the rules - up to 31.16%. When using the machine learning method, higher accuracy of the classification is obtained than in the case of the lexicon-based method: for three classes - up to 69.76% and for two classes - up to 91.15%. However, the method of machine learning produces fewer improvements: for three classes up to 2.65% and for

two classes up to 1.65%. The results showed a statistically significant improvement if the detected rules of negation are included in the short informal text classification method by sentiment. The results showed a statistically significant improvement if the detected rules of negation are included in the short informal text classification method by sentiment.

Scientific Field: Electrical and Computer Engineering (Computer Science)

Scientific Discipline: Natural Language Processing; Text mining

Key Words: sentiment analysis, text mining, negation detection, negation rules, Serbian language, Twitter, machine learning

UDC: (004.738.5+004.774):(81'322+811.163.41)

CERIF Classification: P176: Artificial intelligence

ZAHVALNICA

Ova disertacija predstavlja rezultat višegodišnjeg istraživanja i pronalaženja mogućnosti za rešavanje problema gde najveću zahvalnost dugujem prof. dr Mileni Stanković. Njena, pre svega, prijateljska a posle i stručna pomoć su bile veliki pokretač rada na disertaciji i motivacija i ohrabrenje u trenucima kada je bilo najteže. Veliku zahvalnost dugujem mentoru prof. dr Suzani Stojković, na konstruktivnim sugestijma i predlozima koji su mi pomogli kod pisanja disertacije. Hvala članovima komisije za ocenu i odbranu disertacije. Njihova pitanja i predlozi su mi pomogli da rešim nedoumice i otvorili nove vidike za dalje unapređenje metode koja je u disertaciji korišćena.

Zahvaljujem se kolegama Državnog univerziteta u Novom Pazaru sa studijskog programa Informatika i Matematika i studijskog programa Računarska tehnika, posebno koleginici doc. dr Ulfeti Marovac na prijateljskoj i stručnoj pomoći i idejama i sugestijama koje mi je davala u ključnim momentima izrade disertacije.

Ova disertacija nije samo moj uspeh, već i uspeh svih koji su bili u mojem okruženju jer su posredno uticali na moj rad. Zahvaljujem se mojem suprugu na motivaciji, ljubavi i podršci koja mi je dala snagu da završim ovaj projekat. Hvala mojoj majci na životnoj podršci i ljubavi, kao i ostaloj velikoj porodici na razumevanju koje mi je mnogo značilo.

Disertaciju posvećujem svojoj deci, da im bude pokretač za sve što budu radili i potvrda o tome koliko me je ljubav prema njima motivisala da završim projekat na koji će i oni, nadam se, biti ponosni.

SADRŽAJ

Spisak slika

Spisak tabela

1. UVOD

1.1. Motivacija

Dostupnost i široka uptreba interneta dovode do raspoloživosti ogromne količine podataka na društvenim mrežama, blogovima, forumima i interenet sajtovima. Broj korisnika koji ovakve podatke kreiraju i koriste je u stalnom porastu pa to dovodi do stalnog porasta količine dostupnih podataka. Ove korisničke podatke je moguće iskoristiti u različite svrhe. Istraživanje podataka (*eng. data mining*) je postupak kojim se sirovi podaci svih vrsta obrađuju, analiziraju i kreiraju informacije koje sadrže odgovarajuće znanje. Istraživanje teksta (*eng. text mining*) predstavlja analizu teksta na takav način da se od teksta kreiraju podaci čijom obradom se može doći do kvalitativnih informacija. Istraživanje teksta je mnogo više od običnog pretraživanja teksta jer nastoji da otkrije nepoznate, skrivene veze iz teksta.

U poslednjih desetak godina, uspešne aplikacije za obradu teksta na prirodnim jezicima su postale deo svakodnevnog iskustva ljudi. Primeri su aplikacije za ispravljanje pravopisa i gramatike reči kod procesora reči, mašinsko prevođenje na veb-u, otkrivanje e-mail spama, automatsko odgovaranje na pitanja, otkrivanje mišljenja ljudi o proizvodima ili uslugama i izdvajanje obaveza zabeleženih u e-mailu.

Analiza sentimenta (*eng. Sentiment Analysis, SA*) je podoblast obrade prirodnog jezika (*eng. Natural language processing, NLP*), a u literaturi se sreću još i nazivi: izdvajanje mišljenja, semantička orijentacija. U ovoj disertaciji je analiziran sentiment kratkih neformalnih tekstova - tvitova. Sa obzirom da korisnici socijalnih mreža često javno iznose svoje stavove, mišljenja i osećanja, ovakav tip teksta predstavlja dobar izvor podataka za analizu sentimenta - stava iznetog u njima. Poseban izazov u ovom slučaju predstavljaju tvitovi, koji su po prirodi ograničene dužine (u vreme sakupljanja podataka je tekst tvita bio ograničen na 140 karaktera). Ovako kratke tekstove je teže klasifikovati na bilo koji način pa je i određivanje sentimenta zahtevnije nego u slučaju dugačkih tekstova. Kratki tekstovi

sadrže sažetije informacije pa je ljudima lakše da utvrde sadržaj i poentu ovako kratkih tekstova nego u slučaju dužih tekstova. Međutim, kompjuterska obrada, a pre svega klasifikacija kratkih tekstova je zahtevnija nego u slučaju dužih tekstova. Poteškoće kod klasifikacije ovakvih tekstova se pre svega odnose na problematiku odabira pravih atibuta (*eng. features*) za uspešnu klasifikaciju. Neki od razloga otežane klasifikacije kratih tekstova su sledeći:

- nedovoljno pojavljivanje reči kao u dužim tekstovima pa je nemoguće koristiti metrike kao što su uzajamno pojavljivanje reči,
- oslanjanje na kontekst – nedostatak informacija o kontekstu,
- pristustvo ironije i sarkazma,
- neformalno pisanje i korišćenje slenga i
- često korišćenje skraćenica.

Analiza sentimenta se može primenjivati kod različitih zadataka kao što su određivanje pozitivnih ili negativnih recenzija, određivanje zadovoljstva kupca nekim proizvodom, predviđanje prodaje nekog proizvoda na osnovu sentimenta komentara o njemu. Analiza sentimenta je aktuelna oblast i njena aktuelnost raste. Razvojem tehnika za mašinsko procesiranje prirodnih jezika i algoritama za klasifikaciju, došlo je do velikog progresa u ovoj oblasti. Mišljenja ljudi izražena i zapisana konkretnim jezikom nije lako analizirati i odrediti im sentiment, s obzirom na kompleksnost jezičkih izraza i različitog načina izražavanja. Pomenute poteškoće kod klasifikacije kratkih tekstova nameću potrebu posebne pripreme ovakvih tekstova za klasifikaciju kao i posebno biranje atributa kojima će se kratki tekst predstaviti i uspešno klasifikovati.

Sriram i saradnici su u [1] klasifikovali tvitove na unapred definisani skup klasa: vesti, događaji, mišljenja, ponude i privatne poruke. Kao atribute za klasifikaciju su koristili: nominalni atribut (podatak o autoru) i sedam binarnih atributa o prisustvu odgovarajućih karakteristika u tekstu: skraćenice, korišćenje slenga, reči kojima se označava sentiment, znake valute i procenata i sl.

O karakteristikama kratkih tekstova i teškoćama pri klasifikovanju istog su diskutovali Song i saradnici u [2]. U radu su predstavili popularne modele klasifikacije kratkih tekstova: semantičku analizu, metodu klasifikacije polu-nadgledanim učenjem, kombinovanu

klasifikaciju i klasifikaciju u realnom vremenu. Uradili su analizu evaluacije ovih modela i predložili trendove u klasifikaciji kratkih tekstova.

1.2. **Predmet i cilj naučnog istraživanja**

Istraživanja koja su predstavljena u ovoj doktorskoj disertaciji obuhvataju analizu sentimenta tvitova na srpskom jeziku sa posebnim naglaskom na obradu pravila sintaksičke negacije. Poboljšanje klasifikacije po sentimentu u odnosu na do sada postignute rezultate iz ove oblasti je traženo primenom gramatičkih pravila sintaksičke negacije u srpskom jeziku. Obradom ovih pravila negacije i njihovom integracijom u metod klasifikacije tvitova po sentimentu, može se poboljšati tačnost klasifikacije sentimenta kratkih tekstova – u ovom slučaju tvitova. Pravila sintaksičke negacije predstavljaju podskup gramatičkih pravila za obradu negacije u srpskom jeziku. Primena ovih pravila zahteva posebno normalizovan tekst i postojanje jezičkih resursa za obradu negacije.

Evaluacija detektovanih pravila negacije je rađena pomoću metode za klasifikaciju teksta po sentimentu, tako što su izdvajane karakteristike (atributi) teksta koji detektuju prisustvo ili odsustvo negacije. Ovi atributi služe za klasifikaciju različitim metodama i ti rezultati su poređeni. Za analizu sentimenta su u dosadašnjim istaživanjima karakteristična tri generalna pristupa:

- Analiza metodama zasnovanim na korišćenju rečnika sentimenata
- Analiza bazirana na metodama mašinskog učenja
- Kombinovana analiza metodama baziranim na korišćenju rečnika sentimenata i metodama mašinskog učenja

U disetraciji su korišćene metode klasifikacije bazirane na rečniku sentimenata i bazirane na mašinskom učenju.

Cilj je poboljšanje metode klasifikacije tvitova pre svega obradom pravila negacije u srpskom jeziku. Pokazano je da se obradom ovih pravila povećava tačnost klasifikacije u odnosu na metode koje ne obrađuju negaciju ili je obrađuju heuristički.

Kako bi se podaci pripremili za klasifikaciju, korišćena je normalizacija koja je specifična za ovakav tip teksta (tvit). Od leksičkih resursa su korišćeni rečnik reči kojima se

izražava sentiment, rečnik stop reči, rečnik specifičnih vrsta reči za koje postoje pravila upotrebe u rečeničnim konstrukcijama i koja će biti obrađena kod procesiranja negacije. Dodatna poboljšanja su dobijena kreiranjem atributa teksta koji nose kritičnu količinu informacija. Ovi atributi su dobijeni tehnikama selekcije i redukcije atributa. Ovako kreirana metoda predstavlja originalnu metodu za obradu negacije kod određivanja sentimenta tvitova na srpskom jeziku.

Jedan od ciljeva je bio da se obezbedi skup obeleženih tvitova, zajedno sa ostalim jezičkim resursima koji su kreirani u toku izrade ove disertacije. Pored kreirane napredne metode, plan je da ovi resursi budu javno dostupni za istraživanje. Takođe, planira se i njihovo proširenje u saradnji sa zainteresovanim stranama.

1.3. **Pregled disertacije**

Rad je organizovan u deset poglavlja. Prvo poglavlje je uvodno i u njemu je data motivacija za ovo istraživanje, opisan je predmet istraživanja i postignuti ciljevi disertacije.

U drugom poglavlju su opisane metode istraživanja podataka, dat je pregled tipova podataka koje metode za istraživanje podataka koriste. U drugom poglavlju se posebno daje pregled istaživanja iz oblasti istraživanja teksta. Sa obzirom da se korpus podataka koji je korišćen u disertaciji sastoji od teksta na prirodnom jeziku (srpskom), u drugom poglavlju su opisani pristupi obrade teksta koji koriste tehniku procesiranja prirodnog jezika i pristupi koji obrađuju tekst bez pomoći ove tehnike.

U trećem poglavlju je opisano koja su generalna znanja o jeziku i koja su potrebna za njegovu kompjutersku obradu. Poseban naglasak je dat semantici jezika i njenom značaju za analizu značenja teksta, što je uvod za analizu sentimenta koja je opisana u narednom poglavlju.

Četvrto poglavlje sadrži definicije analize sentimenta, zadatke analize sentimenta i problem i izazove ove oblasti. Prvo potpoglavlje četvrtog podglavlja sadrži opis metoda za klasifikaciju teksta po sentimentu i pregled metrika za ocenu kvaliteta metoda klasifikacije. Drugo potpoglavlje četvrtog podglavlja opisuje ono što je do sada postignuto kod analize

sentimenta za srpski jezik. Treće potpoglavlje četvrtog podglavlja daje pregled pristupa za obradu negacije kod analize sentimenta.

U petom poglavlju je opisana negacija u srpskom jeziku sa lignvističke tačke gledišta. Dat je pregled vrsta negacija i podela na odgovarajuće tipove koji će biti obrađivani.

Korišćena metoda je opisana u šestom poglavlju: motivacija za primenu metode, skup podataka, potrebna normalizacija, uticaj negacije na reči kojima se izražava sentiment, obrada pravila negacije, struktura i funkcionisanje metode. Poredbene metode su takođe opisane na kraju šestog poglavlja.

Korišćeni jezički resursi su opisanu u sedmom poglavlju. Poseban naglasak u poglavlju je na rečniku sentimenata gde je urađena i njegova validacija različitim načinima normalizacije i uticaj ove normalizacije na određivanje sentimenta tvitova. Osim rečnika sentimenata, opisani su i rečnik signala negacije, rečnik odrečnih kvantifikatora, rečnik stop reči i rečnik pojačivača.

Testiranje i rezultati primene metode za klasifikaciju tvitova su prikazani u poglavlju osam. Testirana su dva pristupa:

- Klasifikacija metodom koja se zasniva na rečniku sentimenata
- Klasifikacija metodom mašinskog učenja

Pre primene metode mašinskog učenja, testirana je primena dodatnih atributa transformacijom teksta u vetor reči.

Deveto poglavlje je zaključak. Deseto poglavlje daje pravce daljeg razvoja sa predlozima budućeg unapređenja i proširenja metode.

2. METODE ISTRAŽIVANJA PODATAKA

Istraživanje podataka je proces kojim se sirovi podaci bilo kojeg tipa sređuju, obrađuju i analiziraju kako bi se utvrdile osobine podataka koje mogu nositi skriveno znanje. Istraživanje podataka je proces otkrivanja skrivenih (ne eksplicitnih) veza iz velikih skupova podataka. Istraživači iz različitih oblasti sakupljaju podatke koji se, u poređenju sa tradicionalnim metodama analize podataka, mogu na novi način analizirati i iz njih se može dobiti novo znanje. Velike količine podataka koje su dostupne u poslednje vreme stvaraju potrebe za razvijanjem različitih metoda za njihovo normalizovanje, analizu i obradu. Sofisticirane metode istraživanja podataka daju mogućnost obrade velikog skupa podataka iz različitih oblasti (medicine, molekularne biologije, geografije, astronomije…). Mogućnost obrade velike količine podataka pruža mogućnost boljeg razumevanja fenomena koje proučavaju navedene oblasti nauke.

Modeli metoda istraživanja podataka se u odnosu na cilj istraživanja, po autorima u [3], mogu generalno svrstati u dve grupe:

- prediktivni modeli imaju za cilj da predvide vrednost ciljanog atributa (zavisne/objašnjavajuće promenljive) na osnovu vrednosti ostalih atributa (nezavisnih promenljivih). Prediktivni modeli su modeli klasifikacije i regresije. I kod klasifikacije i kod regresije je cilj da se istrenira model koji predviđa ciljnu vrednost najpribližnije stvarnoj vrednosti promenljive
- deskriptivni modeli imaju za cilj da proizvedu obrazac (patern) koji predstavlja veze između podataka. Deksriptivni modeli su pravila udruživanja, klaster analiza, detektovanje anomalija, računanje korelacije atributa. Deskriptivni modeli zahtevaju primenu tehnika postprocesiranja kako bi se dobijeni rezultati potvrdili i objasnili.

2.1. **Tipovi podataka kod metoda istraživanja podataka**

Za klasifikaciju podataka je mnogo bitno razumevanje podataka koji se obrađuju kako bi se na najbolji način odredile metode kojima će se podaci normalizovati i dalje obrađivati. Sirovi podaci moraju biti pripremljeni za analizu i po formi moraju odgovarati konkretnoj tehnici obrade podataka koja se nad njima primenjuje. Podaci mogu biti različitog tipa:

- numerički,
  - diskretni
  - kontinualni
- kategorijski,
- vremenske serije i
- tekst.

Numerički podaci su svi podaci koji su meriljivi, kao na primer: visina, težina, broj učenika u odeljenju i sl. U statistici se numerički podaci nazivaju i kvantitativni podaci. Numerički podaci se mogu okarakterisati kao kontinualni ili diskretni podaci. Kontinualni podaci mogu imati bilo koju vrednost unutar opsega, dok diskretni podaci imaju različite vrednosti.

Kategorijski podaci predstavljaju karakteristike koje mogu imati numeričke vrednosti ali ovi brojevi nemaju matematičko značenje, npr. ne mogu se sabirati i ne može se računati njihov prosek. U kontekstu klasifikacije, kategorijski podaci bi bili oznake klasa. Na primer, može se koristiti 0 za oznaku negativne, 2 za oznaku neutralne i 4 za oznaku pozitivne klase ali se nad njima ne mogu primenjivati najjednostavnije matematičke operacije. Postoje i redni podaci (*eng. ordinal*) koji su u nekom smislu kombinacija numeričkih i kategorijskih podataka. U rednim podacima, podaci i dalje spadaju u kategorije, ali te kategorije se rangiraju na određeni način.

Vremenske serije predstavljaju niz brojeva koji se prikupljaju u redovnim intervalima u nekom vremenskom periodu. Podaci vremenskih serija imaju vremensku vrednost koja im je pridružena - neka vrsta datuma ili vremenske oznake koja se može pretraživati.

Tekstualni podaci su u osnovi samo reči, rečenice ili dokumenti. Tekstualni podaci se uglavnom pretvaraju u brojeve koristeći neke funkcije kao što je vreća reči (*eng. bag of words*) i sl.

Podaci se opisuju atributima (*eng. faeature/attribute*) koji sadrže osobine koje karakerišu objekat (instancu) u skupu podataka. Atributi imaju svoj tip i vrednost (nekada vrednosti atributa za neke objekte mogu nedostajati (null).

2.2. **Istraživanje teksta**

Istraživanje teksta predstavlja anlizu tekstualnih podataka na takav način da se od njega kreiraju podaci čijom obradom se može doći do kvalitativnih informacija. Tekstualni podaci su uglavnom nestruktuirani pa je otkrivanje znanja iz ovakvog tipa podataka netrivijalan posao.

Istraživanje teksta je novo područje istraživanja koje pokušava da reši problem preopterećenja informacija koristeći tehnike iz istraživanja podataka, mašinskog učenja, obrade prirodnog jezika, pronalaženja informacija (eng. *Information Retrieval, IR*) i upravljanja znanjem. Istraživanje teksta podrazumeva procesiranje prikupljenih dokumenata (kategorizacija teksta, izdvajanje informacija, izdvajanje termina), skladištenje međurezultata, tehnike za analizu tih rezultata (kao što su klasifikacija, klasterovanje, analiza trendova i pravila udruživanja) i vizuelizacija rezultata [4].

2.2.1. Obrada prirodnog jezika

Prirodni jezik je onaj koji ljudi koriste za svakodnevnu komunikaciju. To su, na primer: engleski, francuski, srpski, ruski ili španski jezik. Za razliku od veštačkih jezika, kao što su programski jezici, prirodni jezici su se razvijali sa generacije na generaciju, pa je teško odrediti sva pravila njihovog korišćenja.

Obrada prirodnog jezika (*NLP*) je podoblast veštačke inteligencije (*eng.* A*rtificial Intelligence, AI*) koja se kod istraživanja teksta koristi za inteligentnu obradu teksta. Obrada

prirodnog jezika pokriva svaku vrstu kompjuterske manipulacije prirodnim jezikom. Obrada prirodnog jezika obezbeđuje da se iz slobodnog teksta izvuče potpuna reprezentacija njegovog značenja. Ona obično koristi lingvističke koncepte kao što su vrste reči (imenica, glagol, pridev, itd.) i strukturu rečenice (redosled reči, fraza ili klauzula). Neki od izazova za NLP jesu anafora (na koju prethodnu imenicu ili frazu se zamenica odnosi) ili dvosmislenost (šta se modifikuje određenom rečju ili predlogom). U te svrhe NLP koristi različite reprezentacije znanja, kao što su rečnik reči i njihovih značenja, gramatičke osobine i skup gramatičkih pravila, ontologije entiteta i akcija i rečnike sinonima i skraćenica. U [5] su istraženi pristupi anafori u teorijskoj lingvistici i NLP-u i prikazani su rezultati istraživanja o uticaju anafore na obradu informacija. Autori u [6] uvode rekurzivnu neuronsku mrežnu arhitekturu za zajedničko parsiranje prirodnog jezika i reprezentacije vektorskog prostora za učenje za promenljive veličine ulaza. U srži arhitekture su kontekstno osetljive rekurzivne neuronske mreže (*eng. Context-Sensitive Recursive Neural Networks, CRNN*) koje mogu indukovati raspodele prikazanih atributa za neviđene fraze i pružiti sintaksičku informaciju za tačno predviđanje strukture stabala izraza.

Tehnologije zasnovane na računarskoj obradi prirodnog jezika postaju sve rasprostranjenije. Na primer, telefoni i ručni računari podržavaju prepoznavanje teksta i rukopisa; mašinski prevod omogućava prevod se jednog na drugi jezik; tekstualna analiza omogućava detektovanje sentimenta u tvitovima i blogovima. Porastom broja interfejsa između čoveka i mašine i sofisticiranijeg pristupa tekstualnim informacijama, proces obrade prirodnih jezika dobija centralnu ulogu u višejezičnom informacionom društvu. Aplikacije izvedene iz ovog područja istraživanja su mnogobrojne a neke od aplikacija su sledeće (navedene su od jednostavnijih do složenih):

- provera pravopisa, pretraživanje ključnih reči, pronalaženje sinonima,
- izdvajanje informacija sa veb-a: cena proizvoda, imena firmi i sl.,
- klasifikacija: pozitivan ili negativan sentiment nekog teksta,
- mašinsko prevođenje,
- sistemi za jezički dijalog,
- odgovaranje na pitanja.

Ono što jezik čini prirodnim je upravo ono što otežava njegovu obradu. Pravila kojima se reprezentuju informacije na prirodnim jezicima su u stalnom razvoju. Ova pravila mogu

biti na većem nivou apstrakcije, na primer kako se sarkazam koristi u prenesenom značenju ili prilično niskom nivou, kao što je korišćenje karaktera "s" za označavanje množine imenica na engleskom jeziku. Jedan od zadataka obrade prirodnog jezika je da identifikuje i koristi ova pravila za prevođenje nestruktuiranih jezičkih podataka u struktuirane informacije. Podaci nekog jezika mogu biti formalni i tekstualni, kao što su novinski članci ili neformalni i zvučni (koji zahtevaju dodatni korak prepoznavanja govora i preobraćanja govornih signala u niz reči), kao što su tvitovi ili telefonski razgovori. Jezički izrazi iz različitih konteksta i izvora podataka imaju različita gramatička pravila, sintaksu i semantiku. Strategije za izvlačenje i prikaz informacija iz prirodnih jezika koje rade u jednoj situaciji često ne funkcionišu u drugoj. Zato je za različite skupove podataka nad kojim se primenjuje obrada prirodnog jezika potrebo posebno prilagoditi na primer, tokenizaciju i normalizaciju a i ostale zadatke koji slede posle njih. Obrada prirodnog jezika uključuje korišćenje jezičkih resursa konkretnog jezika kao što su gramatička pravila, morfološki rečnici, i sl. Većina radova koji se bave analizom teksta koriste neku od tehnika procesiranja prirodnog jezika.

2.2.2. Pristupi u istraživanju teksta koji ne koriste tehnike NLP

Ako se pri obradi teksta ne koriste tehnike obrade prirodnog jezika onda se problem rešava statističkim, stohastičkim ili probabilističkim metodama. Najčešće korišćene metode ovog tipa su latentna semantička analiza (*eng. Latent Semantic Analysis, LSA*), probabilistička latentna semantička analiza (*eng. Probabilistic Latent Semantic Analysis, pLSA*), skriveni Markovljevi modeli (*eng. Hidden Markov Model, HMM*) i Markovljeva slučajna polja (*eng. Markov Random Field, MRF*).

Mera PMI predstavlja meru količine informacija koju sadrže reči koje se pojavljuju jedna pored druge. Ako imamo dve reči *r1* i *r2*, onda je verovatnoća da se reči pojavljuju zajedno data u formuli [7]:

$$PMI\ (r1, r2) = \log_2 \left(\frac{p(r1, r2)}{p(r1)p(r1)} \right) \tag{2.1}$$

McNamara i koautori su u [8] uporedili različite metode korišćenja LSA i metode zasnovane na rečima kako bi usmerili odgovore koje učitelj daje učenicima u vezi sa kvalitetom njihovih objašnjenja. Ispituju efikasnost 7 sistema pomoću algoritama zasnovanih na rečima, LSA, i kombinacijama oba, koja se razlikuju u stepenu ručne pripreme ciljnog teksta. Njihova efikasnost se meri u smislu njihove usklađenosti s ljudskim ocenama objašnjenja.

Autori u radu [9] predstavljaju strategiju za učenje semantičke orijentacije na osnovu semantičke asocijacije (*eng. Semantic Orientation from Association, SO-A*) reči koristeći dve strategije: PMI-IR meru (*eng. Pointwise Mutual Information- Information Retrieval ,PMI-IR*) i latentnu semantičku analizu (*eng. Latent Semantic Analysis, LSA*). PMI-IR meru računaju kao kombinaciju PMI mere za izračunavanje jačine semantičke povezanosti između reči a za statistiku uzajamnog pojavljivanja (*eng. co-occurrence*) reči koriste tehniku pronalaženja informacija (IR). LSA merom analiziraju statističku vezu između reči u korpusu koristeći razlaganje na svojstvene vrednosti (*eng. Singular Value Decomposition, SVD*). SVD se može posmatrati kao oblik analize glavnih komponenti (*eng. Principal Components Analysis, PCA*). Umesto pomoću originalne matrice, LSA meri sličnosti reči koristeći kompresovanu matricu tako što sličnost dve reči izražava kosinusom ugla između odgovarajućih redova matrice. Poredeći se sa [10], Turnej i saradnici su zaključili da su njihovi podaci uporedivi.

Pang i saradnici u [11] primenjuju metode mašinskog učenja nad skoro sirovim podacima (tekst je pretprocesiran samo otklanjanjem suvišnih HTML tagova). Atributi koje koriste kod kreiranja modela su klasičan vektor vreće reči (*eng. bag-of-words*).

Boontham i saradnici u [12] raspravljaju o upotrebi tri različita pristupa kategorizaciji besplatnih tekstova odgovora učenika na otvorena pitanja: jednostavno usklađivanje reči, LSA i varijacija na LSA koju nazivaju modeli tema (*eng. topic models*). LSA i modeli tema su i numeričke metode za generisanje novih funkcija zasnovanih na linearnoj algebri i kreću sa predstavljanjem teksta pomoću modela vreće reči. Pored toga, za klasifikaciju koriste diskriminantnu analizu iz statistike. Stemovanje i saundeks algoritmi (metod za ispravljanje pogrešne oznake predstavljanjem reči na način koji približno odgovara njihovom izgovoru) se koriste u komponenti za usklađivanje reči. Stemovanje je jedina NLP tehnika koja se koristi.

McCarthy i saradnici u [13] takođe koriste LSA kao svoju primarnu tehniku, i upoređuju različite delove dokumenta umesto celokupne dokumente da razviju "potpis" dokumenata zasnovanih na korelaciji između različitih sekcija. U [14] se SVM kombinuje sa

Markovljevim lancem kako bi se odredilo kako se razdvajaju sekvence tekstualnih stranica u različite dokumente različitih tipova s obzirom na to da su stranice sa tekstovima vrlo nestruktuirane, jer predstavljaju proizvod optičkog prepoznavanja znakova. Oni istražuju različite načine na koje mogu modelovati sekvencu stranica, u smislu koje se kategorije mogu dodeliti stranama i kako kombinovati sadržaj stranice i informacije o sekvenci. Oni koriste jednostavne tehnike kao što su tokenizacija i stemovanje, ali ne i složenije NLP tehnike.

Atkinson u [15] koristi tehniku koja je vrlo nova za istraživanje teksta - genetske algoritme (GAs). Genetski algoritmi se obično koriste za rešavanje problema gde se funkcije mogu predstaviti kao binarni vektori. Atkinson ovo prilagođava tekstualnim predstavama tako što koristi čitav niz numeričkih i statističkih metoda, uključujući LSA i Markovljeve lance, i razne metrike nadograđuje na njima. Osim nekih ručno konstruisanih konteksta za retoričke uloge, on ne koristi stvarne NLP tehnike.

3. ZNANJE O JEZIKU I NJEGOVA KOMPJUTERSKA OBRADA

Kompjuterska obrada teksta i govora nekog prirodnog sjezika obuhvata sveprisutne aplikacije koje obrađuju pisani i govorni jezik. Aplikacije mogu biti jednostavnije kao što su brojanje reči, automatsko ispravljanje nepravilno napisanih reči ili složenije kao što su autmatsko prevođenje sa jednog na drugi jezik, automatsko odgovaranje na pitanja i slično. Sve ove aplikacije za kompjutersku obradu moraju koristiti neko znanje i/ili resurse konkretnog jezika.

Znanje o jeziku, koje je potrebno kod složenog jezičkog izražavanja, se može svrstati u znanje o šest različitih jezičkih kategorija [16]:

- fonetika i fonologija – proučavaju osobine govornih glasova, nivoe glasovnog sistema i apstraktne glasovne jedinice,
- morfologija - proučava unutrašnju strukturu reči,
- sintaksa - proučava strukturne odnose između reči i pravila kojima se reči kombinuju u rečenice,
- semantika - proučava značenje reči,
- pragmatika - proučava načine kojima se jezik koristi za postizanje različitih ciljeva,
- diskurs - proučava lingvističke jedinice veće od jednog iskaza.

Svaka od ovih diciplina se bavi jezikom na određeni način a imaju jedan zajednički problem a to je dvosmislenost jezičkih izraza. Dvosmislenost u tekstu se, na primer, može demonstrirati u različitom tumačenju jedne iste rečenice.

Način na koji ljudi razumeju tekst se svodi na njihovo poznavanje jezika, znanje o konceptima i objektima i njihovim međusobnim vezama. Na osnovu iskustva, ljudi razumeju sadržaj i skoro nesvesno izvode zaključke o značenju teksta. Mašine ne mogu razumeti tekst na način kako ga ljudi razumeju pa se oslanjaju na statističke pristupe, pristupe zasnovane na

mašinskom učenju, pristupe zasnovane na ključnim rečima, različitim rečnicima iz određene oblasti i slično. Neki od zadataka analize teksta su:

- sintaksička analiza se može odnositi na početnu tačku analize teksta, gde se analizira rečenica da bi se razumele i označile različite vrste reči (*eng. Part of Speech, POS*).
- prepoznavanje imenovanog entiteta (*eng. Named Entity Recognition, NER*) - pronalaženje delova govora koji se odnose na entitet (osobu, lokaciju, organizaciju)
- vektorska reprezentacija reči (*eng. word embedding*) – obrađivanje teksta kako bi se dobila vektorska reprezentacija reči u obliku n-dimenzionalnog vektora. Zatim se može računati mera sličnosti (npr. kosinusne sličnosti) između vektora za određene reči da bi se analiziralo kako su povezane
- lematizacija - ova metoda redukuje različite oblike reči u njihove osnovne oblike, što znači da se pojavljuju češće jer se ne posmatraju npr. različiti glagolski oblici kao posebne reči. Na primer, reči: „voleću“, „volite“, „voleli“ se sve mogu svesti na osnovni oblik „voleti“.
- stemovanje - je proces kojim se različiti oblici (izvedene) reči redukuju na njihov koreni/osnovni oblik - stem.

3.1. **Semantička analiza**

Semantička analiza analizira značenje sadržano u tekstu. Semantička analiza ima zadatak da utvrdi odnose između reči, kako se one kombinuju i koliko često se neke reči pojavljuju zajedno.

Latentna semantička analiza (*LSA*) je teorija i metoda za izvlačenje i predstavljanje kontekstualnog značenja reči pomoću statističkih izračunavanja primenjenih na velikom korpusu. LSA je tehnika pronalaženja informacija bez nadgledanja koja analizira i pronalazi obrazac i vezu između skupa tekstova koji su nestruktuirani. LSA se koristi kod obrade prirodnog jezika (NLP) za analizu odnosa između skupa dokumenata i termina koje oni sadrže.

TF-IDF (*eng*. T*erm Frequency-Inverse Document Frequency*) je tehnika pronalaženja informacija koja meri frekvenciju termina u dokumentu (TF) i njenu inverznu frekvenciju (IDF). Svaka reč ima svoj TF i IDF rezultat i računaju se po datim formulama 3.1 i 3.2, respektivno:

$$TF\ (r) = \frac{\text{broj pojava reči r u dokumentu}}{\text{ukupan broj reči u dokumentu}} \tag{3.1}$$

$$IDF\ (r) = \log\left(\frac{\text{ukupan broj dokumenata}}{\text{broj dokumenata koji sadrže } reč\ r}\right) \tag{3.2}$$

Proizvod TF i IDF težina jedne reči se naziva TF-IDF težina te reči.

U poslednje vreme, veliki značaj dobija podoblast semantičke analize – analiza sentimenta. Neki od razloga popularnosti ove oblasti jesu sledeći:

- najveći broj tekstualnih sadržaja se kreira na socijalnim mrežama, forumima, blogovima – jednom rečju na internetu
- tekstovi koji privlače najviše pažnje jesu oni u kojima je sadržan neki oblik emocija, stava, mišljenja i sl.

Analiza ovakvih tekstova koji sadrže emocije može koristi kompanijama o proizvodima, političkim partijama o popularnosti kandidata, psiholozima o otkrivanju pojedinih duševnih stanja vezano za situaciju i slično. Analiza sentimenta je podoblast semantičke analize teksta koja analizira tekst u kojem je prisutan sentiment. Pošto se analiza sentimenta bavi pre svega utvrđivanjem polariteta teksta (pozitivan, negativan, neutralan) ona ne može da se koristi za otkrivanje skrivenih koncepata koji se nalaze u tekstu.

Semantički pristupi koji se zasnivaju na vektoru reči se mogu koristiti za modelovanje bogatih leksičkih značenja, ali teško uspevaju da obuhvate informacije o sentimentu koje su bitne za veliki broj NLP zadataka. Pristup koji uči vektore reči koji sadrže semantičke informacije o dokumentu kao i bogat sadržaj sentimenta je predstavljen u [17]. Još jedan semantički pristup analizi sentimenta su dali Saif i saradnici u [18] i to tako što su dodali nove atribute kod treniranja klasifikatora tvitova po sentimentu. Ti novi atributi predstavljaju meru korelacije konkretnog semantičkog koncepta i polariteta datog tvita.

4. ANALIZA SENTIMENTA

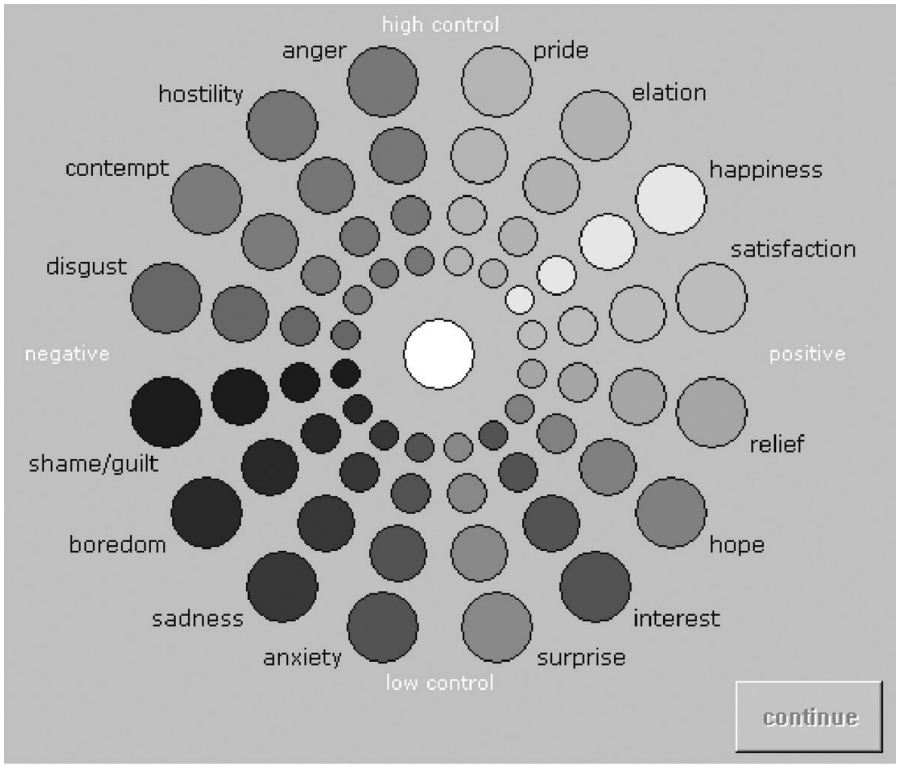
Procesiranje prirodnog jezika je tehnika koja se bavi kompjuterskom obradom prirodnog ljudskog jezika koji se pojavljuje kao tekst na veb stranicama, u opisu proizvoda, novinskim člancima, socijalnim stranicama, forumima, naučnim radovima i sl. Analiza sentimenta je podoblast tehnike procesiranja prirodog jezika i aktuelna je tema istraživanja u poslednje vreme. Ona teži da utvrdi kakav je stav izražen u nekom tekstu, koje su emocije prisutne, kakvo je mišljenje izneseno. Analiza sentimenta je, dakle, vezana za analizu tekstova u kojima postoji subjektivno izraženo osećanje. Tekstovi koji sadrže objektivne činjenice nisu predmet analize ove oblasti.

Mišljenja, stavove i emocije ljudi izražene i napisane na datom jeziku nije lako analizirati i nije lako odrediti sentiment iz teksta, s obzirom na složenost lingvističkih izraza i različite načine izražavanja koje ljudi koriste na određenom jeziku. Neki drugi problemi u analizi sentimenta su dvosmislenost reči ili sintagmi, neformalno pisanje, upotreba neologizama, problem segmentacije, ironije, metafore i negacije. Bez obzira na nemogućnost da se kompjuterski potpuno tačno odredi sentiment, pod određenim pretpostavkama složenost takvog zadatka može biti pojednostavljena. Najjednostavniji sistemi određuju sentiment na osnovu broja pojava pozitivnih i negativnih reči iz rečnika sentimenata. Neki napredniji sistemi uzimaju u obzir i POS oznake (oznake vrste reči, roda, broja, i ostalih gramatičkih karakteristika), primjenjuju pravila negacije i otkrivaju ironiju. Iako je određivanje polariteta, kao jednog od zadataka analize sentimenta, oblast koja je u određenoj meri istražena, neki aspekti koji su lingvistički specifični (negacija, ironija, metafora) i dalje predstavljaju izazove i oblasti u kojima se očekuju poboljšanja. Pojavom mašinskog procesiranja prirodnog jezika i pojavom algoritama za mašinsko učenje, desio se veliki napredak u ovoj oblasti. Iako pristup zasnovan na mašinskom učenju u većini slučajeva daje bolju tačnost kod klasifikacije, pristup koji se zasniva na rečniku sentimenata ima prednost u situacijama kada nije dostupan skup obeleženih podataka i kada je vreme treniranja klasifikatora ključno.

Diskretna klasifikacija emocija je pristup koji se zasniva na organizaciji semantičkih polja za emocije u prirodnim jezicima. Prvi diskretni pristup emocijama je vezan za poreklo jezika i nastanak reči i izraza koji opisuju stanja koja se jasno mogu razdvojiti. Ovaj pristup ima naučnu zasnovanost jer su začetkom bihevioralnih nauka filozofi koristili emocionalne reči za analizu ljudskog emocionalnog iskustva. Darvin [19] je ovaj pristup učinio prihvatljivim za biološke i društvene nauke pokazujući evolucioni kontinuitet niza osnovnih emocija. On je identifikovao vidljive fiziološke i ekspresivne simptome koji odgovaraju osnovnim emocijama. Vundt je još 1905. godine subjektivna osećanja opisivao njihovim položajem u trodimenzionalnom prostoru koji se formira pomoću merenja ocećanja u tri dimenzije:

- polariteta (*eng. valence*) (pozitivno-negativno),
- uzbuđenja (mirno-uzbuđenje) i
- napetosti (napeto-opušteno).

Ženevski točak emocija (*eng. Geneva Emotion Wheel, GEW*) je teoretski izveden i empirijski testiran instrument za merenje emocionalnih reakcija na objekte, događaje i situacije. Prototip verzija ženevskog točka emocija je data na slici 4.1. Raspored termina koji odgovaraju emocijama je dat u dvodimenzionalnom prostoru a intenzitet emocije se meri veličinom kruga na ženevskom točku emocija. Intenzitet emocije je najslabiji kod tačaka najbližim centru a najveći na obodu točka – gde su krugovi najveći. Studija pouzdanost i validnosti ovog instrumenta je i dalje u razvoju.



Slika 4.1: Prototip verzija ženevskog točka emocija. Preuzeto iz [20]

Psihololog Klaus Šeror [20] je pokazao da ima više stanja osećanja:

- emocije: tuga, žalost, sreća, bezbrižnost, posramljenost, ponos,...
- raspoloženje: veselost, depresija, tromost, razdražljivost, sumornost,...
- interpersonalni stavovi: prijateljstvo, distanca, hladnoća, koketiranje, podrška, …
- stav: ljubav, mržnja, želja, sviđanje, ...
- afektivna stanja: nervoza, zabrinutost, ljubomora...

Tabela u kojoj su date osobine navedenih stanja osećanja (intenzitet, trajanje, sinhronizacija, fokus na događaj, ocenjivanje/procena, brzina promene i uticaj na ponašanje) su dati na slici 4.2. Sa slike se vidi da je su karakteristike "intenzitet" i "trajanje" za "stavove" kao jedno od stanja osećanja veće u odnosu na druga stanja osećanja. Pošto analiza sentimenta ima za cilj da utvrdi osećanje koje je trajnije onda se na osnovu ove tabele može zaključiti da se analiza sentimenta najviše bavi analizom stava kao jednog od stanja osećanja. Pored

trajanja, za analizu sentimenta je važno da se osećanja u tekstu izražavaju sa što većim intenzitetom.

Od navedenih osećanja koji se tiču sentimenta, analiza sentimenta se najviše primenjuje za određivanje stava (*eng*. *attitude*) prema nekom fenomenu, objektu ili subjektu. Zadatak analize sentimenta jeste utvrđivanje stava prema izvoru (držaocu) stava, cilju (aspektu) stava i tipu stava (sviđanje, ljubav, mržnja, želja...). Kada je u pitanju tip stava koji se zauzima prema nekom objektu ili osobi, najčešće se radi o prostom merenju polariteta (pozitivan, negativan, neutralan) a nekada se mere samo nivoi pozitvnog ili negativnog. Kada se radi na utvrđivanju sentimenta dokumenata, može se utvrđivati sentiment celog dokumenta ili jednog njegovog dela, npr. rečenice. Jednostavniji zadatak analize sentimenta zahteva određivanje pozitivnog ili negativnog stava prema nekom tekstu. Kompleksniji zadatak analize sentimenta zahteva određivanje stava prema nekom tekstu zahtevajući numeričko kvantifikovanje pomoću skale, na primer od 1 do 10. Najzahtevnija i naprednija analiza sentimenta zahteva utvrđivanje cilja, izvora i kompleksnijih tipova stavova prema odgovarajućim objektima.

Kratki opis afektivnih stanja	Intenzitet	Trajanje	Sinhronizacija	Fokus na događaj	Procena	Brzina promene	Uticaj na ponašanje
Emocije: relativno kratka epizoda sinhronizovanih odgovora svih ili većine organizma na procenu spoljašnjeg ili unutrašnjeg događaja (npr. bes, tuga, radost, strah, stid, ponos, ushićenje, očajanje)	++→++	+	+++	+++	+++	+++	+++
Raspoloženja: difuzno afektivno stanje, najizraženije kao promena subjektivnog osećaja, niskog intenziteta ali relativno dugog trajanja, često bez vidljivog uzroka (npr. vedro, mračno, razdražljivo, ravnodušno, depresivno, plivajuće)	+→++	++	+	+	+	++	+
Interpersonalni stavovi: afektivni stav prema drugoj osobi u specifičnoj interakciji, obojen interpersonalnom razmenom u toj situaciji (npr. udaljenost, hladnoća, toplina, podržavanje, prezir)	+→++	+→++	+	++	+	+++	++
Stavovi: relativno trajna, afektivno obojena uverenja, preferencije i predispozicije prema objektima ili osobama (npr. sviđati, voleti, mrzeti, vrednovati, želeti)	0→++	++→+++	0	0	+	0→+	++
Karakteristike ličnosti: emocionalno opterećene, stabilne dispozicije ličnosti i tendencije ponašanja, tipične za osobu (npr. nervozne, teskobne, bezobzirne, mrzovoljne, neprijateljske, zavidne, ljubomorne)	0→+	+++	0	0	0	0	+

*Simboli označavaju stepen prisutnosti karakteristika, pri čemu 0 označava najnižu (odsutnost) a + + + označava najveću prisutnost; strelice označavaju hipotetičke opsege

Slika 4.2. Osobine različiih stanja osećanja. Tabela je preuzeta iz [21].

Analiza sentimenta se može primenjivati kod različitih zadataka kao što su ocenjivanje pozitivnih ili negativnih recenzija filmova, određivanje mišljenja ljudi o nekom proizvodu, određivanje poverenja kupaca merenjem rasta ili pada zadovoljstva, mišljenja o političkoj partiji u politici, predviđanje rezultata ili trenda prodaje na osnovu izraženog sentimenta.

Analiza sentimenta spada u oblasti tehlologije jezika koja je u usponu, za razliku od oblasti kao što su: detekcija spama, detekcija i označavanje vrste, roda i broja reči u rečenici, koje su uglavnom oblasti gde su dobijena prilično dobra rešenja. Pored analize sentimenta, u oblasti koje su u usponu spadaju i: izdvajanje informacija, mašinsko prevođenje, određivanje smisla tj. nedvosmislenosti reči u rečenici, parsiranje i sl. Oblasti gde je još uvek teško doći do rešenja su one gde je potrebno utvrditi odgovore na pitanja, odrediti koje parafraze imaju isto značenje, doneti zaključak na osnovu nekoliko premisa i dijaloga.

Generalno, problem koji se javlja kod svih podoblasti NLP-a jeste dvosmislenost reči ili sintagmi u zavisnosti u kojem se delu rečenice nalaze ili u zavisnosti od ostalih reči koje se pojavljuju uz njih. Pored dvosmislenosti, problemi koji otežavaju tumačenje prirodnih jezika jesu nestandardno zapisivanje reči (npr. „U“ umesto „you“, „tebra“ umesto „brate“, „navući se“ umesto „postati zavisan“), korišćenje slenga, neologizama (npr. „post“, „tvit“,...), idioma (npr. „Traži hleba preko pogače“), problemi kod segmentacije (Novi Pazar, Novi Sad se može podeliti na reči....).

Analiza sentimenta se može odnositi na određivanje polariteta, ili na složenije – pronalaženje aspekta na koji se tekst (npr. jedna rečenica) odnosi. Kada se radi o polaritetu, algoritam zahteva prvo prikupljanje resursa (teksta) koji će se analizirati. Algoritam se (na osnovu predloga Panga i saradnika [22]) može podeliti na sledeće korake:

- Pretprocesiranje
- Izdvajanje karakteristika (atributa) koji mogu biti numerički ili tekstualni
- Klasifikacija pomoću odgovarajućeg algoritma za klasifikaciju (Naive Bayes, MaxEnt, SVM,...), a koristeći predhodno izdvojene atribute.

Kod procesa tokenizacije potrebno je utvrditi u kojim se mogućim izvornim formatima može pojaviti tekst na koji se tokenizacija primenjuje. Ako se radi o tekstu koji ima HTML ili XML tagove, potrebno je to uzeti u obzir kod tokenizacije. Označavanje postova na Tviteru je praćeno heš tagom(#). Potrebno je voditi računa i o kapitalizaciji slova (nekad kod prevođenja teksta u mala slova treba sačuvati npr. akronime i slično). Posebno treba obraditi

brojeve telefona, datume i to pomoću regularnih izraza. Kod tokenizacije, pre analize sentimenta, posebnu pažnju treba posvetiti emotikonima i njihovom načinu zapisivanja. Postoje posebni regularni izrazi kojima se mogu obuhvatiti poznati emotikoni.

Izdvajanje karakteristika se odnosi na zadatke kao što su: kako obraditi negaciju u rečenicama, koje vrste reči kojima se izražava sentiment koristiti (samo prideve ili sve vrste reči). Pokazano se da je korišćenje svih reči bolje nego korišćenje samo prideva, jer često glagoli, imenice i druge vrste reči daju više informacija nego samo pridevi.

Negacija je karakteristika koju je jako bitno pravilno obraditi kada je u pitanju sentiment. Prvi jednostavni algoritam za obradu negacije je dat u radu [11], a kasnije je često korišćen i u drugim radovima. Zasniva se na dodavanju prefiksa „NE_“ (*eng.* „*NOT_*“) svakoj reči koja se nalazi u tekstu između negacije i prvog znaka intrepunkcije koji se pojavi posle negacije. Na primer rečenica:

„ne volim ovaj kolač….“

se po ovom algoritmu transformiše u rečenicu

„ NE_volim NE_ovaj NE_kolač….“

Na ovaj način se iz reči NE _volim i NE _kolač uči negativan sentiment u konkretnoj rečenici.

Sentiment se može analizirati na nivou celog teksta, rečenice ili jednog aspekta teksta. Stav se može izražavati diskretno (pozitivan, negativan i neutralan) ili na skali od pozitivnog do negativnog. Metode koje se zasnivaju na rečniku za analizu sentimenta koriste rečnike koji imaju diskretne vrednosti polariteta. Ovakvi rečnici su: Bing Liu's Opinion Lexicon [23] i MPQA Subjectivity Lexicon [24]. Sentiment rečnik koji sadrži vrednosti polariteta koje su na određenoj skali je SentiWordNet [25]. Kada su u pitanju tehnike kojima se određuje sentiment, većina radova koristi metode učenja sa nadgledanjem, iako nije mali broj pristupa koji daju analizu metodama nenadgeldanog učenja koje se zasniva na rečniku sentimenata i kombinovanom polu-nadgledanom učenja.

Poslednjih godina postoji veliko interesovanje za istraživanje u oblasti analize sentimenta. Prilikom određivanja sentimenta, istraživači obično pretpostavljaju da osoba ima neki stav ili osećaj prema subjektu, objektu ili osobi, da raspoloženje ima fiksnu vrednost (dobro ili loše) i da je sentiment u tekstu predstavljen rečju ili kombinacijom malog broja reči. Međutim, u literaturi se malo govori o tome šta je zapravo "sentiment" ili "mišljenje/stav"

[26]. U radovima [22] i [27] autori daju pregled različitih pristupa kod određivanja sentimenta. Pošto se analiza sentimenta koristi da se utvrdi subjektivni stav o fenomenu (objektu, osobi), neki sistemi za analiza sentimenta uključuju, pored pozitivnih i negativnih, i neutralne (objektivne) stavove. Prema tome, analiza sentimenta uglavnom uključuje istraživanje unutar tri klase: pozitivne, negativne i neutralne. Većina radova se bavi analizom sentimenta na engleskom jeziku. Višejezična analiza sentimenta je predstavljena u [28] gde autori upoređuju sopstvenu primenu postojećih pristupa za određivanje sentimenta. U ovoj disertaciji je naglasak na metodama koje se koriste za klasifikaciju po sentimentu i specifične su za srpski jezik. Popularnost društvene mreže Tviter raste sa brojem tvitova na njoj, pa je sve veći broj autora odlučio da testira sentiment ovog tipa kratkog teksta. Pregled metoda za analizu sentimenta na Tviteru dat je u [29]. Kod većine istraživanja za određivanje sentimenta se koriste metode nadgledanog učenja, iako nije neznatan broj pristupa koji koriste analizu metodama nenadgledanog ili polu-nadgledanog učenja.

U ovoj disertaciji je akcenat na obradi pravila sintaksičke negacije na srpskom jeziku. Namera je da se utvrdi da li tretiranje ovih pravila negacije i njihovo integrisanje u metodu klasifikacije po sentimentu može poboljšati tačnost klasifikacije kratkih tekstova - tvitova. Pravila koja se obrađuju predstavljaju podgrupe gramatičkih pravila za procesiranje negacije na srpskom jeziku, tj. pravila koja se najčešće javljaju u ovom tipu kratkog teksta. Primenjena pravila i njihov uticaj na određivanje polariteta će biti testirani nenadgledanom metodom zasnovanom na rečniku sentimenata i metodom mašinskog učenja sa nadgledanjem.

4.1. **Metode klasifikacije po sentimentu**

U zavisnosti od resursa koji se koriste za određivanje sentimenta teksta, može se govoriti o metodama koje koriste rečnik sentimenata i metodama koje koriste korpus neke specifične oblasti za određivanje sentimenta. Poređenje metode zasnovane na rečniku sentimenata sa metodom zasnovanom na korpusu su izvršili autori u [30] za analizu sentimenta tekstova na arapskom jeziku. Zaključili su da najbolje rezultate daje prisup zasnovan na korpusu i to ako se koristi SVM metoda mašinskog učenja. Takođe su zaključili da se rastom rečnika koji sadrži reči kojima se označava sentiment, raste i tačnost klasifikacije

ali je potreban mnogo veći rast rečnika da bi se dobilo malo povećanje tačnosti klasifikacije metodama zasnovanim na rečniku.

U radu [31] se metode za klasifikaciju sentimenta svrstavaju u jedan od dva osnovna pristupa: klasifikacija mašinskim učenjem i klasifikacija zasnovana na rečniku sentimenata.

U radu [32] je predstavljena metoda klasifikacije koja kombinuje klasifikaciju koja se zasniva na pravilima i klasifikaciju nadgledanim mašinskim učenjem.

4.1.1. Metode zasnovane na rečniku sentimenata

Pristup zasnovan na rečniku sentimenata se zasniva na određivanju sentimenta na osnovu prisustva reči kojima se izražava sentiment (i možda nekog kraćeg konteksta, na primer negacije) računajući pojave reči kojima se izražava sentiment kako bi se predvideo generalni sentiment teksta. U istraživačkoj literaturi, pristup zasnovan na rečniku sentimenata se obično naziva najjednostavnijim (i stoga manje preciznim) pristupom. Međutim, prednost metoda koje se zasnivaju na rečniku sentimenata jeste jednostavnost i mogućnost korišćenja opšteg rečnika sentimenata za različite domene kao i mogućnost korišćenja domenskih rečnika sentimenata za različite skupove podataka iz istog domena. Postoji dosta radova koji predstavljaju metode koje daju prilično dobre rezultate koriščenjem samo rečnika sentimenata. U [33] je predložena metoda kod koje reči kojima se izražava sentiment imaju različit polaritet u zavisnosti od konteksta i lingvističkog obrazca u kojem se pojavljaju. Dokazano je da njihov metod daje bolje rezultate od postojećih sistema za klasifikaciju rečenica po sentimentu koji se zasnivaju na rečniku sentimenata. Kim i saradnici se u [34] bave utvrđivanjem reči koje nose sentiment i kombinovanjem ovih reči kako bi se utvrdio sentiment rečenice. Oni su kao početni rečnik reči kojima se izražava sentiment koristili ručno kreirani skup malog broja reči koje su proširili koristeći WordNet rečnik - upotrebom veza sinonima i antonima. Jedan od najcitiranijih radova koji klasifikuje tekst po sentimentu na osnovu rečnika reči kojima se izražava sentiment jeste [31]. Doprinos metodama klasifikacije koji se zasnivaju na rečniku reči kojima se izražava sentiment je dat i u sledećim radovima: [35], [23], [36], [37] i dr.

Merenjem sentimenta su se prvi počeli baviti autori iz oblasti sociologije [38]. Merenje sentimenta nekog teksta na osnovu rečnika sentimenata se zasniva na dve

pretpostavke: potrebno je da pojedinačne reči imaju definisan polaritet, to jest, semantičku orijentaciju koja je nezavisna od konteksta; i da se pomenuta semantička orijentacija može izraziti kao numerička vrednost. Ovakav pristup računanja sentimenta su usvojili autori u radovima [31], [39], [23] i [34].

Većina sistema za analizu sentimenta koji se zasnivaju na rečniku sentimenata na nivou rečenice ili dokumenta se suočava sa poteškoćama pri dodeljivanju rezultata sentimenta rečima koje nisu dostupne u njihovim bazama podataka. Domenski rečnici sentimenata se mogu kreirati na način što se od polaznog skupa reči kojima se izražava sentiment (*eng. seed words*) kreira dodatni skup reči iz korpusa, koristeći se nekim pravilima konkretnog jezika ili koristeći sinonime, antonime, hiperonime i slično. Proširenje ovakvih rečnika koristeći WordNet je često i prikazano je u radovima: [40], [41] i [42]. Kao rezultat čestog korišćenja WordNet-a u pomenute svrhe, nastao je rečnik reči kojima se izražava sentiment SentiWordNet. Trenutna službena verzija je 3.0, a bazirana je na WordNet 3.0 i opisana je u radu [43].

Autori u radu [31] predstavljaju pristup zasnovan na rečniku reči kojima se izražava sentiment za dobijanje sentimenta teksta. Njihov "računar semantičke orijentacije" (*Semantic Orientation CALculator, SO-CAL*) koristi rečnike reči označene njihovom semantičkom orijentacijom (polaritet i snaga) i uključuje pojačavanje i negaciju. Evaluaciju su radili nad četiri nezavisna skupa podataka (400 recenzija tekstva o knjigama, automobilima, kompjuterima, hotelima i sl.; 400 tekstova sa sajta epinions.com; 1900 recenzija filmova; 2400 tekstova sa recenzijama kamera, štampača i dečijih kolica) i pokazali su da njihov pristup zasnovan na rečniku reči kojima se izražava sentiment ima prosečnu tačnost za sva četiri navedena skupa podataka od 78.74% tačno klasifikovanih tekstova.

4.1.2. Metode zasnovane na korpusu

Metode analize sentimenta koje se zasnivaju na korpusu teže da utvrde sentiment nekog teksta u specifičnom kontekstu ili specifičnom domenu. Za razliku od metoda koje se zasnivaju na rečniku reči kojima se izražava sentiment (kod kojih je polaritet ovih reči nezavisan od konteksta), kod metoda zasnovanih na korpusu se polaritet jedne iste reči može razlikovati u različitim kontekstima. Svaki korpus nosi određenu specifičnost domena, koji

može da pruži algoritmu informacije o različitosti reči kojima se izražava sentiment u zavisnosti od konteksta/domena. Na primer tekstovi u modnom magazinu „Kako izgledati skupo i prefinjeno“ i u časopisu za ishranu „Organska hrana neopravdano skupa“ oba sadrže reč „skupa/o“ koja u prvom slučaju ima pozitivan a u drugom negativan sentiment. Izrada domenskih rečnika sentimenata je u ovakvim slučajevima jedno od rešenja.

Često se metode analize sentimenta koje se zasnivaju na korpusu koriste i kao dopuna metodama zasnovanim na rečniku sentimenata i to na način da se postojeći opšti rečnici sentimenata dopunjavaju rečima iz domena datog korpusa. U [44] autori predlažu pristup izdvajanja reči koji se zasniva na propagaciji i rečnika sentimenata i izdvojenih svojstava/atributa proizvoda, koji oni nazivaju dvostrukom propagacijom. Algoritam koristi odnose zavisnosti kako bi utvrdio vezu između svojstava/atributa i reči kojima se izražava sentiment i suprotno, reči kojima se izražava sentiment i svojstava/atributa instanci (u njihovom slučaju proizvoda).

Kanajama i saradnici [45] predlažu metodu izgradnje rečnika sentimenta bez nadgledanja, za otkrivanje polarnih klauzula koje sadrže pozitivne ili negativne aspekte u određenom domenu. Leksičke jedinice koje se dobijaju nazivaju polarnim atomima - minimalne ljudski razumljive sintaksičke strukture koje određuju polaritet klauzula. Kao ključ za dobijanje kandidatskih polarnih atoma koriste kontekstnu koherentnost - tendenciju da se isti polariteti pojavljuju sukcesivno u kontekstima.

Izrada rečnika sentimenat na osnovu korpusa tvitova je predstavljena u radu [46]. Ova metoda se sastoji iz dva dela: prvi deo je algoritam učenja reprezentacije za efikasno učenje fraza koje se koriste kao atributi za klasifikaciju a drugi deo je algoritam proširenja početne liste reči kojima se izražava sentiment (*eng*. *sentiment seeds*) kako bi se dobili podaci za treniranje klasifikatora sentimenta na nivou fraza. Treniranje metodom neuronskih mreža je izvršeno nad velikom količinom neobeleženih tvitova, koristeći samo pozitivne i negativne emotikone za klasifikaciju.

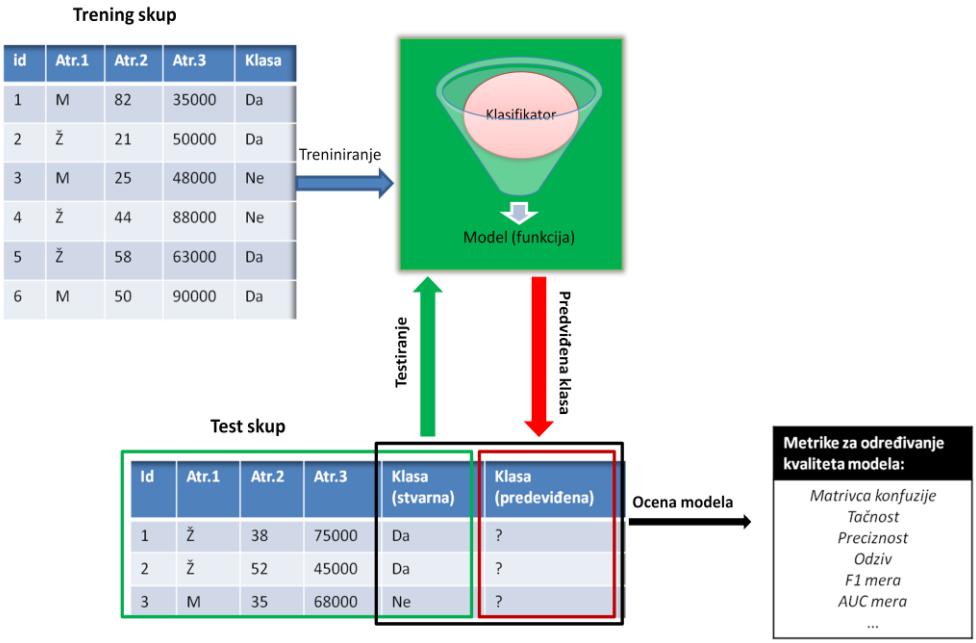
Metoda koja koristi korpus tvitova je predstavljena i u radu [47]. Ova metoda vrši automatsko prikupljanje korpusa koji se može koristiti za treniranje sentiment klasifikatora. Klasifikator koristi multinomijalnu Naive Bayes metodu koja koristi *n*-grame i POS-oznake kao atribute. Da bi povećali tačnost klasifikacije, odbacili su uobičajene *n*-grame, tj. *n*-grame koji ne ukazuju na postojanje bilo kakvog sentimenta niti ukazuju na objektivnost procene. Takvi *n*-grami se ravnomerno pojavljuju u svim skupovima podataka. Kako bi izdvojili ove

uobičajene *n*-grame, uveli dve posebne strategije. Kod kreiranja *n*-grama, autori su posebnu pažnju posvetili negaciji, pa su signale negacije uvek vezali za reč koja prethodi ili sledi. Na taj način su omogućili da se poboljša tačnost klasifikacije, jer negacija igra bitnu ulogu kod izražavanja mišljenja i osećanja.

4.1.3. Metode mašinskog učenja za klasifikaciju po sentimentu

Klasifikacija je metoda mašinskog učenja sa nadgledanjem. Zadatak klasifikacije jeste svrstavanje objekata u jednu od diskretnih kategorija (klasa). Skup podataka nad kojim se primenjuje klasifikacija se sastoji od nezavisnih promenljivih (atributa X) i jedne zavisne (objašnjavajuće) promenljive (kategorički atribut y). Kategorički atributi (atributi klase) moraju biti diskretne vrednosti i skup tih vrednosti čini broj klasa. Diskretne vrednosti atributa klase su karakteristika modela klasifikacije, za razliku od modela regresije kod kojih atributi klase imaju kontinualne vrednosti. Model (funkcija) klasifikacije se kreira u procesu učenja gde se nad skupom podataka za trening, koji sadrži sve atribute i vrednosti kategoričkog atributa, utvrđuje način preslikavanja skupa nezavisnih promenljivih (X) u jednu od vrednosti kategoričke promenljive (y). Model klasifikacije se može testirati nad skupom podaka za test. U ovom slučaju se radi validacija treniranog modela koristeći neku od metrika pogodnu za klasifikaciju. Ilustracija procesa klasifikacije je data na slici 4.3.

Najčešće korišćene metode za evaluaciju klasifikatora su sledeće: Naivni Bajesov algoritam (*Naïve Bayes, NB*), nominalna logistička regresija (*Multinomial Logistic regression, LOG*), metoda potpornih vektora (*Support Vector Machines, SVM*), stabla odlučivanja (*decision tree*), k-najbližih suseda (*k-Nearest Neighbour*), veštačke neuronske mreže (*Artificial Neural Network*).



Slika 4.3: Proces treniranja, testiranja i ocene modela kod klasifikacije

4.1.4. Merenje kvaliteta klasifikacije

Evaluacija modela klasifikacije je sastavni deo procesa razvoja modela. Pomaže da se pronađe najbolji model koji predstavlja podatke i daje informaciju kako će izabrani model raditi u budućnosti. Vrednovanje performansi modela sa podacima koji se koriste za obuku nije prihvatljivo jer može lako da generiše model koji je previše prilagođen podacima (*eng. overfitting*). Da bi se utvrdilo koliko je klasifikacija bila uspešna, potrebno je odrediti podelu podataka za treniranje i testiranje, kao i metriku kojom bi se kvalitet merio. Postoje dva metoda podele podataka za evaluaciju modela: "hold-out" metoda i unakrsna validacija (*eng. cross-validation, CV*).

Kako bi se izbegla preterana obuka modela (model koji je previše prilagođen podacima), obe metode za evaluaciju modela koriste test skup. Test skup sadrži podatke koji u fazi treniranja (učenja modela) nisu korišćeni.

Kod "hold-out" metode se veliki skup podataka nasumično deli na dva ili tri podskupa (podskup za validaciju se nekada i ne koristi):

- Skup za obuku je podskup skupa podataka koji se koristi za izgradnju prediktivnih modela.
- Skup za validaciju je podskup skupa podataka koji se koristi za procenu performansi modela izrađenog u fazi obuke. On pruža testnu platformu za fino podešavanje (*eng. fine tuning*) parametara modela i izbor najboljeg modela. Podskup za validaciju nije uvek neophodan za izradu modela.
- Test skup je podskup skupa podataka koji služi za procenu verovatnoće uspešnosti modela. Test skup sadrži podatke koji se u trening fazi nisu koristili.

Podela podataka koristeći "hold-out" metodu ima prednost što je jednostavnija, fleksibilnija i zahteva manje vreme izvršenja.

Kada je na raspolaganju samo ograničena količina podataka, da bi se postigla nepristrasna procena performansi modela, koristi se unakrsna validacija sa *k*-slojeva (*eng. k-fold cross-validation*). Kod unakrsne validacije sa *k*-slojeva, podaci se dela na *k* podskupova jednake veličine. Izrađuje se model *k* puta, svaki put izostavljajući jedan različit podskup *k* i taj podskup se u konretnoj fazi koristi kao test skup.

Unakrsna validacija je poželjnija metoda jer daje modelu priliku da trenira na višestrukim test skupovima. Ovo daje bolji pokazatelj kako će model raditi na novim (neviđenim) podacima. Pošto podela skupa na trening i test (*eng. hold-out*) zavisi od samo jedne trening-test podele skupa, ono je zavisno od toga kako su podaci podeljeni na trening i test skupove, a ta podela nekada može biti takva da ne odslikava dobro pravo stanje. Bez obzira na prednosti, unakrsna validacija se pokreće *k* puta pa je za *k* put sporija od "hold-out" metode.

Mere kvaliteta modela klasifikacije daju uvid u kvalitet modela. Postoje različite metrike a najčešće korišćene su:

- matrica konfuzije (*eng. confusion matrix*)
- tačnost (*eng. accuracy)*
- preciznost (*eng. precision*),
- odziv ili osetljivost (*eng. recall or sensitivity*),
- specifičnost (*eng. specificity*),
- F1 koeficijent.

Matrica konfuzije je jedna od najintuitivnijih i najjednostavnijih metrika za utvrđivanje ispravnosti i tačnosti modela. Koristi se za problem klasifikacije sa dve ili više klasa. Izgled matrice za dve klase je dat na slici 4.4. Matrica konfuzije sadrži četiti vrednosti koje su rezultat klasifikacije instanci po klasama i te vrednosti predstavljaju:

- broj stvarno pozitivnih (*eng. true positive, TP*) je broj onih instanci koje su pozitivne i klasifikovane su kao pozitivne
- broj stvarno negativnih (*eng. true negative, TN*) je broj instanci koje su negativne i klasifikovane su kao negativne.
- broj lažno pozitivnih (*eng. false positive, FP*) predstavlja broj instanci koje su negativne ali ih je klasifikator svrstao u pozitivne.
- Broj lažno negativnih (*eng. false negative, FN*) je broj onih instanci koje su pozitivne ali su klasifikovane kao negativne.

		Predviđena klasa	
Stvarna klasa		pozitivna	negativna
	pozitivna	TP	FN
	negativna	FP	TN

Slika 4.4: Matrica konfuzije za dve klase

Najčešće korišćena metrika je tačnost i računa se kao odnos tačno klasifikovanih instanci i ukupnih instanci, kao po formuli:

$$\text{Tačnost} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

Preciznost je mera koja govori koliki procenat instanci jedne klase (npr. pozitivne sa slike 4.4) koji su klasifikovani kao kao instace te konkretne klase (pozitivne) zapravo pripada toj klasi (pozitivnoj). Predvidene pozitivne instance su TP i FP, a one koji su stvarno pozitivne su označene sa TP. Formula za računanje preciznosti je sledeća:

$$Preciznost = \frac{TP}{TP + FP} \tag{4.2}$$

Odziv (osetljivost) je merika koja govori koliki je procenat instanci jedne klase (npr. pozitivne sa slike 4.4) zapravo dijagnostikovano algoritmom kao instace te konkretne klase (pozitivne). Stvarni pozitivni rezultati su TP i FN, a instance koje su modelom klasifikovane kao pozitivne (TP). Odziv se računa po formuli:

$$Odziv = \frac{TP}{TP + FN} \tag{4.3}$$

Specifičnost je metrika suprotna odzivu. Specifičnost govori koliki je procenat instanci druge klase (npr. negativne sa slike 4.4), a model ih je predvidio kao da pripadaju toj drugoj klasi (negativnoj). Stvarno negativni (instance koje nisu pozitivne) su FP i TN, a instance koje su klasifikovane kao negativne su TN. Specifičnost se računa po sledećoj formuli:

$$Specifičnost = \frac{TN}{FP + TN} \tag{4.4}$$

F1 metrika kombinovano prikazuje dve metrike - preciznost i odziv. U principu, F1 predstavlja harmonijsku sredinu između preciznosti i odziva. Harmonijska sredina dva broja *x* i *y* teži da bude bliža manjem od dva broja. Dakle, visoka vrednost F1 mere osigurava da su i preciznost i odziv prilično visoki. F1 mera ima sledeću funkciju:

$$F1 = \frac{2 * Preciznost * Odziv}{Preciznost + Odziv} \tag{4.5}$$

4.2. **Analiza sentimenta na srpskom jeziku**

Obrada sentimenta je vezana za jezik. Analiza sentimenta na srpskom jeziku je do sada rađena za skup novinskih članaka [48], recenzija filmova [49] i skupa tvitova [50]. Mladenović i koautori su u [48] koristili pravila morfološkog rečnika za dobijanje flektivnih oblika (*eng. inflected form*) reči radi izgrade rečnika reči kojima se izražava sentiment i

rečnika stop reči koje sadrži leme i flektivne oblike leme kako bi se smanjio broj atributa (*eng. feature space*) za klasifikaciju mašinskim učenjem. Koristili su metodu nominalne logističke regresije (maksimalne entropije) za klasifikaciju teksta na srpskom jeziku po sentimentu koristeći skup novinskih članaka za primenu stratifikovane unakrsne validacije sa 10 slojeva (*eng. 10-fold cross validation*). Da bi potvrdili dobijene rezultate, testirali su metodu na još dva posebna skupa: skup novinskih članaka i skup filmskih recenzija. Autori su koristili bogate leksičke resurse nad skupom od dve klase (pozitivne i negativne). Koristili su metod za redukciju atributa za atribute (reči) koje su pronađene u rečnicima sentimenata i mapirali ih na novi skup atributa. Oni su dobili tačnost stratifikovanom unakrsnom validacijom do 95.6%, do 78,3% koristeći nezavisni test skup recenzija filmova i do 79.2% koristeći test skup novinskih članaka. U njihovom radu negacija nije posebno obrađivana.

Batanović i koautori su u [49] kreirali skup podataka recenzija filmova na srpskom jeziku – SerbMR, koji koristi algoritam za balansiranje podataka koji minimizira pristrasnost izbora uzorka eliminacijom nebitnih sistemskih razlika između sentiment klasa. Predložili su analizu sentimenta koristeći različite kombinacije *n*-grama, stemova, lematizatora i različitih tipova normalizacije atributa. Na kraju, primenom optimalnih parametara atributa pomoću NBSVM (kombinacija polinomijalnog Naivnog Bajesovog klasifikatora i klasifikatora metodom potpornih vektora), postigli su tačnost do 85.55% za dve i do 62.69% za tri klase. Autori u [51] istražuju uticaj dva načina morfološke normalizacije teksta (koristeći stemmer i lemmatizator za srpski jezik) na klasifikaciju sentimenta nad skupom podataka o filmskim recenzijama. Koristeći kombinaciju unigrama i bigrama, postigli su statistički značajno poboljšanje u poređenju sa osnovnom metodom za dve klase (tačnost 86.11%) i za tri klase (tačnost 63.02%). Uporedni pregled metode analize sentimenta za hrvatski jezik dat je u [52]. Zbog velike sličnosti hrvatskog i srpskog jezika u kompleksno-morfološkom smislu i drugim karakteristikama, može se smatrati da su postignuti rezultati slični onima na srpskom jeziku. Autori su upoređivali metodu "word embedding" i "string kernels" nad tri skupa kratkih tekstova (recenzije igara, tvitova iz određenog domena i tvitova o generalnim temama). Oni su pokazali da su tehnikama "word embedding" i "string kernel" postigli poboljšanje u odnosu na osnovnu/poredbenu metodu vreće reči (*eng. bag-of-words, BOW*).

4.3. **Negacija kod analize sentimenta**

Najčešće korišćena metoda za obradu negacije je dodavanje sufiksa „_NE“ na reč kojom se izražava sentiment i koja se javlja u opsegu negacije a negacija se tretira od signala negacije do prvog znaka interpunkcije kao u radovima [11], [53] i [54] ili do prve pronađene pozitivne ili negativne reči kojom se izražava sentiment [55]. Iako većina sistema za tretiranje negacije uzima u obzir sve vrste reči kojima se izražava sentiment, neki uzimaju samo prideve [23] ili prideve i priloge [56]. U [54], autori su ispitali uticaj signala negacije, kao jednog od mnogih menjača polariteta, na promenu polariteta reči u negiranom opsegu. Većina autora jednostavno menja polaritet negirane reči prilikom obrade negacije (od n do -*n* i obratno) kao u [55] i [57]. Sofisticirani sistemi za obradu negacije posebno tretiraju sentiment reči koje se pojavljuju u opsegu negacije i pokazuju da negacija sa različitim intenzitetima menja polaritete pozitivnih i negativnih reči kojima se izražava sentiment [58]. Autori koji su se zalagali za ovaj pristup pokazali su u [59] da kreiranje takvog rečnika za reči koje se javljaju u opsegu negacije daje značajno poboljšanje u sistemu za predviđanje sentimenta. Pored radova koji se bave obradom negacije uz pomoću signala negacije, u [60] se pojavljuju fenomeni koji podsećaju na negaciju, jer menjaju polaritet, ali nisu negacije u klasičnom smislu, jer ne sadrže reči koje su signali negacije (ne, nije ...). U [61] su autori predložili otkrivanje opsega negacije koristeći metod podsticajnog učenja (*eng*. *reinforcement learning*). Evaluaciju uticaja detektovane negacije na analizu sentimenta uradili su merenjem korelacije između sentimenta rečenice i odgovarajuće vrednosti dnevne zarade akcija na tržištu (*eng*. *daily stock market return*) kao mere predikcije, i dobili da postoji korelacija od 10.63% između sentiment vrednosti i vrednsti dnevne zarade akcija na tržištu. U [62] autori detektuju signale negacije i opsege negacije. Rezultate detektovane negacije primenjuju na klasifikator sentimenta i to za ceo skup i skup koji sadrži negacije. Njihova sofisticirana metoda za predviđanje sentimenta (kada se koriste atributi obrađene negacie) na celom skupu postiže poboljšanje u odnosu na osnovnu metodu (preciznost je povecana za 0.01, odziv za 0.004 i F1 za 0.003). Na skupu podataka koji sadrži samo negacije, rezultati su sledeći: sofisticirana metoda u odnosu na osnovnu postiže sledeća poboljšanja: preciznost za 0.033%, odziv za 0.042% i F1 za 0.39%. Iako se većina metoda za evaluaciju uticaja negacije na sentiment teksta zasniva na nadgledanom učenju, u radovima [57] i [62] autori daju predlog računanja intenziteta negacije i njenog uticaja na polaritet primenjujući svoje metode izračunavanja intenziteta

negacije i polariteta teksta. U [57] se autori bave analizom negacije na španskom jeziku i njihovim uticajem na određivanjee sentimenta. Korišćen je korpus tvitova na španskom jeziku i metoda za predviđanje sentimenta koja, pored ostalih resursa, uključuje i obrađenu negaciju. Dobili su rezultate u kojima se vidi da tretiranjem negacije postižu značajno poboljšanje tačnosti kod određivanja polariteta tvita.

U [63] su opisana pravila za identifikovanje negacije i izračunavanje njenog intenziteta. Pravila su kreirana u cilju poboljašanja analize sentimenta teksta. Metoda kojom predviđaju sentiment je metoda bez nadgledanja koja računa polaritet i intenzitet reči i fraza. Pokazali su da postoji pozitivna korelacija između polariteta određenog sistemom i onog dodeljenog od strane pet učesnika eksperimenta.

Teško je reći da postoji unverzalno najbolji klasifikator sentimenta teksta. Sistemi koji postižu dobre rezultate za duge tekstove mogu za kratke postizati loše. Primer je rad autora [64] čija metoda na skupu podataka koji sadrži sarkazme (Tweet Sarcasm 2014) zauzima prvo mesto, dok na SemEval 2015 (Analysis Task 10) skupu zauzima 16. mesto. Pregledni rad [65] daje prikaz većine navedenih pristupa u obradi negacije od najranijih radova; u radu su predstavljeni načini obrade negacije, opsega negacije, izbora atributa za treniranje i navedene granice u modelovanju negacije kod analize sentimenta.

Za srpski jezik, ne postoje radovi koji se bave detektovanjem gramatičkih pravila negacije u tekstu, a samim tim ni uticajem takve obrade na analizu sentimenta. Klasičan način obrade negacije, menjanjem polariteta rečima koje se javljaju posle signala za negaciju primenili su Batanović i koautori u [49]. Tretirajući negaciju na skupu recenzija filmova, postižu sledeća poboljšanja u odnosu na inicijalnu metodu opisanu u [11] : za tri klase postižu najveće poboljšanje metodom MNB (*Multinomial Naive Bayes*) od 0.94 % (tačnost 54.72% na 55.66%) markirajući dve reči posle negacije; za dve klase postiže najveće poboljšanje metodom SVM od 0.66% (75.98% na 76.64%) markirajući samo prvu reč posle negacije.

5. NEGACIJA U SRPSKOM JEZIKU

Negacija je fenomen kojim se bave matematika, logika i sintaksa i to svaka od ovih disciplina proučavajući elemente negacije koji su polje njenog interesovanja. Deo pitanja u vezi sa negacijom dele sve ove tri nauke, a najzanimljivije u tezi će biti pitanje sintaksičke negacije.

U disertaciji će akcenat biti na analizi sintaktičke negacije, tj. na efekatu signala negacije na deo rečenice ili na celu rečenicu. Ako se uporede logički i jezički kriterijumi, zaključuje se da su logičke kategorije pojma u dobroj meri pandami jezičkim kategorijama reči (leksema) i rečenice. Svako sintaksičko i gramatičko razmatranje negacije mora u sebi nužno uključiti i logičko razmatranje negacije. U zavisnosti od vrste jezičke jedinice u kojoj se pojavljuje, negacija u srpskom jeziku može biti leksička, sintaksička i morfološka. Kao što se u logici negacija se može odnositi na izraz ili mišljenje, u gramatici se negacija može posmatrati u odnosu na reč (leksička i morfološka negacija) ili u odnosu na rečenicu (sintaksička negacija) [66].

Morfološka negacija se postiže upotrebom prefiksa kao što su "ne-", "bez-" , "ni-" , "a-", "dis- " i " in- ". Prefiksi kao morfološka sredstva u srpskom jeziku sadrže negaciju u sebi i dopunjavaju značenje primarne reči. Morfološka negacija utiče na negaciju rečenice samo ako se lekseme pojavljuju u ulozi članova rečenice (kao negativne reči, tj. negativne lekseme). U disertaciji je morfološka negacija obrađena na takav način da su negirane lekseme (morfološka negacija) uključene u rečnik sentimenata. Interesantno je da reči koje su rezultat morfološke negacije u srpskom jeziku u većini slučajeva imaju negativan polaritet, pa se ova činjenica može iskoristiti kao statistički podatak u analizi sentimenta tekstova u kojima se ona javlja.

Za razliku od morfološke negacije, leksičko negiranje je uglavnom implicitno - reči nemaju negaciju u svojoj strukturi, već samo u svom značenju. Leksička negacija se odnosi na upotrebu reči čije značenje ima negativnu komponentu ("sumnja", "odsutnost", ”zaboravljeno”). Na osnovu ovoga se može reći da je leksička negacija reči vezana isključivo za jezik i da ne postoji način da se odredi pomoću gramatičkih pravila već samo na osnovu

znanja konkretnog jezika. S obzirom na navedeno, može se reći da su reči koje su leksička negacija uključene u rečnik sentimenata i da je to jedini način da se ova vrsta negacije u srpskom jeziku obuhvati. Ovakve reči koje su leksička negacija imaju negativno značenje i u rečniku sentimenata imaju negativan polaritet. Primeri su sledeći:

„Postoji **sumnja** u ispravnost rada komisije “

„**Odustnost** inteligencije je osobina nekih životinja“

Iako u rečenici nema prisustva negatora a ni negativnih prefiksa u rečima, ipak je negacija prisutna jer se reči „sumnja“ i „odsutnost“ negiraju sadržaj na koji se odnose.

Sintaksička negacija se postiže upotrebom signala negacije (npr. “ne” i “ni”) koji stoje ispred reči (ulgavnom glagola) koju negiraju u rečenici. Primer sintaksičke negacije je dat u rečenici:

“Ne volim kišu”

gde je signal negacije “ne” ispred glagola “volim”.

Kada u sintaksičkoj negaciji učestvuju glagoli “imam”, “biti” i “hteti” onda oni grade nove odreče oblike koji u prezentu imaju formu “nemam”, “nisam” i “neću”. Glagoli koji imaju nepravilnu negaciju se spajaju sa signalima negacije i kreiraju nove signale negacije. Primer je dat u sledećoj rečenici:

“Nisam skijala kad je bilo magle”

gde je negacija nepravilna i signal negacije “ne” se spaja sa glagolim “jesam” u nepravilni oblik sintaksičke negacije i dobija se novi signal negacije “nisam”.

U rečenicama u kojima se javljaju i sintaksička i morfološka negacija se može reći da postoji dvostruka negacija. Međutim, s obzirom da su reči koje su rezultat morfološke negacije u disertaciji uključene u rečnik sentimenata, ovakav slučaj će se svoditi na običnu negaciju. Primer ovakvog slučaja je sledeći:

(1) “Nije nepoštovanje zakona… ”

gde je “nepoštovanje” morfološka negacija koja je nastala kao rezultat prefiksa “ne-” i reči “poštovanje” (ne+poštovanje = nepoštovanje).

S obzirom na opseg negacije (opseg sadržaja rečenice koji je negiran), negacija rečenice (sintaksička) može biti parcijalna ili totalna. U slučaju totalnog negiranja, predikat je uvek negiran (i celokupan sadržaj rečenice je uvek negiran), a u parcijalnoj negacije deo

rečenice je negiran, pa negacija utiče samo na deo sadržaja rečenice. Primer totalne negacije je rečenica (1) a parcijalne rečenica (2).

(1) “Malo kiše neće pokvariti planove za izlet.”

(2) "Došao je ne jedan, nego više predstavnika."

Specijalne vrste parcijalne negacije će biti detaljnije razrađene u nastavku.

5.1. **Parcijalna negacija**

Parcijalna negacija se odnosi na slučajeve gde signal negacije negira samo deo rečenice. Parcijalna negacija negira samo nepredikatski član u rečenici i na srpskom jeziku se sastoji od sledećih slučajeva: “

1. “Ni…ni” konstrukcija (nrp. Nije ni lepo ni ružno)
2. “Ne/nije/… nego/no/već” konstrukcija (nrp. Film nije bio zanimljiv nego dosadan)
3. “Ne/nije/… samo….nego/no/već” konstrukcija (nrp. Očistili su ne samo dvorište, nego i celu ulicu)

5.1.1. “Ni…ni” konstrukcija

Konstrukcija “ni..ni” po [66] ima ulogu katahreze, jer se stavljanjem ni-ni negatora ispred oba leksički ili kontekstualno suprotna pojma zapravo imenuje neutralni pojam za koji ne postoji posebna leksema. Ova vrsta parcijalne negacije je u stvari pseudonegacija jer se negiranim formama izražava potvrdno značenje. Ovaj tip pseudonegacije ne utiče na definisanje opsega negacije već se prosto rečenica u kojoj se javlja posmatra kao da negatori ne postoje (uglavnom su u pitanju rečenice koje nisu ni pozitivne ni negativne). Primer rečenice sa ovom konstrukcijom negacije je sledeći:

“Vreme je ni sunčano ni kišovito”

Pošto se rečenicama sa navedenom konstrukcijom uglavnom ne iznosi nikakav stav već je reč o neutralnom ili neodređenom pojmu, ovakav tip konstrukcija neće biti od interesa.

5.1.2. "Ne/nije/… nego/no/već" konstrukcija

Ovo je tip negacije gde se javljaju kontrarne negirane forme ali se u negiranom delu uvek izražava pravo parcijalno odrečno značenje [66]. Pošto signal negacije uvek stoji ispred sadržaja koji se negira, ova negacija pokazuje da se negira samo dati sadržaj, a ne cela rečenica. Uz negirani sadržaj uvek ide i kontrastna potvrdna forma kojom se izražava realizovani sadržaj. Primer ovog tipa parcijalne negacije je sledeća rečenica:

"Nisu izgubili već pobedili."

Opseg negacije u ovom tipu parcijalne negacije obuhvata negirani deo sadržaja rečenice, tj. onaj deo rečenice koji kreće od signala negacije (ne, nisu, nije,…) do reči "nego/no/već".

5.1.3. "Ne/nije/… samo….nego/no/već" konstrukcija

Ovaj tip parcijalne negacije je, kao i "ni…ni" konstrukcija, u stvari pseudonegacija jer prisustvo signala negacije ne dovodi do negiranja, već nasuprot, dolazi do potvrde i čak gradacije. To znači da negator "ne" (ili bilo koji drugi signal negacije) uvek dolazi ispred priloga "samo". Negacijom priloga "samo", isključuje se njegovo apsolutno delovanje pa je potrebno navođenje uporednog, istorodnog elementa kao poredbenog korelata [66]. Primer ovog tipa parcijalne negacije je dat sledećom rečenicom:

"Cveće ne samo da miriše nego i lepo izgleda."

U ovom sličaju se signalom negacije naglašava da sledi ne samo sadržaj X nego i sadržaj Y. Ovaj tip parcijalne negacije će biti obrađen tako što će se negator (signal negacije) ukloniti, pa će se sadržaj rečenice posmatrati bez uticaja negacije jer se u ovom slučaju signal negacije ne koristi za pravu negaciju već se, uz prilog "samo" želi naglasiti ono što ide posle njega.

5.2. **Dupla (dvostruka) negacija**

Matematika deli sve aritmetičke brojeve na pozitivne i negativne, logika deli sve izraze i presude na pozitivne i negativne, a lingvistika sve leksme i rečenice na pozitivne i negativne. Kada je u pitanju negacija, oblasti u kojima se ove tri nauke (logika, matematika i lingvistika) najviše preklapaju su polja množenja negacija ili dvostruke negacije [66]. Matematika kaže da je proizvod dva broja istog znaka broj čija je apsolutna vrednost proizvod apsolutnih vrijednosti ovih brojeva, a njegov znak je +, proizvod dva negativna broja u matematici je uvek pozitivan broj. Logika kaže: "pod dvostrukim negiranjem neke izjave P se podrazumeva izjava ne ne P" [67].

Dvostruka negacija se može definisati se kao dva puta iskazana negacija u sastavu jedne rečenice. Kod dvostruke negacije je obavezno da jedan od dva negirana člana bude predikat ili subpredikat. Ukoliko su u jednoj rečenici najmanje dvaput upotrebljeni rečenični negatori „ne“ i „ni“, s tim da je jedan od njih subpredikat ili predikat, onda je reč o dvostukoj negaciji [66]. Pojam dvostuke negacije se dakle koristi za najmanje dvostruko negiranje jer se višestruka negacije najčešće i ostvaruje kao dvostruka.

Kod dvostruke/višestruke sintaksičke negacije u srpskom jeziku postoje dve situacije kada pri njenoj upotrebi dolazi do:

- kongruencije negacija, kada formalno odrečni iskaz ima istu takvu semantiku ili
- množenja negacija, kada formalno odrečni iskaz ima afirmativnu semantičku vrednost.

Kod dvostruke negacije se razlikuje ponašanje „ne-“ i „ni-“ negatora. Lekseme negirane „ne-“ i „ni-“ negatorima utiču različito na dvostruku negaciju. Odrečne lekseme sa prefiksom „ne-“ u bilo kojoj sintaksičkoj poziciji ne zahtevaju obavezno i negaciju predikta (tj. ne zahtevaju totalnu negaciju). Primer je sledeća rečenica:

“Nepažnja u saobraćaju je sve češća pojava čak i kod savesnih vozača.”

U ovoj rečenici prefiks “ne-” u kombinaciji sa “pažnja” kreira reč “nepažnja” koja ne zahteva negaciju sledećeg dela rečenice.

Na osnovu ponašanja negativnih prefiksa (negatora) u rečenici, može se zaključiti da lekseme negirane prefiksom „ni-“ imaju drugačiji uticaj na rečeničnu (totalnu) negaciju od

leksema koje su negirane drugim prefiksima. Sve lekseme negirane prefiksom „ni-“ spadaju u odrečne kvantifikatore i nužno zahtevaju i negaciju predikta – znači zahtevaju dvostruku negaciju. Primer lekseme negirane „ni-“ prefiksom je sledeći:

“Nikad nemoj okrenuti leđa onima kojima treba pomoć”

U ovom slučaju je leksema “nikad” negirana prefiksom “ni-” i posle nje obavezno mora doći signal negacije (“nemoj”) pa dolazi i do pojave dvostruke negacije.

U skladu sa ovim, izrađen je rečnik odrečnih kvantifikatora (leksema negiranih prefiksom “ni-”) koji će koristiti metoda detektovanja opsega negacije. Bitnost izdvajanja odrečnih kvantifikatora kao posebnih elemenata negacije je u tome što oni proizvode poseban efekat kod duple negacije (ne poništavaju je, već je pojačavaju).

5.2.1. Pojačavanje (sabiranje) negacije

Dvostruka negacija kojom dolazi do slaganja (podržavanja) negacije je u srpskom jeziku data u sledećim slučajevima:

1. Tip dvostruke negacije gde negativna “ni-“ leksema uslovljava i negativnu formu predikta dovodi do slaganja (sabiranja) ova dva negirana rečenična člana: negativne lekseme (odrečnog kvantifikatora) i negiranog predikta (negacije predikta). U disertaciji su ove negativne lekseme izdvojene u rečnik „odrečni kvantifikatori“ a negacije u rečnik „signali negacije“. Ovaj tip dvostruke negacije u disertaciji se obrađuje tako što odrečni kvantifikatori pojačavaju ili potvrđuju negativnu formu rečenice (npr. nikad nisam povredio životinju.)
2. Gomilanje većeg broja odrečnih kvantifikatora takođe dovodi do njihovog slaganja (sabiranja) i slaganja sa negiranim prediktom koji obavezno sledi iza odrečnih kvantifikatora (nrp. Niko nikom ništa nije ukrao). Znači, nagomilavanja odrečnih kvantifikatora takođe zahteva negiranje predikta i potvrđuje negativnu formu rečenice ili je pojačava.

5.2.2. Množenje negacije

Dvostrukom negacijom predikta u rečenicama u srpskom jeziku se uvek dobija potvrdna vrednost. Ovaj tip negacije se naziva i logičko-matematička negacija jer se negacije unutar predikta ne sabiraju nego se množe**,** tako da predikat nema odrečnu nego potvrdnu vrednost. Kod dvostruke negacije u ovim vrstama rečenica dolazi do množenja (poništavanja) negacije. Sledeći slučajevi dvostruke negacije u srpskom jeziku dovode do množenja negacija, tj. njenog poništavanja, jer se dobija afirmativna vrednost rečenice:

1. Negacijom negacije unutar predikta se dobija afirmacija (npr. Nije ostalo nerešeno pitanje.). Ovaj tip negacije u disertaciji je obrađen na taj način što su negativne lekseme uključene u rečnik sentimenata.
2. Logičko-matematička dvostruka negacija u srpskom jeziku ostvaruje se u okviru složenog glagolskog predikata i to tako što se posebno negira i modalni glagol i glagol koji ga dopunjava u formi infinitiva ili prezenta sa veznikom „da“ (npr. Nemoj da ne dođeš.).
3. Logičko-matematička negacija se takođe ostvaruje i negacijom oba glagola u perifrastičkim izrazima u konstrukciji sa veznikom „da“ (npr. Nije da nemamo vremena.).
4. U jednom tipu prostog glagolskog predikta negacija negacije daje afirmaciju. U srpskom jeziku postoje tri glagola koji prefigirani prefiksom „ne“ daju potvrdnu vrednost ako ispred njih stoji negator „ne“. Ti glagoli su: nemati, nedostajati (npr. Ne nedostaje nam znanje.) [66].

U sva četiri navedena slučaja, uvek se dobija potvrdna vrednost rečenice. Znači negacije se množe i poništavaju kao u logičko matematičkim izrazima.

5.3. Negacija u pitanjima

Upitna rečenica sadrži zahtev za odgovorom i zahtev za nekom vrstom odluke povodom onoga na šta se pitanje odnosi. U upitnoj rečenici se ne sme tvrditi ni istinitost ni neistinitost njenog smisla. Stoga, smisao upitne rečenice nije nešto čije se postojanje sastoji u

istinitosti [68]. U oblasti lingvistike, proučavanje negacije tiče se niza jezičkih nivoa i predmet je mnogih semantičkih i pragmatičnih analiza. Istraživanjem kombinacije ova dva fenomena se može formalizovati kako se značenje pitanja menja kada ono sadrži negaciju.

Upitna rečenica koja sadrži negaciju nekada može izraziti/potvrditi misao koja je i sadržaj pitanja. Ovakva pitanja se nazivaju retoričkim, emocionalno obojenim pitanjima koja u sebi sadrže odgovor. Retoričko pitanje ima za cilj da naglasi sadržaj, a ne da se dobije odgovoer na to postavljeno pitanje. Iako retoričko pitanje ne zahteva direktan odgovor, u mnogim slučajevima može biti namera da se započne diskusija ili bar da se prihvati da slušalac razume nameravanu poruku.

Takav primer upitnih rečenice u srpskom jeziku je "Zar nije lep dan?". Znači, ukoliko pitanje sadrži rečcu "zar" posle koje sledi negacija, takvo pitanje izražava potvrdu onoga što se pita. Naizgled ista pitanja mogu imati sasvim drugaciji rezultat. U slučaju (1)

(1) Da li je bolja socijalizacija kod dece koja idu u vrtić?

se traži odgovor na pitanje koji može biti ili "da" ili "ne". Međutim ako se pitanje postavi na način (2):

(2) Zar nije bolja socijalizacija kod dece koja idu u vrtić?

onda se pitanje svodi na potvrdu ili se bar traži potvrda.

U prvom primeru pitanje ne sadrži negaciju i može se reći da je po formi pitanje u pravom smislu te reči, što nije slučaj u drugom primeru. Zato se pojava negacije u ovakvim specifičnim tipovima pitanja posebno obrađuje jer, u stvari, ne predstavlja negaciju već njenu neutralizaciju. Akcenat se kod ovakvog tipa pitanja stavlja na onaj deo rečenice koji preostaje ako se negacija ukloni. Zato je ovaj tip negacije kod pravila koja su definisana za njenu obradu i obrađen tako što se ovakav tip pitanja svodi na potvrdu, znači negacija se neutrališe.

Na konkretnom skupu podataka (tvitovi) često se, zbog neformalnog govora, nepravilno koristi upitni oblik "jel" kao zamena za gramatički ispravno "je li". Tako se upitni oblik "jel" u svakodnevnom govoru a posebno kod dopisivanja na društvenim mrežama u konstrukciji sa negacijom upotrebljava kako bi se postavilo pitanje na koji se očekuje pozitivan odgovor. Iz ovog razloga se i oblik "jel" uz prisustvo negacije koristi za obradu ovog tipa pravila negacije u pitanjima.

Ne postoji literatura na srpskom jeziku u kojoj se lignvistički prilazi problemu negacije u pitanjima. U radu [69] autori ispituje razliku između pozitivnih pitanja i negativnih

pitanja sa jakom i slabom negacijom na engleskom jeziku. Fokusiraju se na razlike u kontekstima u kojima su različite vrste pitanja pojavljuju i pokazuju da se pozitivna pitanja i različiti tipovi negativnih pitanja zaista razlikuju po kontekstu u kojem se pojavljuju. Primeri pozitivnog pitanja i pitanja sa slabom i jakom negacijom su dati u tabeli 5.1.

Tabela 5.1: Tipovi pitanja sa različitim polaritetom [69]

Pitanje	Polaritet
Did Lucy go to Greece?/Da li je Lucy otišla u grčku?	Pozitivno pitanje
Did Lucy not go to Greece?/Da Lucy nije otišla u Grčku?	Pitanje sa slabom negacijom
Didn’t Lucy go to Greece?/Zar Lucy nije otišla u Grčku?	Pitanje sa jakom negacijom

Tip upitnih rečenica (pitanja) koja će biti predmet analize u disertaciji jesu one koje sadrže jaku negaciju. U odnosu na tipove pitanja iz tabele 5.1, to je pitanje “Zar Lucy nije otišla u Grčku?”.

6. PREDLOŽENA METODA

6.1. Motivacija

U rečenicama koje imaju određenu konstrukciju u kojima se javlja negacija nekad dolazi do neutralizacije negacije, nekad do pojačavanja a nekada do promene polariteta. Ideja je da se obrade pravila kojima će se utvrditi kada se negacijom menja polaritet reči i u kojem delu rečenice a kad je negacija u stvari pseudonegacija i prosto ne menja polaritet već izražava potvrdno značenje. Metoda klasifikacije treba da pokaže da se uzimanjem u obzir specifičnih pravila negacije, detektovanjem opsega u kojem signali negacije deluju ili ignorisanjem signala negacije (kod jednog od pravila), može poboljšati kvalitet klasifikacije tvitova po sentimentu.

Kako bi se uradila evaluacija primene pravila negacije, rađena je klasifikacija teksta po sentimentu, tako što su izdvojene karakteristike/atributi teksta koji detektuju prisustvo ili odsustvo negacije. Ovi atributi služe za klasifikaciju različitim metodama i ti rezultati su poređeni. Korišćena su dva metoda za klasifikaciju sentimenta:

- pristup baziran na rečniku sentimenata i
- pristup baziran na mašinskom učenju.

Metoda je primenjena nad skupom podataka tvitova na srpskom jeziku. Da bi se podaci pripremili za klasifikaciju, korišćena je normalizacija koja je specifična za korpus tvitova. Od leksičkih resursa su korišćeni rečnik reči kojima se izražava sentiment (rečnik sentimenata), rečnik stop reči, rečnik specifičnih vrsta reči koje za koje postoje pravila upotrebe u rečeničnim konstrukcijama i koja će biti obrađena kod procesiranja negacije. Dodatna poboljšanja su dobijena kreiranjem atributa teksta koji nose kritičnu količinu informacija. Ovi atributi su dobijeni tehnikama mašinskog učenja (selekcija i redukcija atributa).

6.2. **Skup podataka**

Dosadašnje analize sentimenta za srpski jezik su rađene nad skupovima podataka dugih tekstova (novinski članci, recenzije filmova). Porast korišćenja socijalnih mreža i raspoloživost teksta na njima, čini ovu vrstu teksta popularnom za istraživanje. Jedna od najpopularnijih društvenih mreža čiji se podaci koriste za analizu sentimenta u poslednje vreme jeste društvena mreža Tviter (*eng. Twitter*). Tekstualna objava na društvenoj mreži Tviter se naziva tvit (*eng. tweet*). U tvitovima ljudi često izražavaju svoje mišljenje o određenim fenomenima, stvarima, ličnostima. Izražavanje sentimenta u tvitu je sažeto zbog ograničenog broja karaktera od kojih se tvit sastoji (u vreme sakupljanja skupa podataka dužina tvita je bila ograničena na 140 karaktera). Za testiranje metode izabran je skup podataka tvitova. Ne postoji označeni korpus kratkih tekstova za analizu sentimenta na srpskom jeziku, pa su za ovaj eksperiment tvitovi posebno sakupljani. U nastavku će biti opisani načini prikupljanja podata i njihova struktura.

6.3. **Priprema skupa podataka**

Tvitovi su sakupljani koristeći Twitter Streaming API u periodu od 30.11.2016 do 30.6.2017.godine. Skup je ručno ozačen od strane tri osobe od kojih su dva muškarca i jedna žena sa zanimanjem lekara, elektroinženjera i studenta srpskog jezika i književnosti. U slučaju neslaganja bilo koja dva ili sva tri anotatora kod označavanja tvita, takav tvit je izbačen iz skupa. Konačni skup podatka se sastoji od 7636 tvitova a struktura tvitova po klasama sentimenata je data u tabeli 6.1.

Tabela 6.1: Broj tvitova po klasama

Sentiment	Broj tvitova
Negativni	4193
Neutralni	2625
Pozitivni	818

6.4. Normalizacija skupa podataka

Tvitovi su kratki, uglavnom neformalno napisani tekstovi koji često sadrže nepravilno napisane reči, neformalno napisane reči, ironiju i sarkazam. Ako sve ovo uzme u obzir, zaključak je da je normalizacija i klasifikacija ovakvih neformalno pisanih tekstova vrlo zahtevan zadatak. Autori su u [70] kreirali posebne sisteme za normalizaciju ovakvih tekstova.

Metode klasifikacije po sentimentu zahtevaju normalizovan skup podataka, kao i jezičke resurse koji se koriste u metodi. Proces normalizacije zahteva odgovarajući redosled primene alata kojima će se početni skup podataka i jezičkih resursa postepeno normalizovati u oblik koji će biti ulaz u metodu klasifikacije po sentimentu. Normalizcija tvitova zahteva sledeći redosled primene alata:

- Tokenizacija
- Prebacivanje teksta u jedno pismo - latinično
- Podela na rečenice
- Izbacivanje stop reči
- Stemovanje.

6.4.1. Tokenizacija i svođenje na jedno pismo

Tokenizacija je rađena tokenizatorom izdrađenim u Python programskom jeziku koristeći nltk.tokenize modul; RegexpTokenizer je korišćen za obradu datuma, raznih formata brojeva, oznaka termina ili fraza (*eng. hashtags),* oznaka korisnika (*eng. mentions*) i izraza koje Tviter koristi a koji nisu od začaja za konkretnu analizu (via, RT,…). Tvit je podeljen na reči korišćenjem nltk.tokenize metode word_tokenize.

Pošto srpski jezik ima dva zvanična pisma (ćirilicu i latinicu) bilo je potrebno obraditi posebno tvitove napisane ćirilicom i posebno napisane latinicom. Kako bi se izbegla dupla analiza, tvitovi koji su napisani ćirilicom su prebačeni u latinicu. Posle prevođenja na latinično pismo i podele na rečenice, odrađeno je stemovanje. Korišćeni stemer je opisan u

sledećem potpoglavlju. Na ovaj način je odrađena normalizacija teksta kako bi tekst bio pogodan za dalju analizu sentimenta.

6.4.2. Stemovanje

Za svođenje različitih oblika reči na jedinstven oblik je korišćen stemer opisan u radu [71]. Stemer kodira specijalne karaktere ‘č’, ‘ć’, ‘š’, ‘đ’, ‘ž’ u ‘cx’, ‘cy’, ‘sx’, ‘dx’, ‘zx’. Stemer kreira stemove duže od 2 karaktera i ima rečnik nepravilnih glagola i njihovih fleksija. Ostale reči se stemuju generalizovanim pravilima stemera. Zadaci koje navedeni stemer ne obrađuje jesu nepravilne fleksije, glasovne promene i kratke reči. Pored navedenih nedostataka, preciznost stemera je 90% pa je odlučeno da je to dovoljno za normalizovanje skupa podataka korišćenog u disertaciji. Prilagođavanje stemera za različite eksperimente je moguća jer je kôd javno dostupan u Python programskom jeziku. U disertaciji je testirano stemovanje na stem duži od 2 slova bez obzira na dužinu reči i stemovanje na stem duži od 2 slova ali samo za reči koje su duže od 4 slova. Eksperiment je pokazao da se stemovanjem na stem duži od 2 slova ali samo za reči duže od 4 slova dobija bolji rezultat. Srpski jezik je morfološki bogat, stoga se stemovanjem očekuje poboljšanje u klasifikaciji sentimenta. Stemovanjem se različite forme jedne iste reči svode na isti stem pa se i broj pojavljivanja tog stema povećava. U [49] je Batanović pokazao da se korićenjem stemera na njihovom skupu podataka povećava prosečan broj pojavljivanja svake reči približno za 50% i da se smanjuje veličina sentiment rečnika za svaku klasu za približno 30-35%. Posle stemovanja, skup podatka je spreman za primenu pravila za obradu negacije.

6.5. Uticaj negacije na reči kojima se izražava sentiment

Kako bi se utvrdila veza negacije i reči kojima se izražava sentiment, posmatrano je ponašanje ovakvih reči koje se pojavljuju u korpusu tvitova. Reči kojima se izražava sentiment su posmatrane u dva slučaja: kada se javljaju u afirmativnom kontekstu (nema negacije) i kada se javljaju u negiranom kontekstu (postoji negacija). Reči kojima se izražava sentiment su klasifikovane u pozitivne ili negativne reči, na osnovu njihovog pojavljivanja u

pozitivnim ili negativnim tvitovima. Za svaku reč kojom se izražava sentiment su izračunati sledeći parametri:

• *br_u_poz_tv_ak* - broj pojava konkretne reči u pozitivnim tvitovima u afirmativnom kontekstu

• *br_u_neg_tv_ak* - broj pojava konkretne reči u negativnim tvitovima u afirmativnom kontekstu

• *br_u_ poz_tv_nk* - broj pojava konkretne reči u pozitivnim tvitovima u negiranom kontekstu (reč se nalazi u opsegu negacije).

• *br_u_neg_tv_nk* - broj pojava konkretne reči u negativnim tvitovima u negiranom kontekstu (reč se nalazi u opsegu negacije).

Očekivano je se da će se pozitivne reči pojavljivati češće u pozitivnim tvitovima, a negativne reči češće u negativnim tvitovima. Kada se reč kojom se izražava sentiment pojavi u opsegu negacije, onda se njen polaritet menja, stoga se očekuje da će se negirane pozitivne reči češće pojavljivati u negativnim tvitovima a da će se negirane negativne reči češće pojavljivati u pozitivnim tweetovima. Klasifikacija reči kojima se izražava sentiment na osnovu njihovog pojavljivanja u tvitovima je izvršena u dve klase na sledeći način:

• reč je klasifikovana kao pozitivna ako važi da je *br_u_poz_tv*> *br_u_neg_tv*

• reč je klasifikovana kao negativna ako važi da je *br_u_poz_tv*< *br_u_neg_tv*

Vrednosti promenljivih *br_u_poz_tv* i *br_u_neg_tv* su računate na tri različita načina (tri metode):

Metoda M0 - ova metoda klasifikuje reči na osnovu svih njihovih pojavljivanja bez obzira da li se pojavljuju uz signal negacije ili ne:

br_u_poz_tv =br_u_poz_tv_ak + br_u_poz_tv_nk

br_u_neg_tv =br_u_neg_tv_ak + br_u_neg_tv_nk

Metoda M1 - Ova metoda klasifikuje reči na osnovu svih njihovih pojavljivanja samo u afirmativnom kontekstu (samo ako se nisu javljale pored signala negacije):

br_u_poz_tv = br_u_poz_tv_ak

br_u_neg_tv = br_u_neg_tv_ak

Metoda M2 - Ova metoda klasifikuje reči na osnovu svih njihovih pojavljivanja uzimajući u obzir promenu polariteta sentimenta ako su se reči kojima se izražava sentiment pojavile uz signal negacije:

br_u_poz_tv = br_u_poz_tv_ak + br_u_neg_tv_nk

br_u_neg_tv = br_u_neg_tv_ak + br_u_poz_tv_nk

Reči kojima se izražava sentiment koje su se pojavljivale jednako u pozitivnim i negativnim tvitovima ili su se pojavljivale samo u neutralnim tvitovima nisu klasifikovane. Tabela 6.1 prikazuje rezultate klasifikacije reči iz rečnika kada se primenjuju navedene tri metode klasifikacije. Formula (6.1) predstavlja preciznost (*Pre*), kao broj ispravno klasifikovanih reči (br_tačno_klasifikovanih) podeljenih ukupnim brojem sentiment reči (br_klasifikovanih). Formula (6.2) predstavlja odziv (*Rcall*) kao broj ispravno klasifikovanih reči (br_tačno_klasifikovanih) podeljeno sa ukupnim brojem reči koje se pojavljuju u korpusu (br_ukupno). Mera F1 koristi kombinaciju Pre i Rcall predstavljenu u formuli (6.3), dajući relevantnije rezultate sa neuravnoteženim skupom podataka.

$$Preciznost = \text{br\_tačno\_klasifikovanih} / \text{br\_klasifikovanih} \tag{6.1}$$

$$Odziv = \text{br\_tačno\_klasifikovanih} / \text{br\_ukupno} \tag{6.2}$$

$$F1 = 2 * \text{Preciznost} * \text{Odziv} / (\text{Preciznost} + \text{Odziv}) \tag{6.3}$$

Tabela 6.2: Klasifikacija reči kojima se izražava sentiment na osnovu njihovog pojavljivanja u pozitivnim i negativnim tvitovima

Metoda	M0	M1	M2
Preciznost	74.41%	76.43%	76.37%
Odziv	67.45%	66.41%	67.87%
F1mera	70.76%	71.07%	71.87%

Na osnovu pojavljivanja sentiment reči samo u afirmativnom kontekstu (M1), dobijena je najbolja preciznost u klasifikaciji, što potvrđuje očekivanu distribuciju sentiment

reči u tvitovima. S druge strane, pojava ovih reči uz signal negacije je poremetila tačnost klasifikacije ako se negacija nije obradila (M0). U obradi negacije jednostavnom promenom polariteta, dobijeni su bolji rezultati (M2 je imao najbolje rezultate kod mere F1), ali je preciznost bila nezadovoljavajuća, što je rezultiralo potrebom za definisanjem pravila obrade negacije.

6.6. **Obrada pravila negacije**

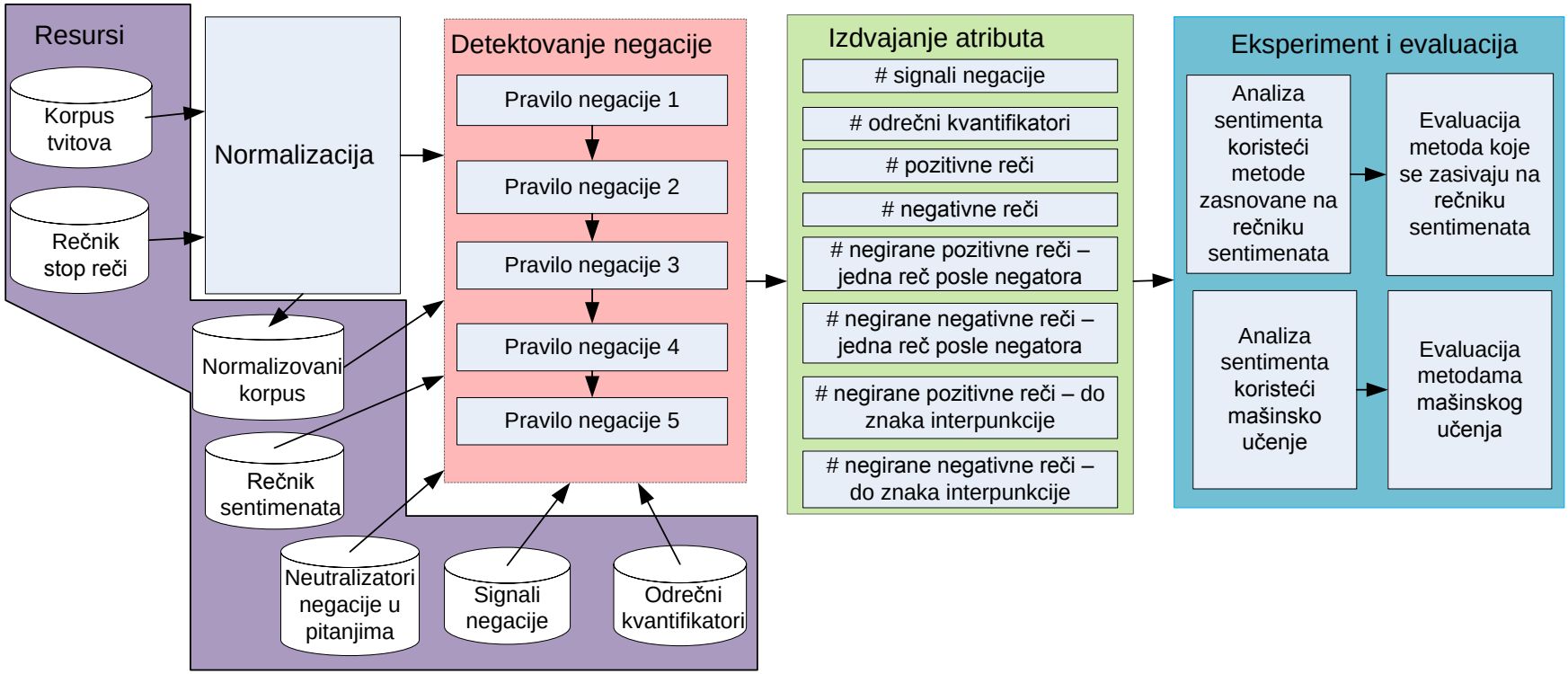
Analizom skupa podataka izdvojena su pravila negacije koja se najčešće javljaju kod tvitova na srpskom jeziku pa su ta pravila obrađena. Utvrđeno je da se primenom pravila negacije poboljšava detektovanje opsega na koji se negacija odnosi. Obrađena pravila negacije su:

- Obrada negacije tipa "Nije…. nego(već)" – definiše se opseg negacije od signala negacije do reči "nego/već"
- Obrada negacije tipa "Nije samo …. nego/već" – izbacuje se reč koja je signal negacije a koja stoji ispred reči „samo"
- Obrada negacije u pitanjima tipa "Zar nije lepo?", "Zar to nije previše prosto?" – u ovom slučaju dolazi do poništavanja (neutralizacije) negacije jer sentiment reč posle signala negacije ne menja polaritet sentiment reči.
- Obrada prve reči sa sentimentom posle negacije ali do znaka interpunkcije. Ako se posle negirane reči javi pojačivač onda negirati i sledeći deo rečenice.
- Pojačavanje negacije odrečnim kvantifikatorima - slaganje odrečnih kvantifikatora sa negacijama. Odrečni kvantifikatori pojačavaju intenzitet negacije.

Na slici 6.1 je data arhitektura sistema za klasifikaciju tvitova po sentimentu koji uključuje specifična pravila obrade negacije. Na slici su date sve komponente sistema uključujući i detektovanje negacije. Resursi se sastoje od korpusa tvitova, liste stop reči, rečnika reči kojima se izražava sentiment, signala negacije, odrečnih kvantifikatora i negatora koji se javljaju u specifičnim pitanjima a koji neutralizuju negaciju. Korpus tvitova prolazi

kroz komponentu pretprocesiranja koja kao izlaz daje korpus koji je normalizovan. Komponenta za detektovanje negacije se sastoji od skupa pravila koja su prethodno objašnjena. Da bi se primenila metoda za klasifikaciju tvitova po sentimentu, potrebno je izdvojiti osobine/atribute teksta tvita na osnovu kojih je moguće izvršiti eksperiment metodama klasifikacije koje će biti u nastavku detaljno opisane. Eksperiment i evaluacija predstavljaju poslednju komponentu sistema i one su rađene u dva slučaja: metodama klasifikacije zasnovanim na rečniku reči kojima se izražava sentiment i metodama mašinskog učenja. Rezultati kvaliteta klasifikacije će se kod evaluacije izraziti merama tačnosti, preciznosti i odziva.

Kod komponente za detekciju negacije, pravilo1 i pravilo2 su pravila za parcijalnu negaciju pa se njima negira samo deo rečenice. Pravilo1 isključuje pravilo2 i potrebno je prvo primeniti pravilo1. Ispravan redosled primene pravila (pravilo1 pre pravila2) je bitan jer je pravilo1 u stvari pseudonegacija (negiranim delom izražava potvrdno značenje). Pravilo3 se odnosi na posebna pritanja koja sadrže negaciju kojoj prethodi rečca “Zar” ili rečca “Jel”. Ove rečce čine rečnik neutralizatora negacije u pitanjima. Pravilo4 je opštije i može se primeniti za određivanje opsega negacije nezavisno od jezika. Njegovom primenom obavezno dolazi do promene polariteta negirane reči. U pravilu5 se identifikuju odrečni kvantifikatori koji se slažu sa negacijom (potvrđuju je) ili je pojačavaju.



Slika 6.1: Arhitektura sistema za klasifikaciju tvitova po sentimentu koji uključuje specifična pravila obrade negacije

U tabeli 6.3 su dati primeri po jednog tvita na kojima su primenjena definisana pravila.

Tabela 6.3: Primeri primene pravila negacije

Pravilo	Primer	Šta je urađeno
1	“ Ovo nas čeka kad nam stignu rate ovih kredita koje dižu a novac ne investiraju nego ga troše”	Definisan opseg negacije: “…ne investiraju nego..”
2	“Ne samo bezobrazluk, vec i gazenje i ismevanje naroda.”	Negacija neutralisana
3	“Zar vas @KurirVesti @Blic_online i ostale nije sramota da ne pratite uzivo šta se desava...?”	Negacija neutralisana
4	"I kad budem odlazio, otići ću sam, neće me pobediti"	Promena polariteta reči “pobediti”
5	””Taj čovek ništa drugo i ne zna da radi”	Pojačan sentiment negiranoj reči “zna” , zbog prisustva odrečog kvantifikatora “ništa”

6.7. **Struktura i rad korišćene metode**

S obzirom da su značajna poboljšanja u određivanju sentimenta teksta generalno postignuta bez uzimanja u obzir lingvističih pravila jezika na kojem je tekst (posebno za engleski jezik), zalaženje u procesiranje posebnih lingvistička pravila specifičnog jezika daje podlogu za mogućnost poboljšanja određivanja sentimenta. Prisustvo negacije kod neformalnog i kratkog teksta, kao u slučaju tvitova, otežava generanlo predviđanje sentimenta. Korišćena metoda treba da odredi sentiment tvita uzimajući u obzir obrađena pravila negacije. Metoda koristi sledeće resurse: rečnik sentimenata, rečnik signala negacije, rečnik neutralizatora u pitanjima i rečnik odrečnih kvantifikatora.

Jezici koji su morfološki bogati (*eng. highly inflected*), kao što je i srpski jezik, zahtevaju primenu ligvističkih pravila za kvalitetnu obradu i klasifikaciju teksta. Da bi se tekst klasifikovao, prvo mora proći kroz proces normalizacije. Svaka klasifikacija teksta zahteva i specifičnu normalizaciju teksta. Normalizacija tvita je rađena primenom tokenizatora prilagođenog specifičnoj stukturi tvitova, prebacivanjem teksta u latinično pismo, uklanjanjem stop reči i stemovanjem. Posle normalizacije teksta utvrđuje se da li tekst sadrži negaciju i ako je sadrži kojoj grupi pravila negacije pripada. Korišćena pravila za obradu negacije su opisana u poglavlju 6.6. Posle obrade pravila negacije, pristupa se izdvajanju atributa značajnih za određivanje sentimenta, a to su:

- broj signala negacije
- broj odrečnih kvantifikatora
- broj pozitivnih reči
- broj negativnih reči
- broj negiranih pozitivnih reči - jedna reč nakon negacije
- broj negiranih negativnih reči - jedna reč nakon negacije
- broj negiranih pozitivnih reči - do prvog znaka interpunkcije
- broj negiranih negativnih reči - do prvog znaka interpunkcije

Pored atributa koji se dobijaju primenom rečnika sentimenata i pravila negacije, biće izdvojeni tekstualni atributi iz normalizovanog teksta i izbor ovih dodatnih tekstualnih atributa će biti prikazan u poglavlju u kojem su dati rezultati testiranja korišćene metode. Korišćenjem izdvojenih atributa se vrši predviđanje sentimenta tvitova pomoću dva pristupa:

- korišćenjem metode klasifikacije koja se zasniva na rečniku sentimenata (LBM metode)
- korišćenjem metoda nadgledanog mašinskog učenja (MLM metode)

Efekat metode će biti prikazan na primeru klasifikacije tvitova u odnosu na dve poredbene metode. Prva poredbedna metoda ne obrađuje negaciju a druga obrađuje negaciju prostom promenom polariteta reči posle negacije. Ove dve metode su detaljno opisane u sledećem potpoglavlju.

6.8. **Poredbene metode**

Metoda klasifikacije po sentimentu koja uključuje obrađena pravila negacije je upoređivana sa dve poredbene metode. Prva poredbena metoda (Metoda0) klasifikuje tvit po sentimentu bez obrade negacije, samo na osnovu prostog brojanja pozitivnih i negativnih reči iz rečnika sentimenata. Ova metoda je uzeta u obzir jer će biti korisno uporediti je i sa metodom koja obrađuje negaciju na klasičan način (Metoda1), kako radi većina sistema za obradu negacije.

Druga poredbena metoda (Metoda1) klasifikuje tvit po sentimentu obrađujući negaciju na način tako što prosto menja polaritet prve reči kojim se izražava sentiment a koja se nalaze iza negatora (signala negacije). Ovaj pristup obrade negacije koristi većina sistema za obradu negacije kod analize sentimenta. Metoda1 je interesantna za poređenje jer se očekuje da će raditi bolje od metode koja uopšte ne obrađuje negaciju (Metoda0). Međutim, nekada može doći i do situacije da obrada negacije prostom promenom polariteta rečima kojima se izražava sentiment u opsegu negacije daje lošije rezultate nego u slučaju kada se uopšte ne obrađuje.

Metoda2 je metoda koja je u disertaciji korišćena i koja je opisana u prethodnom potpoglavlju. Očekuje se da će obrađena pravila sintaksičke negacije koja su uzeta u obzir kod korišćene metode lepo pokazati kako se obradom pravila utiče na kvalitet klasifikacije po sentimentu. Poređenjem sa dve prethodno opisane poredbene metode će moći da se vidi efekat primene ovih pravila.

7. JEZIČKI RESURSI

U ovom poglavlju su opisani jezički resursi koji su korišćeni kod obrade negacije i analize sentimenta. Analiza sentimenta zahteva upotrebu rečnika sa označenim sentimentom reči (rečnik sentimenata). Signali negacije, negativni kvantifikatori, pojačivači i stop reči su neophodni jezički resursi u procesu analize sentimenta sa obradom negacije. Upotreba rečnika stop reči je široko rasprostranjena u analizi teksta i mora se prilagoditi analizi negacije i generalno sentimenta kako se neka reč koja je potrebna (pripada opštem rečniku stop reči) ne bi izbacila.

7.1. Rečnik sentimenata

Rečnik sentimenata je resurs koji je potreban većini metoda za određivanje sentimenta a direktnu upotrebu rečnika sentimenata imamo kod metoda određivanja sentimenta koje se zasnivaju na ovim rečnicima. Uglavnom se sastoji od spiska reči kojima se izražava sentiment sa dodeljenim polaritetom ili intenzitetom polariteta. Nekada reči u rečnicima sentimenata mogu sadržati i dodatne karakteristike.

Većina gotovih rečnika sentimenata je na englenskom jeziku i ima opštu namenu. Rečnik sentimenata „General Inquirer" potiče još iz 1966. godine [72] i sastoji se od 1915 pozitivnih i 2291 negativnih reči. Rečnik Linguistic Inquiry and WordCount (LIWC) [73] sadrži 2300 reči i 70 klasa kao što su negativne emocije (mržnja, bes, problematičnost…) i pozitivne emocije (ljubav, lepota, dobrota…). MPQA rečnik [24] sadrži 6885 reči od 8221 lema, od čega 2718 pozitivnih i 4912 negativnih. Opinion Lexicon se sastoji od 6786 reči, od čega 2006 pozitivnih i 4783 negativnih i prva verzija je data u [23]. SentiWordNet [43] je rečnik gde svaka reči ima pridružen koficijent koji određuje koliko je pozitivna, negativna ili objektivna. Još uvek ne postoji javno dostupan rečnik sentimenata za srpski jezik. Za potrebe rada na disertaciji je izrađen rečnik sentimenata [74] koji kao polaznu verziju imaj prevod rečnika Opinion Lexicon [30].

Prisustvo sentimenta u tvitovima podrazumeva pojavu reči kojima se izražava sentiment. Pošto su tvitovi kratki tekstovi, broj pojava reči kojima se izražava sentiment u njima je mali, što čini njihovo prisustvo značajnijim.

Da bi se sentiment utvrdio samo na osnovu takvih reči, postoje otežavajuće okolnosti kao što je prisustvo negacije, upotreba ironije i sl.

7.1.1. Normalizacija rečnika sentimenata

Normalizacija sentiment rečnika je rađena da bi se utvrdilo koji oblik normalizacije rečnika najbolje koristi metodi klasifikacije tvitova po sentimentu. Kao polazni rečnik je korišćen rečnik sentimenata koji se sastoji od 5632 reči (svedene na morfološku osnovu), od toga je 4058 negativnih i 1574 pozitivnih reči [75]. Šest različitih vrsta normalizacija je primenjeno nad njim i kao rezultat su dobijeni normalizovani rečnici koji su korišćeni kod skupa podataka koji je takođe normalizovan jednom od normalizacija (stemovanje, normalizacija morfološkim rečnikom, odsecanje na 4-grame, odsecanje na 5-grame, odsecanje na 6-grame, odsecanje na 7-grame).

Primena normalizacije na rečnike sentimenata utiče na ukupan broj reči u rečniku kao i na njihov kvalitet. Primenom jezički zavisnih normalizatora (stemera i morfološkog rečnika), reči sa istim ili sličnim značenjem se svode na zajedničku osnovu. Primenom normalizacije odsecanjem na *n*-grame, reči iz rečnika se odsecaju na prve *n*-grame pri čemu se ne vodi računa o značenju reči. Zbog karakteristika odgovrajućih normalizatora dobijaju se rečnici sa značajno različitim brojem reči [76].

Obređeni broj reči iz rečnika sentimenta koje imaju različit sentiment se normalizacijom transformišu u isti koren, zbog čega ove reči postaju kontradiktorne. Normalizacije zasnovane na jezičkim pravilima daju manji broj takvih reči. Kontradiktorni podaci će biti isključeni iz rečnika sentimenata.

U tabeli 7.1 su prikazani rezultati nakon normalizacije reči u rečnicima i odstranjivanja kontradiktornih reči. Prikazan je broj reči u rečniku kao i ukupan broj različitih osnova dobijenih nakon normalizacije. Radi boljeg poređenja rezultata prikazani su rezultati dobijeni kada se ne primenjuje nijedna normalizacija (NN) i rezultati primene različitih

normalizacija: stemovanja (ST), normalizacija morfološkim rečnikom (NM) i odsecanjem na 4-grame (4G), 5-grame (5G), 6-grame (6G), 7-grame (7G).

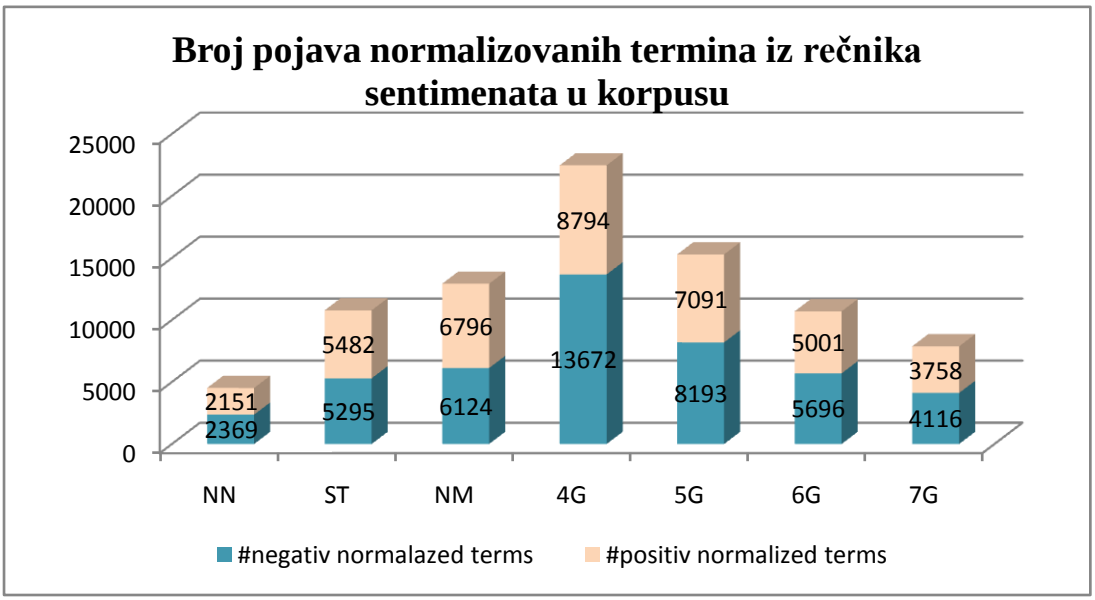
Tabela 7.1: Broj reči u normalizovanim rečnicima sentimenata posle izbacivanja kontradiktornih i broj različitih osnova na koje su svedene normalizacijama

Tip normalizcije	NN	ST	NM	4G	5G	6G	7G
Broj različitih reči	5632	5596	5632	4139	5116	5481	5576
Broj različitih osnova	5632	5218	5632	2271	3506	4283	4803

Broj različitnih osnovnih oblika u nomalizovanim rečnicima se smanjuje usled isključivanja kontradiktornih reči što je najizraženije kod normalizacije odsecanjem na 4-grame, a opada kako raste dužina *n*. Stemovanjem je takođe značajan broj reči izbačen.

Za svaki oblik normalizacije, izračunato je pojavljivanje ovako normalizvanih reči kojima se izražava sentiment nad skupom podataka tvitova. Ukupan broj pojava reči kojima se izražava sentiment je sledeći: za stemovanje 10777; za lematizaciju 12920; za 4-gram 22466; za 5-gram 15284; za 6-gram 10697; za 7-gram 7874. Raspodela broja pojava sentiment reci u korpusu po polaritetu je prikazana na slici 7.1. Sa slike se vidi da je broj pojava termina iz normalizovanih rečnika u tvitovima iz korpusa najveći za 4-grame. Sledeći po broju pojava jesu 5-grami a posle njih slede reči normalizovane morfološkim rečnikom. Ova raspodela ukazuje da su reči koje su odsecane na 4-grame najbolje mapirane u tvitovima, što je i očekivano zbog velikog broja različitih oblika reči koje počinju istim 4-gramom.

Ono što je vidljivo sa slike 7.1 i iz tabele 7.1 je da sa porastom dužine *n* dolazi do smanjenja uticaja normalizacije pa se odsecanjem na 7-grame dobijaju podaci koji teže rezultatima bez normalizacije.

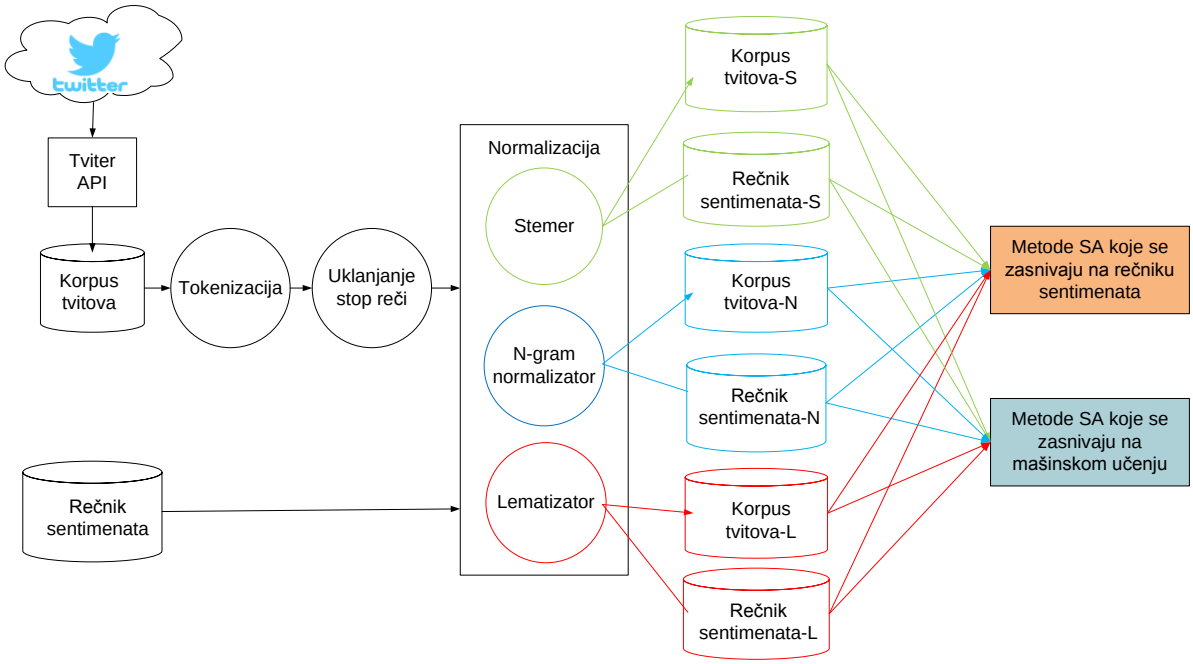


Slika 7.1: Broj pojava normalizovanih termina iz rečnika sentimenata u korpusu

U narednom poglavlju, kvalitet dobijenih rečnika je proveren ispitivanjem da li se ove reči iz normalizovanih rečnika pojavljuju u tvitovima sa odgovarajućim polaritetom i kako ovako normalizovani rečnici utiču na klasifikaciju po sentimentu.

7.1.2. Validacija rečnika sentimenata

Da bi se ispitalo kako svođenje reči na odgovarajuće osnove primenom različitih tipova normalizacije utiče na klasifikaciju tvitova, urađena je validacija rečnika sentimenata nad skupom tvitova koji su normalizovani istim vrstama normalizacija. Na slici 7.2 je dat postupak normalizacije rečnika i skupa podataka kako bi se uradila validacija rečnika sentimenata nad datim skupom podataka. Pre primene konkretnog oblika normalizacije, potrebno je da korpus tvitova prođe kroz faze tokenizacije i uklanjanja stop reči. Posle toga, i korpus i rečnik prolaze kroz komponentu normalizacije koja se sastoji od tri normalizatora: stemer, n-gram normalizator i normalizator morfološkim rečnikom. Na izlazu iz komponenente normalizacije imamo normalizovane rečnike i korpuse za svaki tip normalizacije. Sa ovim podacima se dalje vriši klasifikacija po sentimentu.



Slika 7.2: Proces normalizacije i validacije korpusa i rečnika sentimenata

Za svaki oblik normalizovane reči iz rečnika sentimenata (stem, lema ili *n*-gram) izračunava se rezultat (broj pojavljivanja tog oblika reči u tvitovima sa pozitivnim i negativnim sentimentom). Ako se reč javi u opsegu negacije, njena pojava se računa kao da se pojavila u tvitu sa suprotnim polaritetom. Normalizacija rezultata se vrši dijeljenjem broja pojavljivanja s brojem tvitova iz te klase. Rezultat se izračunava po formuli (7.1),

$$rezultat = \frac{p - np}{br\_poz\_tv} - \frac{n - np}{br\_neg\_tv} \tag{7.1}$$

gde je:

- *p* - broj pojavljivanja reči u pozitivnim tvitovima
- *np* - broj pojavljivanja reči sa negacijom u pozitivnim tvitovima
- *n* - broj pojavljivanja reči u negativnim tvitovima
- *nn* - broj pojavljivanja reči sa negacijom u negativnim tvitovima
- *br_poz_tv* – broj pozitivnih tvitova
- *br_neg_tv* – broj negativnih tvitova

Klasifikacijom reči kojima se izražava sentiment na osnovu toga da li se pojavljuju više u pozitivnim ili negativnim tweetovima, testira se efekat normalizacije na analizu sentimenta. Rečima u normalizovanim rečnicima je dodeljen sentiment na sledeći način:

- pozitivan – ako se pojavljuju više u pozitivnim nego u negativnim tvitovima, tj. ako je rezultat>0
- negativan – ako se pojavljuju više u negativnim nego u pozitivnim tvitovima, tj. ako je rezultat<0

Tabela 7.2 prikazuje rezultate klasifikacije: reči sa pozitivnim sentimentom, reči sa negativnim sentimentom i svih reči iz rečnika sentimenata i to kada se primjenjuju različite vrste normalizacije. Prikazana je preciznost, tj. procenat tačno klasifikovanih u odnosu na ukupan broj klasifikovanih za tu klasu i odziv, tj. procenat tačno klasifikovanih u odnosu na celu populaciju odgovarajuće klase. Mera F1 koristi kombinaciju preciznosti i odziva i kod nebalansiranog skupa podataka daje relevantnije rezultate. Opisane mere su računate po formulama koje su date u odeljku 4.1.4.

Tabela 7.2: Dobijena preciznost, odziv i F1 klasifikacije sentiment reči na osnovu korpusa za različite tipove normalizacija. Podebljani su odgovarajuci maksimumi.

		NN	**ST**	**NM**	**4G**	**5G**	**6G**	**7G**
negativne sentiment reči	Pre	82%	81%	80%	84%	82%	81%	80%
	Rcall	16%	27%	25%	60%	45%	35%	28%
	F1	26%	41%	38%	**70%**	58%	49%	41%
pozitivne sentiment reči	Pre	79%	67%	66%	52%	60%	64%	69%
	Rcall	13%	24%	19%	47%	37%	28%	22%
	F1	22%	35%	30%	**49%**	46%	39%	34%
sve sentiment reči	Pre	81%	77%	77%	75%	76%	76%	77%
	Rcall	15%	26%	23%	57%	43%	33%	26%
	F1	25%	39%	36%	**65%**	55%	46%	39%

Na osnovu dobijenih rezultata prikazanih u tabeli 7.2 se može zaključiti da su *n*-grami, a posebno 4-grami, dobro klasifikovani po sentimentu (klasifikacija ima najbolji F1 rezultat). Razlog tome je što je veći broj *n*-grama pronađen u skupu tvitova u poređenju sa stemovima i lemama. Kao neformalni tekstovi, tvitovi često sadrže pogrešno napisanu reč. S druge strane,

srpski jezik kao morfološki bogat jezik je teško obraditi, i veliki broj reči se nalazi u oblicima koji nisu adekvatno obrađeni od strane stemera i lematizatora, tako da se takve reči kojima se izražava sentiment ne mogu naći u rečniku sentimenata.

Testiranje poboljšanja klasifikacije reči kojima se izražava sentiment odsecanjem na 4-grame naspram svođenja na stemove i leme je vršeno je primenom McNemar-ovog testa [77]. Napravljena je korelaciona matrica za klasifikaciju poredeći 4-gram normalizaciju i lematizaciju i poredeći 4-gram normalizaciju i stemovanje. U oba slučaja je utvrđeno da je vrednost p <0.0001, što znači da odsecanje na 4-grame značajno utiče na poboljšanje klasifikacije sentimentalnih reči.

7.1.2.1. Uticaj vrste normalizacije na određivanje sentimenta tvitova

Uticaj svih navedenih načina normalizacije je testiran na određivanju sentimenta tvitova. Eksperiment uticaja načina normalizacije rečnika sentimenata i skupa podataka na određivanje sentimenta je urađen na dva načina. Prvi eksperiment klasifikuje tvit samo na osnovu reči koje su pronađene u sentiment rečniku (LBM). U drugom eksperimentu su za klasifikaciju po sentmentu korišćena metoda učenja sa nadgledanjem: nominalna logistička regresija ili maksimalna entropija (MLM).

Kvalitet predviđanja odgovarajućeg načina normalizacije je u prvom eksperimentu rađen metodom klasifikacije koja se zasniva na rečniku sentimenata. Sentiment je računat po formuli (7.2). Suma pozitivnih termina iz rečnika sentimenata koji se pojavljuju u tvitu je data u atributu *sumPos*, sa tim što se ovom broju dodaju negativni termini koji se nalaze u opsegu negacije. Suma negativnih termina iz rečnika sentimenata koji se pojavljuju je data u atributu *sumNeg* s tim što se ovome broju dodaju pozitivni termini koji se nalaze u opsegu negacije [78].

$$Sentiment\ tvita = \begin{cases} pozitivan\ ako\ sumPos > sumNeg \\ neutralan\ ako\ sumPos = sumNeg \\ negativan\ ako\ sumPos < sumNeg \end{cases} \quad (7.2)$$

Rezultati dobijeni ovom metodom (LBM) su dati u tabeli 7.3 i to upoređujući rezultate bez primene normalizacije sa rezultatima kada se primeni normalizacija stemovanjem, normalizacija morfološkim rečnikom ili normalizacija odsecanjem na *n*-grame (n = 4,5,6 i 7).

Prikazani su rezultati klasifikacije tvitova za sve tri klase (3K) kao i tačnost klasifikacije za dve klase - pozitivne i negativne (2K).

Tabela 7.3: Tačnost klasifikacije tvitova metodom koja se zasniva na rečniku sentimenata u zavisnosti od normalizacije

Normalizacija	NN	ST	NM	4G	5G	6G	7G
LBM 3K	47.09%	**50.51%**	48.98%	49.25%	49.10%	50.13%	49.15%
LBM 2K	33.45%	**52.17%**	50.73%	**59.69%**	53.54%	49.51%	43.76%

Ako se posmatra samo klasifikaciju pozitivnih i negativnih tvitova, dobija se da 4-gram normalizacija daje najbolju tačnost. Međutim, neutralni tvitovi dosta remete klasifikaciju, pa kod klasifikacije skupa tvitova sa tri klase normalizacija stemovanjem daje najbolje rezultate. U ovom slučaju, odsecanje na *n*-grame pronalazi veliki broj *n*-grama sa sentimentom u tvitovim, pa i u neutralnim čime ih klasifikuje u pozitivne ili negativne. Neutralni tvitovi neretko nose deo sentimenta koji nije jasno definisan pa se ovo može rešiti tek uvođenjem klasifikacije tvitova na više sentiment grupa. S druge strane, odsecanjem na *n*-grame je izgubljen jedan deo reči koje nose sentiment ako su nakon normalizacije upale u grupu kontradiktornih. Izostavljanjem ovih reči gubi se preciznost određivanja sentimenta, pa to može biti razlog slabije tačnosti normalizacije odsecanjem na 4-grame za 3K (49.25%).

U drugom eksperimentu je primenjeno nadgledano mašinsko učenje metodom nominalne logističke regresije. Korišćena je stratifikovana unakrsna validacija sa 10 slojeva (10-fold cross-validation). Rezultati ovog eksperimenta su prikazani u tabeli 7.4. Korišćeni su isti atributi kao i u prvom eksperimentu: broj pozitivnih reči, broj negativnih reči, broj negiranih pozitivnih reči - jedna reč nakon negacije, broj negiranih negativnih reči - jedna reč nakon negacije.

Tabela 7.4: Tačnost klasifikacije tvitova korišćenjem metode mašinskog učenja u zavisnosti od normalizacije

Normalizacija	NN	ST	NM	4G	5G	6G	7G
MLM-3K	59.86%	**64.17%**	62.45%	59.23%	60.61%	62.49%	61.80%
MLM -2K	84.71%	**85.27%**	84.25%	83.96%	84.45%	84.47%	84.71%

Ovde se vidi značajno povećanje tačnosti klasifikacije kod svakog načina normalizacije. Stemovanjem se dobija najbolja tačnost. Normalizacija odsecanjem na *n*-grame u ovom slučaju je najbolja za 6-grame i ima veću tačnost od normalizacije morfološkim rečnikom.

7.2. **Rečnik signala negacije**

Postoje reči u srpskom jeziku koje su uvek pristutne u kreiranju sintaktičke negacije, to su takozvani negatori. Ovim rečima je u disertaciji dat naziv "signali negacije". Rečnik signala negacije dobijen je izdvajanjem svih oblika reči na srpskom jeziku koji učeststvuju u kreiranju sintaktičke negacije. Generalni oblici signala negacije "ne" i "ni" su prvi uključeni u rečnik signala negacije. Pored njih, u srpskom jeziku postoje glagoli koji imaju nepravilnu formu kod građenja negacije pa se signal negacije kod njih spaja sa glagolom. Ti glagoli su „jesam“, „biti“, „hteti“ i „imati“ i imaju nepravilnu formu negacije, pa je novi negator dobijen spajanjem nekog od ovih glagola sa osnovnim negatorom. Oblici ovih glagola su: "ne jesam" u sadašnjem vremenu: "nisam", "nisi", "nije", "nismo", "niste", "nisu"; "ne biti" u imperativu: "nemoj" , "nemojmo", "nemojte"; "ne hteti" (u sadašnjem vremenu): "neću", "nećeš", "neće", "nećemo", "nećete", "neće"; "ne imati" u sadašnjem vremenu: "nemam", "nemaš", "nema", "nemamo", "nemate", "nemaju".

Tabela 7.5: Rečnik signala negacije

nema	nemamo	neću	ni	necete
nemate	nemoj	nećavši	nisam	necemo
nemaš	nemojmo	nisu	nisi	neces
nemaju	neće	niste	stop	necu
nemam	nećete	nismo	bez	necavsi
nemojte	nećemo	nije	nece	nit
niti	nećeš	ne	nemas	

Ovi glagoli se u negativnom obliku (negacija) spajaju sa signalima negacije i postaju posebni signali negacije. Svi oblici negacije glagola "jesam" “biti”, "hteti" i "imati" uključeni su u rečnik signala negacije. U signale negacije su uključeni i negatori bez dijakritičkih znakova (“necu, neces,…” označavajući “neću, nećeš,…”), zbog potrebe identifikovanja negacije u neformalnim tekstovima – kakvi su i tvitovi. Reči koje pripadaju rečniku signala negacije su date u tabeli 7.5. Ovaj rečnik će biti korišćen kod detektovanja pravila sintaksičke negacije.

7.3. **Rečnik odrečnih kvantifikatora**

Odrečni kvanitifikatori u pravilima negacije u srpskom jeziku igraju totalno različitu ulogu u odnosu na signala negacije, pa bi bilo pogrešno izjednačavati ih. Ova vrsta reči koja podseća na signale negacije (negatore) je izdvojena u poseban rečnik odrečnih kvantifikatora. Univerzalni odrečni kvantifikatori su leksičke „ni-„ negacije odrečnih zamenica (niko, ništa, nikakav...), odrečnih zameničkih priloga (nikad, nigde, nikako...) i malog broja nezameničkih priloga (nimalo, nijedanput...) [66]. Odrečni kvantifikatori uvek stoje uz negaciju (skoro uvek ispred) i potvrđuju negaciju koja stoji uz njih.

Tabela 7.6. Rečnik odrečnih kvantifikatora

nicija	nijedno	nikakve	nikakvoj	niko	nikoliki	niodkuda
nicije	nijednog	nikakvi	nikakvom	nikog	nikoliko	niotkuda
niciji	nijednu	nikakvih	nikakvome	nikoga	nikom	ničija
nicijoj	nikad	nikakvim	nikakvomu	nikoja	nikome	ničije
nicim	nikada	nikakvima	nikakvu	nikoje	nikomu	ničiji
nigde	nikakav	nikakvo	nikamo	nikoji	nikuda	ničijoj
nijedan	nikako	nikakvog	nikim	nikolika	nimalo	ničim
nijedanput	nikakva	nikakvoga	nikime	nikolike	niodakle	ništa

Npr. u rečenici “Nikad nije lagao, ”“Nikad” je odrečni kvantifikator koji se slaže sa negatorom „nije“ koja menja polaritet reči „lagao“ iz negativne u pozitivni. Slaganje odrečnih

kvantifikatora sa negacijom znači potvrdu ili pojačavanje dela koji se negira. U tabeli 7.6 je dat spisak reči koje pripadaju rečniku odrečnih kvantifikatora. Ovaj rečnik će biti korišćen kod detektovanja pravila sintaksičke negacije.

7.4. **Rečnik pojačivača**

U srpskom jeziku postoje vrste reči koje se zovu partikule i njihov opseg delovanja je takav da se njima iskazuje odnos samo prema sadržaju jednog pojma u rečenici [66]. Razmatrane su partikule koje utiču na negaciju tako što je pojačavaju ili produžavaju opseg važenja. Partikule koje pojačavaju negaciju su korišćene tako što se opseg važenja negacije produžavao i posle znaka interpunkcije ukoliko iza njega stoji partikula pojačivač. Rečnik partikula pojačivača čine ‘ni’, ‘nit’ i ‘niti’.

7.5. **Rečnik stop reči**

Stop reči predstavljaju skup reči nekog prirodnog jezika koje se često pojavljuju u tekstu i obično ne nose nikakvo značenje. Uloga kreiranja rečnika stop reči jeste da se reči koje pripadaju ovom rečniku uklone iz teksta koji se analizira. Ove stop reči predstavljaju potencijalne atribute koji ne donose smisao tekstu pa se njihovim izbacivanjem bitno smanjuje veličina vektorskog prostora atributa. Rečnici stop reči mogu biti opšti (univerzalni) ili specifični za oblast koju se koriste. Specifični rečnici se kreiraju uglavnom izdvajanjem iz skupa podataka reči koje imaju najveći TF koeficijent ili najmanji TF-IDF koeficijent (koeficijenti su opisanu u poglavlju 3.1).

Autori su u [48] eksperimentalno potvrdili da univerzalni rečnik stop reči daje približno iste rezultate kao i specifični rečnici. U radu je korišćen opšti rečnik stop reči [79] koji sadrži uglavnom priloge, rečce i veznike koji ne utiču na smisao rečenice i ne učestvuju ni u jednom pravilu sintaksičke negacije. Rečnik stop reči je korišćen već kod normalizacije dokumenta, pre tokenizacije. Neke stop reči su predstavljene u tabeli 8.7 (231 od ukupno 1008 stop reči).

Tabela 7.7: Neke od stop reči iz rečnika stop reči

a	ali	al	bi	bih	bijah	bijahu
bijaše	bijasmo	bijaste	bila	bile	bili	bilo
bio	biše	bismo	biste	biti	biva	bivaju
bivajući	bivam	bivamo	bivaš	bivate	bivati	bivavši
čija	čije	čijeg	čijega	čijem	čijemu	čiji
čijih	čijim	čijima	čijime	čijoj	čijom	čiju
čim	čime	ću	g.	ga	gđa.	gde
gdečega	gdečem	gdečemu	gdečim	gdečime	gdegde	gdekad
gdekakav	gdekakva	gdekakve	gdekakvi	gdekakvih	gdekakvim	gdekakvima
gdekakvo	gdekakvog	gdekakvoga	gdekakvoj	gdekakvom	gdekakvome	gdekakvomu
iako	ičega	ičem	ičemu	ičim	ičime	igde
ih	ikoga	ikome	ikada	ikakav	ikako	ikakva
ikakve	ikakvi	ikakvih	ikakvim	ikakvima	ikakvo	ikakvog
ikakvoga	ikakvoj	ikakvom	ikakvome	ikakvomu	ikakvu	ikamo
gdekakvu	gdekim	gdekime	gdekoga	gdekoja	gdekoje	gdekojeg
gdekojega	gdekojem	gdekojemu	gdekoji	gdekojih	kakav	kakavgod
kako	kakva	kakve	kakvi	kakvih	kakvim	kakvima
kakvo	kakvog	kakvoga	kakvoj	kakvom	kakvome	kakvomu
kakvu	kamo	kao	kasno	katkad	kim	ko
kod	koga	koja	koje	koječega	koječem	koječemu
koječim	koječime	kojeg	kojega	kojegde	kojekakav	kojekako
kojekakva	kojekakve	kojekakvi	kojekakvih	kojekakvim	kojekakvima	kojekakvo
kojekakvog	kojekakvoga	kojekakvoj	nekako	nekakva	nekakve	nekakvi
nekakvih	nekakvim	nekakvima	nekakvo	nekakvog	nekakvoga	nekakvoj
nekakvom	nekakvome	nekakvomu	nekakvu	nekamo	neke	neki
nekih	nekim	nekima	nekime	neko	nekoga	nekoliko
nekom	nekome	nekomu	nekuda	nešto	nje	njega
njen	njena	njeni	onako	onamo	onda	onde
one	oni	onih	onim	onaj	ona	ono
onog	onoga	onoj	onolik	onolika	onolike	onoliki
onolikih	onolikim	onolikia	onoliko	onolikog	onolikoga	onolikoj
onolikom	onolikome	onolikomu	za	zacijelo	zaista	je
zasigurno	zatim	zato	zbilja	zbog	onoliku	uz

8. TESTIRANJE I REZULTATI PRIMENE METODA

Iako jako nestruktuirani tekstovi, često napisani sa mnogo neformalnosti, tvitovi se ipak mogu obraditi i primenom specifičnih pravila negacije se može poboljšati određivanje sentimenta ovakvih tekstova.

U ovom poglavlju će biti cilj da se utvrdi koliko obrada negacije korišćenim pravilima može da poboljša sentiment analizu kratkih tekstova – u konkretnom slučaju tvitova. U skladu sa ovim urađeno je poređenje tačnosti metoda za analizu sentimenta teksta koja negaciju obrađuje na predloženi način (primenom pravila) sa poredbenim metodama koje su opisane prethodno.

Negacija se moze obrađivati samo promenom polariteta reči iza negacije pa najdalje do prvog znaka interpunkcije [11]. Batanović u [49] vrši analizu utacija negacije na reči iza nje i dobija najbolji rezultat kada se promeni polaritet prvoj reči iza negacije. Za osnovnu metodu za poređenje koja obrađuje negaciju uzeta metoda koja menja polaritet sentimenta samo prvoj reči posle negacije.

Za dve poredbene metode (Metoda0, Metoda1) i korišćenu (Metoda2), primenjeni su različiti pristupi klasifikaciji sentimenta:

- Metodom klasifikacije zasnovanom na sentiment rečniku (LBM)
- Metodeom mašinskog učenja (MLM)

Dodatno, obe tehnike klasifikacije po sentimentu (LBM, MLM) su urađene na:

- celom skupu (ALL)
- skupu koji obuhvata samo negacije (OnlyNeg)
- skupu koji obuhvata samo negacije koje su obuhvaćene korišćenim pravilima negacije (OnlyRuleNeg)

Statistička analiza metodom LBM je rađena na skupu od tri klase (3K). Metode mašinskog učenja su primenjene i na skup od tri klase (3K) i na skup od dve klase (2K). Skup

3K sadrži pozitivne, neutralne i negativne tvitove. Skup 2K ne sadrži neutralne tvitove, već samo pozitivne i negativne.

Kako bi se izveo planirani eksperiment, bilo je potrebno iz skupa podataka izdvojiti one tvitove u kojima se negacija pojavljuje (OnlyNeg) i one u kojima se javlja negacija ali samo ona koja je obrađena pravilima obrade negacije koja su korišćena (OnlyRuleNeg). U skup OnlyNeg su svrstani svi tvitovi u kojima se pojavljuje bilo koji od negatora (signala negacije). U skup OnlyRuleNeg su svrstani svi tvitovi koje je program za detektovanje pravila negacije svrstao u neko od 5 pravila (pravilo1, pravilo2, pravilo3, pravilo4 i pravilo5). U tabeli 8.1 je data raspodela broj tvitova za slučaj 2K i 3K i za račličite skupove ALL i OnlyNeg i OnlyRuleNeg.

Tabela 8.1: Broj tvitova na celom skupu (ALL), skupu negacija (OnlyNeg) i skupu negacija obuhvaćenim pravilima (OnlyRuleNeg) za slučaj 2K i 3K

	3K	2K
ALL	7636	5011
OnlyNeg	3747	2726
OnlyRuleNeg	2313	1733

8.1. Korišćena metoda pristupom zasnovanim na rečniku sentimenta

Kako bi se opravdala primena metode koja uključuje obrađena pravila negacije, kvalitet klasifikacije sentimenta tvitova je analiziran primenom metode klasifikacije zasnovane na sentiment rečniku (*eng. lexicon-based method, LBM*). Korišćena metoda (LBM2) je poređena sa dve poredbene (LBM0 i LBM1). Prva poredbena metoda (LBM0) vrši klasifikaciju samo na osnovu pozitivnih i negativnih sentiment reči iz rečnika sentimenta. Druga metoda (LBM1) klasifikuje tvitove uključujući negaciju samo prve reči posle signala za negaciju. LBM2 uključuje sve što i LBM1 i plus pravila koja su uvedena za detektovanje i obradu negacije. Polaritet tvita za sve metode se određuje po formuli 7.2.

Suma pozitivnih (sumPos) i suma negativnih (sumNeg) atributa se za svaku metodu određuje različito, po pravilima datim u tabeli 8.2.

Tabela 8.2: Izračunavanje atributa za svaku od metoda

Metoda	sumPos=	sumNeg=
LBM0	broj pozitivnih reči iz rečnika sentimenta	broj negativnih reči iz rečnika sentiment
LBM1	broj pozitivnih reči iz rečnika sentimenta + broj negativnih reči koje se javljaju kao prva reč posle signala negacije pa menjaju polaritet u pozitivan	broj negativnih reči iz rečnika sentimenta + broj pozitivnih reči koje se javljaju kao prva reč posle signala negacije pa menjaju polaritet u negativan
LBM2	broj pozitivnih reči iz rečnika sentimenta + broj negativnih reči koje se javljaju u novom opsegu negacije koji je određen posle primene pravila (pravila menjaju opseg negacije ili je neutrališu)	broj negativnih reči iz rečnika sentimenta + broj pozitivnih reči koje se javljaju u novom opsegu negacije koji je određen posle primene pravila (pravila menjaju opseg negacije ili je neutrališu)

Analiza je rađena za sve tri metode i u tri slučaja: za ceo skup (ALL), za skup u kojima se negacija javlja bar jednom (OnlyNeg) i za skup gde se javlja negacija koja je obuhvaćena pravilima koja su obrađena (OnlyRuleNeg). Apsolutna poboljšanja i relativna poboljšanja za LBM1 i LBM2 izračunata su u odnosu na LBM0 (prva poredbena metoda) i izračunata su prema formulama 8.1 i 8.2 , respektivno.

$$\text{poboljšanje} = \text{poređena metoda} - \text{poredbena metoda} \tag{8.1}$$

$$\text{rel. poboljš.} = \frac{\text{poređena met.} - \text{poredbena met.}}{\text{poredbena metoda}} * 100 \tag{8.2}$$

U tabeliTabela 8.3 su dati rezultati korišćene metode LBM2 za sva tri skupa i prikazana su apsolutna i relativna poboljšanja u odnosu na dve poredbene metode: LBM1 i LBM2. Ovi rezultati su rezultat primene metoda na skupu 3K.

Tabela 8.3: Rezultati klasifikacije metodom koja se zasniva na rečniku sentimenata (LBM)

		LBM0	LBM1	LBM2
ALL	Tačnost	48.57%	51.07%	53.73%
	Poboljšanje		2.50%	5.16%
	Rel. poboljšanje		5.15%	10.62%
OnlyNeg	Tačnost	39.66%	44.76%	50.23%
	Poboljšanje.		5.09%	10.56%
	Rel. poboljšanje		12.86%	26.63%
OnlyRuleNeg	Tačnost	38.86%	46.89%	50.97%
	Poboljšanje.		8.03%	12.11%
	Rel. poboljšanje		20.66%	31.16%

Tabela 8.3 ukazuje na značajno poboljšanje nakon primene metode koja uključuje obrađena pravila negacije (LBM2). Ovi rezultati su ohrabrujući i opravdavaju primenu metoda mašinskog učenja za klasifikaciju po sentimentu koja će u nastavku biti sprovedena. Ono što je od velikog značaja je veliko poboljšanje koje daje metoda LBM2 kod skupa OnlyNeg (26.63%), i još važnije kod skupa OnlyRuleNeg (31.16%). OnlyRuleNeg podskup sadrži samo tvitove koji imaju bar jedan oblik negacije koji spada u specifična obrađena pravila, pa se na ovom podskupu najbolje vidi i koliko je teško dobro obraditi negaciju ako se ne uključe sintaksička pravila negacije.

Mali procenti u tačnosti klasifikacije kod ove vrste analize su rezultat korišćenja malog broja atributa za klasifikaciju koji su korišćeni. Ovde je cilj bio da se pokaže opravdanost korišćenja pristupa koji uključuje specifična pravila negacije za klasifikaciju po sentimentu.

Veća preciznost je očekivana i postignuta za skup ALL (53.73%) nego za OnlyNeg skup (50.23%) i OnlyRuleNeg skup (50.97%), jer je klasifikaciju po sentimentu jednostavnije izvršiti ako skup podataka ne sadrži negaciju. Ovo je pokazatelj lošeg uticaja prisustva negacije na predviđanje sentimenta ako se negacija ne obrađuje ili ako se ne obrađuje na adekvatan način.

Međutim, iako su procenti tačnosti klasifikacije za skupove OnlyNeg i OnlyRuleNeg manji nego za skup ALL, LBM1 metoda (osnovna metoda obrade negacije), a posebno LBM2 metoda koja uključuje obrađeno negiranje, osigurava veća poboljšanja za skupove OnlyNeg i OnlyRuleNeg nego sa skup ALL.

8.1.1. Statistička opravdanost rezultata primenom metoda koje se zasnivaju na rečniku sentimenata

Da bi se proverila statistička značajnost dobijenog poboljšanja metodom klasifikacije koja se zasniva na rečniku sentimenata, biće testirana hipoteza da korišćena metoda klasifikacije teksta po sentimentu koja uključuje obrađena pravila sintaksičke negacije (LBM2) značajno bolje klasifikuje tekst u odnosu na metode LBM0 i LBM1.

Da bi se to dokazalo, biće primenjen McNemar-ov test. Koriščen je McNemar-ov test jer su podaci kvalitativni i kad god je moguće bolje je raditi s podacima u izvornom obliku. Za primenu ove metode konstruiše se matrica dimenzija 2x2 u kojoj se nalazi odnos tačno i netačno klasifikovanih tvitova primenom dve različite metode. Dobijeni rezultati zadovoljavaju *hi* kvadrat raspodelu prvog stepena slobode pa za nivo značajnosti od 0.05, granična $\chi 2$ vrednost iznosi 13.44. U tabeli 8.4 su predstavljeni rezultati McNemar-ovog testa za poređenje metoda klasifikacije koje se zasnivaju na rečniku sentimenata.

Tabela 8.4: Statistička značajnost posmatrane metode u odnosu na poredbene metode

Skup podataka	Posmatrana metoda	Poredbena metoda	$\chi 2$	P
Ceo skup	LBM2	LBM0	104.569	<0.0001
Ceo skup	LBM2	LBM1	44.644	<0.0001
Samo oni koji sadrže negaciju	LBM2	LBM0	106.2259	<0.0001
Samo oni koji sadrže negaciju	LBM2	LBM1	45.63	<0.0001

U svim slučajevima poređenja rezultata metode LBM2 i metoda LBM1 i LBM0, dokazano je statistički značajno poboljšanje.

8.2. Različiti načini izbora atributa za metodu mašinkog učenja

U prethodnom poglavlju je urađena klasifikacija tvitova metodom koja se zasniva na rečniku sentimenata i pokazano je da korišćena metoda (LBM2) daje značajno poboljšanje u odnosu na dve poredbene metode. Sledeći korak jeste da se isti skup atributa iskoristi i za primenu metoda mašinskog učenja za klasifikaciju tvitova po sentimentu. Kako je broj atributa mali i numeričkog su tipa, a klasifikatori kod metoda ML zahtevaju složeniji skup atributa za što bolje treniranje, rešeno je da se na osnovni skup atributa dodaju: broj pozitivnih reči (uključujući broj pozitivnih od negacije), broj negativnih reči (uključujući broj negativnih od negacije), broj negacija umanjen posle primene pravila negacije, broj odrečnih kvantifikatora, i još atributa koji su dobijeni transformacijom normalizovanog teskta u vektor reči.

Za testiranje uticaja selektovanih tekstualnih atributa na klasifikaciju po sentimentu, korišćene su tri metode ML: Naïve Bayes, nominalna logistička regresja i SVM. Razlog izbora ove tri metode je što se međusobno dosta razlikuju a pokazale su dosta dobre rezultate u klasifikaciji teksta i određivanju sentimenta. Da bi se smanjila količina (dimenzija) podataka, za svaki slučaj transformacije teksta u vektor reči je urađena redukciju atributa primenom tehnike “information gain”. “Information gain” uzima vrednost od 0 do 1 u zavisnosti koliku informaciju atribut donosi. Korišćeni su svi atributi koji imaju vrednost veću od 0, tj. svi koji donose ikakvu informaciju će se naći u lisi atributa. Rezultati uticaja različitih načina transformacije teksta u vektor reči i selektovanja (redukcije) atributa iz teksta na tačnost predviđanja sentimenta su dati u tabeli 8.5.

Kod evaluacije je korišćena unakrsna validacija sa 5 slojeva - zbog vremenske zahtevnosti metode nominalne logističke regresije i SVM metode nije korišćeno 10 slojeva. U koloni „atributi“ su dati atributi koji su korišćeni, kolona „filter“ predstavlja primenjene filtere kod normalizacije teksta tvita, kolona „#atributa“ daje broj atributa posle selekcije i redukcije i kolone „NB“, „LOG“ i „SVM“ označavaju metode mašinkog učenja koje su primenjene.

Tabela 8.5: Tačnost klasifikacije metoda korišćenjem atributa transformacije u vektor reči; unigram(U), bigram(B), trigram(T) i različite filtere

	Atributi	Filter	# atributa	NB	LOG	SVM
1	U	prisustvo reči	408	57.54%	64.67%	63.26%
2	U	broj pojava reči	404	56.36%	64.00%	62.83%
3	U	TF	404	54.54%	64.21%	63.01%
4	U	TF-IDF	404	54.54%	64.21%	63.00%
5	U	TF + normalizovana dužina tvita	265	39.26%	62.31%	60.92%
6	U	TF-IDF+ normalizovana dužina tvita	255	28.77%	62.04%	60.82%
7	U+B+T	prisustvo reči	1023	57.40%	64.87%	63.88%
8	U+B+T	broj pojava reči	1015	55.94%	64.52%	63.61%
9	U+B	prisustvo reči	756	57.49%	65.30%	64.31%
10	U+B	broj pojava reči	750	56.21%	65.16%	63.97%

Iz tabele se može zaključiti da atributi za treniranje koji donose najveći doprinos jesu unigrami (za NB-red 1) i to u slučaju kad se detektuje samo prisustvo ili odsustvo atributa reči (*eng. word presence*) u tvitu. Ako vektor reči koji sadrži prisustvo ili odsustvo reči zamenimo brojem pojava reči (word count) – filter "broj pojava reči", smanjuje se tačnost predviđanja što se može objasniti time da je tekst tvita dosta kratak i da se reči retko ponavljaju više puta kako u samom tvitu tako i u celom skupu. Takođe, normalizacija IDF ne donosi poboljšanje-naprotiv; razlog je isti kao i kod primene TF normalizacije. Normalizacija dužine tvita ima negativan efekat na predviđanje kod sve tri metode. Negativan uticaj normalozacije dužine tvita se posebno se vidi kod primene Naïve Bayes (NB) metode, zbog pretpostavke o nezavisnosti promenljivih koja doprinosi lošim rezultatima.

Unigrami u kombinaciji sa bigramima i trigramima daju manje poboljšanje u odnosu na unigrame u kominaciji sa bigramima, osim za NB metodu (red 7 NB 57.40%). Unigrami u kombinaciji sa bigramima (Tabela 8.5, red 9) daju drugo po redu poboljšanje u odnosu na unigrame (tačnosti za LOG 65.30% i SVM 64.31%) pa će se i ovaj slučaj izbora atributa testirati kod evaluacije, zajedno sa unigramima (Tabela 8.5, red 1).

Rezultati predstavljeni u tabeli 8.5 nemaju velike vrednosti tačnosti (*eng. accuracy*) ali cilj ove analize je da se utvrdi koja selekcija atributa iz teksta je najpogodnija kao dodatak osnovnim atributima koji su opisani na početku ovog poglavlja. U ovom eksperimentu su korišćeni samo tekstualni atributi dobijeni transformacijom teksta u vektor reči jer cilj nije bio da se prezentuje najbolji sistem za klasifikaciju tvitova po sentimentu korišćenjem samo atributa vektora reči, već da se utvrdi koji od ovih atributa uključiti u metodu koja koristi i ostale atribute.

8.3. Korišćena metoda pristupom zasnovanim na mašinskom učenju

Metodama mašinskog učenja sa nadgledanjem se trenira algoritam na označenom skupu podataka u kojem svaka stavka sadrži informaciju o ishodu. To omogućava algoritmu da zaključi obrasce i identifikuje odnose između ciljne promenljive (oznaka klase – u ovom slučaju dodeljeni sentiment) i ostatka skupa podataka na osnovu informacija iz atributa koje ima. Za evaluaciju skupa metodama ML, korišćene su iste metode kao i za izbor atributa transformacijom teksta u vektor reči: Naïve Bayes, Multinomial Logistic Regresion i SVM i još plus dodatna metoda J48-decision tree. Atributi korišćeni za treniranje sadrže sve atribute kao i one korišćene kod metoda klasifikacije koje se zasnivaju na rečniku sentimenta (LBM0, LBM1, LBM2) i dodatno tekstualne atribute izabrane na način koji je prikazan u podsekciji 8.2.

Korišćena metoda MLM2 koristi različite kombinacije podskupova i selektovanih atributa (ovi atributi su dodatak osnovnim):

- na celom skupu: sa unigramima (ALL U) ; sa unigramima i bigramima (ALL U+B),
- skupu koji sadrži samo negacije: sa unigramima (OnlyNeg U); sa unigramima i bigramima (OnlyNeg U+B),
- skupu koji sadrži samo negacije obuhvaćene pravilima: sa unigramima (OnlyRuleNeg U) ; sa unigramima i bigramima (OnlyRuleNeg U+B)

Rezultati predložene MLM2 metode na skupu 3K (negativni, neutralni i pozitivni tvitovi) su prikazani u tabeli 8.6 a u tabeli 8.7 su prikazani rezultati ove metode ali za skup 2K (pozitivni i negativni tvitovi).

Tabela 8.6: Tačnost korišćene metode za različite skupove i ML metode, za 3K

MLM2	ALL U	ALL U+B	OnlyNeg U	OnlyNeg U+B	OnlyRuleNeg U	OnlyRuleNeg U+B
NB	57.82%	57.80%	62.53%	61.97%	62.17%	62.77%
LOG	**68.46%**	**68.84%**	**69.25%**	**69.76%**	**68.96%**	**69.69%**
J48	61.88%	61.92%	61.57%	61.65%	61.95%	62.04%
SVM	67.45%	65.98%	65.86%	67.65%	67.83%	68.61%

Tabela 8.7: Tačnost korišćene metode za različite skupove i ML metode, za 2K

MLM2	ALL U	ALL U+B	OnlyNeg U	OnlyNeg U+B	OnlyRuleNeg U	OnlyRuleNeg U+B
NB	86.49%	86.47%	84.63%	85.03%	85.46%	85.98%
LOG	**91.15%**	**91.13%**	**90.46%**	**90.53%**	**90.82%**	**90.45%**
J48	86.88%	86.80%	85.29%	84.92%	84.30%	83.32%
SVM	89.56%	89.36%	88.74%	88.92%	88.29%	88.75%

U obe tabele su podebljani najbolji rezultati i vidi se da je u svakom slučaju najveća tačnost dobijena primenom metode nominalne logističke regresije, pa će rezultati u nastavku biti prikazani samo primenom ove metode mašinskog učenja.

Kao i u prethodnom slučaju, kod pristupa koji se zasniva na rečniku sentimenata, i kod ovog pristupa metodama mašinskog učenja će korišćena metoda biti upoređena sa dve poredbene:

- Prva poredbena metoda mašinskog učenja (MLM0) klasifikuje tvitove navedenim metodama ML na osnovu atributa koji predstavljaju broj pozitivnih i negativnih reči iz rečnika sentimenata.
- Druga metoda mašinskog učenja (MLM1) klasifikuje tvitove koristeći atribute koji predstavljaju broj pozitivnih i negativnih reči iz rečnika sentimenata i još dodatne attribute o negaciji samo prve reči nakon signala negacije.

Metoda mašinskog učenja (MLM2) uključuje sve što i prethodne dve i još dodatne atribute o opisanim pravilima za otkrivanje i obradu negacije.

Rezultati koji su postignuti primenom tri navedene metode nad skupovima 3K i 2K su dati u sledećim tabelama. U tabelama kojima se prikazuju rezultati primenom metoda mašinskog učenje će u nastavku biti prikazani i rezultati klasifikacije metodom mašinskog učenja koja kao atribute koristi samo vektore reči (OnlyWords).

Tabela 8.8 daje rezultate primene metoda na celom skupu za 3K. Poboljšanja su izračunata u odnosu na MLM0 metodu. Poboljšanje je prikazano kao relativno i računato je po formuli 8.2.

Tabela 8.8: Rezultati i poboljšanje korišćene metode za ceo skup, za 3K

	ALL U		ALL U+B	
	Tačnost	Poboljšanje	Tačnost	Poboljšanje
OnlyWords	64.6660%		65.3053%	
MLM0	67.4843%		68.3586%	
MLM1	68.2281%	1.1022%	68.7500%	0.5726%
MLM2	68.4629%	1.4501%	68.8413%	0.7061%

Iz tabele se može zaključiti da korišćena metoda koja uključuje pravila negacije koja su obrađena (MLM2) daje bolje rezultate od obe poredbene metode nad celim skupom podataka.

Kako bi se bolje prikazao efekat primene obrađenih pravila, u tabeli 8.9 su prikazani rezultati na skupu tvitova koji obavezno sadrže bar jednu negaciju (OnlyNeg skup).

Tabela 8.9: Rezultati i poboljšanje korišćene metode za skup OnlyNeg, za 3K

	OnlyNeg U		OnlyNeg U+B	
	Tačnost	Poboljšanje	Tačnost	Poboljšanje
OnlyWords	67.0224%		68.7033%	
MLM0	67.4673%		68.6683%	
MLM1	68.5615%	1.6218%	69.7892%	1.6323%
MLM2	69.2554%	2.6503%	69.7625%	1.5935%

Iz tabele se vidi da nad skupom OnlyNeg postoji znatno veće poboljšanje korišćene metode. Kada se koriste dodatni tekstualni atributi unigrami (U) poboljšanje korišćene metode je čak 2.6503% u odnosu na metodu MLM0. Kada se koristi kobminacija unigrama i bigrama kao dodatnih atributa (U+B) onda se generalno postižu veće tačnosti klasifikacije iako poboljšanje korišćene metode u odnosu na MLM0 iznosi 1.5935%, što je manje nego u slučaju kada se koriste samo unigrami.

Prethodna tabela sadrži rezultate u slučaju kada se metoda primeni nad podskupom tvitova koji sadrži negaciju bilo kojeg tipa. Međutim, opisana pravila obrade sintaksičke negacije su detektovana na mnogo manjem skupu tvitova, na podskupu OnlyRuleNeg. Kako bi se video efekat metode na OnlyRuleNeg skupu, u tabeli 8.10 je prikazana tačnost predviđanja sentimenta samo tvitova u kojima se javlja negacija i to onaj tip negacije koji je obrađen korišćenim pravilima.

Tabela 8.10: Rezultati i poboljšanje korišćene metode za skup OnlyRuleNeg, za 3K

	OnlyRuleNeg U		OnlyRuleNeg U+B	
	Tačnost	Poboljšanje	Tačnost	Poboljšanje
OnlyWords	67.7182%		68.6690%	
MLM0	68.4825%		69.4336%	
MLM1	68.2663%	-0.3157%	68.9148%	-0.7472%
MLM2	68.9581%	0.6945%	69.6930%	1.1208%

Iz tabele se vidi da su tačnost klasifikacije na prilično visokom nivou, ako se uzme u obzir da je klasifikacija tvitova koji pripadaju skupu OnlyRuleNeg zahtevnija jer ovaj skup sadrži i najproblematičnije tvitove za klasifikaciju – posebno kada je negacija u pitanju. Negativna poboljšanja koja su vidiljiva kod metode MLM1 (u slučaju U i U+B) govore da se, na navedenom skupu koji sadrži specifične oblike negacije, prostom promenom polariteta reči posle signala negacije (MLM1) postižu gori rezultati nego u slučaju kada se negacija uopšte ne uzima u obzir (kao kod MLM0). Skup OnlyRuleNeg sadrži upravo i najproblematičnije tvitove pa se na njemu mogu bolje uočiti nedostaci i prednosti pojedinih pristupa obrade negacije. Korišćena MLM2 metoda daje poboljšanje i u slučaju kada se koriste samo unigrami i u slučaju korišćenja unigarama i bigrama zajedno.

U nastavku su dati rezultati nad korpusom koji sadrži samo pozitivne i negativne tvitove (2K). Rezultati koji su postignuti primenom tri navedene metode nad celim skupom podataka za dve klase (ALL 2K) su dati u tabeli 8.11.

Tabela 8.11: Rezultati i poboljšanje korišćene metode za ceo skup, za 2K

	ALL U		ALL U+B	
	Tačnost	Poboljšanje	Tačnost	Poboljšanje
OnlyWords	90.1474%		90.8362%	
MLM0	90.4134%		90.9179%	
MLM1	91.0926%	0.7512%	91.1120%	0.2135%
MLM2	91.1508%	0.8156%	91.1314%	0.2348%

Iz tabele se vidi da klasifikacija tvitova, u slučaju kada imamo samo pozitivne i negativne tvitove u skupu podataka, daje mnogo veću tačnost nego u slučaju kada se radi na celom skupu (3K). Ovo generalno ukazuje na problematičnost neutralnih tvitova kod klasifikacije na skupu 3K. Poboljšanje korišćene metode je najbolje za slučaj kada se kao dodatni atributi koriste unigrami (0.8156%).

Na skupu koji obuhvata samo negacije, u tabeli 8.12 se vidi manji procenat tačnosti klasifikacije za sve tri metode nego nad celim skupom podataka -ALL. Razlog tome je što su tvitovi koji sadrže negacije generalno problematičniji za klasifikaciju. Međutim, poboljšanje koje je postignuto metodom MLM2 je veće (1.648%) u odnosu na poboljšanje iste metode postignuto na celom skupu tvitova (0.8156%).

Tabela 8.12: Rezultati i poboljšanja za skup OnlyNeg, za 2K

	OnlyNeg U		OnlyNeg U+B	
	Tačnost	Poboljšanje	Tačnost	Poboljšanje
Only words	89.2268%		90.3994%	
MLM0	88.9949%		90.2421%	
MLM1	89.3984%	0.4534%	90.4622%	0.2439%
MLM2	90.4622%	1.6487%	90.5356%	0.3252%

Može se uočiti manji rast poboljšanja u slučaju kada se razmatraju samo tvitovi sa negacijama koje su obuhvaćene pravilima (Tabela 8.13) u odnosu na slučaj kad se razmatraju tvitovi sa svim negacijama (tabela 8.12). Tvitovi koji sadrže negaciju a nisu obuhvaćeni pravilima uglavnom i nije trebalo da budu obrađeni jer se njihova negacija vezuje za neutralne reči i nije je bitna za obradu sentimenta.

Tabela 8.13: Rezultati i poboljšanja za skup OnlyRuleNeg, za 2K

	OnlyRuleNeg U		OnlyRuleNeg U+B	
	Tačnost	Poboljšanje	Tačnost	Poboljšanje
Only words	88.9273%		89.5040%	
MLM0	89.6711%		89.4403%	
MLM1	90.3058%	0.7078%	89.5557%	0.1290%
MLM2	90.8252%	1.2870%	90.4512%	1.1303%

Skup neutralnih tvitova značajno remeti kvalitet klasifikacije kod sentimenta kako na celom skupu tako i na skupu sa negacijama. To se vidi poređenjem rezultata klasifikacije na skupu sa 3K i 2K. Neutralni tvitovi sadrže reči kojima se izražava sentiment i one utiču na generalni sentiment tog tvita. Smanjenje uticaja sentiment reči u neutralnim tvitovima se može rešiti uvođenjem stepena sentimenta, što bi značajno uticalo na kvalitet klasifikacije neutralnih tvitva. Preciznost dobijena na skupu 2K je zadovoljavajuća, ali bi se smanjenjem uticaja reči kojima se izražava sentiment u neutralnim tvitovima, značajno povećala tačnost klasifikacije na 3K.

Metode klasifikacije zasnovane na rečniku daju manju tačanost klasifikacije od metoda mašinskog učenja jer su pored atributa koji sadrže broj pojava pozitivnih i negativnih termina, korišćeni samo atributi koji su direktna posledica obrade pravila negacije. Međutim, poboljšanja Metode1 i Metode2 u odnosu na Metodu0 su bolje izražena poređenjem metoda klasifikacije koje se zasnivaju na rečniku sentimenta (LBM0, LBM1, LBM2), upravo iz gore navedenog razloga.

8.3.1. Statistička opravdanost rezultata primenom metoda mašinskog učenja

Primenom T-testa ispitivano je da li su rezultati dobijeni različitim metodama (Only words, MLM0, MLM1, MLM2) koristeći nominalnu logističku regresiju kao metodu mašinskog učenja statistički značajni.

Tabela 8.14: Rezultati prmene t-testa nad metodama klasifikacije mašinskim učenjem

Posmatrana metoda	Uporedna metoda	P (3K)	P (2K)
Only words	MLM0	0.028954	0.250274
Only words	MLM1	0.015784	0.042767
Only words	MLM2	0.004331	0.008634
MLM0	MLM1	0.089519	0.006044
MLM0	MLM2	0.007086	0.004997
MLM1	MLM2	0.017721	0.033535

Test je primenjen na svim skupovima (All, OnlyNeg i OnlyRuleNeg) i sve to u dva slučaja: u slučaju kada se posmatraju 3 klase i u slučaju kada se posmatraju 2 klase. Rezultati primene t-testa su dati u tabeli 8.14.

U svim posmatranim slučajevima, osim kod poređenja MLM0 i MLM1 za 3 klase i Only words i MLM0 za 2 klase, postoji statistička značajnost. U svim, osim u dva navedena slučaja, dobija se $p<0.05$. Posebno je značajno što kod poređenja korišćene MLM2 metode sa Only words i MLM0 metodama (u slučaju 3 klase i u slučaju 2 klase) dobija da je $p<0.01$, na osnovu čega može da se tvrdi da je korišćenom MLM2 metodom postignuto statistički značajno poboljšanje u odnosu na druge dve metode. Poboljšanje postignuto MLM2 metodom je statistički značajno i u odnosu na MLM1 metodu ($p<0.05$).

9. ZAKLJUČAK

U tezi je kreirana metoda za analizu sentimenta kratkih neformalnih tekstova na srpskom jeziku koja uključuje specifična pravila obrade sintaksičke negacije. Korišćena metoda je prva metoda koja klasifikuje kratke tekstove po sentimentu na sprskom jeziku uključujući primenjena pravila negacije. Rezultati metode su prikazani za dva slučaja: kada se koristi pristup zasnovan na rečniku sentimenata i pristup koji koristi metode mašinsko učenja.

Kada se koristi pristup zasnovan na rečniku sentimenata, utvrđeno je statistički značajno poboljšanje predložene metode kod klasifikacije tvitova po sentimentu. Ovaj pristup se koristio pre svega da bi se opravdala primena metoda mašinskog učenja. Atributi koji su korišćeni kod ovog pristupa su numeričkog tipa a klasifikacija se svodila na računanje sentimenta funkcijom koja kombinuje date atribute i daje numeričku vrednost na osnovu koje se i određuje sentiment.

Kod pristupa koji se zasniva na mašinskom učenju je korišćeno više metoda mašinskog učenja i eksperimentalno je utvrđeno da metoda nominalne logističke regresije daje najbolje rezultate. Pošto je pristup zasnovan na rečniku sentimenta koristio samo numeričke atribute, kako bi se poboljšao klasifikator metodom mašinskog učenja, testirana je primena dodatnih atributa koji su dobijeni transformacijom teksta tvita u vektor reči. S obzirom da je dobijen veliki broj tekstualnih atributa, izvršena je selekcija atributa metodom redukcije na atribute koji imaju kritičnu količinu informacija – koji najviše utiču na kvalitet klasifikacije. Primenom različitih filtera za selektovanje atributa, utvrđeno je da se najbolji rezultati dobijaju kada se koriste unigrami i unigrami u kombinaciji sa bigramima. Što se tiče filtera, utvrđeno je i da je najbolje računati samo pojavu reči (bez TF, IDF, brojanja broja pojava reči...) – razlozi su vezani za specifičnost tipa teksta tvitova i detaljno su opisani u poglavlju za selekciju dodatnih tekstualnih atributa.

Značajan uticaj kod primene metode, i u pristupu zasnovanom na rečniku i u pristupu korišćenjem mašinskog učenja, imaju jezički resursi. Pored obaveznog rečnika sentimenta, to su i rečnik signala negacije, rečnik odrečnih kvantifikatora, rečnik pojačivača i neutralizatora.

Navedeni jezički resursi su kreirani u toku izrade disertacije i rezultat su rada na njoj. Od eksternih jezičkih resursa je korišćen morfološki rečnik [80].

Za kvalitetnu primenu korišćene metode, bilo je potrebno primeniti kvalitetnu normalizaciju svih resursa koji su korišćeni. Obavezni koraci kod normalzacije korpusa koji su primenjeni jesu: tokenizacija, svođenje na jedno znvanično pismo (latinicu) i podela na rečenice. Testirani su različiti načini normalizacije i sentiment rečnika i korpusa tvitova: normalizacija stemerom, normalizacija morfološkim rečnikom i normalizacija odsecanjem na različite dužine *n*-grama. Utvrđeno je da normalizacija primenom stemera daje najbolje rezultate kod klasifikacije po sentimentu.

Za detaljniju obradu pravila negacije, bilo je potrebno je postojanje obeleženog korpusa tvitova. Za tu potrebu je kreiran korpus tvitova koji su ručno obeleženi od stane tri osobe. Korpus je za potrebe analize negacije i detektovanja pravila sintaksičke negacije u fazi eksperimena korišćen na tri načina: u prvom slučaju je korišćen skup svih tvitova, u drugom su korišćeni samo tvitovi koji sadrže bar jednu negaciju (u tvitu je prisutan bar jedan negator-signal negacije) i u trećem slučaju je korišćen podskup tvitova koji sadrže bar jedno pravilo negacije koje je obrađeno. Testiranje metode na ovako različitim skupovima je omogućilo bolji uvid u to koliko se uvođenjem definisanih pravila negacije utiče na kvalitet klasifikacije sentimenta.

Značaj dobijenih rezultata je veći ako se uzme u obzir priroda teksta od kojih se sastoji korpus koji je korišćen. Tvitovi su kratki tekstovi, kod kojih je sadržaj (pa i sentiment ako postoji) dosta sažeto izražen, sadrže dosta neformalno napisanih reči, skraćenica, sadrže i gramatičke greške u pisanju i reči su često napisane bez dijakritičkih znakova (npr. “nece više” umesto “neće više”) i slično. Sve ovo, i dodatno kompleksnost obrade teksta na morfološki bogatom srpskom jeziku, otežava proces obrade ovakvog tipa teksta. Dodatno, složenost pravila negacije, otežava i način njenog detektovanja i uključivanja u metodu za klasifikaciju tvitova po sentimentu. Ako se uzme sve ovo u obzir, dobijeni rezultati daju zadovoljavajuću tačnost klasifikacije ovako složenih tekstova po sentimentu. Što je još značajnije, poboljšanje koje se dobija obradom određenog skupa pravila sintaksičke negacije u odnosu na tradicionalne načine obrade negacije značajno doprinosi kvalitetu klasifikacije po sentimentu.

Poznato je da analiza sentimenta u poslednje vreme predstavlja interesantu oblast sa primenom u poslovanju firmi i marketingu radi analize zadovoljstva kupaca i davanja pravca

za reagovanje u skladu sa zadovoljstvom kupaca. Negacija je neizostavna stavka kojoj treba posvetiti pažnju, pored ostalih fenomena kao što su sarkazam i ironija.

Značaj obrade negacije sve više dobija na važnosti kod obrade teksta u medicinskim izveštajima, sociološkim i psihološkim izveštajima i anketama koje se sprovode nad ispitanicima, jer ovakvi tekstovi sadrže dosta negacije.

10. PRAVCI DALJEG RAZVOJA

Sledeći planirani zadatak jeste kreiranje posebnog korpusa koji sadrži ručno obeležene opsege negacije (opseg delovanja signala negacije u delu teksta). Postojanje ovakvog resursa bi omogućilo dodatnu analizu fenomena negacije i utvrđivanje dodatnih pravila koja bi pomogla da se obrada negacije još bolje uklopi u sistem određivanja sentimenta kratkih teksova koji je sadrže.

Dodatno, u planu je proširenje rečinka sentimenta sinonimima iz Srpskog wordNet-a [81]. Takođe, plan je i da se posveti veća pažnja indentifikovanju pojačivača polariteta kako u potvrdnom tako i u negiranom kontekstu, a sve u cilju bolje klasifikacije po sentimentu. Upotrebom ovih dodatnih resursa očekuje se poboljšanje u treriranju same negacije, kao i poboljšanje generalnog sistema za detektovanje sentimenta kratkih tekstova.

Još jedan od budućih izazova jeste prepoznavanje stilske figure litote (stilska figura u kojoj se pravi izraz umanjuje jer se zamenjuje slabijim izrazom, koji je suprotan i negativan) koja u sebi sadrži negaciju ali je potrebno prepoznati je na bolji način i posebno obraditi. Slovenska antiteza je stilska figura u srpskom jeziku koja se sastoji od tri dela: slikovitih pitanja, negativnih odgovora i tačnog odgovora (teze i antiteze). Može se reći da litota i slovenska antiteza imaju delove koji sadrže negaciju koji su delimično ili potpuno obuhvaćeni pravilima sintaksičke negacije koja su u disertaciji definisana. Međutim, zbog tačnije interpretacije, potrebno im je detaljnije posvetiti pažnju kao posebnim fenomenima.

Ironija je stilska figura u kojoj se rečima daje suprotan smisao od onoga koji imaju kao svoje osnovno značenje (npr. Toplo je kao na severnom polu). Oksimoron je posebna vrsta antiteze u kojoj se spajaju dva nespojiva, semantički nespojiva, protivrečna pojma (antonimi). Iz definicije ironije i oksimorona se vidi da značajno mogu remetiti utvrđivanje sentimenta teksta na osnovu reči kojima se izražava sentiment, ako se ova dva fenomena ne prepoznaju i ne obrade.

LITERATURA

[1] Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H. & Demirbas, M. (2010). Short Text Classification in Twitter to Improve Information Filtering. *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 841--842, New York, NY, USA: ACM. ISBN: 978-1-4503-0153-4

[2] Song, G., Ye, Y., Du, X., Huang, X. & Bie, S. (2014). Short Text Classification: A Survey.. *Journal of Multimedia*, 9, 635-643.

[3] Tan, P.N., Steinbach, M. , & Kumar,V. (2007). Introduction to Data Mining , Pearson Education, 2007, ISBN: 9788131714720.

[4] Feldman, R., & Sanger, J. (2007). *The text mining handbook : advanced approaches in analyzing unstructured data*. Cambridge; New York: Cambridge University Press. ISBN: 0521836573 9780521836579.

[5] Liddy, E. D. (1990). Anaphora in natural language processing and information retrieval. *Information Processing & Management 26*, 39-52. DOI=http://dx.doi.org/10.1016/0306-4573(90)90008-P

[6] Richard Socher, Cliff Chiung-Yu Lin, Andrew Y. Ng, & Christopher D. Manning. (2011). Parsing natural scenes and natural language with recursive neural networks. *In Proceedings of the 28th International Conference on International Conference on Machine Learning (ICML'11), Lise Getoor and Tobias Scheffer (Eds.)*. Omnipress, USA, 129-136.

[7] Church, K.W., & Hanks, P. (1989). Word association norms, mutual information and lexicography. In *Proceedings of the 27th Annual Conference of the Association of Computational Linguistics*, (pp. 76-83), New Brunswick, NJ: Association for Computational Linguistics

[8] McNamara, D. S., Boonthum, C., Levinstein, I., & Millis, K. (2007). Evaluating self-explanations in iSTART: Comparing word-based and LSA algorithms.

In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), Handbook of latent semantic analysis, 227-241. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

[9] Turney, P. & Littman, M. (2002). Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus (Technical Report NRC Technical Report ERB-1094). *Institute for Information Technology, National Research Council Canada.*

[10] Hatzivassiloglou, V., & McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. *In proceedings of the 35th Annual Meeting on Association for Computational Linguistics.* DOI:10.3115/976909.979640

[11] Pang, B., Lee, L., Rd, H., & Jose, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. I*n EMNLP '02 Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, Vol. 10, 79–86.

[12] Boonthum, C., Levinstein, I. B., & McNamara, D. (2007). Evaluating self-explanations in iSTART: Word matching, latent semantic analysis, and topic models. In *Natural Language Processing and Text Mining*, 91-106. Springer London. https://doi.org/10.1007/978-1-84628-754-1_6

[13] McCarthy, P. M., Briner, S. W., Rus, V., & McNamara, D. (2007). Textual signatures: Identifying text-types using latent semantic analysis to measure the cohesion of text structures. *In Natural Language Processing and Text Mining*, 107-122. Springer London. https://doi.org/10.1007/978-1-84628-754-1_7

[14] Schmidtle, M. A. R., & Amtrup, J. W. (2007). Automatic Document Separation: A Combination of Probabilistic Classification and Finite-State Sequence Modeling. *In Natural Language Processing and Text Mining*, 123-144. Springer London. https://doi.org/10.1007/978-1-84628-754-1_8

[15] Atkinson, J. (2007). Evolving Explanatory Novel Patterns for Semantically-Based Text Mining. *In Natural Language Processing and Text Mining*, 145-169. Springer London. https://doi.org/10.1007/978-1-84628-754-1_9

[16] Jurafsky, D., & Martin, J. H. (2009). Speech and language processing : an introduction to natural language processing, computational linguistics, and speech

recognition. Upper Saddle River, N.J.: Pearson Prentice Hall. ISBN: 9780131873216 0131873210

[17] Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning Word Vectors for Sentiment Analysis. In D. Lin, Y. Matsumoto & R. Mihalcea (eds.), *ACL* (p./pp. 142-150), : The Association for Computer Linguistics. ISBN: 978-1-932432-87-9

[18] Saif, H., He, Y., & Alani, H. (2012). Semantic sentiment analysis of twitter. In Proceedings of the 11th international conference on The Semantic Web - Volume Part I (ISWC'12), Philippe Cudré-Mauroux, Jeff Heflin, Evren Sirin, Tania Tudorache, and Jérôme Euzenat (Eds.), Vol. Part I. Springer-Verlag, Berlin, Heidelberg, 508-524. DOI=http://dx.doi.org/10.1007/978-3-642-35176-1_32

[19] Darwin, C. (1998). The Expression of Emotions in Man and Animals, ed. P. Ekman. London: HarperCollins. (Orig. published 1872.)

[20] Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*, 44(4), 695–729. DOI:10.1177/0539018405058216

[21] Scherer, K. R. (2000). Psychological Models of Emotion In Borod, J. C. (Ed.). *Series in affective science. The neuropsychology of emotion*, 137-162. New York, NY, US: Oxford University Press.

[22] Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. Found. Trends Inf. Retr., Vol. 2, No. 1-2, 1-135.

[23] Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '04). ACM, New York, NY, USA, 168-177.

[24] Wilson, T., Wiebe, J. & Hoffmann, P. (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, 347–354. Vancouver, Canada.

[25] Baccianella, S., Esuli, A. & Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D.

Tapias (eds.), *LREC*, : European Language Resources Association. ISBN: 2-9517408-6-7

[26] Hovy, E. H. (2015). What are Sentiment, Affect, and Emotion? Applying the Methodology of Michael Zock to Sentiment Analysis. Text, Speech and Language Technology, 1*3–24*.doi:10.1007/978-3-319-08043-7_2

[27] Liu, B., Zhang L. (2012). A Survey of Opinion Mining and Sentiment Analysis. In: Aggarwal C., Zhai C. (eds) Mining Text Data. Springer, Boston, MA.

[28] Dashtipour, K., Poria, S., Hussain, A., Cambria, E., Hawalah, A.Y., Gelbukh, A.F., & Zhou, Q. (2016). Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques. *Cognitive Computation*.

[29] Giahanou, A., & Crestani, F.A. (2016). Like It or Not: A Survey of Twitter Sentiment Analysis Methods. *ACM Comput. Surv., 49*, 28:1-28:41.

[30] Abdulla, N. A., Ahmed, N. A., Shehab, M. A., & Al-Ayyoub, M. (2013). Arabic sentiment analysis: Lexicon-based and corpus-based. 2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT). https://doi.org/10.1109/AEECT.2013.6716448

[31] Taboada, M., Brooke, J., Tofiloski, M., Voll, K.D., & Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics, 37*, 267-307, doi:10.1162/COLI_a_00049.

[32] Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined approach. *Journal of Informetrics, Elsevier*, 3(2), 143-157, https://EconPapers.repec.org/RePEc:eee:infome:v:3:y:2009:i:2:p:143-157.

[33] Ding, X., Liu, B., & Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. *Proceedings of the International Conference on Web Search and Web Data Mining - WSDM '08 (pp.231-239).* DOI:10.1145/1341531.1341561

[34] Kim, S.-M. & Hovy, E. (2004). Determining the Sentiment of Opinions. In Proceedings of COLING 2004, 1367–1373, Geneva, Switzerland.

[35] Turney, P. D. (2002). Thumbs up or thumbs down? *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02.* doi:10.3115/1073083.1073153

[36] Abdulla, N. A., Ahmed, N. A., Shehab, M. A., Al-Ayyoub, M., Al-Kabi, M. N., & Al-rifai Saleh. (2014). Towards Improving the Lexicon-Based Approach for Arabic Sentiment Analysis. *International Journal of Information Technology and Web Engineering, 9(3), 55–71.*doi:10.4018/ijitwe.2014070104

[37] Mohammad, S. & Turney, P. D. (2013). Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence*, 29, 436-465.

[38] Osgood, C., Suci, G., & Tenenbaum, P. (1957). The Measurement of meaning. Urbana: University of Illinois Press.

[39] Bruce, R. & Wiebe, J. (1999). Recognizing subjectivity: A case study in manual tagging. *Natural Language Engineering*, *5*(2), 187-205.

[40] Godbole, N., Srinivasaiah, M. & Skiena, S. (2007). Large-Scale Sentiment Analysis for News and Blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM).*

[41] Strapparava, C. & Valitutti, A. (2004). WordNet-Affect: An affective extension of WordNet. *Proceedings of the 4th International Conference on Language Resources and Evaluation* (p./pp. 1083-1086).

[42] Kamps, J., Marx, M., Mokken, R. J. & de Rijke, M. (2004). Using WordNet to Measure Semantic Orientations of Adjectives. In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04), European Language Resources Association.

[43] Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. *In Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, European Language Resources Association. ISBN: 2-9517408-6-7

[44] Qiu, G., Liu, B., Bu, J. & Chen, C. (2009). Expanding Domain Sentiment Lexicon through Double Propagation. In C. Boutilier (ed.), *IJCAI* , 1199-1204.

[45] Kanayama, H., & Nasukawa, T. (2006). Fully Automatic Lexicon Expansion for Domain-oriented Sentiment Analysis, 355-363, Sydney, Australia.

[46] Tang, D., Wei, F., Qin, B., Zhou, M., & Liu, T. (2014). Building Large-Scale Twitter-Specific Sentiment Lexicon: A Representation Learning Approach. *COLING*.

[47] Pak, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10).

[48] Mladenović, M., Mitrović, J., Krstev, C., & Vitas, D. (2016). Hybrid Sentiment Analysis Framework for A Morphologically Rich Language. Journal of Intelligent Information Systems, 46(3), 599–620.

[49] Batanović, V., Nikolić, B., & Milosavljević, M. (2016). Reliable Baselines for Sentiment Analysis in Resource-Limited Languages: The Serbian Movie Review Dataset. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, 2688-2696.

[50] Ljajić, A., & Marovac, U. (2019). Improving Sentiment Analysis for Twitter Data by Handling Negation Rules in the Serbian Language. Computer Science and Information Systems, 16(1), 289-311. https://doi.org/10.2298/CSIS180122013L

[51] Batanović, V., & Nikolić, B. (2016). Sentiment classification of documents in Serbian: The effects of morphological normalization. In Proceedings of the 24th Telecommunications Forum (TELFOR), Belgrade, Serbia, 1-4.

[52] Rotim, L., & Šnajder, J. (2017). Comparison of Short-Text Sentiment Analysis Methods for Croatian. In Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing. Association for Computational Linguistics, 69-75.

[53] Das, S. & Chen, M. (2001). Yahoo! for Amazon: Extracting Market Sentiment from Stock Messageoards. In *Proceedings of the Asia Pacific Finance Association Annual Conference (APFA).*

[54] Polanyi, L. & Zaenen, A. (2006). Contextual Valence Shifters. In: Shanahan J.G., Qu Y., Wiebe J. (eds) Computing Attitude and Affect in Text: Theory and Applications. The Information Retrieval Series, Vol 20. Springer, Dordrecht.

[55] Kennedy, A. & Inkpen, D. (2005). Sentiment classification of movie and product reviews using contextual valence shifters. In Proceedings of the Workshop on the Analysis of Informal and Formal Information Exchange during Negotiations, Ottawa, Ontario, Canada.

[56] Benamara, F., Cesarano, C., Picariello, A., Recupero, D.R., & Subrahmanian, V.S. (2007). Sentiment Analysis: Adjectives and Adverbs are Better than Adjectives Alone. ICWSM.

[57] Jimenez-Zafra, S. M., Martin Valdivia, M. T., Martinez Camara, E., & Urena-Lopez, L. A. (2017). Studying the Scope of Negation for Spanish Sentiment Analysis on Twitter. IEEE Transactions on Affective Computing, IEEE, Vol. PP, Issue: 99.

[58] Zhu, X., Guo, H., Mohammad, S. & Kiritchenko, S. (2014). An Empirical Study on the Effect of Negation Words on Sentiment. In ACL (1), 304-313.

[59] Kiritchenko, S., Mohammad, S. & Zhu, X. (2014). Sentiment Analysis of Short Informal Texts. J. Artif. Intell. Res., Vol. 50, 723-762.

[60] Sharif, W., Samsudin, N. A., Deris, M. M. & Naseem, R. (2016). Effect of negation in sentiment analysis. 2016 Sixth International Conference on Innovative Computing Technology (INTECH), 718-723. IEEE.

[61] Pröllochs, N., Feuerriegel, S. & Neumann, D. (2016). Negation scope detection in sentiment analysis, Decis. Support Syst., Vol. 88, No. C, 67-75.

[62] Reitan, J., Faret, J., Gambäck, B. & Bungum, L. (2016). Negation Scope Detection for Twitter Sentiment Analysis. Association for Computational Linguistics. x

[63] Asmi, A. & Ishaya, T. (2012). Negation identification and calculation in sentiment analysis. In The Second International Conference on Advances in Information Mining and Management, 1-7.

[64] Cerezo-Costas, H. & Celix-Salgado, D. (2015). Gradiant-analytics: training polarity shifters with CRFs for message level polarity detection. In Proceedings of the 9th international workshop on semantic evaluation (SemEval-2015), 539–544. Denver, Colorado: Association for Computational Linguistics.

[65] Wiegand, M., Balahur, A., Roth, B., Klakow, D. & Montoyo, A. (2010). A survey on the role of negation in sentiment analysis. In Proceedings of the Workshop on Negation and Speculation in Natural Language Processing (NeSp-NLP '10), Roser Morante and Caroline Sporleder (Eds.). Association for Computational Linguistics, Stroudsburg, PA, USA, 60-68.

[66] Kovačević, M. (2002). Sintaksička negacija u srpskome jeziku. Izdavačka jedinica Univerziteta u Nišu.

[67] Petrović, G. (1990). Logika. Školska knjiga. Zagreb.

[68] Frege, G. (1918–1919). Die Verneinung. Eine logische Untersuchung [Negacija. Logičko istraživanje], *Beiträge zur Philosophie des deutschen Idealismus* I, 143-157.

[69] Roelofsen, F., Venhuizen, N., & Weidman Sassoon, G. (2013). Positive and negative polar questions in discourse. In E. Chemla, V. Homer, & G. Winterstein (Eds.), Proceedings of Sinn und Bedeutung 17, 455-472. Paris, France.

[70] Liu, F., Weng, F. & Jiang, X. (2012). A broad-coverage normalization system for social media language. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, 1035-1044.

[71] Milosevic, N. (2012). Stemmer for Serbian language. arXiv preprint arXiv:1209.4471.

[72] Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). The General Inquirer: A Computer Approach to Content Analysis. M.I.T. press.

[73] Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). The Development and Psychometric Properties of LIWC2007. The University of Texas at Austin, 1–22.

[74] Ljajic, A., Marovac, U., & Avdic, A. (2017). Processing of Negation in Sentiment Analysis for the Serbian Language, IcETRAN 2017 Conference proceedings at (Serbia), June 2017.

[75] Marovac, U., Ljaić, A., Kajan, E., & Avdić, A. (2018). "Towards the Lexical Resources for Sentiment-Reach Informal Texts-The Serbian Language Case", 5th International Conference CONTEMPORARY PROBLEMS OF MATHEMATICS, MECHANICS AND INFORMATICS. State University of Novi Pazar.

[76] Marovac, U., Ljaić, A., Kajan, E., & Avdić, A. (2013). Similarity Search in Text Data for the Serbian language. In: Proceedings of ICEST (2013), Ohrid, Macedonia, 2013, pp. 607–610

[77] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12(2), 153–157. doi:10.1007/bf02295996

[78] Ljajić, A., Stanković, M., & Marovac, U. (2018). *Detection of Negation in the Serbian Language. Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics - WIMS '18.*doi:10.1145/3227609.3227660

[79] Marovac, U., Pljaskovic, A., Crnisanin, A., & Kajan, E. (2012). N-gram analysis of text documents in Serbian language. 2012 20th Telecommunications Forum (TELFOR).doi:10.1109/telfor.2012.6419476

[80] Krstev, C. (2008). Processing of Serbian. Automata, Texts and Electronic Dictionaries. University of Belgrade, Faculty of Philology.

[81] Krstev, C., Pavlović-Lažetić, G., Vitas, D., & Obradović, I. (2004). Using Textual and Lexical Resources in Developing Serbian Wordnet. Romanian Journal of Information Science and Technology, Vol.7, No. 1-2, 147—161.

SKRAĆENICE KORIŠĆENE U RADU

SA	Sentiment Analysis
NLP	Natural Language Processing
IR	Information Retrieval
AI	Artificial Intelligence
CRNN	Context-Sensitive Recursive Neural Networks
LSA	Latent Semantic Analysis
HMM	Hidden Markov Model
MRF	Markov Random Field
PCA	Principal Components Analysis
PMI	Pointwise Mutual Information
PMI-IR	Pointwise Mutual Information- Information Retrieval
SVD	Singular Value Decomposition
NER	Named Entity Recognition
POS	Part of Speech
TF	Term Frequency
IDF	Inverse Document Frequency
ML	Machine Learning
NB	Naïve Bayes
LOG	Nominal Logistic Regression
CV	Cross-validation
J48	Open source Java implementation of the C4.5 decision tree algorithm in the WEKA data mining tool
SVM	Support Vector Machines
TP	True positive

TN	True negative
FP	False positive
FN	False negative
2K	Označeni skup podataka koji sadrži pozitivnu i negativnu klasu
3K	Označeni skup podataka koji sadrži pozitivnu, negativnu i neutralnu klasu
Pre	Precision
Rcall	Recall
Acc	Accuracy
NN	Bez normalizacije
ST	Normalizacija stemovanjem
NM	Normalizacija pomoću morfološkog rečnika
4G	Normalizacija odsecanjem na 4-grame
5G	Normalizacija odsecanjem na 5-grame
6G	Normalizacija odsecanjem na 6-grame
7G	Normalizacija odsecanjem na 7-grame
ALL	Ceo skup podataka
OnlyNeg	Skup podataka koji obuhvata samo negacije
OnlyRuleNeg	Skup podataka koji obuhvata samo negacije koje su obuhvaćene korišćenim pravilima negacije
LBM	Lexicon-based method
MLM	Machine learning methods

BIOGRAFIJA AUTORA

Adela Ljajić je rođena 28. avgusta 1982. godine u Novom Pazaru. Osnovnu školu i gimnaziju završila je u Novom Pazaru. Školske godine 2000/2001. godine je upisala studije na Fakultetu organizacijonih nauka u Beogradu (studijski program Informacioni sistemi), i diplomirala je 2005/2006. godine. Master studije je upisala je na istom fakultetu 2007/2008. godine (studijski program Informacioni sistemi i tehnologije) i diplomu master studija stekla 2009/2010. godine. Školske 2012/2013 je upisala doktorske studije na Elektronskom fakultetu u Nišu, studijski program Elektrotehnika i računarstvo (modul Računarstvo i informatika). Od 2007. godine je zaposlena kao saradnik u nastavi na Državnom univerzitetu u Novom Pazaru. U periodu 2008-2011. godine je učestvovala na projektima TR-13012 i TR-35023 i u period 2011-2018 na projektu III-44007. Njeni trenutni istraživački interesi odnose se na obradu prirodnog jezika (*Natural Language Processing*), mašinsku obradu teksta, mašinsko učenje i analizu sentimenta s posebnim naglaskom na obradu negacije. Autor/koautor je značajnog broja publikacija na međunarodnim konferencijama i časopisima u oblasti obrade prirodnog jezika i analize sentimenta.

IZJAVA O AUTORSTVU

Izjavljujem da je doktorska disertacija, pod naslovom

OBRADA NEGACIJE U KRATKIM NEFORMALNIM TEKSTOVIMA U CILJU POBOLJŠANJA KLASIFIKACIJE SENTIMENTA

koja je odbranjena na Elektronskom fakultetu Univerziteta u Nišu:

- rezultat sopstvenog istraživačkog rada;
- da ovu disertaciju, ni u celini, niti u delovima, nisam prijavljivala na drugim fakultetima, niti univerzitetima;
- da nisam povredila autorska prava, niti zloupotrebila intelektualnu svojinu drugih lica.

Dozvoljavam da se objave moji lični podaci, koji su u vezi sa autorstvom i dobijanjem akademskog zvanja doktora nauka, kao što su ime i prezime, godina i mesto rođenja i datum odbrane rada, i to u katalogu Biblioteke, Digitalnom repozitorijumu Univerziteta u Nišu, kao i u publikacijama Univerziteta u Nišu.

U Nišu, 13.05.2019. godine

Potpis autora disertacije:

A. Ljajić

Adela B. Ljajić

IZJAVA O ISTOVETNOSTI ELEKTRONSKOG I ŠTAMPANOG OBLIKA DOKTORSKE DISERTACIJE

Naslov disertacije:

OBRADA NEGACIJE U KRATKIM NEFORMALNIM TEKSTOVIMA U CILJU POBOLJŠANJA KLASIFIKACIJE SENTIMENTA

Izjavljujem da je elektronski oblik moje doktorske disertacije, koju sam predala za unošenje u **Digitalni repozitorijum Univerziteta u Nišu**, istovetan štampanom obliku.

U Nišu, 13.05.2019. godine

Potpis autora disertacije:

A. Ljajić

Adela B. Ljajić

IZJAVA O KORIŠĆENJU

Ovlašćujem Univerzitetsku biblioteku „Nikola Tesla“ da u Digitalni repozitorijum Univerziteta u Nišu unese moju doktorsku disertaciju, pod naslovom:

OBRADA NEGACIJE U KRATKIM NEFORMALNIM TEKSTOVIMA U CILJU POBOLJŠANJA KLASIFIKACIJE SENTIMENTA

Disertaciju sa svim prilozima predala sam u elektronskom obliku, pogodnom za trajno arhiviranje.

Moju doktorsku disertaciju, unetu u Digitalni repozitorijum Univerziteta u Nišu, mogu koristiti svi koji poštuju odredbe sadržane u odabranom tipu licence Kreativne zajednice (Creative Commons), za koju sam se odlučila.

1. Autorstvo **(CC BY)**
2. Autorstvo – nekomercijalno **(CC BY-NC)**
3. Autorstvo – nekomercijalno – bez prerade **(CC BY-NC-ND)**
4. Autorstvo – nekomercijalno – deliti pod istim uslovima **(CC BY-NC-SA)**
5. Autorstvo – bez prerade **(CC BY-ND)**
6. Autorstvo – deliti pod istim uslovima **(CC BY-SA)**

U Nišu, 13.05.2019. godine

Potpis autora disertacije:

A. Ljajić

Adela B. Ljajić