UNIVERZITET U BEOGRADU
ELEKTROTEHNIČKI FAKULTET

Vuk V. Batanović

METODOLOGIJA REŠAVANJA SEMANTIČKIH PROBLEMA U OBRADI KRATKIH TEKSTOVA NAPISANIH NA PRIRODNIM JEZICIMA SA OGRANIČENIM RESURSIMA

doktorska disertacija

Beograd, 2020

UNIVERSITY OF BELGRADE
SCHOOL OF ELECTRICAL ENGINEERING

Vuk V. Batanović

A METHODOLOGY FOR SOLVING SEMANTIC TASKS IN THE PROCESSING OF SHORT TEXTS WRITTEN IN NATURAL LANGUAGES WITH LIMITED RESOURCES

Doctoral Dissertation

Belgrade, 2020

Mentori:

dr Boško Nikolić, redovni profesor

Univerzitet u Beogradu – Elektrotehnički fakultet

dr Miloš Cvetanović, vanredni profesor

Univerzitet u Beogradu – Elektrotehnički fakultet

Članovi komisije:

dr Boško Nikolić, redovni profesor

Univerzitet u Beogradu – Elektrotehnički fakultet

dr Miloš Cvetanović, vanredni profesor

Univerzitet u Beogradu – Elektrotehnički fakultet

dr Dragan Bojić, vanredni profesor

Univerzitet u Beogradu – Elektrotehnički fakultet

dr Zoran Ševarac, vanredni profesor

Univerzitet u Beogradu – Fakultet organizacionih nauka

dr Dražen Drašković, docent

Univerzitet u Beogradu – Elektrotehnički fakultet

Datum odbrane:

_____

Zahvalnice

Ova disertacija je plod višegodišnjih naučnih istraživanja i ne bi bila moguća bez pomoći i podrške mnogih. Najpre bih želeo da se zahvalim mentorima, prof. dr Bošku Nikoliću i prof. dr Milošu Cvetanoviću, na velikom broju saveta i sugestija, kao i na podršci da istrajem u ovom pravcu istraživanja. Profesoru dr Bošku Nikoliću i dr Bojanu Furlanu se takođe zahvaljujem što su mi pružili priliku da se upoznam sa oblašću obrade prirodnih jezika i uveli me u problematiku semantičke obrade tekstova.

Izuzetnu zahvalnost dugujem članovima *Regional Linguistic Data Initiative* (ReLDI) projekta – dr Tanji Samardžić, prof. dr Nikoli Ljubešiću i prof. dr Maji Miličević Petrović – na projektnoj podršci u sprovođenju anotacije skupova podataka kreiranih u ovoj disertaciji, kao i na svim stručnim smernicama i preporukama. Dr Tanji Samardžić sam takođe zahvalan na konsultacijama u vezi sa sadržajem uvodnih poglavlja disertacije.

Kao što je u ovom radu pokazano, mogućnost razvoja statističkih rešenja u obradi prirodnih jezika sa ograničenim resursima umnogome zavisi od kvaliteta anotacije podataka, te se stoga duboko zahvaljujem sjajnim anotatorima – Ognjenu Krešiću, koji mi je pomogao i u formulisanju uputstava za anotaciju sentimenta kratkih tekstova, Ani Bjelogrlić i Jeleni Bošnjak, koje su učestvovale u uobličavanju uputstava za anotaciju semantičke sličnosti kratkih tekstova, kao i Marku Jankoviću i Filipu Maljkoviću. Ani Bjelogrlić takođe hvala za pomoć oko prevoda lingvističke terminologije i za lekturu ovog teksta. Aleksandru Milinkoviću sam zahvalan na pomoći u sprovođenju lematizacije *srWaC* korpusa tekstova na srpskom jeziku.

Svim bliskim ljudima, prijateljima i široj porodici, a naročito tetki Mirjani, hvala na razumevanju i podršci. Na kraju, najveću zahvalnost dugujem majci Slavici i ocu Vladanu, za svu ljubav, neizmernu požrtvovanost i veru u mene. Njima posvećujem ovaj rad.

Naslov doktorske disertacije: Metodologija rešavanja semantičkih problema u obradi kratkih tekstova napisanih na prirodnim jezicima sa ograničenim resursima

Sažetak: Statistički pristupi obradi prirodnih jezika tipično zahtevaju značajne količine anotiranih podataka, a često i različite pomoćne jezičke alate, što ograničava njihovu primenu u resursno ograničenim situacijama. U ovoj disertaciji predstavljena je metodologija razvoja statističkih rešenja u semantičkoj obradi prirodnih jezika sa ograničenim resursima. Ovakvi jezici se odlikuju ne samo limitiranim brojem postojećih jezičkih resursa, već i ograničenim mogućnostima za razvoj novih skupova podataka i namenskih alata i algoritama. Predložena metodologija je usredsređena na kratke tekstove zbog njihove rasprostranjenosti u digitalnoj komunikaciji i zbog veće složenosti njihove semantičke obrade.

Metodologija obuhvata sve faze izrade statističkih rešenja, od prikupljanja tekstualnog sadržaja, preko anotacije podataka, do formulisanja, obučavanja i evaluacije modela mašinskog učenja. Njena upotreba je detaljno ilustrovana na dva semantička problema – analizi sentimenta i određivanju semantičke sličnosti. Kao primer jezika sa ograničenim resursima korišćen je srpski jezik, ali se predložena metodologija može primeniti i na druge jezike iz ove kategorije.

Pored opšte metodologije, u doprinose ove disertacije spada razvoj novog, fleksibilnog sistema označavanja sentimenta kratkih tekstova, nove metrike za utvrđivanje ekonomičnosti anotacije, kao i nekoliko novih modela za određivanje semantičke sličnosti kratkih tekstova. Rezultati disertacije uključuju i kreiranje prvih javno dostupnih anotiranih skupova podataka za probleme analize sentimenta i određivanja semantičke sličnosti kratkih tekstova na srpskom jeziku, razvoj i evaluaciju većeg broja modela na ovim problemima, i prvu komparativnu evaluaciju većeg broja alata za morfološku normalizaciju na kratkim tekstovima na srpskom jeziku.

Ključne reči: obrada prirodnih jezika, računarska lingvistika, semantička sličnost tekstova, analiza sentimenta, morfološka normalizacija, lingvistička anotacija, mašinsko učenje

Naučna oblast: Elektrotehnika i računarstvo

Uža naučna oblast: Softversko inženjerstvo

UDK broj: 621.3

Title of the doctoral dissertation: A methodology for solving semantic tasks in the processing of short texts written in natural languages with limited resources

Abstract: Statistical approaches to natural language processing typically require considerable amounts of labeled data, and often various auxiliary language tools as well, limiting their applicability in resource-limited settings. This thesis presents a methodology for developing statistical solutions in the semantic processing of natural languages with limited resources. In these languages, not only are existing language resources limited, but so are the capabilities for developing new datasets and dedicated tools and algorithms. The proposed methodology focuses on short texts due to their prevalence in digital communication, as well as the greater complexity regarding their semantic processing.

The methodology encompasses all phases in the creation of statistical solutions, from the collection of textual content, to data annotation, to the formulation, training, and evaluation of machine learning models. Its use is illustrated in detail on two semantic tasks – sentiment analysis and semantic textual similarity. The Serbian language is utilized as an example of a language with limited resources, but the proposed methodology can also be applied to other languages in this category.

In addition to the general methodology, the contributions of this thesis consist of the development of a new, flexible short-text sentiment annotation system, a new annotation cost-effectiveness metric, as well as several new semantic textual similarity models. The thesis results also include the creation of the first publicly available annotated datasets of short texts in Serbian for the tasks of sentiment analysis and semantic textual similarity, the development and evaluation of numerous models on these tasks, and the first comparative evaluation of multiple morphological normalization tools on short texts in Serbian.

Key words: natural language processing, computational linguistics, semantic textual similarity, sentiment analysis, morphological normalization, linguistic annotation, machine learning

Scientific field: Electrical engineering and computer science

Scientific subfield: Software engineering

UDC number: 621.3

Sadržaj

1 Uvod

Obrada prirodnih jezika (engl. *Natural Language Processing* – NLP) se bavi proučavanjem metoda za računarsku obradu i interpretaciju tekstualnih podataka napisanih na nekom od prirodnih jezika, tj. jezika koje ljudi koriste u međusobnoj komunikaciji (npr. srpski, engleski, itd.). Za razliku od programskih jezika, kod kojih su načini izražavanja unapred strogo definisani, prirodni jezici nemaju takvih inherentnih ograničenja. Iako tekstovi na prirodnim jezicima gotovo uvek imaju svoju internu uređenost, njihova struktura i semantika nisu direktno mašinski čitljivi, što dovodi do potrebe za posebnim rešenjima u cilju obrade i analize takvih podataka. Dodatan problem predstavlja činjenica da se tekstovi na prirodnim jezicima često odlikuju velikim stepenom kompleksnosti, kao i nejasnoćama i dvosmislenostima/višesmislenostima u izražavanju. Osim toga, za pravilno razumevanje određenih iskaza često je neophodno i šire znanje o svetu (engl. *world knowledge*), što sve čini obradu prirodnih jezika izuzetno izazovnom.

Uprkos tome, ova oblast se razvija ubrzanim tempom, naročito od 2000. godine do danas (Mohammad 2020). Podaci iz korpusa radova prikupljenih iz 34 najvažnija NLP časopisa i zbornika konferencija pokazuju da je u toku 1985. godine objavljeno manje od 500 radova iz ove oblasti, u toku 2000. godine oko 2000 radova, a u toku 2014. godine skoro 4000 radova (Mariani et al. 2019). Slično tome, podaci izdavača Elsevier pokazuju da je u toku 2000. godine objavljeno oko 2400 radova iz obrade prirodnih jezika indeksiranih u Scopus bazi, da bi 2016. godine taj broj prestigao 9800 (Elsevier 2018).

Ovom tehnološkom ubrzanju je doprinelo više faktora. Raniji razvoj NLP rešenja je uglavnom bio zasnovan na ručno sastavljenim skupovima pravila (engl. *rule-based NLP*). Njihova izrada i održavanje podrazumeva obimno angažovanje eksperata za određenu problematiku, čiji je zadatak da eksplicitno formulišu svoje znanje u formi pravila koje bi računarski sistem pratio, što je za mnoge probleme obrade jezika izuzetno težak i spor proces. Od devedesetih godina prošlog veka razvoj obrade prirodnih jezika se sve više orijentiše ka statističkim pristupima (engl. *statistical NLP*), zasnovanim na mašinskom učenju, i to pre svega nadgledanom (Hirschberg & Manning 2015). Obučavanje nadgledanih statističkih modela zahteva samo skup primera u obliku parova (ulaz, tačan izlaz), tj. anotiranje tačnih izlaza za određeni skup podataka, bez eksplicitnog objašnjavanja veza između konkretnih podataka i željenih predikcija. Ovo je uglavnom daleko lakše i brže ostvarivo od formulisanja pravila. Dodatna prednost statističkih pristupa jeste što je često iste ili slične algoritme moguće koristiti za rešavanje različitih problema, ako postoje adekvatni podaci za dati problem. Međutim, slaba dostupnost adekvatnih podataka i skromni hardverski kapaciteti računara su dugo bili ograničavajući faktori u primeni statističkih rešenja. Ipak, dostupnost tekstualnog sadržaja je eksponencijalno porasla sa razvojem interneta i širenjem Web 2.0 koncepta, jer su podaci na internetu uglavnom dostupni u obliku tekstova na nekom od prirodnih jezika, pošto su namenjeni međuljudskoj komunikaciji. Simultani stalni rast hardverskih kapaciteta i performansi računara je olakšao efikasno procesiranje većih skupova podataka i pospešio raširenost statističkih modela, koji su na većini NLP problema nadmašili performanse rešenja zasnovanih na ručno sastavljenim skupovima pravila. Brz razvoj celokupne oblasti obrade prirodnih jezika je doveo i do sve većeg interesovanja za komercijalizaciju NLP rešenja, dodatno stimulišući dalja istraživanja.

Znatna komercijalna primenjivost učinila je da problemi semantičke obrade tekstova budu među najpopularnijim u aktuelnim istraživanjima. U semantičke probleme u obradi prirodnih jezika spada veliki broj zadataka koji podrazumevaju ili za cilj imaju pravilno razumevanje značenja tekstova, kao što su: analiza sentimenta i emocija u tekstu, određivanje semantičke sličnosti, detekcija parafraza, odgovaranje na pitanja, dohvatanje informacija, izrada sažetaka teksta, uprošćavanje teksta, mašinsko

prevođenje, zaključivanje na prirodnom jeziku itd. U okviru toga, naročito se ističu problemi obrade kratkih tekstova, kod kojih je, zbog ograničene dužine raspoloživog tekstualnog sadržaja, semantička obrada primetno teža nego kod dužih dokumenata. Iako ne postoji čvrsto utvrđena definicija, pod kratkim tekstovima se obično podrazumevaju oni dužine jedne ili dve rečenice, tj. tekstovi u opsegu od nekoliko reči do jednog pasusa. Tekstovi ove dužine se često sreću na internetu u vidu opisa proizvoda, naslova i sažetaka vesti i članaka, komentara posetilaca na sajtovima, forumima ili društvenim mrežama, itd.

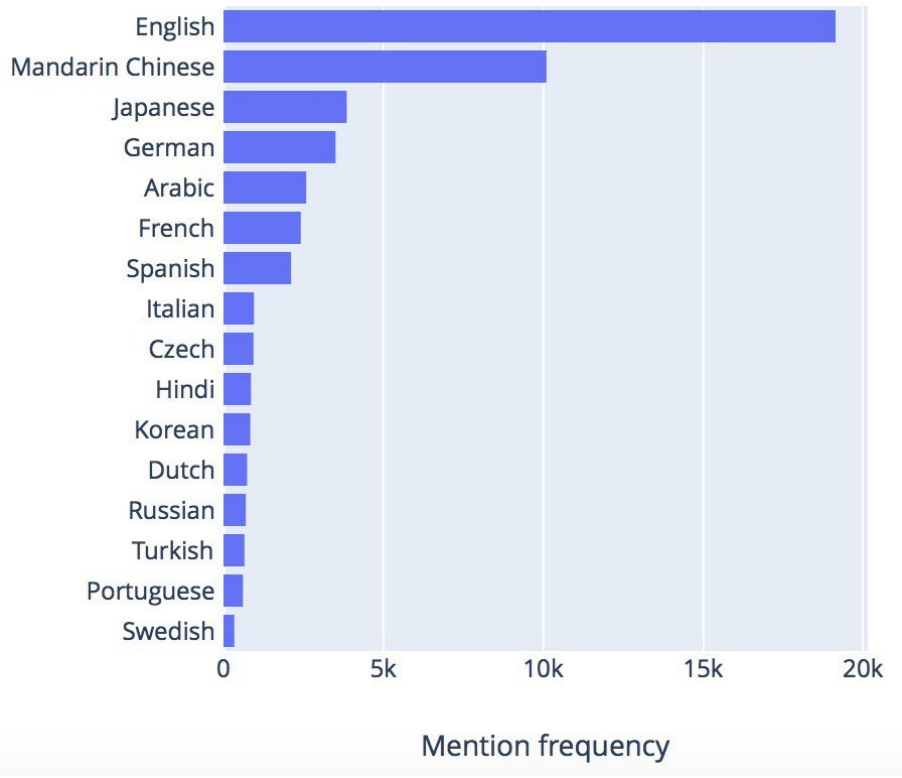
1.1 Neravnomernost NLP istraživanja po jezicima

Razvoj NLP tehnologija je dosta dugo bio izrazito usmeren samo na engleski jezik, a čak je i danas engleski i dalje dominantan u NLP istraživanjima (Benjamin 2018). Ovakva situacija je delimično uzrokovana statusom engleskog kao vodećeg jezika u međunarodnoj komunikaciji i posledičnom velikom komercijalnom atraktivnošću NLP rešenja za englesko govorno područje. Sve ovo je vremenom uticalo na daleko veću raširenost i javnu dostupnost NLP resursa za engleski, što je omogućilo sve lakši i brži dalji razvoj rešenja za ovaj jezik (El-Haj et al. 2015).

Iako se u novije vreme uočava nešto veće interesovanje naučne zajednice za računarsku obradu drugih jezika, to interesovanje je najviše usmereno ka jezicima sa velikim brojem govornika, poput kineskog, japanskog, nemačkog ili arapskog (Maxwell & Hughes 2006; Streiter et al. 2006). Ovo je dodatno akcentovano prodorom NLP modela zasnovanih na dubokom mašinskom učenju (engl. *deep learning*), jer takvi modeli tipično zahtevaju jako velike količine podataka za obučavanje, koje je teško prikupiti za manje jezike. Tek se u poslednjih par godina može primetiti pomak u razvoju i promovisanju višejezičnih modela, pri čemu broj jezika koje takvi modeli podržavaju može znatno varirati i tipično se kreće od samo nekoliko do maksimalno stotinak. Slika 1, preuzeta iz nedavne analize radova objavljenih u poslednjih 20 godina na ACL (*Association for Computational Linguistics*) konferencijama[1], koje spadaju u najprestižnije u oblasti obrade prirodnih jezika, jasno pokazuje navedene trendove i dramatičnu razliku u zastupljenosti u NLP istraživanjima između malog broja velikih svetskih jezika i svih ostalih.

Magnitudu ovog disbalansa dodatno potkrepljuju podaci iz baze *Ethnologue*[2], koji pokazuju da se trenutno na planeti govori oko 7100 jezika, od čega blizu 4000 njih ima pismo. Naravno, znatno manje jezika je prisutno i u digitalnom obliku, ali njihov tačan broj je teško proceniti. U studiji o digitalnom izumiranju jezika (Kornai 2013) utvrđeno je da nije moguće pronaći bilo kakvo prisustvo na internetu za 6541 jezik, dok su u okviru *Crúbadán* projekta (Scannell 2007) najpre prikupljeni veb korpusi za preko 400 jezika, da bi taj broj kasnije porastao na preko 2000[3]. Kao dobar indikator šire zastupljenosti može poslužiti enciklopedija Vikipedija, koja je trenutno aktivna na oko 300 jezika[4], dok je Opšta deklaracija o pravima čoveka trenutno dostupna na sajtu Ujedinjenih nacija na preko 500 jezika[5]. Ipak, i u široj digitalnoj sferi dominantnost engleskog je izrazita – procenjuje se da je danas blizu 60% sadržaja na internetu napisano na ovom prirodnom jeziku[6].

[1] http://towardsdatascience.com/major-trends-in-nlp-a-review-of-20-years-of-acl-research-56f5520d473
[2] http://www.ethnologue.com/
[3] http://crubadan.org/
[4] http://en.wikipedia.org/wiki/List_of_Wikipedias
[5] http://www.ohchr.org/EN/UDHR/Pages/SearchByLang.aspx
[6] http://w3techs.com/technologies/overview/content_language



Slika 1. Zastupljenost različitih prirodnih jezika u naučnim radovima predstavljenim u poslednjih 20 godina na ACL konferencijama. Slika preuzeta sa interneta[1]

Implicitan stav u najvećem delu statističkih NLP istraživanja sprovedenih nad engleskim jezikom jeste da ona zapravo i nisu vezana za taj konkretni jezik, već da su predstavljena rešenja univerzalno primenjiva na bilo koji prirodni jezik (Streiter et al. 2006; Bender 2011). Ipak, čest uslov za primenu semantičkih NLP modela je postojanje adekvatnih NLP alata nižeg nivoa (npr. tokenizatora, modula za obeležavanje vrsta reči i sl.) radi pretprocesiranja podataka, ali takvi alati ne postoje za mnoge prirodne jezike. Pored toga, čak i ako neko NLP rešenje jeste principijelno primenjivo na različite prirodne jezike, to ne znači da će dobijeni rezultati biti podjednako dobri na svim jezicima. Emily Bender (Bender 2009, 2011) razmatra primer *n*-gramskih modela koji koriste sekvence reči i koji na prvi pogled ne zavise od svojstava jezika, jer se njihova upotreba zasniva samo na računanju relativne frekventnosti različitih sekvenca reči. Njena pažljivija analiza pokazuje da je efektivnost ovih jednostavnih modela u rešavanju određenih NLP problema na engleskom jeziku zapravo posledica dva svojstva tog jezika – ograničene morfološke kompleksnosti i relativno fiksnog redosleda reči u rečenici (subjekat-predikat-objekat). Oba ova svojstva smanjuju proređenost podataka (engl. *data sparsity*) tj. broj mogućih kombinacija različitih reči/oblika reči u sekvenci određene dužine. Samim tim, nije moguće *a priori* garantovati da će se ovakvi modeli ponašati podjednako kvalitetno u brojnim jezicima sa bogatom morfologijom i slobodnijim redosledom reči u rečenici, kao što je npr. srpski, gde je proređenost podataka znatno veća.

Dodatna pretpostavka implicitno vezana za ideju opšte primenjivosti statističkih rešenja razvijenih za engleski jezik jeste da je i za druge prirodne jezike moguće prikupiti dovoljne količine adekvatnih podataka za obučavanje i evaluaciju NLP modela. Ova pretpostavka takođe često nije adekvatna. Najpre, imajući u vidu navedenu dominantnost engleskog jezika u sadržajima na internetu, pitanje je da li za odabrani manji prirodni jezik postoji dovoljno dostupnog tekstualnog sadržaja za potrebe NLP modela. Čak i ako taj uslov jeste ispunjen, tekstualne podatke je obično neophodno ručno anotirati na određeni način, da bi se označili ispravni odgovori za NLP problem koji se razmatra.

Složenost postupka anotacije može dramatično da se razlikuje od jednog do drugog NLP problema, kao i od konceptualnih odluka vezanih za rešavanje problema. U zavisnosti od složenosti potrebnih anotacija zavisi koliko osoba je neophodno angažovati na ovom poslu, koji je neophodan nivo njihovog lingvističkog predznanja, kao i koliko vremena će anotatorima biti potrebno za realizaciju zadatka. Ovi faktori zajedno utiču na finansijske zahteve za sprovođenje anotacije, koji su čest problem u razvoju NLP resursa i tehnologija za manje jezike. Razlog za to jeste što su zbog nižeg tržišnog interesovanja i ograničene snage lokalnih ekonomija i javni i privatni izvori finansiranja znatno redukovani kod manjih jezika (Krauwer 2003; Streiter et al. 2006; Scannell 2007).

Crowdsourcing pristupi (Injac-Malbaša 2012) se često navode u naučnoj literaturi kao način za anotaciju i izgradnju NLP resursa koji zahteva manje finansijske utroške (Sabou et al. 2014; Callison-Burch et al. 2015). Ove metode podrazumevaju da se anotacija celokupnog skupa podataka izdeli na veliki broj malih zadataka, gde svaki zadatak pokriva anotaciju samo nekoliko podataka. Takvi zadaci se zatim obično proizvoljno raspodeljuju radnicima na *crowdsourcing* platformi, uz dodavanje kontrolnih zadataka radi provere kvaliteta njihovog rada. Međutim, ovakvi pristupi su uglavnom neprimenjivi na manje jezike, jer je broj govornika tih jezika na *crowdsourcing* servisima poput *Amazon Mechanical Turk* ili *Crowdflower* vrlo ograničen. Sem toga, za neke NLP probleme je neophodno kulturološko i/ili lingvističko predznanje radi valjanog sprovođenja anotacije (Maxwell & Hughes 2006). Spoj ovih faktora znači da je kod manjih prirodnih jezika u većini situacija klasična organizacija procesa anotacije, sa grupom direktno angažovanih anotatora, jedina opcija.

1.2 Jezici sa ograničenim resursima

U literaturi o NLP istraživanjima ne postoji šire prihvaćena jasna definicija jezika sa ograničenim resursima (Berment 2002; Krauwer 2003; Maxwell & Hughes 2006; Streiter et al. 2006; Scannell 2007; King 2015; Duong 2017; Benjamin 2018). Stoga će se, imajući sve ranije navedene faktore u vidu, u ovoj disertaciji pod jezicima sa ograničenim resursima podrazumevati oni koji odgovaraju sledećem opisu:

- Umereno su raširenи u digitalnom obliku, te je stoga za njih moguće na internetu pronaći ograničene količine tekstualnih sadržaja željenog tipa na osnovu kojih bi se, kroz proces prikupljanja i anotacije podataka, mogli obučiti i evaluirati NLP modeli i rešenja.
- NLP alati za ove jezike ili nisu dostupni, ili su dostupni samo osnovni alati za pretprocesiranje teksta, poput tokenizatora, modula za obeležavanje vrsta reči, ili alata za morfološku normalizaciju. Takođe, za njih ne postoje odgovarajući specijalizovani korpusi za dati NLP problem, ali mogu da postoje neanotirani korpusi opšteg tipa, poput veb korpusa.
- Razvojem NLP alata i resursa za ovakve jezike se tipično bavi vrlo uzak krug ljudi, što znači da su mogućnosti dizajniranja specijalizovanih algoritama za posmatrani jezik dosta ograničene.
- Mogućnosti finansiranja anotacionih projekata za izgradnju NLP resursa na ovim jezicima su obično izuzetno limitirane, kao i mogućnosti finansiranja pristupa naprednim hardverskim resursima u vidu klastera memorijski bogatih grafičkih (engl. *Graphics Processing Unit – GPU*) i tenzorskih procesorskih jedinica (engl. *Tensor Processing Unit – TPU*).

Drugim rečima, pod pojmom *resursa* u ovom kontekstu treba imati u vidu ne samo korpuse tekstova i alate za procesiranje tekstualnih podataka, već i ljudske i finansijske resurse, koji su takođe ograničeni. Navedenim kriterijumima odgovara znatan broj manjih jezika, koji spadaju u prostor između ugroženih jezika sa jedne strane, i velikih, globalno raširenih jezika sa druge strane. Kao dobar primer jezika sa ograničenim resursima u ovoj disertaciji biće korišćen srpski jezik, koji jeste

umereno zastupljen u pogledu sadržaja na internetu i za koji jesu dostupni neki osnovni NLP alati i resursi, ali za koji nije moguće pribaviti velike količine anotiranih podataka za izradu NLP modela, kako zbog ograničenosti dostupnog relevantnog tekstualnog sadržaja, tako i zbog finansijskih ograničenja. Naravno, kategorizacija određenog jezika kao jezika sa ograničenim resursima se može promeniti vremenom, kako se ljudski, jezički i finansijski resursi vezani za NLP tehnologije za taj jezik razvijaju.

1.3 Polazne hipoteze i naučni doprinosi i ciljevi disertacije

Širi cilj i naučni doprinos ove disertacije jeste definisanje metodologije rešavanja semantičkih problema u obradi kratkih tekstova koja je pogodna za jezike sa ograničenim resursima. U sklopu toga, proces razvoja rešenja je detaljno analiziran na dva konkretna semantička problema – određivanju semantičke sličnosti i analizi sentimenta kratkih tekstova.

Određivanje semantičke sličnosti kratkih tekstova (engl. *semantic textual similarity – STS* ili *short-text semantic similarity – STSS*) predstavlja problem automatskog ocenjivanja zadatog para tekstova na osnovu sličnosti njihovih značenja (Corley & Mihalcea 2005; Agirre et al. 2012). Za razliku od proste leksičke sličnosti, kod koje se posmatra samo prisustvo podudarnih reči u zadatim tekstovima, cilj semantičke sličnosti jeste da prepozna semantičku bliskost čak i u slučajevima potpune leksičke različitosti. Na primer, rečenice *Uskoro sledi poskupljenje struje* i *Cena električne energije će ubrzo porasti* su semantički praktično ekvivalentne, iako ne dele niti jednu istu reč. Analogno ovome, leksički bliski tekstovi mogu biti primetno semantički različiti. Semantička sličnost kratkih tekstova je jedan od fundamentalnih problema u semantičkoj obradi prirodnih jezika, i predstavlja integralni deo drugih NLP zadataka, kao što su izrada sažetaka tekstova, odgovaranje na pitanja, dohvatanje informacija, itd.

Analiza sentimenta (engl. *sentiment analysis*), takođe poznata i kao i istraživanje mišljenja (engl. *opinion mining*), predstavlja problem automatske detekcije i obrade stavova, procena i mišljenja koje ljudi iskazuju prema određenim entitetima, osobama, događajima, pitanjima, temama i njihovim svojstvima (Pang & Lee 2008; Liu & Zhang 2012). U analizu sentimenta spadaju uži problemi poput određivanja polarnosti teksta (engl. *polarity detection*), gde je cilj binarna podela tekstova na pozitivne i negativne, određivanje subjektivnosti teksta (engl. *subjectivity detection*), gde je cilj razlikovanje objektivnih tekstova od subjektivnih, tj. onih koji izražavaju neko mišljenje, detekcija sarkazma u tekstu (engl. *sarcasm detection*), itd. Zbog velike komercijalne potrebe za automatskom detekcijom i obradom stavova ljudi, analiza sentimenta predstavlja jedan od NLP problema sa najširom direktnom praktičnom primenom.

Određivanje semantičke sličnosti i analiza sentimenta kratkih tekstova su odabrani u ovoj disertaciji kao primeri semantičkih problema kako zbog svog značaja i velike zastupljenosti u aktuelnim trendovima istraživanja u okviru obrade prirodnih jezika, tako i zbog svoje pogodnosti za resursno ograničene jezike. Naime, određivanje semantičke sličnosti para tekstova se najčešće formuliše kao problem predikcije numeričke vrednosti na nekoj skali sličnosti, gde najniža vrednost predstavlja njihovu potpunu semantičku različitost, a najviša potpuno semantičko poklapanje. Analiza sentimenta, s druge strane, obično se konceptualizuje u vidu problema klasifikacije. Samim tim, anotacija tačnih vrednosti izlaza nad određenim skupom tekstova je za ove probleme, strukturno gledano, jednostavna, jer se svodi na dodeljivanje samo jedne oznake svakom tekstu. Pored toga, ni za jedan od navedenih problema nije neophodno lingvističko predznanje anotatora, što dodatno olakšava proces anotacije. Konačno, istraživanja na engleskom jeziku su pokazala da su oba problema takva da se barem osnovna rešenja za njih mogu kreirati i bez sintaktičke analize tekstova (Pang et al. 2002; Islam & Inkpen 2008), što je od značaja za jezike sa ograničenom dostupnošću NLP alata.

Naime, u trenutku otpočinjanja rada na ovoj disertaciji, javno dostupni parser za srpski jezik nije postojao. Iako je u toku izrade disertacije došlo do objavljivanja prvih alata ovog tipa (Samardžić et al. 2017), kao i referentnog korpusa tekstova na srpskom jeziku sa sintaktičkim anotacijama (Batanović et al. 2018b), ovakvi alati i resursi su i dalje nedostupni za mnoge jezike sa ograničenim resursima.

Imajući navedene faktore u vidu, u izradi metodologije rešavanja semantičkih problema u obradi kratkih tekstova napisanih na jezicima sa ograničenim resursima usvojene su sledeće polazne hipoteze:

- Moguće je pribaviti dovoljne količine neophodnih tekstualnih resursa na srpskom jeziku za razvoj i evaluaciju statističkih rešenja za posmatrane semantičke probleme.
- Osnovna referentna rešenja (engl. *baselines*) za posmatrane probleme mogu se unaprediti korišćenjem dostupnih korpusa opšteg tipa i osnovnih NLP alata.
- Problem proređenosti podataka koji je prouzrokovan morfološkom kompleksnošću jezika poput srpskog može se minimizovati primenom alata za morfološku normalizaciju.
- Određivanje semantičke sličnosti i analiza sentimenta kratkih tekstova se mogu realizovati i bez potpune sintaktičke analize teksta, čak i u jezicima sa složenijim sintaktičkim pravilima kao što je srpski.
- Iako su neuralni modeli tipično veoma zahtevni u pogledu količine podataka potrebnih za njihovo obučavanje, moguće je primeniti određene varijante takvih modela i na rešavanje semantičkih problema u jezicima sa ograničenim resursima.

Za razliku od glavnih trendova dosadašnjih istraživanja u obradi resursno ograničenih jezika, koji su se uglavnom bavili isključivo razvojem različitih računarskih modela, u ovoj disertaciji dodatna pažnja je data i aspektu izgradnje i anotiranja tekstualnih resursa, radi formulisanja sveobuhvatnije metodologije razvoja statističkih rešenja. U naučne doprinose i ciljeve ove disertacije stoga spadaju:

1. Formulisanje šire metodologije statističkog rešavanja semantičkih problema u obradi kratkih tekstova napisanih na jezicima sa ograničenim resursima, koja obuhvata sve faze razvoja;
2. Identifikacija postojećih pristupa za rešavanje semantičkih problema u obradi kratkih tekstova koji su pogodni za primenu u jezicima sa ograničenim resursima;
3. Izrada metrike za merenje ekonomičnosti anotacije, tj. ekonomičnosti upotrebe uputstava za anotaciju u okviru određenog sistema označavanja podataka;
4. Izrada novog, fleksibilnog i ekonomičnog sistema označavanja sentimenta tekstova, koji omogućava više nivoa interpretacije za oznake sentimenta i pogodan je za primenu u jezicima sa ograničenim resursima;
5. Stvaranje prvih referentnih javno dostupnih anotiranih skupova podataka za određivanje semantičke sličnosti i analizu sentimenta kratkih tekstova na srpskom jeziku;
6. Kreiranje i evaluacija osnovnih referentnih statističkih rešenja za određivanje semantičke sličnosti i analizu sentimenta kratkih tekstova na srpskom jeziku i njihovih varijanti i unapređenja;
7. Prva komparativna evaluacija efekata morfološke normalizacije kratkih tekstova na srpskom jeziku i različitih rešenja razvijenih za ove potrebe, u kontekstu problema određivanja semantičke sličnosti i analize sentimenta kratkih tekstova;
8. Razvoj i evaluacija novih modela za određivanje semantičke sličnosti kratkih tekstova koji su adekvatni za upotrebu u jezicima sa ograničenim resursima, a postižu bolje performanse od osnovnih modela i njihovih unapređenja;
9. Prilagođavanje i evaluacija najnovijih neuralnih modela zasnovanih na *Transformer* arhitekturama (Vaswani et al. 2017) – koji su na širokom spektru semantičkih problema pokazali primetno bolje rezultate na engleskom jeziku od prethodnih pristupa zasnovanih na

neuralnim mrežama (Conneau & Lample 2019; Devlin et al. 2019) – na rešavanju posmatranih problema na srpskom jeziku.

Doprinosi vezani za ciljeve 1, 2, 4 i 8 su direktno primenjivi na širi spektar jezika sa ograničenim resursima, dok je doprinos vezan za cilj 3 primenjiv na bilo koji jezik. Naučni doprinosi vezani za preostale ciljeve su prevashodno usmereni na srpski jezik, ali su postupci i zaključci vezani za njihovo ostvarivanje korisni i za druge jezike sa ograničenim resursima.

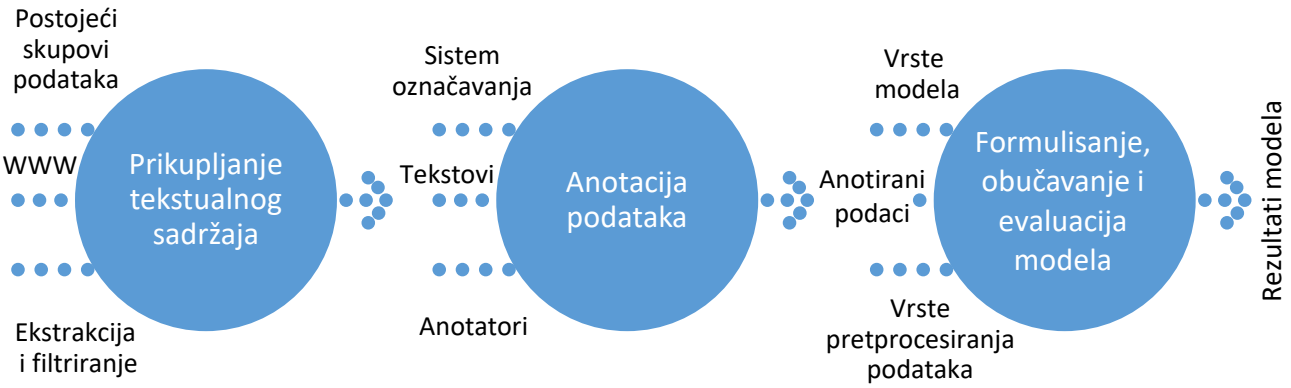
Ova disertacija je strukturirana na sledeći način: u glavi 2 je dat pregled procesa razvoja statističkih rešenja u semantičkoj obradi prirodnih jezika i načelno je diskutovana svaka od faza tog procesa, a zatim su razmotreni postojeći pristupi koji se koriste za rešavanje semantičkih problema u jezicima sa ograničenim resursima. Sve faze razvoja statističkih rešenja su obuhvaćene predloženom metodologijom i detaljno su analizirane i ilustrovane kroz razvoj rešenja za određivanje semantičke sličnosti i analizu sentimenta kratkih tekstova na srpskom jeziku. U glavi 3 prikazana je faza prikupljanja tekstualnog sadržaja, dok se faza anotacije podataka razmatra u glavi 4. Glava 5 opisuje fazu formulisanja, obučavanja i evaluacije modela i sadrži rezultate evaluacije. Uzimajući u obzir iskustva prikupljena tokom izrade ovih rešenja, u glavi 6 je dat pregled metodologije razvoja rešenja semantičkih problema koja je pogodna za jezike sa ograničenim resursima, dok su zaključak i mogući pravci daljih istraživanja izneti u glavi 7.

2 Proces razvoja statističkih rešenja u semantičkoj obradi prirodnih jezika

U ovoj glavi je najpre prikazan tipičan proces razvoja statističkih rešenja u semantičkoj obradi prirodnih jezika i načelno je diskutovana svaka od faza u ovom procesu. Nakon toga su razmotrene specifičnosti razvoja rešenja za jezike sa ograničenim resursima. Na kraju je dat uporedni pregled po fazama razvoja za rešenja problema određivanja semantičke sličnosti i analize sentimenta kratkih tekstova koja su razvijena u ovoj disertaciji.

2.1 Faze razvoja statističkih rešenja u semantičkoj obradi prirodnih jezika

Razvoj statističkih NLP rešenja se može podeliti u tri karakteristične faze, prikazane na dijagramu na slici 2.



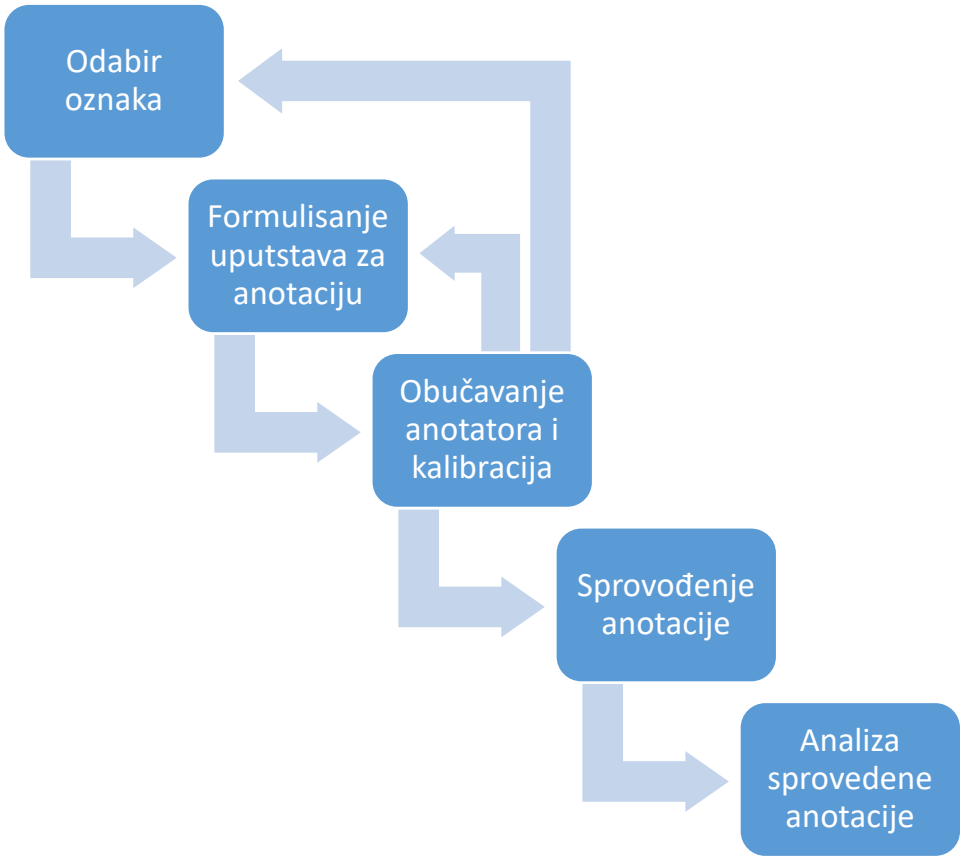
Slika 2. Dijagram faza u procesu razvoja statističkih rešenja za semantičku obradu prirodnih jezika

U fazi prikupljanja tekstualnog sadržaja neophodno je najpre identifikovati pogodne izvore tekstova koji bi se mogli koristiti za obučavanje i evaluaciju modela za NLP problem koji se razmatra. Ponekad je moguće upotrebiti postojeće skupove podataka koji su napravljeni za rešavanje nekih srodnih problema ili raspoložive digitalizovane kolekcije tekstova koje se direktno mogu dalje obrađivati, ali je ovo relativno retka situacija u jezicima sa ograničenim resursima (Streiter et al. 2006). Prikupljanje tekstualnog sadržaja stoga najčešće podrazumeva pronalaženje pogodnih izvora na internetu i pravljenje modula za ekstrakciju željenog sadržaja. Ovakvi moduli se tipično izrađuju korišćenjem neke od biblioteka za parsiranje i obradu HTML (i JavaScript) koda, kao što su *JSoup*[7] i *HTMLUnit*[8] biblioteke za programski jezik Java, ili *BeautifulSoup*[9] i *Scrapy*[10] biblioteke za Python. Iako broj podataka u prikupljenom skupu zavisi od konkretnog problema koji se razmatra i konkretnog jezika, u jezicima sa ograničenim resursima on se uglavnom kreće u opsegu od više stotina do nekoliko

[7] http://jsoup.org/
[8] http://htmlunit.sourceforge.net/
[9] http://www.crummy.com/software/BeautifulSoup/
[10] http://scrapy.org/

hiljada primera. Bez obzira na način prikupljanja tekstova, njih je obično neophodno na neki način isfiltrirati radi kontrole kvaliteta, pre prelaska na fazu anotacije podataka.

U fazi anotacije skup anotatora obeležava prikupljene tekstove koristeći određeni sistem označavanja. U literaturi ne postoji jedan standardizovan pristup anotaciji podataka, zbog velike raznorodnosti u pogledu prirode i složenosti različitih anotacionih projekata. Ipak, postoje radovi koji su pokušali da sistematizuju aspekte ove faze razvoja statističkih NLP rešenja i da formulišu širi metodološki okvir anotacije (Hovy & Lavid 2010; Pustejovsky & Stubbs 2012; Pustejovsky et al. 2017). Iako se tačna podela anotacije na etape menja od rada do rada i od projekta do projekta, u većini anotacionih poduhvata se može uočiti pet karakterističnih koraka, prikazanih u vidu dijagrama na slici 3.



Slika 3. Dijagram koraka u fazi anotacije podataka

Prvi korak predstavlja odabir oznaka koje će biti korišćene u obeležavanju podataka, zajedno sa njihovim definicijama i interpretacijama. Složenost ovog koraka umnogome zavisi od toga da li za dati NLP problem već postoji dobro definisan standard označavanja podataka koji se može koristiti, ili se za potrebe konkretnog anotacionog projekta razvija nov sistem oznaka. U slučaju kreiranja novog skupa oznaka, kompleksnost odabira zavisi i od toga da li je potrebno da označavanje bude zasnovano na određenoj lingvističkoj teoriji.

Drugi korak čini formulisanje uputstava/smernica za anotaciju koja će biti data anotatorima pre otpočinjanja samog rada. Tačan izgled i obim ovih uputstava može primetno varirati u zavisnosti od konkretnog NLP problema koji se razmatra, usvojenog sistema oznaka, kao i konceptualnih odluka vezanih za željeni nivo detaljnosti uputstava. U principu, detaljnija uputstva najčešće omogućavaju viši stepen saglasnosti anotacionih odluka između različitih anotatora, tj. veću konzistentnost

anotacije, što podiže kvalitet anotiranih resursa, ali po cenu nešto sporijeg procesa anotacije. Veći stepen konzistentnosti anotacija je od značaja i pri obučavanju i evaluaciji NLP modela, zato što pruža viši gornji prag za performanse bilo kog računarskog modela (npr. ako se ni ljudi međusobno ne slažu oko ispravne oznake u 10% slučajeva, nema preteranog smisla isticati da model ima tačnost preko 90%).

Treći korak u fazi anotacije predstavlja upoznavanje anotatora sa izrađenim uputstvima i njihovo obučavanje. To obično podrazumeva da anotatori pokušaju da označe manji podskup podataka i da zatim razreše eventualne nedoumice kroz komunikaciju sa koordinatorom procesa anotacije. U toku ovog koraka moguće su i dorade uputstava za anotaciju, pošto inicijalni set uputstava retko kada adekvatno obuhvati sve situacije koje se mogu sresti u praksi. Iz tog razloga se ova procedura ponekad naziva kalibracijom. Ukoliko se za sprovođenje anotacije koristi neki namenski alat, u okviru ovog koraka se anotatori obučavaju za njegovu upotrebu (Finlayson & Erjavec 2017). U toku ovog koraka takođe treba utvrditi sva tehnička pitanja vezana za sprovođenje anotacije, poput formata čuvanja podataka.

Četvrti korak jeste sprovođenje anotacije. U ovom koraku, svaki anotator obično radi individualno. U zavisnosti od zahteva vezanih za NLP problem koji se razmatra, kao i raspoloživog budžeta, anotaciju je moguće sprovoditi jednostruko ili višestruko. Jednostruka anotacija znači da svaki podatak dobija oznaku od strane samo jednog anotatora, tj. da između anotatora ne postoji preklapanje u podacima koje obeležavaju. Višestruka anotacija pak znači da više anotatora obeležava svaki podatak. U slučaju da su oznake koje anotatori koriste numeričke ocene, finalne oznake za svaki podatak se kod višestruke anotacije obično dobijaju uprosečavanjem individualnih ocena anotatora. Ako su oznake koje anotatori koriste kategoričkog ili strukturiranog tipa, finalne oznake za podatke kod kojih postoji razlika mišljenja između anotatora se dobijaju putem razrešavanja neslaganja (engl. *curation*). Ovaj proces obično podrazumeva da, na osnovu oznaka koje su individualno zadali svi anotatori, koordinator anotacije ili posebna osoba zadužena za ovaj deo procesa donese finalnu odluku o ispravnoj oznaci.

Peti i poslednji korak jeste analiza sprovedene anotacije, u koju spada razmatranje konzistentnosti anotacije tj. stepena saglasnosti anotatora, statistička analiza raspodele anotiranih podataka po oznakama, kao i poređenje rezultata sprovedene anotacije sa eventualnim prethodnim anotacionim projektima istog tipa i sa istim skupom korišćenih oznaka, ako takvi postoje. U slučajevima jednostruke anotacije, radi utvrđivanja saglasnosti anotatora obično se koristi manji podskup podataka (npr. oko 5-10% svih primera) koje paralelno i zasebno anotiraju svi anotatori. Poželjno je da se saglasnosti anotatora ne izražavaju samo procentualno, već i korišćenjem neke od metrika koje ne uračunavaju slučajne saglasnosti, kao što su κ (*kappa*) koeficijent (engl. *Cohen's kappa*; *Fleiss' kappa*) (Cohen 1960; Fleiss 1971) ili Kripendorfov α (*alfa*) koeficijent (engl. *Krippendorff's alpha*) (Krippendorff 2004). Naime, anotatori će nekada nekom primeru zadati istu oznaku pukom slučajnošću, a ne zato što se zaista slažu u proceni. Za razliku od κ koeficijenata, koji su namenjeni pre svega kategoričkim oznakama, Kripendorfov α koeficijent je izuzetno fleksibilan i primenjiv je na veliki broj različitih tipova oznaka (numeričke, kategoričke, redne, itd.) i proizvoljan broj anotatora, što ga čini preporučenom univerzalnom metrikom za izražavanje saglasnosti u projektima anotacije (Artstein & Poesio 2008). Nulta vrednost α koeficijenta označava nepostojanje bilo kakve saglasnosti koja nije slučajna, dok vrednost 1 označava potpunu saglasnost. Iako u literaturi ne postoji sasvim ujednačen stav oko toga kako tačno treba interpretirati različite vrednosti α koeficijenta (Siegert et al. 2014), svi izvori se slažu da su saglasnosti preko 0,8 odlične. Sam Kripendorf sugeriše da se saglasnosti preko 0,8 smatraju pouzdanim, one između 0,8 i 0,667 prihvatljivim, a one ispod 0,667 neprihvatljivim (Krippendorff 2004), te će se to tumačenje koristiti i u ovoj disertaciji.

Treba istaći da u praksi neretko dolazi do preklapanja između nekih od navedenih koraka u fazi anotacije. Ovo se obično dešava kod složenijih anotacionih projekata, kod kojih je teško unapred, pre detaljnog uvida u podatke, dizajnirati sasvim adekvatan skup oznaka i dovoljno sveobuhvatan set uputstava za anotaciju. Tada čak može doći do cikličnog povezivanja prva tri koraka (Pustejovsky & Stubbs 2012; Finlayson & Erjavec 2017), dokle god odabrane oznake i uputstva ne postanu adekvatni za kompleksnost problema koji se u anotaciji razmatra.

Ponekad se dešava i da je moguće da se celokupna faza anotacije preskoči ili automatizuje, tako što se određeni već postojeći, jasno izdvojeni indikatori u prikupljenom sadržaju koriste kao oznake za potrebe obučavanja i evaluacije modela (tipičan primer ovoga jeste korišćenje numeričkih ocena ili broja zvezdica u okviru opisa/recenzije kao indikatora sentimenta tog teksta (Pang et al. 2002; Pang & Lee 2005)). Međutim, ovakve situacije se ređe sreću, pogotovo kod jezika sa ograničenim resursima, gde je dostupnost tekstualnog sadržaja sa takvim indikatorima upitna. One takođe nose sa sobom ozbiljna ograničenja u pogledu fleksibilnosti oznaka koje će se koristiti, pošto je sistem oznaka tada u potpunosti uslovljen indikatorskim informacijama sadržanim u tekstualnom sadržaju. Osim toga, automatizovano dobijanje oznaka može da unese znatan šum u anotaciju (El-Haj et al. 2015). To u praksi može značiti da automatizacija nekada ne može da u potpunosti zameni ručnu anotaciju, već samo da olakša posao anotatorima, čiji zadatak postaje kontrola automatski generisanih oznaka.

Nakon što su prikupljeni podaci adekvatno anotirani, moguće je iskoristiti ih u obučavanju i evaluaciji različitih vrsta statističkih modela. Dobra praksa jeste da se najpre evaluiraju osnovna referentna rešenja (engl. *baselines*), radi lakše i bolje procene performansi složenijih modela. Kakvog će tačno oblika biti modeli, kako osnovni tako i kompleksniji, zavisi od konkretnog NLP problema koji se razmatra. Takođe, u ovoj fazi se obično razmatraju i različite tehnike pretprocesiranja podataka.

2.2 Statistička rešenja u semantičkoj obradi prirodnih jezika sa ograničenim resursima

Različiti pristupi su predloženi za poboljšanje i olakšanje razvoja statističkih rešenja u kontekstu ograničenih anotiranih resursa, od kojih je najpopularnije transferno učenje (engl. *transfer learning*) (Pan & Yang 2010). Transferno učenje obuhvata veliki broj metoda čija je osnovna ideja da se model ili neki njegov deo obuče na nekom problemu/domenu za koji postoji dovoljno podataka za uspešno obučavanje, i da se zatim takav model iskoristi za rešavanje konkretnog problema od interesa, za koji postoji ograničena količina dostupnih podataka. Ovo je moguće sprovesti bilo ekstrakcijom odlika (engl. *feature extraction*) tj. korišćenjem (delova) obučenog modela kao ulaza za drugi model koji rešava problem od interesa, bilo finim podešavanjem (engl. *fine-tuning*) postojećeg, već inicijalizovanog modela, tako da on nauči rešavanje problema od interesa (Howard & Ruder 2018; Peters et al. 2019).

Najrašireniji oblik transfernog učenja u semantičkoj obradi prirodnih jezika su vektori značenja reči (engl. *word embeddings*) (Kale et al. 2019). Ovi vektori fiksne dužine na numerički način predstavljaju značenje reči, i zasnivaju se na distribucionoj hipotezi (Harris 1954; Firth 1957) po kojoj se reči sa sličnim značenjima upotrebljavaju u sličnim kontekstima. To znači da je semantiku reči moguće izvesti iz posmatranja njene upotrebe u nekoj kolekciji tekstova, pri čemu za to nije neophodna bilo kakva vrsta anotacije, jer se kontekst date reči uočava prostim uvidom u njoj okolne reči. Pri tome je poželjno da odabrana kolekcija tekstova bude što veća, radi uočavanja što više primera upotrebe svake reči. Vektorska reprezentacija značenja reči dovodi do efekta da se semantički bliske reči nalaze blizu jedna drugoj u semantičkom vektorskom prostoru i omogućava da se značenja reči porede preko sličnosti njihovih vektora.

Vektori značenja reči se koriste još od devedesetih godina prošlog veka, kada su se dominantne metode njihovog konstruisanja zasnivale na formiranju matrice uzajamnog pojavljivanja reči (engl. *co-occurrence matrix*) (Deerwester et al. 1990; Lund & Burgess 1996). Takva matrica je zatim najčešće minimizovana korišćenjem *Singular Value Decomposition* (SVD) operacije (Deerwester et al. 1990; Rohde et al. 2005), da bi se vektori značenja sveli na dimenzije koje omogućavaju lakše manipulisanje. Ipak, u zavisnosti od implementacije SVD procedure, minimizacija ogromnih matrica može da bude veoma memorijski zahtevna. Od 2013. godine primat su jasno preuzela neuralna rešenja za kreiranje vektora značenja reči, u vidu varijanti *word2vec* algoritma (Mikolov et al. 2013a, b, c) i njemu sličnih modela, kao što su *GloVe* (Pennington et al. 2014) i *fastText* (Bojanowski et al. 2017). Ovi algoritmi omogućavaju iterativnu izradu niskodimenzionalnih, tj. zgusnutih vektorskih reprezentacija značenja (engl. *dense representation*) uz niže zauzeće memorijskih resursa, što je značajno pri radu sa jako velikim korpusima (Levy et al. 2015).

Snaga vektorskih reprezentacija značenja reči leži u tome što je njih moguće naučiti jednom, na datom velikom neanotiranom korpusu tekstova, a zatim koristiti proizvoljno puno puta u okviru različitih NLP problema, čak (ili posebno) i onih za koje postoje dosta ograničene količine anotiranih podataka. Drugim rečima, vektori značenja reči predstavljaju način za prenošenje ranije stečenog znanja modela o semantici reči na rešavanje konkretnog NLP problema.

Međutim, sve navedene metode vektorskog predstavljanja značenja reči pružaju samo jednu globalnu/statičku reprezentaciju za sva moguća značenja reči. Stoga će kod polisemičnih reči i homonima jedan jedinstven semantički vektor predstavljati amalgam svih njihovih različitih značenja. Kao rešenje ovog problema, u poslednjih par godina sve aktuelnije su kontekstno osetljive reprezentacije značenja, kod kojih se tačan vektorski oblik predstavljanja reči dobija tek uvidom u njenu upotrebu tj. uvidom u konkretan kontekst u kome se reč javlja (Peters et al. 2018; Ethayarajh 2019). Kontekstno osetljive reprezentacije značenja se dobijaju kroz proces obučavanja jezičkih modela (engl. *language modeling*), tj. rešavanja problema predikcije narednih, prethodnih i/ili nedostajućih reči u nekoj sekvenci, za šta se takođe koriste veliki neanotirani korpusi tekstova. Ovakvi jezički modeli, uglavnom zasnovani na varijacijama *Transformer* neuralne arhitekture (Vaswani et al. 2017), predstavljaju korak dalje u transfernom učenju u obradi prirodnih jezika i uz pomoć procedura finog podešavanja dovode do primetno boljih performansi na širokom spektru NLP problema, kako u odnosu na statičke vektore značenja reči, tako i u odnosu na ranije neuralne pristupe bazirane na konvolucionim ili rekurzivnim neuralnim mrežama (Radford et al. 2018, 2019; Devlin et al. 2019).

Njihov nedostatak u pogledu primene u jezicima sa ograničenim resursima jeste izuzetna zahtevnost u pogledu potrebne količine podataka za obučavanje. Na primer, BERT (*Bidirectional Encoder Representations from Transformers*) model za engleski jezik (Devlin et al. 2019) je obučen na kombinovanom korpusu sačinjenom od 2,5 milijarde reči sa Vikipedije na engleskom jeziku i 800 miliona reči iz književnog korpusa (Zhu et al. 2015). Poređenja radi, najveći javno dostupni korpus tekstova na srpskom, srpski veb korpus *srWaC* (Ljubešić & Klubička 2014), sadrži oko 555 miliona tokena uključujući interpunkciju, odnosno oko 470 miliona reči. Dodatan problem u primeni ovih modela u jezicima sa ograničenim resursima jeste i velika hardverska zahtevnost koja ih karakteriše u fazi obučavanja, zbog potrebe optimizacije ogromnog broja parametara. Stoga njihovo obučavanje zahteva klastere posebnih tenzorskih procesorskih jedinica (TPU) ili grafičkih procesora (GPU). Na primer, obučavanje BERT modela za engleski jezik, koji, u zavisnosti od varijante, sadrži od 110 do 340 miliona parametara, traje oko 4 dana na TPU klasteru od 4-16 jedinica (Devlin et al. 2019), dok obučavanje GPT modela koji sadrži 110 miliona parametara traje oko mesec dana na klasteru od 8 GPU jedinica (Radford et al. 2018). Iako su navedeni hardverski resursi dostupni i preko *cloud computing* servisa, cena njihovog dugoročnog iznajmljivanja ih i dalje čini nepristupačnim za razvoj rešenja u jezicima sa ograničenim resursima.

Ipak, neki od modela ovog tipa su obučavani na višejezičnim velikim korpusima, poput sadržaja celokupne Vikipedije na preko 100 jezika (Conneau & Lample 2019; Devlin et al. 2019; Sanh et al. 2019; Conneau et al. 2020a), tako da podržavaju i bar neke od jezika sa ograničenim resursima, uključujući srpski. Ovakva višejezičnost modela je delom moguća (Conneau et al. 2020b) zahvaljujući tokenizaciji tekstova na jedinice manje od reči (engl. *subwords*), korišćenjem tehnika kao što su *Byte-Pair Encoding* (BPE) (Gage 1994; Sennrich et al. 2016; Conneau & Lample 2019) i *Wordpiece* segmentacija (Schuster & Nakajima 2012; Devlin et al. 2019), što čini da više jezika deli isti vokabular i kreirani semantički prostor (Pires et al. 2019; Wu & Dredze 2019). U ovoj disertaciji će stoga biti razmotrene mogućnosti transfernog učenja kako na osnovu statičkih vektora značenja reči, tako i pomoću kontekstno osetljivih reprezentacija značenja u vidu neuralnih jezičkih modela.

Za engleski jezik je razvijen i veći broj modela transfernog učenja čiji je cilj pravljenje vektora značenja rečenica (engl. *sentence embeddings*), koji se zatim mogu upotrebljavati u okviru rešavanja raznih semantičkih problema. Ovakvi modeli se takođe tipično obučavaju na jako velikim korpusima neanotiranih tekstova, ali se za te potrebe neretko (dodatno) koriste i obimni anotirani korpusi vezani za određene NLP probleme, i to najčešće za problem zaključivanja na prirodnom jeziku (engl. *natural language inference*). Iako ovi pristupi dostižu dobre performanse na širokom spektru semantičkih problema, oni nisu primenjivi na većinu jezika sa ograničenim resursima upravo zbog obima tekstualnih resursa potrebnih za njihovo obučavanje, bilo anotiranih (Wieting et al. 2016; Conneau et al. 2017; Cer et al. 2018; Subramanian et al. 2018), bilo neanotiranih (Kiros et al. 2015; Hill et al. 2016; Cer et al. 2018; Logeswaran & Lee 2018).

Ilustracije radi, *skip-thoughts* (Kiros et al. 2015), *fastSent* (Hill et al. 2016) i *sent2Vec* (Pagliardini et al. 2018) modeli za engleski su nenadgledano obučeni na književnom korpusu od oko 74 miliona rečenica i skoro milijardu reči (Zhu et al. 2015), dok je za potrebe *quick thoughts* modela (Logeswaran & Lee 2018) korišćen i još veći korpus od 129 miliona rečenica. U poređenju sa tim, pomenuti veb korpus srpskog jezika *srWaC* sadrži duplo manje tokena (uključujući interpunkciju) i tri puta manje rečenica, a pritom je zbog svoje prirode u znatnoj meri sastavljen od tekstova koji nisu napisani standardnim jezikom, što neretko uključuje pravopisne i gramatičke greške. Što se anotiranih resursa tiče, *InferSent* (Conneau et al. 2017) i *Universal Sentence Encoder* (Cer et al. 2018) modeli za engleski su obučeni na *Stanford Natural Language Inference* korpusu koji sadrži 570 hiljada anotiranih primera (Bowman et al. 2015), ali za ogromnu većinu drugih jezika, uključujući srpski, ne postoji bilo kakav resurs ove prirode. Slično tome, (Wieting et al. 2016) su obučili model na korpusu od preko 220 miliona parova parafraza na engleskom (Ganitkevitch et al. 2013). Neki od modela ovog tipa zahtevaju i punu sintaktičku analizu tekstova (Pham et al. 2015; Subramanian et al. 2018), što u mnogim jezicima sa ograničenim resursima nije automatski izvodljivo zbog nepostojanja sintaktičkih parsera. Konačno, iako mnogi od navedenih modela pružaju izuzetno obećavajuće performanse na problemima određivanja semantičke sličnosti i detekcije parafraza, neki od njih su na drugim semantičkim problemima, poput analize sentimenta, lošiji od znatno jednostavnijih pristupa (Wang & Manning 2012).

Podvrsta transfernog učenja koja je takođe dosta popularna u NLP istraživanjima jeste istovremeno učenje za više problema (engl. *multi-task learning*) (Caruana 1993; Collobert & Weston 2008). Ovaj vid učenja se zasniva na ideji da je neke kompleksne probleme lakše rešiti simultano nego zasebno, tj. da model koji se simultano obučava za rešavanje više problema postiže bolje rezultate od modela pravljenih za svaki problem zasebno. Iako je u kontekstu ograničene količine anotiranih podataka ovo privlačan koncept, on u praksi uglavnom zahteva pažljivo balansiranje funkcije greške između individualnih optimizacionih ciljeva vezanih za probleme koji se razmatraju, što komplikuje njegovu širu primenu (Chen et al. 2018). Sem toga, ovakvo učenje je po prirodi vezano za konkretne probleme koji se grupišu u modelu i za konkretne domene u koje spadaju podaci koji se koriste za obučavanje modela, što ga u okviru ciljeva ove disertacije čini manje adekvatnim u odnosu na fleksibilnije metode

transfernog učenja kao što su semantički vektori. Naime, problemi određivanja semantičke sličnosti i analize sentimenta su u ovoj disertaciji tretirani sa stanovišta primera šire porodice semantičkih problema. Sa druge strane, zaključci vezani za eventualni uspeh ili neuspeh primene *multi-task* učenja nad ova dva problema i nad konkretnim kreiranim skupovima podataka se ne bi nužno mogli generalizovati na druge semantičke probleme i metodologiju njihovog rešavanja.

U jezicima sa većom dostupnošću tekstualnih sadržaja, često se koriste i metode polunadgledanog učenja (engl. *semi-supervised learning*) (Søgaard 2013), bilo samostalno (npr. (Täckström & Mcdonald 2011)) bilo u kombinaciji sa drugim pristupima kao što je istovremeno učenje na više zadataka (npr. (Collobert & Weston 2008; Rei 2017)). Ova porodica metoda je od velike koristi u situacijama kada količina anotiranih podataka jeste mala, ali je zato dostupna veća količina neanotiranih primera istog tipa. Međutim, u jezicima sa ograničenim resursima količina svih dostupnih primera je limitirana, te stoga metode polunadgledanog učenja uglavnom nisu primerene. Ipak, gledano u širem smislu, i pomenute metode transfernog učenja korišćenjem semantičkih vektora ili prethodno obučenih jezičkih modela se mogu smatrati varijantom polunadgledanog učenja, jer se zasnivaju na korišćenju većih neanotiranih korpusa tekstova. Razlika u odnosu na polunadgledano učenje u užem smislu jeste što navedeni korpusi tekstova ne moraju biti istog tipa kao primeri podataka za koje postoje anotacije (npr. ako je problem koji se razmatra analiza sentimenta u domenu proizvoda, a semantički vektori su dobijeni na osnovu korpusa opšteg tipa, kao što je veb korpus).

Dodatan aspekt transfernog učenja kome se pridaje pažnja u novijim istraživanjima vezanim za semantičku obradu resursno siromašnih jezika jeste transfer znanja iz resursno bogatijih jezika. U te svrhe se najčešće koriste višejezični i međujezični vektori značenja (engl. *multilingual/cross-lingual embeddings*) čiji je cilj da u jedan zajednički semantički prostor ugrađuju pojmove iz više jezika (Widdows 2004; Ruder et al. 2019). Pored višejezičnih/međujezičnih vektora značenja reči, predloženi su i pristupi za pravljenje višejezičnih vektora značenja rečenica (Singla et al. 2018; Chidambaram et al. 2019; Artetxe & Schwenk 2019; Reimers & Gurevych 2020). Privlačnost višejezičnih/međujezičnih vektora značenja leži u tome što je uz pomoć njih moguće obučiti model za rešavanje nekog konkretnog NLP problema korišćenjem višejezičnog semantičkog prostora i to oslanjajući se na anotirane podatke iz nekog od resursno bogatih jezika (obično engleskog), a zatim takav model koristiti u obradi tekstova na nekom od jezika sa ograničenim resursima. Ipak, radi izgradnje ovog zajedničkog semantičkog prostora uglavnom su potrebni ili paralelni ili uporedivi resursi određene granularnosti – na nivou pojedinačnih reči, rečenica ili celih dokumenata – kakvi nisu dostupni u mnogim jezicima sa ograničenim resursima. Paralelni resursi podrazumevaju prevode istih tekstualnih sadržaja između razmatranih jezika u kojima su reči, rečenice, odnosno dokumenti međusobno poravnani (npr. bilingvalni rečnici ili paralelni transkripti skupštinskih sednica na više jezika), što ih čini teže dostupnim. Uporedivi resursi, sa druge strane, mogu da budu fleksibilnijeg tipa i da se samo odnose na istu temu (npr. opisi iste slike na različitim jezicima). Međutim, najnoviji kontekstno osetljivi modeli zasnovani na *Transformer* arhitekturama pokazuju da se, uz obučavanje na izuzetno velikim količinama tekstualnog sadržaja na više jezika, višejezičnost modela može ostvariti bez eksplicitne paralelnosti/uporedivosti resursa i bez formulisanja procedure koja bi poravnala različite jezike u jedinstven semantički prostor (Conneau & Lample 2019; Devlin et al. 2019; Pires et al. 2019; Wu & Dredze 2019; Conneau et al. 2020b; Karthikeyan et al. 2020).

Iako može delovati da navedene višejezične/međujezične metode rešavaju probleme vezane za semantičku obradu tekstova na jezicima sa ograničenim resursima, njihov kvalitet u praksi može da znatno varira od jednog do drugog NLP problema i od jezika do jezika, pri čemu je tipično lakše u okviru modela izvršiti transfer znanja između strukturno bliskih jezika (Ruder et al. 2019; Conneau et al. 2020b; Karthikeyan et al. 2020). Takođe, rezultati mogu varirati i u zavisnosti od usvojenog protokola obučavanja i evaluacije modela (Kann et al. 2019). Još važniju prepreku predstavlja činjenica da se pomenuti pristupi zapravo svi fokusiraju samo na obučavanje modela. Da bi se

višejezični model kvalitetno evaluirao na određenom jeziku i dalje su neophodni adekvatni anotirani podaci za razmatrani problem na tom jeziku (El-Haj et al. 2015).

Prevođenje postojećih anotiranih resursa sa drugih jezika, bilo mašinsko bilo ručno, nije univerzalno rešenje, jer su neke jezičke pojave specifične za posmatrani jezik i ne javljaju se u resursno bogatim jezicima poput engleskog (npr. upotreba deminutiva i augmentativa u jezicima koji imaju razvijenu derivaciju). Osim toga, potpuno očuvanje značenja tekstova pri prevođenju je često nemoguće, kako zbog jezičkih osobenosti tako i zbog kulturoloških razlika, što u nekim problemima, poput analize sentimenta, može dramatično uticati na odluku o pravilnoj oznaci za određeni tekst (Mohammad et al. 2016). Konačno, ručno prevođenje većih korpusa je dugotrajan i skup postupak, što ga za većinu jezika sa ograničenim resursima čini neizvodljivim. Alternativa u vidu mašinskog prevođenja često nije dostupna za posmatrani jezik, a čak i ako jeste, pitanje je kakvog je kvaliteta, pošto izrada dovoljno kvalitetnog sistema za mašinsko prevođenje sama po sebi zahteva veće količine paralelnih tekstova na datim jezicima.

Širih metodoloških okvira za razvoj NLP rešenja za jezike sa ograničenim resursima nema mnogo u literaturi, i uglavnom su usredsređeni na konkretne NLP probleme, najčešće sintaktičkog tipa. Na primer, (Duong 2017) razmatra razvoj sintaktičkih NLP alata u resursno siromašnim jezicima pomoću mašinskog učenja, i to pre svega transfernog učenja. (King 2015) se takođe bavi transfernim učenjem u sintaktičkim zadacima, poput obeležavanja vrsta reči i parsiranja rečenica, ali u okviru integrisanog sistema za automatsko prepoznavanje jezika sa ograničenim resursima na kome je zadati tekst napisan i pronalaženje dodatnog tekstualnog sadržaja na tom jeziku na internetu. Sa druge strane, (Ruder 2019) razmatra širok spektar neuralnih metoda transfernog učenja u obradi prirodnih jezika, koje su od velike koristi u jezicima sa ograničenim resursima, i demonstrira njihovu efektivnost na različitim NLP problemima i u različitim kontekstima.

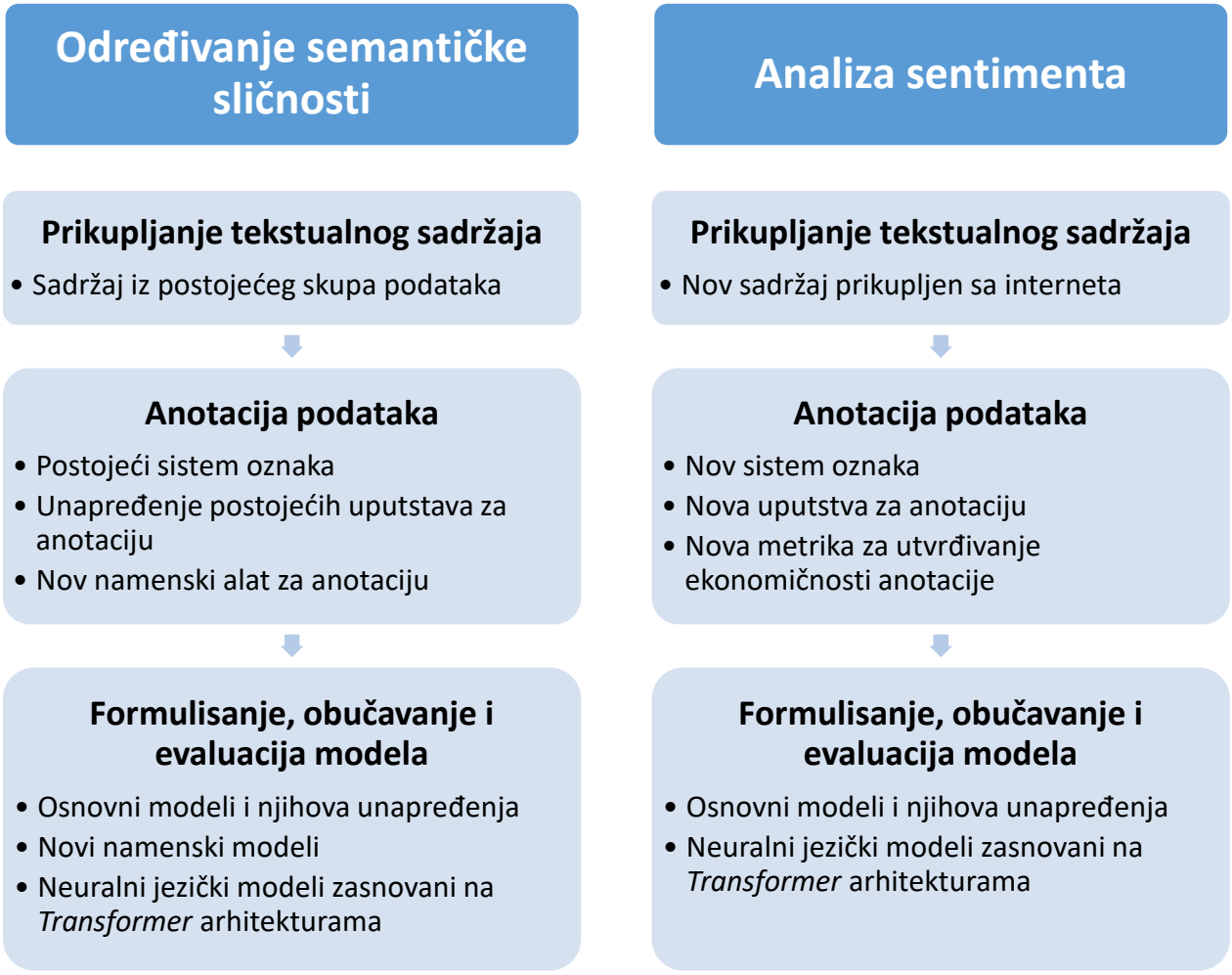
Pregled literature pokazuje da je u dosadašnjim istraživanjima vezanim za statističko rešavanje NLP problema u jezicima sa ograničenim resursima pitanje anotacije podataka često zapostavljano, dok se ogromna pažnja pridaje izradi modela. Iz ovih razloga, u ovoj disertaciji se i faza anotacije podataka detaljno razmatra, dok se u fazi formulisanja i izrade modela evaluira paleta pristupa podeljenih u nekoliko nivoa kompleksnosti, počevši od osnovnih, najšire primenjivih rešenja, preko njihovih unapređenja koja koriste dostupne NLP resurse za konkretan jezik, uključujući statičke vektore značenja reči, do neuralnih kontekstno osetljivih višejezičnih modela zasnovanih na *Transformer* arhitekturama (Conneau & Lample 2019; Devlin et al. 2019; Sanh et al. 2019).

2.3 Uporedni pregled kreiranih rešenja za probleme određivanja semantičke sličnosti i analize sentimenta kratkih tekstova na srpskom jeziku po fazama razvoja

U ovoj disertaciji su kroz rešavanje problema određivanja semantičke sličnosti i analize sentimenta kratkih tekstova prikazane sve faze razvoja statističkih rešenja. Ipak, određene odlike ili varijante svake od faza su jasnije ili detaljnije ilustrovane na jednom od razmatranih problema, zbog različitih objektivnih mogućnosti vezanih za njihovo rešavanje. Stoga su na dijagramu, prikazanom na slici 4, po fazama upoređena kreirana rešenja za oba razmotrena semantička problema.

Sve navedene razlike u rešavanju razmotrenih problema su detaljno diskutovane u odgovarajućim poglavljima disertacije, ali su ovde ukratko navedene radi lakšeg praćenja sadržaja. Kao što se sa dijagrama može videti, u fazi prikupljanja tekstualnog sadržaja za problem određivanja semantičke sličnosti kratkih tekstova bilo je moguće iskoristiti sadržaj iz postojećeg skupa podataka pravljenog

za srodnu problematiku, dok za problem analize sentimenta prethodni resursi tog tipa nisu postojali, te se sadržaj morao prikupljati sa interneta. Razlike u fazi anotacije podataka proističu iz toga što za problem određivanja semantičke sličnosti kratkih tekstova postoji standardizovana međunarodna metodologija anotacije podataka, koja je u ovoj disertaciji i upotrebljena, uz dopune postojećih smernica za anotaciju. Sa druge strane, takav standardizovan pristup ne postoji za označavanje sentimenta, te je pri izradi rešenja za analizu sentimenta razvijen nov sistem označavanja, uključujući oznake i njihove definicije, kao i detaljna uputstva za anotaciju. Radi razmatranja ekonomičnosti upotrebe sastavljenih uputstava za anotaciju razvijena je i nova metrika za utvrđivanje ekonomičnosti anotacije. Nasuprot tome, pri anotaciji semantičke sličnosti razvijen je nov namenski alat za anotaciju. Konačno, u fazi formulisanja i izrade modela, osnovni modeli, njihova unapređenja i neuralni jezički modeli su razmatrani za oba problema. Za problem određivanja semantičke sličnosti je takođe razvijeno nekoliko novih namenskih modela. Kao što je u glavi 5 prikazano, ovi novi modeli, iako bolji od osnovnih varijanti i njihovih unapređenja, ipak primetno zaostaju za najboljim neuralnim pristupima, zbog čega za problem analize sentimenta nisu razvijana namenska rešenja tog tipa.



Slika 4. Uporedni pregled kreiranih rešenja za razmotrene semantičke probleme na srpskom jeziku po fazama razvoja

3 Prikupljanje tekstualnog sadržaja

U ovoj glavi je opisana faza prikupljanja tekstualnih podataka i njihovog čišćenja i formatiranja za dalju upotrebu. Ova faza je najpre diskutovana u okviru problema određivanja semantičke sličnosti, a zatim i u okviru problema analize sentimenta. Na problemu određivanja semantičke sličnosti demonstrirana je upotreba već postojećih resursa u kontekstu pripreme podataka za rešavanje novog, ali srodnog zadatka. S druge strane, problem analize sentimenta služi kao primer pronalaženja adekvatnih izvora tekstualnog sadržaja na internetu, analize i odabira najbolje opcije među dostupnim izvorima, i prikupljanja i formatiranja podataka dobijenih iz odabranih izvora.

3.1 Prikupljanje tekstualnog sadržaja za problem određivanja semantičke sličnosti kratkih tekstova

Odgovarajući anotirani i javno dostupni skupovi podataka za problem određivanja semantičke sličnosti kratkih tekstova, u kojima se koriste granulirane ocene sličnosti na određenoj numeričkoj skali, su i za engleski jezik razvijeni tek relativno skoro, u periodu od 2012. do 2017. godine. Oni su izrađeni u okviru serije tzv. zajedničkih zadataka (engl. *shared tasks*), međunarodnih istraživačkih takmičenja u rešavanju ovog problema, održanih u okviru *SemEval* serije konferencija (Agirre et al. 2012, 2013, 2014, 2015, 2016; Xu et al. 2015; Cer et al. 2017).

Pre toga, najistaknutiji skup podataka koji se koristio za ovu problematiku jeste *Microsoft Research Paraphrase Corpus* (Dolan et al. 2004; Dolan & Brockett 2005). Ovaj korpus od 5801 para rečenica na engleskom jeziku, prikupljenih iz novinskog sadržaja, svojevremeno je sastavljen za problem detekcije parafraza. Detekcija parafraza je zadatak binarne klasifikacije u kome je neophodno razlikovati parove tekstova koji su semantički dovoljno ekvivalentni da se mogu smatrati parafrazama, od onih koji su semantički različiti, te stoga ne stoje u parafraznom odnosu. Po svojoj prirodi, problem detekcije parafraza je veoma srodan problemu određivanja semantičke sličnosti i zapravo predstavlja jednostavniji oblik ovog zadatka.

U trenutku otpočinjanja izrade ove disertacije nije postojao nijedan javno dostupan skup podataka na srpskom jeziku za problem određivanja semantičke sličnosti kratkih tekstova, ali jeste bio dostupan *Srpski korpus parafraza (paraphrase.sr)*[11] (Batanović et al. 2011). Ovaj korpus je autor disertacije napravio u okviru svog završnog rada na master akademskim studijama (Batanović 2011) i on sadrži 1194 para rečenica na srpskom jeziku i latiničnom pismu. Parovi su prikupljeni iz novinskog sadržaja i ručno anotirani u pogledu postojanja ili nepostojanja parafraznog odnosa između uparenih rečenica. Kako je ovaj skup podataka pravljen za potrebe rešavanja problema detekcije parafraza, zaključeno je da on predstavlja pogodan izvor tekstualnog sadržaja i za problem određivanja semantičke sličnosti kratkih tekstova na srpskom, i da se postojeće binarne ocene sličnosti mogu u novom postupku anotacije zameniti granularnijim ocenama. Ovaj postupak je već ranije primenjen na engleskom jeziku, gde je deo *Microsoft Research Paraphrase Corpus* skupa naknadno anotiran granularnijim ocenama semantičke sličnosti (Agirre et al. 2012).

Radi potpunosti, ovde je ukratko predstavljen proces prikupljanja podataka uključenih u *paraphrase.sr* korpus. *Paraphrase.sr* je izrađen po uzoru na *Microsoft Research Paraphrase Corpus*. Oba skupa podataka su konstruisana oslanjanjem na novinarsku konvenciju po kojoj prvih par rečenica novinskog članka predstavljaju njegov sažetak. Korišćenjem ove konvencije i većeg broja

[11] https://vukbatanovic.github.io/paraphrase.sr/

novinskih izvora, moguće je upariti početne rečenice članaka iz različitih izvora koji se odnose na istu temu/događaj, i na taj način dobiti skup sa nezanemarljivim brojem semantički bliskih parova rečenica. Pri izgradnji *paraphrase.sr* korpusa kao izvor podataka korišćen je sajt *vesti.rs*, agregator vesti svih većih medijskih kuća u Srbiji, uključujući televizijske stanice, štampane medije, kao i internet portale. Tokom prikupljanja podataka u obzir su uzimane samo najvažnije vesti za svaki dan, jer je za njih najverovatnije da ih je prenelo više medijskih izvora. Pored toga, na ovaj način je izbegnuto da prikupljeni skup podataka bude usko fokusiran na neku određenu tematiku. Takođe, da bi se održala tematska raznovrsnost, za svaku vest je odabran maksimalno po jedan par rečenica za uključivanje u skup. Svi mogući kandidat-parovi su rangirani na osnovu seta heuristika za procenu njihovog kvaliteta, gde je cilj bilo prikupljanje što većeg broja primera u kojima su parafraze zasnovane na dubljoj semantičkoj sličnosti, a ne na prostom leksičkom poklapanju. Prvobitna, binarna anotacija je pokazala da *paraphrase.sr* sadrži 553 para rečenica (odnosno 46,31%) za koje je u binarnoj kategorizaciji procenjeno da su semantički ekvivalentni tj. parafraze, i 641 par rečenica (odnosno 53,69%) za koje je u binarnoj kategorizaciji procenjeno da su semantički različiti, tj. da nisu parafraze.

Tekstovi iz skupa *paraphrase.sr* su ručno očišćeni od slovnih grešaka i nedostajućih dijakritičkih znakova. U sklopu toga, dva para tekstova su uklonjena iz skupa – jedan koji je predstavljao duplikat, i drugi koji je sadržao tekst duži od jedne rečenice. Stoga je finalni skup podataka nakon ove faze sadržao 1192 para rečenica. Ovaj korpus je označen kao *Srpski novinski STS korpus* ili *STS.news.sr*.

U tabeli 1 prikazani su osnovni statistički podaci o veličini skupa prikupljenih podataka za određivanje semantičke sličnosti kratkih tekstova. Za tokenizaciju u ovom pregledu je korišćeno razdvajanje na osnovu blanko znakova (engl. *whitespace tokenization*).

Tabela 1. Osnovni statistički prikaz prikupljenih podataka za određivanje semantičke sličnosti kratkih tekstova na srpskom jeziku

Skup podataka	Broj parova tekstova	Broj tokena	Prosečna dužina teksta u tokenima	Veličina vokabulara
STS.news.sr	1192	64K	~27	17K

3.2 Prikupljanje tekstualnog sadržaja za problem analize sentimenta kratkih tekstova

U okviru problema analize sentimenta vrlo često se kao referentni koriste skupovi podataka iz domena filmova i filmskih recenzija (Pang et al. 2002; Pang & Lee 2004, 2005; Maas et al. 2011; Socher et al. 2013). Razlog za to jeste što je ovaj domen među najtežima za problem analize sentimenta (Turney 2002), jer svaki tekst o filmu uključuje dva nezavisna sloja značenja – jedan koji se odnosi na sadržaj filma, tj. na zaplet i likove u filmu, i drugi koji se odnosi na sam film kao delo. Zbog toga je rešeno da je i za analizu sentimenta kratkih tekstova na srpskom jeziku najbolje koristiti podatke vezane za ovaj domen.

Među najpoznatijim i najčešće korišćenim skupovima kratkih tekstova kreiranim za problem analize sentimenta na engleskom jeziku jesu dva skupa koja su predstavili Pang i Lee (Pang & Lee 2004, 2005). Jedan od njih jeste skup za problem određivanja polarnosti teksta (Pang & Lee 2005), gde je cilj razlikovati pozitivne od negativnih tekstova. Ovaj balansirani skup sadrži 10662 primera, tj. 5331

pozitivan primer i isto toliko negativnih. Kao izvor podataka za ovaj skup korišćen je poznati agregator filmskih recenzija i kritika *Rotten Tomatoes* (www.rottentomatoes.com). Na tom sajtu uz svaku recenziju stoji oznaka da li je globalni utisak recenzenta pozitivan (engl. *fresh*) ili negativan (engl. *rotten*), kao i kratak tekstualni isečak, obično dužine jedne rečenice, koji predstavlja sažetak stava recenzenta. Uparivanjem tekstualnih isečaka sa oznakama o pozitivnosti/negativnosti, Pang i Lee su mogli da automatski izgrade anotirani skup podataka. Tekstovi iz ovog skupa su naknadno ručno svrstani u pet klasa na osnovu jačine sentimenta, zarad dobijanja granularnijih ocena sentimenta na skali od najnegativnijeg do najpozitivnijeg (Socher et al. 2013).

Drugi referentni skup kratkih tekstova koji su sakupili navedeni autori jeste skup za problem određivanja subjektivnosti kratkih tekstova (Pang & Lee 2004), gde je cilj razlikovati tekstove koji su subjektivni tj. koji izražavaju određeni sentiment, od objektivnih, koji ne izražavaju bilo kakav sentiment. Ovaj balansirani skup obuhvata 5000 objektivnih i 5000 subjektivnih kratkih tekstova i takođe je automatski izgrađen. Naime, tekstualni izvodi sa sajta *Rotten Tomatoes* su korišćeni kao primeri subjektivnih tekstova, dok su sa sajta *Internet Movie Database* (www.imdb.com) uzeti sinopsisi filmova kao primeri objektivnih tekstova iz istog domena.

U trenutku otpočinjanja izrade ove disertacije nije postojao nijedan kvalitetan javno dostupan skup kratkih tekstova na srpskom jeziku anotiranih u pogledu njihovog sentimenta. Naime, jedini prethodni javno dostupni korpus ovog tipa na srpskom jeziku jeste onaj koji su predstavili (Mozetič et al. 2016) u okviru rada gde su anotirani i analizirani skupovi tvitova na 13 evropskih jezika, uključujući srpski. Tvitovi su na osnovu sentimenta svrstavani u tri klase – pozitivnu, negativnu i neutralnu. Međutim, kreirani skup tvitova na srpskom jeziku ima dosta nisku međusobnu saglasnost anotatora (Kripendorfov α koeficijent je jednak 0,329). Osim toga, prikupljeni tvitovi ne pripadaju svi istom domenu/tematici, što može negativno uticati na ponašanje statističkih klasifikatora obučavanih na takvim podacima i navesti ih da klasifikaciju vrše preko informacija nevezanih za sam sentiment, poput tematike tekstova (Batanović et al. 2016). Iz navedenih razloga je procenjeno da se ovaj skup podataka ne može koristiti u izradi rešenja za analizu sentimenta kratkih tekstova na srpskom jeziku.

Postoji još nekoliko ranijih radova u kojima je razmatrana analiza sentimenta tekstova na srpskom. (Mladenović et al. 2015; Mladenović 2016) su evaluirali pristup zasnovan na kombinovanju leksikona sentimenata reči sa statističkim metodama koristeći sopstvene korpuse tekstova koji takođe ne pripadaju svi istom domenu, a nisu ni javno dostupni. (Batanović et al. 2016) su predstavili korpus dokumenata *SerbMR* – javno dostupan skup filmskih recenzija automatski raspodeljenih u pozitivnu, neutralnu i negativnu klasu na osnovu numeričkih ocena zadatih od strane recenzenata. Ovaj korpus je upotrebljen za obučavanje i evaluaciju nekoliko klasifikatora zasnovanih na mašinskom učenju, ali je zbog dužine prikupljenih dokumenata neodgovarajući za potrebe ove disertacije. (Grljević 2016) je razmatrala analizu sentimenta u komentarima studenata o nastavnom osoblju univerziteta u Srbiji i okviru toga je predstavila korpus anotiranih kratkih tekstova na srpskom, ali on nije javno dostupan. Ljajić i Maraovac (Ljajić 2019; Ljajić & Marovac 2019) su istraživali različite načine za obradu negacije u kontekstu analize sentimenta tvitova na srpskom jeziku, i u te svrhe su sastavili i anotirali sopstveni korpus tvitova podeljenih u negativnu, neutralnu i pozitivnu klasu. Međutim, ni ovaj skup podataka nije objavljen.

U svetlu ograničene dostupnosti i/ili neodgovarajućeg tipa ili kvaliteta prethodno razvijenih skupova podataka, odlučeno je da je za potrebe ove disertacije neophodno napraviti nov, javno dostupan referentan skup kratkih tekstova na srpskom jeziku koji bi bio anotiran u pogledu sentimenta. Kako na srpskom ne postoji sajt kao što je *Rotten Tomatoes*, sa koga bi se mogli prikupiti i na osnovu postojećih indikatora automatski označiti podaci, bilo je neophodno pronaći adekvatan izvor kratkih tekstova na srpskom jeziku iz domena filmova, pri čemu bi se podaci nakon prikupljanja ručno anotirali. Zaključeno je da se za ove potrebe mogu potencijalno koristiti dva tipa izvora sadržaja. Prvi

tip čine komentari posetilaca na recenzirane filmove, pošto na srpskom postoji nekoliko sajtova sa recenzijama filmova koji omogućavaju posetiocima da ostave komentare. Drugi tip izvora jeste društvena mreža Tviter, tj. prikupljanje tvitova na osnovu određenog broja odabranih *hashtag*-ova.

Procenjeno je da korišćenje komentara posetilaca ima višestruke prednosti u odnosu na prikupljanje podataka sa Tvitera. Najpre, skupljanjem svih komentara koje su posetioci ostavili osigurava se reprezentativnost skupa podataka, dok pri korišćenju Tvitera kao izvora uvek ostaje mogućnost da su neki relevantni tvitovi greškom izostavljeni iz skupa, tako što nije razmotren neki *hashtag* ili ključna reč, i da je time poremećena prirodna distribucija komentara u pogledu njihovog sentimenta. Pored toga, iako se i u komentarima posetilaca i u tvitovima često koriste neformalni i nestandardni jezički izrazi, jezik korišćen na Tviteru u još većoj meri odskače od jezičkih normi i ispoljava još veći broj nestandardnih osobenosti (*hashtag*-ovi, nestandardna skraćivanja raznih reči i izraza, itd.), zbog ograničenije dužine poruka. Kako za srpski jezik kvalitetni javno dostupni skupovi podataka za analizu sentimenta kratkih tekstova ne postoje, procenjeno je da je primerenije, za početak, usmeriti se na donekle jednostavniji jezik korišćen u komentarima. Konačno, zbog pravnih propisa koji se odnose na privatnost korisnika Tvitera i uslove pristupanja podacima sa te društvene mreže[12], bilo koji skup prikupljenih tvitova nije moguće distribuirati u izvornom obliku, tj. u obliku tekstova, već samo u vidu jedinstvenih identifikatora tih tvitova. Kako korisnici Tvitera imaju potpunu slobodu da obrišu svoje ranije poruke ili čak i same naloge, to znači da ne bi bilo moguće ubuduće garantovati integritet i potpunost anotiranog skupa podataka koji bi bio zasnovan na sadržaju prikupljenom sa Tvitera.

Pretraživanjem interneta pronađeno je desetak veb sajtova na srpskom koji se bave filmskom tematikom. Međutim, ne nude svi među njima mogućnost posetiocima da ostave komentare. Čak i u slučajevima gde ta mogućnost postoji, aktivnost posetilaca na sajtu je uglavnom vrlo skromna. U okviru podskupa sajtova sa većom aktivnošću korisnika, primetno je da su na nekim sajtovima (npr. 2kokice.com) komentari posetilaca više usmereni na čisto tehnička pitanja vezana za pronalaženje i gledanje filmova nego na same filmove, čime izlaze iz odabranog domena. Uzimajući u obzir navedena pitanja, kao najbolji izvor istakao se sajt *Kakav je film?* (kakavfilm.com), najveći sajt na srpskom jeziku sa recenzijama filmova, koji sadrži i najveći broj komentara posetilaca. Pri tome, komentari sa ovog sajta jesu domenski relevantni i, za razliku od nekih drugih sajtova usmerenih samo na filmove određenog žanra, takođe su i žanrovski raznovrsni.

Radi prikupljanja komentara sa sajta *Kakav je film?* napisan je namenski veb indekser i parser (engl. *scraper*) pomoću koga je najpre sastavljen spisak svih strana sa filmskim recenzijama, a zatim su sa svake takve strane ekstraktovani komentari posetilaca. Ovaj modul je implementiran u programskom jeziku Java, korišćenjem biblioteka *JSoup*[13] i *HTMLUnit*[14]. *JSoup* biblioteka omogućava parsiranje HTML koda i ekstrakciju željenog sadržaja iz njega kretanjem kroz DOM (*Document Object Model*) stablo. Za razliku od *JSoup* biblioteke, koja je namenjena samo obradi statičkog sadržaja, *HTMLUnit* biblioteka omogućava sličnu obradu i za HTML kod dinamički generisan pomoću JavaScript-a.

U trenutku prikupljanja podataka, na sajtu *Kakav je film?* korišćen je *IntenseDebate*[15] modul za ostavljanje i prikazivanje komentara (u međuvremenu je sistem upravljanjima komentarima posetilaca na ovom sajtu izmenjen). *IntenseDebate* je jedno od popularnih rešenja za upravljanje komentarima kod kojeg se komentari ne smeštaju u okviru samog sajta/strane, već na serverima firme *IntenseDebate* odakle se dinamički učitavaju. Komentari su u *IntenseDebate* modulu prikazani u vidu skupa stabala odgovora. Korenovi stabala su komentari najvišeg nivoa. Prvi odgovor na neki

[12] http://twitter.com/en/tos
[13] http://jsoup.org/
[14] http://htmlunit.sourceforge.net/
[15] http://www.intensedebate.com/

komentar najvišeg nivoa znači prelazak u nov, niži nivo stabla tj. povećava dubinu stabla, a svi naredni odgovori na taj isti komentar najvišeg nivoa znače povećavanje širine nižeg nivoa. Kako je i na odgovore moguće odgovarati, stablo može biti proizvoljne dubine a svaki nivo može biti proizvoljne širine.

Prilikom prikupljanja podataka svakom komentaru je dodeljen jedinstveni identifikator. Osnovna forma tog identifikatora, koja se koristi za komentare najvišeg nivoa, jeste oblika *ABC-D*, gde je *ABC* kod koji identifikuje stranu tj. recenziju sa koje je komentar prikupljen, a *D* je redni broj posmatranog komentara najvišeg nivoa. Ukoliko postoje odgovori na zadati komentar, tada se njihovi identifikatori dobijaju tako što se na identifikator njima nadređenog komentara najvišeg nivoa dodaju nastavci. Za svako spuštanje na niži nivo stabla na identifikator nadređenog komentara se dodaje znak „–" iza koga se navodi redni broj posmatranog komentara u okviru datog nižeg nivoa stabla. Na primer, ako je kod strane *123* i ako drugi komentar najvišeg nivoa ima dva odgovora, tada će identifikator drugog odgovora na drugi komentar biti *123-2-2*. Ovakav sistem identifikacije omogućava da se lako uoči kom podstablu komentarisanja neki komentar pripada. Zbog toga, iako to nije bio fokus u ovoj disertaciji, kreirani skup komentara može kasnije poslužiti i za analizu diskursa, jer su informacije o redosledu komentara i njihovoj međusobnoj povezanosti u potpunosti sačuvane.

Koristeći opisanu proceduru sa sajta *Kakav je film?* prikupljeno je ukupno 4660 komentara. Uvidom u prikupljene komentare uočeno je da je nemali broj njih prilično dugačak te da stoga izlaze iz okvira kratkih tekstova. Stoga je odlučeno da se u daljem radu razmatraju samo kratki komentari, pod kojima se podrazumevaju oni kod kojih je broj tokena manji ili jednak 50. Alternativna definicija kratkih komentara koja bi uzimala u obzir samo komentare dužine jedne rečenice nije korišćena, jer je tekst prikupljenih komentara često vrlo slobodno formatiran u smislu interpunkcije, te bi pravilno automatsko određivanje granica rečenica bilo gotovo nemoguće. Nakon uklanjanja dugačkih komentara, skup je brojao 3563 primera.

Pre samog anotiranja skup kratkih komentara je dodatno ručno pročišćen. Iz razmatranja su izbačena dupliranja komentara (uključujući ponavljanja zbog grešaka u kucanju), kao i svi komentari koji ne sadrže nijednu reč na srpskom – tu se uglavnom radilo ili o prostim numeričkim ocenama ili emotikonima (za čiju interpretaciju nije ni potreban sistem za analizu sentimenta), ili o komentarima napisanim isključivo na engleskom (za čiju analizu se mogu koristiti sistemi pravljeni za engleski jezik). Konačno, izbačen je i vrlo mali broj izuzetno vulgarnih komentara, iako su, generalno govoreći, komentari koji sadrže neke vulgarne izraze zadržani. Procesom ručnog pročišćavanja eliminisana su ukupno 73 komentara, tako da je za fazu anotacije pripremljeno 3490 kratkih komentara. Ovaj korpus je nazvan *SentiComments.SR*.

U toku faze anotacije podataka za problem analize sentimenta zaključeno je da je radi verifikacije kvaliteta anotacije neophodno prikupiti još dva manja skupa kratkih tekstova, od kojih bi jedan takođe bio iz domena filmova, a drugi iz nekog drugačijeg domena, ali slične kompleksnosti. Kao pogodan drugi domen usvojen je domen književnosti, pri čemu su kao primeri podataka prikupljeni komentari posetilaca na recenzije knjiga sa sajta *Happy Novi Sad* (happynovisad.com), jednog od većih portala na srpskom jeziku o kulturi. Komentari posetilaca na recenzije filmova sa istog sajta su iskorišćeni za formiranje manjeg skupa kratkih tekstova iz domena filmova, ali pošto njih nije bilo puno, pridodati su im i komentari posetilaca sa sajta *Gledaj me* (gledajme.rs – sajt više nije aktivan), koji je sadržao top liste i preporuke filmova. Sa sajta *Gledaj me* je prikupljeno 330 komentara, dok je filmskih komentara sa sajta *Happy Novi Sad* bilo 134, te verifikacioni skup komentara iz domena filmova, nazvan *SentiComments.SR.verif.movies*, ukupno sadrži 464 primera. Verifikacioni skup komentara iz domena književnosti, izgrađen na osnovu komentara sa sajta *Happy Novi Sad* i nazvan *SentiComments.SR.verif.books*, sadrži 173 komentara.

U tabeli 2 prikazani su osnovni statistički podaci o veličini prikupljenih skupova podataka za analizu sentimenta kratkih tekstova. Za tokenizaciju u ovom pregledu je korišćeno razdvajanje na osnovu blanko znakova. Glavni skup podataka, *SentiComments.SR*, prikazan je u dve varijante – jednoj koja sadrži izvorne tekstove komentara (označena sa *SentiComments.SR.orig*), i drugoj u kojoj su tekstovi komentara ručno korigovani u smislu slovnih ispravki i unosa nedostajućih dijakritičkih znakova (označena sa *SentiComments.SR.corr*).

Tabela 2. Osnovni statistički prikaz prikupljenih podataka za analizu sentimenta kratkih tekstova na srpskom jeziku

Skup podataka	Broj tekstova	Broj tokena	Prosečna dužina teksta u tokenima	Veličina vokabulara
SentiComments.SR.orig	3490	63K	~18	19K
SentiComments.SR.corr	3490	65K	~19	17K
SentiComments.SR.verif.movies	464	8K	~16	3K
SentiComments.SR.verif.books	173	2K	~13	1K

4 Anotacija podataka

U ovoj glavi je opisana faza anotacije prikupljenih tekstualnih podataka. Najpre je prikazana anotacija podataka za problem određivanja semantičke sličnosti, a nakon toga anotacija podataka za problem analize sentimenta. U oba slučaja, anotacija je sprovedena u potpunosti ručno, zbog nepostojanja pogodnih oznaka koje bi se mogle automatski preslikati u ocene semantičke sličnosti, odnosno sentimenta. Kroz anotaciju semantičke sličnosti je prikazano prilagođavanje i poboljšanje postojećeg sistema oznaka i uputstava za anotaciju, dok je na anotaciji sentimenta demonstriran razvoj novog sistema oznaka i detaljnih uputstava, kao i metrika na osnovu koje se može utvrditi ekonomičnost njihove primene. Pored toga, anotacija semantičke sličnosti je primer jasnog razdvajanja svih ranije navedenih koraka u fazi anotacije podataka, dok anotacija sentimenta ilustruje kako su u situacijama razvoja novog sistema označavanja koraci u anotaciji često isprepleteni.

4.1 Anotacija podataka za problem određivanja semantičke sličnosti kratkih tekstova

Do 2012. godine anotirani podaci za problem određivanja semantičke sličnosti kratkih tekstova su bili vrlo ograničeni čak i za engleski jezik. Kao najistaknutiji skup podataka za ovu problematiku koristio se već pomenuti *Microsoft Research Paraphrase Corpus* (Dolan et al. 2004; Dolan & Brockett 2005), koji sadrži 5801 parova rečenica kojima su pridružene binarne ocene sličnosti (1 – jesu parafraze, 0 – nisu parafraze). Ocene sličnosti u ovom skupu su paralelno zadala dva anotatora, pri čemu je u slučajevima njihovog neslaganja treći anotator donosio finalnu odluku. Prosečan procenat saglasnosti anotatora na ovom skupu je bio 83%. Uprkos ograničenosti na binarne ocene sličnosti, ovaj skup podataka se zbog svoje veličine bio nametnuo kao dominantan u istraživanjima semantičke sličnosti kratkih tekstova. Naime, drugi skupovi podataka, kao što su korpus od 65 parova rečenica koji su predstavili (Li et al. 2006), ili njegovo proširenje na 131 par rečenica u vidu STSS-131 skupa (O'Shea et al. 2013), jesu bili obeleženi granularnijim numeričkim ocenama semantičke sličnosti, ali je izuzetno ograničen broj primera u tim skupovima bio prepreka za njihovu primenu u algoritmima mašinskog učenja.

U periodu od 2012. do 2017. godine u sklopu *SemEval* konferencija održavana su godišnja takmičenja istraživačkih timova u rešavanju problema određivanja semantičke sličnosti kratkih tekstova (Agirre et al. 2012, 2013, 2014, 2015, 2016; Cer et al. 2017). U okviru pripreme ovih takmičenja napravljeni su novi anotirani skupovi podataka, korišćenjem standardizovane metodologije za anotaciju i većeg broja raznorodnih izvora podataka, uključujući novinske vesti, rečničke definicije, itd. Ovi skupovi podataka su anotirani preko *Amazon Mechanical Turk crowdsourcing* servisa, korišćenjem granularnijih ocena sličnosti na Likertovoj skali od 0 do 5. Ocene na ovoj skali su definisane na sledeći način:

- Ocena **0** – upareni tekstovi se odnose na različite teme;
- Ocena **1** – upareni tekstovi nisu ekvivalentni, ali se odnose na istu temu;
- Ocena **2** – upareni tekstovi nisu ekvivalentni, ali dele neke detalje;
- Ocena **3** – upareni tekstovi su ekvivalentni grubo govoreći, ali se neke važne informacije razlikuju ili nedostaju u jednom od njih;
- Ocena **4** – upareni tekstovi su uglavnom ekvivalentni, ali se neki nevažni detalji razlikuju;
- Ocena **5** – upareni tekstovi su potpuno ekvivalentni jer znače istu stvar.

Od 2013. godine, ovom skupu definicija za svaku oznaku je pridružen po jedan reprezentativan primer za svaku ocenu, kao i kratak set uvodnih uputstava koja objašnjavaju problem semantičke sličnosti i anotacioni zadatak. U okviru toga su date i opšte smernice koje anotatorima govore na koje pojave treba obraćati pažnju prilikom obeležavanja tekstova, a koje pojave se mogu ignorisati.

Korišćenjem ove metodologije, svake godine je pravljen nov skup anotiranih podataka na engleskom jeziku, pri čemu se broj primera kretao od nekoliko stotina do nekoliko hiljada po godini. Pritom, ista metodologija je primenjena i za izradu skupova podataka na španskom (Agirre et al. 2014, 2015; Cer et al. 2017) i arapskom jeziku (Cer et al. 2017), kao i za izradu višejezičnih skupova na engleskom i španskom (Agirre et al. 2016; Cer et al. 2017) i na engleskom i arapskom (Cer et al. 2017), čime je dokazana njena široka primenjivost. Zahvaljujući velikim količinama novih anotiranih podataka, kao i nedostacima ranije razvijenih resursa ovog tipa, metodologija anotacije semantičke sličnosti kratkih tekstova razvijena u okviru *SemEval* konferencija se nametnula kao standard u ovoj oblasti. Ona je stoga korišćena kao polazna tačka i u nezavisnim anotacionim projektima vezanim za rešavanje ovog problema na drugim jezicima, kao što su portugalski (Fonseca et al. 2016), japanski i kineski (Hayashi & Luo 2016) i hindi (Agarwal et al. 2017).

Ostatak ovog poglavlja je podeljen po koracima u fazi anotacije podataka. Najpre je prikazan odabir oznaka, kao i proces formulisanja uputstava za anotaciju semantičke sličnosti kratkih tekstova i njihov sadržaj. Zatim je opisano obučavanje anotatora i sprovođenje anotacije. Na kraju je izneta analiza konzistentnosti sprovedene anotacije, a anotirani *STS.news.sr* korpus je statistički prikazan i upoređen sa ranije predstavljenim skupovima podataka ovog tipa na drugim jezicima.

4.1.1 Odabir oznaka i formulisanje uputstava za anotaciju semantičke sličnosti kratkih tekstova

Korišćenje postojećeg standardizovanog seta oznaka i anotacione metodologije omogućava uporedivost kreiranih resursa i štedi vreme u početnim koracima faze anotacije. Stoga je odlučeno da se pristup anotaciji semantičke sličnosti kratkih tekstova razvijen u okviru *SemEval* konferencija primeni i za izradu anotiranog skupa podataka za ovaj problem na srpskom jeziku. Iako su podaci sa *SemEval* takmičenja anotirani putem *crowdsourcing* servisa, oznake i uputstva korišćeni u ovom procesu su podjednako primenjivi i na klasičnu organizaciju anotacije koja je, kao što je već objašnjeno, obično jedina opcija za izvođenje anotacije u manjim prirodnim jezicima. U skupovima podataka sa *SemEval* takmičenja finalne ocene sličnosti parova tekstova su dobijane uprosečavanjem individualnih ocena petoro anotatora koji su zasebno anotirali svaki par. Stoga je rešeno da se isti pristup i broj anotatora koristi i pri anotaciji podataka na srpskom jeziku. Angažovani anotatori nisu imali ranijeg iskustva sa ovim tipom anotacionih zadataka.

Radi sprovođenja anotacije na srpskom jeziku najpre su prevedena uputstva za anotaciju sa engleskog na srpski jezik. U toku analize i inicijalnih konsultacija sa anotatorima zaključeno je da je skup oznaka i njihovih definicija jasan i intuitivan, kao i da su opšta uputstva razumljiva i pregledna, čime je završen prvi korak u fazi anotacije – korak odabira oznaka. S druge strane, pokazalo se da su primeri koji su korišćeni u anotaciji podataka na engleskom jeziku nedovoljno jasni i da mogu dovesti do nekonzistentnih anotacionih odluka. Ovo se naročito odnosilo na primere navedene za ocene **2**, **3** i **4**. Stoga je odlučeno da se svi postojeći primeri zamene novima, kao i da se umesto po jednog primera para tekstova za svaku ocenu, u uputstva uključe po tri primera za svaku ocenu.

Da bi se minimizovala subjektivnost i pristrasnost u izboru takvih tekstova, pogodni primeri su pronađeni među parovima rečenica na engleskom anotiranih u okviru *SemEval* konferencija, i zatim profesionalno prevedeni na srpski. Pošto *STS.news.sr* skup koji je trebalo anotirati pripada domenu

vesti, odabrani su primeri koji su takođe iz ovog domena, izabrani među parovima iz *MSRPar* korpusa iz 2012. godine (Agirre et al. 2012) i *Headlines* (*HDL)* korpusa iz perioda 2013-2016. godine (Agirre et al. 2013, 2014, 2015, 2016). Pri izboru primera uzimani su u obzir samo parovi čije su ocene sličnosti bile celobrojne vrednosti, pošto to obično znači da su svi anotatori dodelili istu ocenu takvim parovima, što je dobar pokazatelj njihove nedvosmislenosti. Finalni odabir primera je obavljen uz konsultacije sa anotatorima, da bi se osigurala reprezentativnost izabranih tekstova.

Ovako modifikovana uputstva za anotaciju su data anotatorima u verziji na srpskom jeziku, ali su sačuvani i izvorni tekstovi primera na engleskom radi njihovog potencijalnog korišćenja u budućim anotacionim projektima ovog tipa na tom jeziku. Obe verzije su javno dostupne na *GitHub* repozitorijumu[16]. Time je završen drugi korak u fazi anotacije – formulisanje uputstava za anotaciju. U nastavku je prikazan tekst uputstava na srpskom jeziku.

4.1.1.1 Uputstva za anotaciju semantičke sličnosti kratkih tekstova

Uputstva – Uporedite značenja dva iskaza

Dva iskaza mogu da znače istu stvar čak i ako koriste vrlo drugačije reči ili fraze. Nasuprot tome, dva iskaza koja su na prvi pogled veoma slična u pogledu izbora reči, fraza i ukupne strukture mogu imati veoma različita značenja.

Vaš zadatak je da uporedite dva iskaza i odredite vrstu odnosa koji postoje između njihovih značenja/poruka (tj. između onoga šta oni kažu o/na šta se odnose u svetu). Da biste uspešno sproveli ovaj zadatak zamislite situacije koje iskazi opisuju i kontrastirajte ono što je sadržano u prvom sa onim što je sadržano u drugom iskazu.

Da li se iskazi odnose na istu osobu, radnju, događaj, ideju ili stvar? Da li su iskazi slični ali se razlikuju u pogledu bilo manjih bilo većih detalja?

Saveti:

- Budite precizni u ocenjivanju i pokušajte da izbegnete da se previše oslanjate na samo jednu od kategorija (npr. nemojte da samo većinu parova obeležite kao „uglavnom ekvivalentne" ili „grubo govoreći ekvivalentne").
- Vodite računa o suptilnim razlikama između parova koje imaju važan uticaj na to šta se opisuje ili govori.
- Ignorišite gramatičke greške i loše sročene izraze dokle god ti faktori ne zamagljuju značenje koje iskaz treba da prenese.

Ocene i primeri

(5) – Dve rečenice su potpuno ekvivalentne jer znače istu stvar.

- o Buš planira da se u sredu sastane sa izraelskim premijerom Arijelom Šaronom i novim palestinskim premijerom Mahmudom Abasom u jordanskoj luci Akaba.
- o Naredne srede, Buš će se sastati sa izraelskim premijerom Arijelom Šaronom i novim palestinskim liderom Mahmudom Abasom u Akabi, Jordanu.

[16] Uputstva na srpskom jeziku: https://vukbatanovic.github.io/STS.news.sr/Annotation%20instructions%20-%20Serbian
Uputstva na engleskom jeziku: https://vukbatanovic.github.io/STS.news.sr/Annotation%20instructions%20-%20English

- Kažu da će zarade u predstojećem drugom kvartalu biti naročito važne u pružanju smernica investitorima.
  - Kažu da će izveštaji o zaradama u drugom kvartalu biti ključni u davanju tih smernica investitorima.

- Kasnije je saznao da je incident izazvalo Konkordovo probijanje zvučnog zida.
  - Kasnije je otkrio da je zabrinjavajući incident izazvalo Konkordovo snažno probijanje zvučnog zida.

(4) – Dve rečenice su uglavnom ekvivalentne, ali se neki nevažni detalji razlikuju.

- Kratka, problematična košarkaška karijera Rikija Klemonsa u Misuriju je završena.
  - Misuri je izbacio Rikija Klemonsa, okončavajući njegovu problematičnu karijeru u tom timu.

- Nekoliko država i savezna vlada kasnije su usvojile slične ili strože zabrane.
  - Sledeći primer Kalifornije, nekoliko država i savezna vlada usvojile su slične ili oštrije zabrane.

- „Na osnovu našeg prvobitnog izveštaja, čini se da je ovo bio školski primer sletanja imajući u vidu date uslove", rekao je Berk.
  - Berk je rekao: „Bio je to školski primer sletanja imajući u vidu date uslove."

(3) – Dve rečenice su ekvivalentne grubo govoreći, ali se neke važne informacije razlikuju/nedostaju u jednoj od rečenica.

- Po treći put tokom četiri godine, šumski požari su primorali park na zatvaranje.
  - Po treći put za četiri godine, zbog šumskih požara zatvoren je nacionalni park Mesa verde, jedini park u zemlji posvećen drevnim ruševinama.

- Parovi, američka verzija britanske hit komedije, dobiće udarni termin četvrtkom u 9:30 uveče.
  - Parovi, američka verzija britanske hit komedije, biće prikazivana četvrtkom, u paru sa Prijateljima.

- Selenski se spustio niz zid i iskoristio dušek da se prebaci preko bodljikave žice.
  - Selenski je iskoristio dušek da se uspenje uz tri metra visoku bodljikavu žicu, rekao je Fiši.

(2) – Dve rečenice nisu ekvivalentne, ali dele neke detalje.

- Kompanija nije iznela detalje o troškovima zamene i popravki.
  - Ali zvaničnici kompanije očekuju da troškovi radova na zameni dostignu milione dolara.

- Prema izvorima, u preliminarnim izveštajima navodi se da ti muškarci nisu viđeni zajedno na aerodromu.
  - Ovi muškarci su imali pakistanske pasoše i navodno su viđeni zajedno na aerodromu ranije iste večeri, rekli su izvori iz policije.

- Kineski ledolomac je skrenuo sa kursa prema sumnjivim objektima.
  - Kineski patrolni avion je pronašao „sumnjive objekte".

(1) – Dve rečenice nisu ekvivalentne, ali se odnose na istu temu.

- Sirija je apelovala na izbeglice da se vrate.
  - Krhko primirje je stupilo na snagu u Siriji.

- Islamista Morsi je pobedio na egipatskim predsedničkim izborima.
  - Al-Šater će se kandidovati na predsedničkim izborima u Egiptu.

- Japanski premijer je pozvao na bržu obradu otpada od cunamija.
  - Japan obeležava godinu dana od nesreće prouzrokovane zemljotresom i cunamijem.

(0) – Dve rečenice se odnose na različite teme.

- Potres umerene jačine zabeležen je u centralnom Sredozemlju.
  - Kandidat umerene struje pobedio je na predsedničkim izborima u Iranu.

- Jedna osoba je ubijena u verskim sukobima u Libanu.
  - 180 ljudi je poginulo u zemljotresu u Iranu.

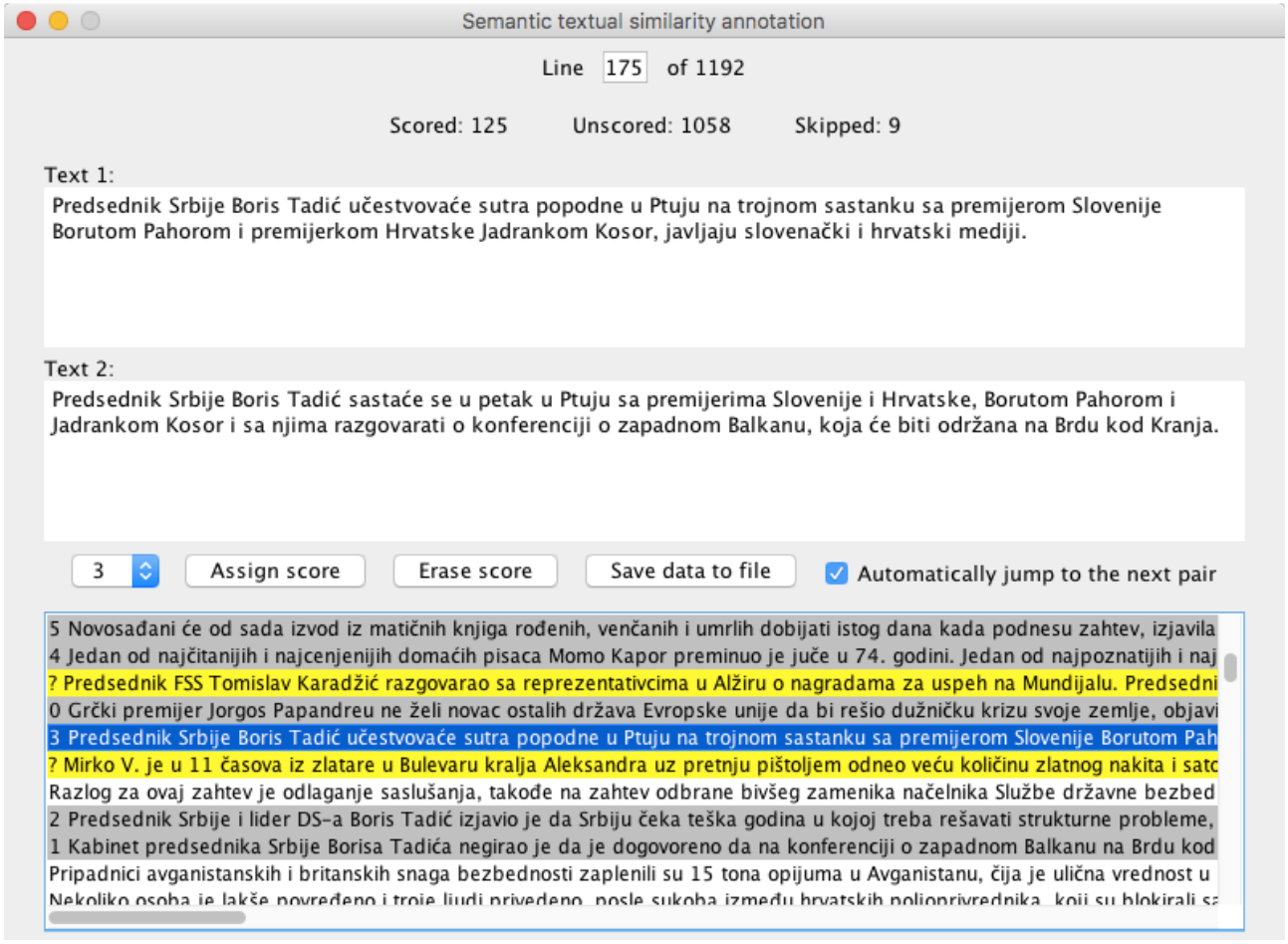
- Avion Malezija erlajnsa srušio se na rusko-ukrajinskoj granici.
  - SAD tvrde da Rusija grupiše vojsku na ukrajinskoj granici.

4.1.2 Obučavanje anotatora i sprovođenje anotacije semantičke sličnosti kratkih tekstova iz *STS.news.sr* korpusa

Radi lakšeg i bržeg sprovođenja anotacije, kreiran je namenski alat nazvan *STSAnno*, čiji je interfejs prikazan na slici 5 i koji je objavljen na posebnom *GitHub* repozitorijumu[17]. *STSAnno* je *offline* alat napisan u Javi koji omogućava anotatoru paralelni pregled tekstova čiji stepen semantičke sličnosti treba odrediti, dodeljivanje ocena semantičke sličnosti, kao i brisanje i menjanje ranije dodeljenih ocena. Pored numeričkih ocena sličnosti u opsegu od **0** do **5**, anotator ima mogućnost da paru tekstova dodeli i posebnu oznaku **?**, kojom se privremeno mogu obeležiti i time preskočiti oni parovi čije je ocenjivanje teže i na koje anotator planira da se naknadno vrati. Različite boje se koriste za isticanje ocenjenih (siva), neocenjenih (bela) i preskočenih parova (žuta), čime se omogućava lak vizuelni pregled. U svakom trenutku program pri vrhu prozora prikazuje statistički uvid u napredak anotacije, u vidu trenutnog broja ocenjenih, neocenjenih i preskočenih parova.

Pri pokretanju programa, *STSAnno* traži od korisnika da odabere ulazni TXT fajl sa UTF-8 kodiranjem koji sadrži skup tekstova koje treba anotirati. Očekivani format tekstova je jedan par tekstova u svakom redu fajla, pri čemu tekstovi u okviru para treba da budu razdvojeni znakom za tabulaciju. Tekstove je moguće ocenjivati redom kojim se javljaju u fajlu, za šta postoji opcija za automatski prelazak na naredni neocenjeni (ili, ako takvih nema, preskočeni) par tekstova nakon unosa ocene za trenutno posmatrani, ili proizvoljnim redosledom, korišćenjem tabelarnog pregleda svih tekstova u donjem delu prozora. Izlaz programa ima istu TSV (engl. *tab-separated value*) strukturu kao i ulazni fajl, osim što se u svakom redu pre samih tekstova zapisuje njima dodeljena ocena sličnosti, razdvojena od tekstova znakom za tabulaciju. *STSAnno* snima svoj izlaz u zadati ulazni fajl, što omogućava anotatoru da koristi jedan fajl za smeštanje podataka tokom cele anotacije. Zadate ocene se mogu upisati u fajl izborom korisnika putem odgovarajućeg tastera, ili automatski, prilikom izlaska iz programa.

[17] https://vukbatanovic.github.io/STSAnno/



Slika 5. Interfejs *STSAnno* alata za anotaciju semantičke sličnosti kratkih tekstova

Kada su uputstva za anotaciju finalizovana, svi anotatori su upoznati sa *STSAnno* alatom i načinom njegove upotrebe, a zatim su označili podskup od 60 nasumično izabranih parova rečenica iz skupa *STS.news.sr*, što je oko 5% od ukupnog broja. Anotacija ovog podskupa je omogućila da se razreše sve eventualne nedoumice anotatora u tumačenju i primeni uputstava, čime je završen treći korak u fazi anotacije – obučavanje anotatora i kalibracija. Posle toga se pristupilo anotaciji celokupnog skupa podataka, pri čemu su anotatori radili samostalno i anotirali svih 1192 parova rečenica, čime je zaokružen i korak sprovođenja anotacije. Oko 37 radnih sati po anotatoru je skupa utrošeno na obučavanje, kalibraciju i sprovođenje anotacije. Anotirani skup *STS.news.sr* je objavljen na *GitHub* repozitorijumu[18] (Batanović et al. 2018a).

4.1.3 Analiza anotacije semantičke sličnosti kratkih tekstova iz *STS.news.sr* korpusa

U ovom odeljku najpre je prikazana analiza konzistentnosti anotacije semantičke sličnosti kratkih tekstova iz *STS.news.sr* korpusa, a zatim je dat statistički prikaz anotiranog korpusa. Na kraju je izneto poređenje *STS.news.sr* korpusa sa postojećim skupovima podataka sličnog tipa na drugim jezicima, koji su anotirani korišćenjem istog sistema oznaka.

[18] https://vukbatanovic.github.io/STS.news.sr/

4.1.3.1 *Analiza konzistentnosti anotacije semantičke sličnosti kratkih tekstova iz STS.news.sr korpusa*

Međusobne saglasnosti anotatora izračunate su korišćenjem Pirsonovog koeficijenta korelacije r (engl. *Pearson correlation coefficient*), jer se ta binarna metrika koristila kao standard u okviru *SemEval* takmičenja (Agirre et al. 2012, 2013, 2014, 2015, 2016; Cer et al. 2017), kako za izražavanje stepena saglasnosti anotatora, tako i za merenje performansi NLP modela. Vrednost +1 kod Pirsonovog koeficijenta korelacije označava savršenu pozitivnu korelaciju, dok vrednost 0 označava nepostojanje korelacije. Za računanje Pirsonovog koeficijenta korelacije korišćena je *SciPy* biblioteka (Virtanen et al. 2020). Saglasnosti su takođe izražene i preko Kripendorfovog α koeficijenta. Za računanje α koeficijenta u ovoj disertaciji korišćena je *Krippendorff* Python biblioteka[19].

Tabela 3. Međusobne saglasnosti anotatora u označavanju semantičke sličnosti kratkih tekstova na srpskom jeziku iz *STS.news.sr* korpusa, izražene u vidu Pirsonovog koeficijenta korelacije r

Anotator	1	2	3	4	5	Prosek
1	/					
2	0,899	/				
3	0,893	0,867	/			
4	0,885	0,844	0,850	/		
5	0,898	0,878	0,863	0,856	/	
Prosek ocena ostalih anotatora	0,945	0,916	0,912	0,900	0,918	0,918

Tabela 4. Međusobne saglasnosti anotatora u označavanju semantičke sličnosti kratkih tekstova na srpskom jeziku iz *STS.news.sr* korpusa, izražene u vidu Kripendorfovog α koeficijenta

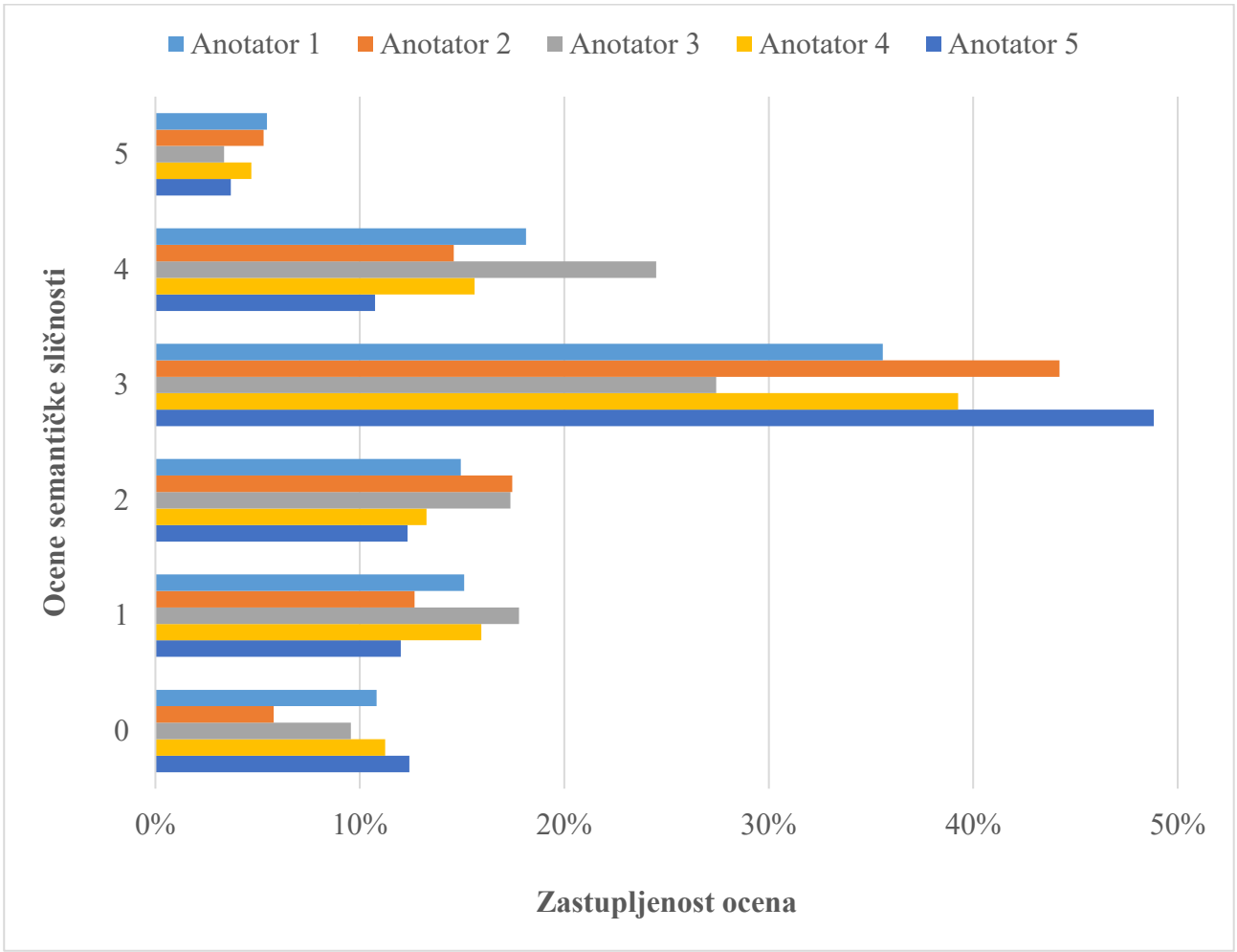
Binarna saglasnost							Globalna saglasnost
Anotator	1	2	3	4	5	Prosek	
1	/						0,868
2	0,886	/					
3	0,893	0,853	/				
4	0,885	0,828	0,850	/			
5	0,895	0,863	0,861	0,855	/		
Prosek ocena ostalih anotatora	0,939	0,905	0,907	0,896	0,916	0,913	

[19] https://github.com/pln-fing-udelar/fast-krippendorff

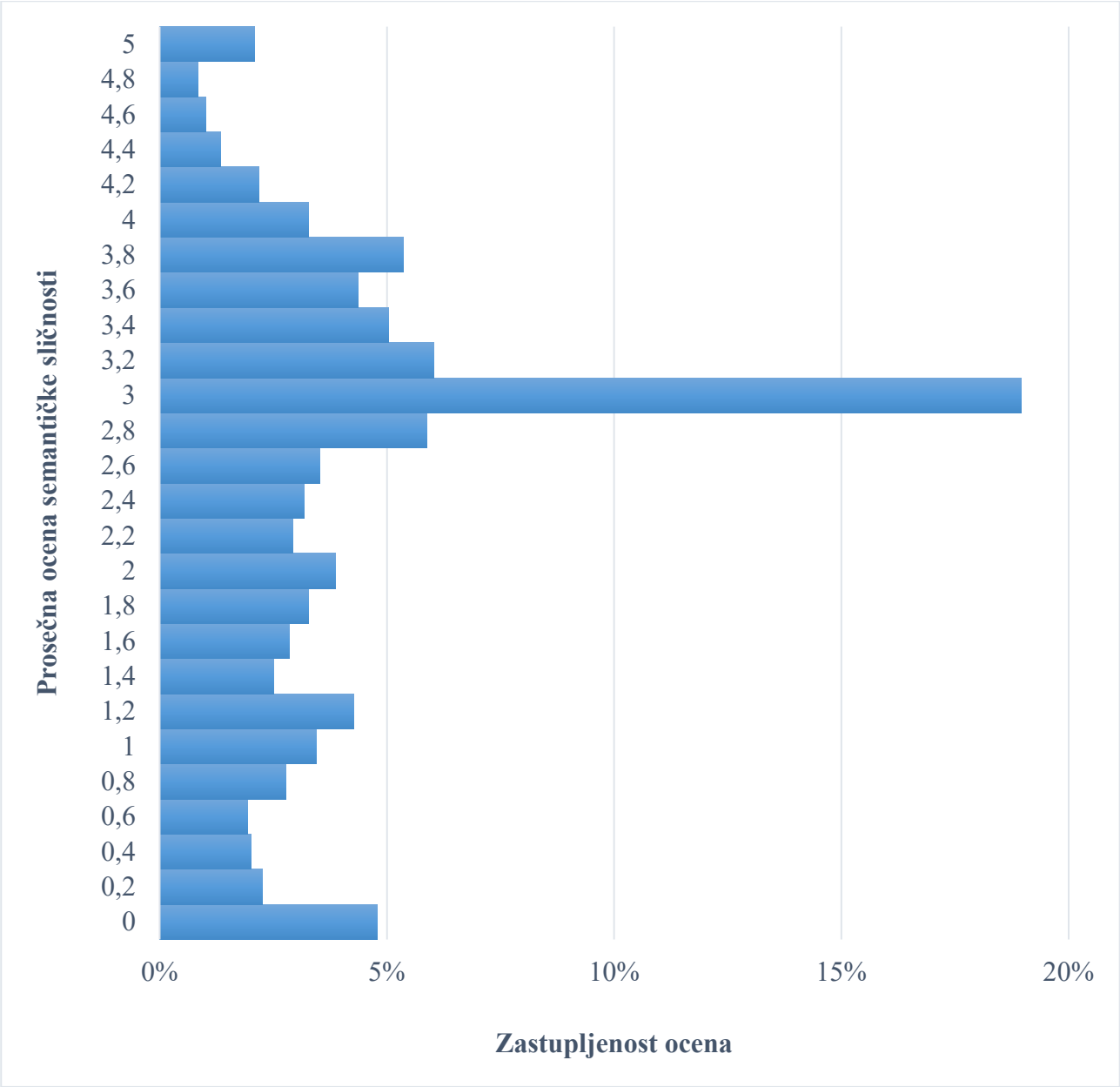
Tabele 3 i 4 prikazuju međusobne saglasnosti anotatora, gde su pored saglasnosti između svakog para anotatora izračunate i saglasnosti svakog anotatora sa prosekom ocena svih ostalih. Za Kripendorfov α koeficijent je takođe izračunata vrednost globalne tj. ne-binarne saglasnosti. Može se primetiti da su dobijene saglasnosti vrlo visoke u pogledu obe razmotrene metrike i da nema značajnijih razlika između anotatora u postignutim stepenima saglasnosti. Sve dobijene vrednosti Kripendorfovog α koeficijenta su jasno iznad granice od 0,8, što znači da su sve saglasnosti anotatora odlične. Prosečan koeficijent korelacije anotatora sa prosekom ocena drugih anotatora na *STS.news.sr* korpusu je jednak 0,918, te stoga ta vrednost predstavlja gornji prag za performanse NLP modela na ovom skupu podataka. Ovaj prosečan stepen saglasnosti je za 0,05 – 0,1 veći od prosečnog Pirsonovog koeficijenta korelacije na skupovima podataka iz novinskog domena kreiranih u okviru *SemEval* takmičenja (Agirre et al. 2013, 2014, 2015), čemu je najverovatniji uzrok povećan broj i kvalitet primera tekstova uključenih u modifikovana uputstva za anotaciju predstavljena u odeljku 4.1.1.1.

4.1.3.2 Statistički prikaz anotiranog korpusa kratkih tekstova STS.news.sr

Finalne ocene sličnosti za svaki par rečenica su dobijene uprosečavanjem individualnih ocena svih petoro anotatora. Na slici 6 prikazana je raspodela parova rečenica iz *STS.news.sr* korpusa po individualnim ocenama sličnosti za svakog anotatora, dok je na slici 7 prikazana raspodela po finalnim, uprosečenim ocenama sličnosti.



Slika 6. Raspodela parova rečenica iz *STS.news.sr* korpusa po individualnim ocenama semantičke sličnosti za svakog anotatora



Slika 7. Raspodela parova rečenica iz *STS.news.sr* korpusa po finalnim, uprosečenim ocenama semantičke sličnosti

Između raspodela individualnih ocena anotatora se mogu uočiti određena odstupanja, ali ona uglavnom nisu značajnijeg obima. Najprimetnija razlika je ona između anotatora 3 sa jedne strane i anotatora 2 i 5 sa druge strane i tiče se izbora između ocena **3** i **4** – dok je anotator 3 dosta ravnomerno koristio ove dve ocene sličnosti, druga dva navedena anotatora su daleko češće dodeljivala ocenu **3**. S tim u vezi, razmatranjem raspodele finalnih ocena sličnosti može se primetiti da je ona umereno izbalansirana, sa izuzetkom velikog odskakanja u pogledu broja parova čija je prosečna ocena sličnosti 3,0. Međutim, slične nepravilnosti u raspodeli parova tekstova po ocenama sličnosti su prisutne i u drugim skupovima podataka za ovu problematiku koji pripadaju domenu novinskog sadržaja, što će biti prikazano u narednom odeljku. Detaljni numerički podaci o raspodelama parova rečenica iz *STS.news.sr* korpusa po ocenama sličnosti su navedeni u tabelama P1 i P2 u prilozima ove disertacije, gde su pored procentualnih zastupljenosti za svaku ocenu date i apsolutne numeričke vrednosti.

4.1.3.3 Poređenje anotiranog korpusa kratkih tekstova STS.news.sr sa prethodnim skupovima podataka sa istim sistemom oznaka

Pored skupova podataka anotiranih za problem semantičke sličnosti sa *SemEval* takmičenja, koji su napravljeni za engleski, španski i arapski jezik, jedino su još skupovi podataka na portugalskom jeziku (na evropskom i na brazilskom portugalskom) javno dostupni (Fonseca et al. 2016). Tabela 5 sadrži uporedni pregled *STS.news.sr* skupa na srpskom i drugih javno dostupnih korpusa kratkih tekstova iz novinskog domena anotiranih za problem određivanja semantičke sličnosti. U te korpuse spadaju:

- Na engleskom jeziku: *SemEval MSRPar* korpus sa konferencije *SemEval* 2012 (Agirre et al. 2012), spojeni *SemEval HDL* korpusi iz perioda 2013-2016 (Agirre et al. 2013, 2014, 2015, 2016) i *SemEval Deft-news* korpus iz 2014. godine (Agirre et al. 2014);
- Na španskom jeziku: spojeni *SemEval News* korpusi iz 2014. i 2015. godine (Agirre et al. 2014, 2015);
- Na portugalskom jeziku: ASSIN korpus iz 2016. godine, podeljen na deo na evropskom portugalskom i deo na brazilskom portugalskom (Fonseca et al. 2016);
- Na arapskom jeziku: prevod dela *SemEval MSRPar* korpusa na arapski jezik, predstavljen 2017. godine u okviru *SemEval* konferencije (Cer et al. 2017).

STS.news.sr korpus je prosečne veličine u poređenju sa ostalim javno dostupnim korpusima ovog tipa, kako u pogledu broja parova rečenica, tako i u pogledu broja tokena (za tokenizaciju je korišćeno razdvajanje na osnovu blanko znakova). Rečenice iz *STS.news.sr* skupa su u proseku duže tj. sadrže više tokena nego rečenice iz većine drugih razmatranih korpusa. Prosečna ocena sličnosti od 2,51 u *STS.news.sr* korpusu je gotovo idealna uzimajući u obzir skalu ocena od **0** do **5**.

U poređenju sa *SemEval MSRPar* korpusom na engleskom jeziku, koji mu je najbliži po tipu, veličini i vrsti izvornog materijala, *STS.news.sr* skup je znatno izbalansiraniji u smislu raspodele parova po ocenama sličnosti. Ipak, skoro svi razmotreni korpusi ispoljavaju primetne skokove u broju parova sa ocenama sličnosti oko **3** i **4**, pri korišćenju skale ocena **0 – 5**, odnosno u broju parova sa ocenama sličnosti oko **2** i **3**, pri korišćenju skale ocena **0 – 4**. Slično tome, raspodela parova iz ASSIN korpusa po prosečnim ocenama pokazuje izraženo grupisanje u opsegu centralnih vrednosti ocena od **2** do **4**.

Jedini korpus sa uniformnijom raspodelom parova jeste *SemEval HDL* korpus na engleskom jeziku, koji je sastavljen od naslova vesti. Naslovi su po prirodi kraći od punih rečenica kakve se javljaju u ostalim korpusima, uključujući *STS.news.sr*, što je verovatno bar delimičan uzrok veće ravnomernosti ocena. Naime, pri uparivanju dužih rečenica, verovatnoća susretanja parova koji sadrže rečenice skoro identičnog značenja (ocene sličnosti blizu **5** na skali **0 – 5**) je prirodno manja. Slično tome, kod dužih rečenica prirodno postoji veća šansa da uparene rečenice sadrže bar neki stepen semantičkog poklapanja, što snižava verovatnoću minimalnih ocena semantičke sličnosti.

Koliko je autoru disertacije poznato, pre izrade *STS.news.sr* anotiranog skupa na srpskom jeziku nije postojao nijedan javno dostupni resurs ovog tipa na nekom od manjih jezika i jezika sa ograničenim resursima. Pored toga, iako je za poljski jezik prethodno predstavljen skup podataka za širi problem semantičke srodnosti (engl. *semantic relatedness*) (Wróblewska & Krasnowska-Kieraś 2017), *STS.news.sr* korpus ujedno je i prvi skup podataka anotiran za problem semantičke sličnosti kratkih tekstova na nekom slovenskom jeziku.

Tabela 5. Uporedni pregled javno dostupnih korpusa kratkih tekstova iz novinskog domena anotiranih za problem određivanja semantičke sličnosti

Korpus		*STS.news.sr*	*SemEval MSRPar*	*SemEval HDL*	*SemEval Deft-news*	*SemEval News*	ASSIN (PT)	ASSIN (BR)	*SemEval MSRPar*
Jezik		SR	EN	EN	EN	ES	PT-PT	PT-BR	AR
Skala ocena		**0 – 5**	**0 – 5**	**0 – 5**	**0 – 5**	**0 – 4**	**1 – 5**	**1 – 5**	**0 – 5**
Parova		1192	1500	2499	300	980	5000	5000	510
Tokena		64K	54K	37K	9K	68K	145K	130K	18K
Prosečno tokena u rečenici		~27	~18	~7	~16	~35	~14	~13	~18
Prosečna ocena sličnosti		2,51	3,30	2,62	3,03	2,20	3,04	3,04	3,36
Procenat parova po ocenama (zaokruž.)	**0**	9,06	0,13	10,56	4,00	6,33	/	/	0,20
	1	14,93	4,47	17,89	11,00	16,33	2,72	1,74	3,33
	2	16,11	13,87	17,29	16,00	36,94	35,14	33,94	13,14
	3	39,43	36,47	19,93	27,33	29,08	24,16	28,84	34,90
	4	16,53	36,20	22,09	31,33	11,32	31,56	30,76	40,39
	5	3,94	8,86	12,24	10,33	/	6,42	4,72	8,04

4.2 Anotacija podataka za problem analize sentimenta kratkih tekstova

Kao što je već pomenuto u uvodnoj glavi, u analizu sentimenta spada veći broj užih problema, kao što je određivanje polarnosti teksta, određivanje subjektivnosti teksta, detekcija sarkazma, itd. Ovi problemi se često predstavljaju u obliku zadataka binarne klasifikacije, gde je cilj razlikovati pozitivne od negativnih tekstova, subjektivne od objektivnih, odnosno sarkastične od nesarkastičnih. Shodno tome, u literaturi postoji veliki broj radova koji razmatraju samo jedan od navedenih užih problema, i u okviru toga koriste sisteme binarnog označavanja podataka (Pang et al. 2002; Turney 2002; Pang & Lee 2004; Maas et al. 2011; Maynard & Greenwood 2014; Ptáček et al. 2014).

Osnovni i verovatno najrašireniji zadatak u analizi sentimenta jeste određivanje polarnosti teksta (Pang et al. 2002; Turney 2002). Naravno, binarna podela sentimenta na pozitivan i negativan je dosta rigidna i može da dovede do teškoća u određivanju sentimenta tekstova koji ne potpadaju jasno pod jednu od binarnih kategorija. Iz tog razloga se skup oznaka često proširuje na neki od sledećih načina:

- Uvođenjem neutralne klase u skup oznaka (Koppel & Schler 2006) – iako naizgled intuitivno, ovo proširenje je ipak donekle nejasno, zbog dvosmislenog tumačenja koncepta neutralnosti. Naime, kao što navode (Koppel & Schler 2006), postoje dve različite definicije neutralnog sentimenta. Prva podrazumeva da neutralni tekstovi budu oni koji ne izražavaju bilo kakav sentiment, tj. izražavaju samo objektivne činjenice. Nasuprot tome, moguća je i definicija po kojoj neutralni tekstovi sadrže i pozitivne i negativne iskaze, tj. mešavinu oba sentimenta. Naravno, moguće je i napraviti posebne klase/oznake za obe definicije neutralnosti.
- Zamenom binarnih klasa skalom ocena (Pang & Lee 2005; Thelwall et al. 2010) – skala ocena sentimenta omogućava da se izrazi ne samo polarnost, nego i snaga sentimenta. Pri korišćenju skale ocena, neutralnost po pitanju sentimenta se prirodno reprezentuje kroz srednje vrednosti na skali, ali i dalje ostaje otvoreno pitanje koju od navedenih interpretacija neutralnosti te srednje vrednosti predstavljaju. Pored toga, u ovakvom sistemu ocena ostaje nejasno i kako označiti mešavinu pozitivnog i negativnog sentimenta ako srednje vrednosti na skali reprezentuju izostanak bilo kakvog sentimenta, i obratno.
- Prelaskom na granularnije posmatranje teksta tako što se odvojenim stavkama u tekstu dodeljuju odvojene ocene sentimenta (Pontiki et al. 2014, 2015, 2016) – ovaj pristup, poznat kao aspektna analiza sentimenta (engl. *aspect-based sentiment analysis*), omogućava da se izbegne upotreba jedne globalne ocene sentimenta za ceo tekst. Na prvi pogled, povećanje granularnosti deluje kao elegantno rešenje za tekstove u kojima se izražava i pozitivan i negativan sentiment. Na primer, ako se u tekstu pomene da je *A* dobro, a *B* loše, aspektni pristup bi jednostavno dodelio pozitivan sentiment aspektu *A*, a negativan sentiment aspektu *B*. Ipak, postoje problematične situacije u kojima veća granularnost teksta nije od pomoći za određivanje sentimenta. Dobar primer ovoga su tekstovi koji su po prirodi dvosmisleni, poput *U pravu je, ovako nešto definitivno niste videli ranije.*

Kao što se može videti, sva navedena proširenja skupa oznaka nose sa sobom određene nedostatke. Naravno, njih je moguće i međusobno kombinovati, na primer u vidu aspektne analize sentimenta koja koristi skalu ocena (Grljević 2016). Ipak, čak i takvi, relativno komplikovani sistemi oznaka ne mogu da pruže prirodan način za obeležavanje svih mogućnosti izražavanja sentimenta u jeziku (npr. označavanje dvosmislenih ili sarkastičnih iskaza i dalje ostaje nerešeno). Sistem oznaka u kome bi to bilo moguće bi stoga nužno bio izuzetno složen, a njegova složenost bi posledično povećala kompleksnost dizajniranja i izrade računarskog modela za analizu sentimenta. Osim toga, korišćenje složenog sistema oznaka bi verovatno dovelo do češćih nesaglasnosti između ljudi u pogledu određivanja ispravnih ocena sentimenta, čime bi se snizili i sami dometi automatske analize sentimenta.

Navedena razmatranja su naročito relevantna za jezike sa ograničenim resursima, u kojima je pažljivo osmišljavanje sistema oznaka nužno zbog limita u pogledu ljudskih i materijalnih resursa koji se mogu utrošiti na anotacione projekte. Shodno tome, u ovakvim jezicima nema smisla ni primenjivati pomenuto izolovano posmatranje svakog od užih problema u analizi sentimenta, jer bi ono značilo da anotaciju podataka za svaki uži problem treba sprovesti zasebno, sa posebnim setom oznaka.

Pored pomenutih čestih proširenja sistema označavanja polarnosti, u literaturi je predložen veliki broj različitih pristupa za označavanje sentimenta koji prevazilaze binarne sisteme oznaka, počevši od višeklasnih pristupa (Abdul-Mageed & Diab 2011), preko pristupa zasnovanih na označavanju sintaktičkih (pod)stabala (Socher et al. 2013) ili koreferentnosti entiteta (Kessler et al. 2010), do kompleksnih sistema, često delimično zasnovanih na određenim lingvističkim teorijama (Wiebe et al. 2005; Deng & Wiebe 2015; Van de Kauter et al. 2015). Postojeći sistemi za označavanje sentimenta kratkih tekstova su uglavnom razvijeni za potrebe *crowdsourcing* anotacije podataka sa Tvitera, i koriste tri klase – pozitivnu, neutralnu i negativnu. Zbog *crowdsourcing* pristupa anotaciji, ovi sistemi označavanja teže što većoj jednostavnosti i umnogome se oslanjaju na ličnu intuiciju

anotatora. Kod ovakvih sistema, posebno problematični tekstovi oko čijih se oznaka anotatori ne slažu se često prosto izbacuju iz skupa podataka (npr. (Nabil et al. 2015)).

Postizanju kvalitetne anotacije podataka u resursno ograničenim jezicima je do sada u literaturi pridavano malo pažnje, što je evidentno i kod problema analize sentimenta. Najrelevantniji prethodni rad na ovu temu jeste (Mohammad 2016). U njemu je identifikovano nekoliko tipičnih situacija koje otežavaju označavanje sentimenta, uključujući iskaze u okviru kojih se ispoljavaju različiti sentimenti prema različitim pojmovima, iskaze u kojima se prenosi uspeh ili neuspeh jedne strane u odnosu na drugu, sarkastične iskaze, citate, retorička pitanja, itd. Mohammad je predložio i dva odvojena sistema oznaka – jedan jednostavniji i primereniji za jezike sa ograničenim resursima, i drugi, složeniji. Jednostavniji sistem sadrži pet kategorija zasnovanih na iskazima koje govornik koristi. To su: kategorija za pozitivne iskaze, kategorija za negativne iskaze, kategorija za sarkastične iskaze, kategorija za iskaze koji su delom pozitivni a delom negativni, i kategorija za iskaze koji nisu ni pozitivni ni negativni. Sa izuzetkom kratkih opisa ovih kategorija, Mohammad nije pružio detaljnija uputstva vezana za ovaj sistem oznaka, ali ga jeste primenio u *crowdsourcing* anotaciji tvitova na engleskom jeziku (Mohammad et al. 2017). U ovom projektu, kategorija za negativne iskaze je nakon završene anotacije spojena sa kategorijom za sarkastične iskaze, dok je kategorija za iskaze koji nisu ni pozitivni ni negativni spojena sa kategorijom za iskaze koji su delom pozitivni, a delom negativni. Sa ovako dobijene tri klase, autori su postigli međusobnu saglasnost anotatora od 85,6%.

U literaturi je predloženo još nekoliko sistema za označavanje sentimenta kratkih tekstova koji bi se mogli primeniti u resursno ograničenim situacijama. (Abdul-Mageed & Diab 2011) su predložili sistem osmišljen za domen vesti koji sadrži četiri klase: objektivnu, subjektivno-pozitivnu, subjektivno-negativnu i subjektivno-neutralnu. Ovaj sistem je primenjen u anotaciji sentimenta rečenica iz novinskih izvora napisanih na standardnom arapskom jeziku, pri čemu je postignuta međusobna saglasnost anotatora od 88,06%, odnosno $\kappa = 0{,}823$. U kasnijem radu istih autora (Abdul-Mageed & Diab 2012), navedeni sistem je izmenjen tako što je izbačena objektivna kategorija, a uvedena kategorija za mešavinu sentimenata. Uputstva za anotaciju su takođe proširena napomenama zasnovanim na teoriji učtivosti (engl. *politeness theory*) (Brown & Levinson 1987), kao i smernicama koje se odnose na tretiranje iskaza slaganja i neslaganja. Ovaj modifikovani sistem je primenjen na anotaciju tekstova iz drugih domena, i to strana za razgovor sa Vikipedije na arapskom jeziku i sadržaja arapskih veb foruma. Na prvom domenu je postignuta saglasnost anotatora od $\kappa = 0{,}790$, dok je na drugom domenu $\kappa = 0{,}793$. Isti autori su takođe eksperimentisali sa jednostavnim sistemom označavanja koji sadrži samo tri klase – pozitivnu, negativnu i neutralnu – i vrlo oskudna uputstva. Ovaj pristup je primenjen kako u klasičnoj organizaciji anotacije, tako i preko *crowdsourcing* servisa, ali su u oba slučaja dobijene jako niske međusobne saglasnosti anotatora – $\kappa = 0{,}19$ za klasičnu anotaciju, odnosno $\kappa = 0{,}065$ za *crowdsourcing* anotaciju.

(Al-Twairesh et al. 2017) su predložili sistem označavanja koji sadrži pet klasa – za pozitivan, negativan, neutralan i neodređen sentiment, kao i za mešavinu sentimenata – i sedam kratkih smernica za anotaciju. Troje anotatora, čije je obučavanje trajalo po sat vremena, je korišćenjem ovog sistema anotiralo skup tvitova na saudijskom dijalektu arapskog jezika, pri čemu su dobijene umerene međusobne saglasnosti ($\kappa = 0{,}60$).

Nažalost, nijedan od pomenutih radova ne pruža podatke o vremenu potrebnom anotatorima da završe označavanje sentimenta tekstova, zbog čega je nemoguće utvrditi ekonomičnost predloženih pristupa. Jedini pronađeni rad koji je predstavio bilo kakvu analizu troškova i koristi (engl. *cost-benefit analysis*) vezanih za anotaciju sentimenta jeste (Balamurali et al. 2012). U tom radu je razmotreno pitanje anotiranja značenja reči na osnovu *WordNet* grafa (Miller et al. 1990; Miller 1995), i korišćenja tako dobijenih značenja umesto samih reči kao odlika pri binarnom određivanju polarnosti tekstova. U radu (Joshi et al. 2014) je predložena mera za kompleksnost anotacije sentimenta, ali je

ona zasnovana na uvidu u kretanje pogleda anotatora po ekranu, što je vrsta podataka koja je retko kada dostupna pri anotaciji. Osim toga, predložena metrika se oslanja i na skup lingvističkih obeležja tekstova koji se anotiraju, od kojih mnoga (npr. udaljenost koreferentnih pominjanja) nije moguće automatski pribaviti u jezicima sa ograničenim resursima.

Od ranijih radova o analizi sentimenta u srpskom jeziku pomenutih u poglavlju 3.2, jedino je (Grljević 2016) detaljnije razmatrala ručnu anotaciju podataka i prikazala korišćena uputstva za anotaciju. Sistem označavanja izložen u tom radu je dizajniran za aspektnu analizu sentimenta u domenu ocena nastavnog osoblja na univerzitetima u Srbiji. Predstavljeni sistem je dosta kompleksan, jer podrazumeva ne samo identifikaciju aspekata u tekstu i obeležavanje njihove polarnosti, već i zadavanje numeričkih ocena jačine za negativni i pozitivni sentiment. Četvoro anotatora je korišćenjem ovog sistema ostvarilo izuzetno dobre saglasnosti u anotaciji polarnosti (κ = 0,964) kao i vrlo dobre saglasnosti u anotaciji aspekata (κ = 0,792), ali su saglasnosti u anotaciji jačine sentimenta značajno niže (κ = 0,443), i pored dosta obimnih i detaljnih uputstava za anotaciju.

Imajući u vidu sva navedena prethodna istraživanja i njihove prednosti i mane, zaključeno je da do sada nije razvijen nijedan sistem označavanja koji bi bio, sa jedne strane, lako primenjiv na različite domene, i sa druge, prilagođen jezicima sa ograničenim resursima. Glavna odlika takvog sistema bi trebalo da bude fleksibilnost u pogledu interpretacije oznaka, tako da se jednom anotirani skup podataka može koristiti za različite uže probleme u analizi sentimenta, a ne samo za jedan, poput određivanja polarnosti. Pored toga, takav sistem bi morao da obezbedi ne samo visok kvalitet anotacije, već i njenu ekonomičnost, što je pitanje koje se u postojećoj literaturi gotovo uopšte ne razmatra. Iz tog razloga se u izradi ove disertacije pristupilo izradi takvog sistema označavanja, kao i predlaganju metrike pomoću koje bi se mogla proceniti ekonomičnost anotacije sentimenta tekstova.

Ostatak ovog poglavlja je podeljen po koracima u fazi anotacije podataka. Najpre je prikazan kreirani sistem oznaka i opisan proces formulisanja uputstava za anotaciju i njihov sadržaj. Za razliku od anotacije semantičke sličnosti, na zadatku anotacije sentimenta kalibracija i dopunjavanje uputstava su tekli paralelno sa sprovođenjem anotacije nad glavnim skupom podataka, zbog čega ove faze nisu jasno razdvojene. Nakon toga je izvršena analiza konzistentnosti, efikasnosti i ekonomičnosti sprovedene anotacije sentimenta na *SentiComments.SR* korpusu, pri čemu je iznet statistički prikaz svih anotiranih podataka. Konačno, nov sistem označavanja sentimenta tekstova je detaljno upoređen sa najrelevantnijim prethodno predloženim pristupima koji su predstavljeni u ovom poglavlju.

4.2.1 Odabir oznaka, formulisanje uputstava i sprovođenje anotacije sentimenta kratkih tekstova

Razvijeni sistem označavanja sentimenta se sastoji iz sledećih šest oznaka:

- **+1** – za tekstove koji su potpuno ili predominantno pozitivni;
- **-1** – za tekstove koji su potpuno ili predominantno negativni;
- **+NS** – za tekstove koji su objektivni, ali u binarnoj klasifikaciji na pozitivne/negativne ipak naginju ka pozitivnim;
- **-NS** – za tekstove koji su objektivni, ali u binarnoj klasifikaciji na pozitivne/negativne ipak naginju ka negativnim;
- **+M** – za tekstove koji su dvosmisleni u pogledu sentimenta ili izražavaju mešavinu sentimenata, ali u binarnoj klasifikaciji na pozitivne/negativne ipak naginju ka pozitivnim;
- **-M** – za tekstove koji su dvosmisleni u pogledu sentimenta ili izražavaju mešavinu sentimenata, ali u binarnoj klasifikaciji na pozitivne/negativne ipak naginju ka negativnim.

Pored ovih ocena, oni tekstovi za koje je ili sigurno ili dosta izvesno da su sarkastični su obeležavani tako što je na kraj njihove oznake dodavan nastavak **s**. Pošto sarkazam uvek implicira izražavanje sentimenta, nastavak **s** se može dodati samo na oznake +/**-1** i +/-**M**.

Ovakav sistem oznaka je usvojen zbog svoje fleksibilnosti u pogledu interpretacije značenja oznaka, jer pruža četiri načina za redukovanje broja klasa sentimenta u naknadnom procesiranju i modelovanju:

- Redukovanje na binarno određivanje polarnosti teksta, gde se gleda samo +/- deo oznake;
- Redukovanje na određivanje subjektivnosti teksta, gde +/-**NS** oznake predstavljaju objektivne tekstove, a preostale četiri oznake subjektivne tekstove;
- Redukovanje na četvoroklasnu klasifikaciju, sa pozitivnom klasom (oznaka +**1**), negativnom klasom (oznaka **-1**), klasom za dvosmislene tekstove ili tekstove koji izražavaju mešavinu sentimenata (oznake +/-**M**) i klasom za objektivne tekstove tj. tekstove koji ne izražavaju bilo kakav sentiment (oznake +/-**NS**);
- Redukovanje na detekciju sarkazma, gde se gleda samo prisustvo ili odsustvo nastavka **s** u oznaci.

Ovakva adaptivnost oznaka sentimenta omogućava da se kroz samo jedan projekat anotacije označi više slojeva kompleksnosti izražavanja sentimenta, čime se povećava vrednost i primenjivost dobijenih anotiranih podataka. Ipak, usvajanje novog sistema oznaka je nužno zahtevalo i izradu novih, namenskih uputstava za anotaciju.

Počevši od inicijalnih smernica za anotaciju, tekstove iz glavnog skupa kratkih komentara *SentiComments.SR* je ručno zajedno anotiralo dvoje anotatora, pri čemu su tokom anotacije uputstva za anotaciju dopunjavana. Iako je na ovaj način izbrisana jasna granica između različitih koraka anotacije, ovakav pristup se pokazao neophodnim radi izrade što jasnijih i nedvosmislenijih uputstava za nov sistem označavanja. Shodno ranijim zaključcima o neadekvatnosti *crowdsourcing* metoda u ovom kontekstu, izrađena uputstva su prevashodno prilagođena klasičnoj organizaciji anotacije, sa manjom grupom direktno angažovanih anotatora.

Anotacija glavnog skupa kratkih komentara *SentiComments.SR* je sprovedena nesekvencijalno, u 4 prolaza kroz ceo skup. Prvi prolaz je omogućio upoznavanje sa raznim jezičkim fenomenima prisutnim u skupu komentara, kao i ručno ispravljanje slovnih grešaka i nedostajućih dijakritičkih znakova u tekstovima komentara. Ove greške su ručno ispravljane jer posmatrani tekstovi pripadaju neformalnom registru, sa velikim brojem nestandardnih reči za koje postojeći automatski alati za srpski ne bi bili adekvatni. Pri tome namerno nije vršena korekcija u slučajevima kada se jedno isto slovo ili grupa slova ponavlja više puta kao način za naglašavanje te reči, jer se taj fenomen često sreće u neformalnim tekstovima i može biti od koristi za automatsko detektovanje emocionalnog naboja u tekstu. Originalni tekstovi su takođe sačuvani da bi se ispitalo u kojoj meri tipografska i dijakritička ispravnost teksta utiče na kvalitet rada sistema za analizu sentimenta na srpskom jeziku.

U drugom prolazu su ocenjeni komentari za koje je odmah jasno kojoj kategoriji pripadaju – primeri takvih komentara su *Odličan film!* ili *Film je potpuno isprazan.* Shodno preporukama iz (Hovy & Lavid 2010), za treći prolaz su ostavljeni problematičniji komentari, čija je anotacija teža. Najveći deo dorada uputstava za anotaciju je urađen u ovom prolazu, kroz diskusije i konsultacije između anotatora. U slučajevima razlike mišljenja do približavanja stavova se dolazilo traženjem sličnih primera u ostatku skupa komentara i ustanovljavanjem kriterijuma za tretiranje karakterističnih iskaza i situacija. Finalni, četvrti prolaz je služio za verifikovanje konzistentnosti prethodno donetih

anotacionih odluka shodno svim utvrđenim kriterijumima, kao i za obeležavanje sarkastičnih komentara. Po završetku anotacije, skup *SentiComments.SR* je objavljen na *GitHub* repozitorijumu[20].

Anotacija sentimenta kratkih tekstova je sprovođena uz pomoć popularnih programa za unos i izmenu tekstualnog sadržaja, kao što su *Microsoft Word*, *Wordpad*, *Notepad* i sl, po izboru anotatora. Pri tome je korišćen TSV format zapisivanja oznaka sentimenta, koje su od teksta komentara odvojene znakom za tabulaciju.

4.2.1.1 Uputstva za anotaciju sentimenta kratkih tekstova

Uputstva za anotaciju sentimenta kratkih tekstova se sastoje iz 21 kriterijuma i pravila za anotaciju, sa primerima kratkih komentara na koje se ti kriterijumi odnose i pravilnim oznakama za svaki primer. Kriterijumi su navedeni u nastavku ovog odeljka, propraćeni obrazloženjima vezanim za njihovo usvajanje tamo gde je to potrebno radi lakšeg razumevanja.

K1 – Kontekst

Kriterijum: Svaki komentar treba anotirati sam za sebe, bez korišćenja okolnih komentara kao kontekstnih informacija.

Obrazloženje: Ovakav pristup anotiranju je bio neophodan jer su dugački komentari izostavljeni iz skupa prikupljenih tekstova, kao što je opisano u poglavlju 3.2, te stoga kontekst javljanja nije ni bio dostupan pri anotaciji svih komentara. Međutim, kontekstno agnostički pristup je bio poželjan i sa stanovišta pojednostavljivanja anotacije sentimenta i kasnije izrade računarskih modela. Pored toga, u mnogim realnim situacijama redosled javljanja komentara ne dovodi do njihove međusobne povezanosti. Kao što je opisano u prethodnom odeljku, komentari nisu označavani sekvencijalno, što je potpomoglo da se informacije o širem kontekstu ignorišu pri anotiranju. Iako ignorisanje konteksta može da, principijelno govoreći, dovede do grešaka u razumevanju teksta, takve situacije su se vrlo retko javljale tokom anotacije i uglavnom su se odnosile na nesigurnost procene sarkastičnosti određenog komentara.

K2 – Kompozitno ocenjivanje

Kriterijum: Pri anotiranju komentara koji sadrže veći broj odvojenih iskaza, svaki iskaz treba evaluirati za sebe, a do konačne ocene treba doći pomoću pravila za kombinovanje takvih individualnih iskaza.

Najpre, kada komentar sadrži veći broj iskaza iste polarnosti, bez obzira na njihovu temu, ceo komentar dobija ocenu shodno toj polarnosti:

- **+1** .. po meni jedan od boljih filmova u 2012… i moram dodati da me Joseph Gordon-Levitt prijatno iznenadio svojom glumom.. što je uspeo takođe potvrditi u The Dark Knight Rises !! 10/10
- **-1** Odavno nisam gledao gori film, totalno razočarenje, očekivao sam nešto dobro, ali već posle pola sata sam počeo da se vrpoljim na sedištu u bioskopskoj sali. moja preporuka - zaobiđite ovaj film..

Kada komentar sadrži dva ili više iskaza suprotnih polarnosti o različitim temama ili o različitim aspektima/delovima/viđenjima jedne iste teme, adekvatna ocena je **M**. Polarnost ocene ne zavisi samo

[20] https://vukbatanovic.github.io/SentiComments.SR/

od broja pozitivnih i negativnih iskaza, već i od njihove jačine. Ocenu **M** treba koristiti i kada je preovlađujući sentiment negativan, ali postoji pozitivno mišljenje o bar jednom aspektu, i obratno:

- **+M** woody-jeva uloga je genijalna, šteta što režiju nisu mogli da prilagode, pa da ovaj bude "total blast" !!!
- **+M** Odličan, super priča i super animacija, odlično se uklapa :) jedina stvar koju bih zamerio to je korišćenje pesme eye of tiger, malo suvišno, dok su ostale reference na mestu :)
- **-M** Za ljubitelje horora ovo je jedno veoma veliko razočarenje. Ljudi koji ne prate i ne gledaju filmove ovog žanra mogu biti zadovoljni.
- **-M** Jedino što ga vadi je to što je 3D, da ga gledaš na DVD-u ocena bi otišla u minus, barem tako ja mislim. :)
- **+M** predvidivo pomalo na kraju ali ne može se reći da je loše odrađeno sve od početka do kraja...
- **-M** Pogledao sam tek sad film, film je dobar, sviđa mi se ali mi fali nešto da mi ostavi potpuni utisak, jeste istinita priča i ne može se izmišljati ali fali mi nešto

U principu, globalni stav ima veću jačinu od užeg stava o nekom pojedinačnom aspektu. Ipak, veći broj slabijih/manje globalnih iskaza jedne polarnosti može da nadjača mali broj jačih/globalnijih iskaza druge polarnosti. Takođe, ako globalni stav nije više validan u posmatranom trenutku, onda prevagu odnosi jače izražen uži stav:

- **+M** Dobar film.. ali su efekti loši
- **-M** Ok gluma, ok tema, ali film kao film ništa specijalno!
- **-M** Odličan film i pored bednog izdanja Krisa Evansa. Stvarno je neupečatljiv i bez ikakvog pokušaja da dublje prikaže pravo psihološko stanje glavnog lika.
- **-M** Ovo je ekstremno glup film! Volim horor, o Kopoli ne treba trošiti reči, ali ovo… kakvo razočaranje…
- **-M** Upravo tako… Udavi nas inače nekad sjajni Kamerun.

Međutim, kada komentar sadrži mešavinu sentimenata prema istoj temi/aspektu, ali je jasno koji je sentiment jači, dodeljuje se ocena shodno tom jačem sentimentu, na primer (relevantni delovi komentara su podvučeni):

- **+1** Dobra recenzija. Prilično dobar film uz sve navedene nedostatke.
- **-1** Nije loš ali je za moj ukus previše mračan, počevši od priče pa sve do scenografije.
- **-1** Jer im nije trebao neko ko će dobro da odglumi već, svojom trenutnom popularnošću, da privuče pažnju. Nije film loš ali je poražavajuće prosečan.

Ovakve situacije se često dešavaju i kada se prvobitno izneta pohvala/kritika u ostatku komentara označi kao pogrešna ili nebitna, na primer:

- **-1** Jedan veoma bledi film. Izlizani kliše koji se ne razlikuje od jeftinih akcionih filmova iz devedesetih. Scenografija je lepa, ali to ne pomaže filmu. Bond je mrtav.
- **-1** Dosadan, čudan (ne na dobar način "čudan"), neinventivan, usiljen, par lepih kadrova, ali nedovoljno za pozitivan utisak, veliko razočarenje...
- **+1** poznata formula, poznati segmenti, a ipak odlično i efektno ukombinovani
- **+1** Izuzetan film, iako ima sitnijih propusta.. naučio sam da se 'konju ne treba gledati u zube' cela ideja mi se jako sviđa, i što je najbitnije film ima 'dušu'. =D
- **+1** Potpuno pogubljeno... Ali opet mi se sviđa… Nadam se da će ovaj režiser nastaviti u ovom stilu… u tom slučaju jedva čekam da vidim šta je sledeće…

Sličan efekat se može javiti i kada se prvo negira raniji iskaz, a zatim se navede još jači iskaz iste polarnosti, čime se stilski pojačava intenzitet celog komentara, na primer:

- **+1** Skorseze nije talentovan reditelj. On je legenda.
- **-1** Užas, jedan od retkih filmova koje nisam mogao da pogledam do kraja. Ovo definitivno nije osrednje - ovo je sramota…

Kada komentar sadrži i iskaze objektivne prirode i iskaze kojima se izražava neki vrednosni stav, ukupna ocena komentara zavisi samo od iskaza koji izražavaju vrednosni stav. U sledećim primerima takvi iskazi su podvučeni:

- **+1** Mogao bi da ispraviš ono "ujak" u "stric", jeste da englezi sve to svode pod "uncle" ali kod nas ujak nije isto što i stric i to, s obzirom na rodbinske šeme u filmu utiče na stvaranje slike kod nekoga ko film nije gledao. Inače, dobra kritika odličnog filma!
- **-1** Tristan je srednji sin a Alfred najstariji…. malo više studioznosti bi bilo OK, s obzirom da opisuješ filmove i daješ ocene i kritiku..
- **-M** prikaz onoga što očekujemo u video igrama budućnosti na PC-u koji neće biti sputavan hardverskim zahtevima konzola - real time ray tracing rendering. u svetu filma ovo je Crysis 1, super prikaz ali slaba priča i sve ostalo

Shodno tome, komentari anotirani prema vrednosnom stavu koji izražavaju (ocene **+1**/**-1**/**+M**/**-M**) mogu da sadrže i neke objektivne informacije, ali komentari koji se označavaju kao objektivni (ocene **+NS**/**-NS**) ne mogu da sadrže vrednosne stavove.

K3 – Tematska ravnopravnost

Kriterijum: Pri anotiranju sentimenta, tematika nekog dela komentara ne treba da ima uticaja na njegovu važnost. Drugim rečima, ne treba davati posebnu/veću težinu sentimentu izraženom prema filmovima na uštrb sentimenta izraženom prema pojmovima druge tematike. Na primer, u prvom sledećem komentaru se pohvaljuje film a kritikuje recenzija, dok je u naredna dva komentara prisutna obrnuta situacija:

- **+M** Najrealniji i najiskreniji film koji se bavi tematikom muško-ženskih odnosa koji sam gledala do sada. Perfektna gluma glavnih likova. Gorane omašio si ovaj put :)
- **-M** Film je smešno loše urađen – bez trunke duha i umetnosti, odlična recenzija
- **+M** Koliko si dobro opisao moje razočarenje ovim filmom da nemam ništa da dodam.

U slučajevima poput navedenih, izbor između oznaka **-M** ili **+M** treba da zavisi samo od toga da li je veći akcenat u globalu stavljen na pohvale ili na kritike, a ne od toga na koje elemente komentara se odnose kritike, a na koje pohvale.

Jedini izuzetak od navedenog pravila se odnosi na razdvajanje elemenata diskursa koji se odnose na realne aspekte, koji su predominantni, od onih koji se odnose na fiktivne, poput likova u filmu. Sentiment izražen prema fiktivnim aspektima se može koristiti samo ako daje indikaciju sentimenta prema realnim, a u suprotnom se ignoriše. Naravno, iz sadržaja komentara mora da bude jasno da se određeni deo komentara odnosi na fiktivne aspekte da bi se taj deo komentara ignorisao prilikom anotacije. Ako nije sasvim sigurno da je to slučaj, usvaja se konzervativan pristup i pretpostavlja se da se posmatrani deo komentara ipak odnosi na neki realan aspekt te se ravnopravno tretira pri anotaciji komentara.

Obrazloženje: Navedeni pristup je usvojen radi opšte primenjivosti uputstava za anotaciju, nezavisno od domena tekstova koji se razmatraju. Ignorisanje sentimenta izraženog prema fiktivnim aspektima

je primereno jer radnja nekvalitetnog filma često može da sadrži likove ili situacije koje se mogu smatrati dobrim, a radnja kvalitetnog filma likove ili situacije koje se mogu smatrati lošim, što se može videti na primeru sledećeg komentara:

- **+1** Odličan film! Majstorski odrađen! Upravo (i) zbog toga što nam nisu holivudski unapred opisali likove i dali nam njihove psiho karakteristike! Šta je sve jadna žena morala da trpi i da preživljava! Strava i užas! Vrhunski film.

K4 – Dualnost izražavanja sentimenta

Kriterijum: Iskazima koji istovremeno izražavaju i pohvalu i kritiku, upućenu različitim akterima ili aspektima, treba dodeliti ocenu **M**, pri čemu polarnost zavisi od toga da li iskaz više akcentuje pohvalu ili kritiku, na primer:

- **+M** Svaka čast svima koji su opljuvali recenziju!
- **+M** Zaslužuje veću ocenu.
- **-M** pa da, samo što mislim da nikada nije dobio pažnju koju zaslužuje
- **-M** Ovaj film ne može da ukapira svako!!! Pošto tera na razmišljanje on nije ni namenjen plitkim ljudima!!!

K5 – Jačina sentimenta

Kriterijum: Između tekstova iste polarnosti ne treba praviti razliku u ocenama na osnovu jačine izraženog sentimenta. Dakle, dva iskaza iste polarnosti ali sasvim različitih jačina sentimenta treba označiti istom ocenom. U skladu sa tim, stav da objekat komentarisanja jeste dobar ali nije idealan ne treba smatrati, samo po sebi, za njegov nedostatak. Numeričke ocene u opsegu 1–4 na skali od 1 do 10 treba tretirati kao negativne, ocene u opsegu 7–10 kao pozitivne, dok ocene 5 i 6 (a nekad i 7) mogu biti indikatori bilo jedne bilo druge polarnosti, u zavisnosti od preostalog sadržaja komentara.

- **+1** film je solidan
- **+1** Remek delo. Tačka!
- **+1** Daleko od toga da mi je loš, čak sam mu dao i visoku ocenu… ali mi nije najbolji.
- **+1** Ne znam generalno mi je cool, razumem taj igrica platforma build up, ali mi je možda samo još malo zapleta i neizvesnosti falilo da bude savršen.
- **-1** Film je apsolutna nula u svim aspektima.
- **-1** Film je mogao umesto sat i 30 min, da se lepo spakuje u jedan sat i bilo bi mnogo bolje.
- **-1** Realna ocena ovom filmu je 3/10, omanuo si totalno...
- **+1** ovo zaslužuje 10/10
- **-1** 6/10. To je dovoljno.
- **+1** Pogledao sam ga sada na Cinema City-ju i meni je film samo OK. Ocena 6/10
- **+1** Hvala na preporuci za ovaj film... dobro me nasmejao 7/10
- **-1** Ništa posebno. Pogotovo poslednja trećina filma gde Django glumi Ramba u fazonu "sve ću vas sam pobiti". 7/10

K6 – Autorstvo vrednosnog stava

Kriterijum: Pri anotaciji sentimenta kao relevantne vrednosne stavove treba uzimati samo one čiji je autor sam govornik, dok vrednosne stavove drugih lica treba tretirati samo kao objektivnu informaciju, pošto se često dešava da se govornik poziva na tuđe stavove, ali ih ne deli. Na primer, prva klauza narednog komentara izražava pozitivan vrednosni stav drugih osoba, ali se sa njim autor

komentara ne slaže. Slično tome, u drugom navedenom komentaru se navodi negativan vrednosni stav drugih osoba, ali on ne treba da utiče na ukupnu pozitivnu ocenu sentimenta komentara:

- **-1** Po filmskim forumima posvećenim horor žanru čuo sam dosta dobrog o ovom filmu, ali po mom mišljenju nikako ne zaslužuje te hvalospeve. Od mene ocena 4/10.
- **+1** Stvarno odličan film.. gledao sam više puta.. ostale delove nisam.. mada sam čuo da nisu nešto...

K7 – Slaganje i podrška / neslaganje

Kriterijum: Iskaze uopštenog slaganja/podržavanja nekog ranije izraženog mišljenja koje nije jasno formulisano treba, u odsustvu drugih indikatora stava, označiti ocenom **+M**, dok izraze uopštenog neslaganja treba označiti ocenom **-M**. Ipak, iskazi slaganja i podrške konkretnim idejama/predlozima su nedvosmisleno pozitivni, i treba ih označiti sa **+1**. U retkim slučajevima, slaganje sa nekim prethodnim iskazom može da se ne odnosi na bilo kakav stav, već samo na činjenice, kada ga treba označiti ocenom **NS**.

- **+M** potpuno se slažem sa tvojom konstatacijom i ja sam doneo isti zaključak posle sinoćnog gledanja filma.
- **+M** I ja sam ga isto tako doživela !
- **+M** Amin od prve do poslednje reči.
- **+1** slažem se sa ovom idejom o top listi!
- **+NS** Složio bih se tu… nije doslovce prikazan film, poseduje primetne alternacije u priči… ali sadrži isti duh i atmosferu.

Kada je pored iskaza slaganja ili neslaganja stav komentatora jasno prezentovan, bilo direktno bilo preko navođenja tuđeg stava sa kojim se komentator slaže/ne slaže, slaganje/neslaganje ne treba da, samo po sebi, utiče na ocenu sentimenta:

- **-1** uf! skroz si u pravu, nisam mogla da ga odgledam do kraja koliko me je smorio.
- **-1** Šokirati i tako privući pažnju, šok radi šoka. U pravu si ocena 1/10
- **+1** Baš tako, mene je isto oduševila. Ona i Dženifer Lorens su dva najveća talenta u poslednjih desetak godina sigurno.
- **-M** Slažem se sa svime izuzev sa komentarom 3D tehnike, 3D je ovde izvukao film, jedina svetla tačka u produkciji.
- **+M** slažem se da je dobar, ali Spilberg često zna da pretera i da napravi baš duge "evo vam plačite" scene, pa bih mu dao možda 8/10, nikako čistu desetku... moje mišljenje naravno ;)
- **-1** Uh ja ne mogu da se složim, ja sam iz bioskopa izašla mnogooo iznervirana jer pored načina snimanja i efekata cela priča je jedan blagi užas... ali kraj, kraj mi je bio vrhunac sa svom svojom patetikom. Mnogo sam očekivala a veoma malo dobila od filma, za mene veliko razočarenje.
- **+1** Ne slažem se sa ovom kritikom, smatram da je film perfektan.
- **+M** Uopšte se ne slažem. Jedan od najboljih filmova godine. Odličan Clooney (kojeg inače ne volim) i Shailene u ulozi starije ćerke. Dirljiv, tužan, smešan.... Odličan!

Obrazloženje: Iako za iskaze uopštenog slaganja/neslaganja nije moguće zasigurno znati da li se odnose na neki pozitivan ili negativan komentar, ipak se najčešće radi o podržavanju ili slaganju sa nekim vrednosnim stavom, te je stoga ocena **+/-M** najadekvatnija.

K8 – Potvrde i negiranja

Kriterijum: Za razliku od iskaza slaganja i neslaganja, potvrde i negiranja se često odnose na objektivne činjenice ili pitanja, te ih stoga, ako nisu propraćeni nekim jasnim vrednosnim stavom, treba označavati ocenama **NS**, na primer:

- **+NS** jeste, u Petak je bila premijera.
- **+NS** da, naravno da jesam.
- **-NS** Ovo nije režirao Shyamalan
- **-NS** Nisu isti autori na oba teksta.

K9 – Dvosmisleni iskazi

Kriterijum: Komentarima koji su dvosmisleni, tj. čija polarnost može biti i pozitivna i negativna u zavisnosti od šireg konteksta, treba zbog neodređenosti dodeliti ocenu **M**. Polarnost ocene **M** zavisi od procene toga koji je širi kontekst (pozitivan/negativan) verovatniji tj. prirodniji za dati komentar, što ponekad zavisi od stava autora komentara prema nekom drugom entitetu, na primer:

- **-M** Goran je u pravu, ovako nešto do sada sigurno niste gledali.
- **-M** Šta reći…
- **-M** E bukvalno je to to... Into the Wild 2
- **+M** Fala na recenziji :) ovo je još više potvrdilo moja očekivanja o filmu.
- **+M** jednom rečju ŠVEĐANI
- **+M** Dao bih sedmicu, ipak.

K10 – Citati

Kriterijum: Citati u okviru komentara se obično odnose na delove recenzija posmatranog filma i najčešće sadrže vrednosne stavove čiji autor nije autor komentara. Shodno kriterijumu K6 o autorstvu vrednosnog stava, sadržaj citiranih tekstova najčešće treba tretirati samo kao dodatnu objektivnu informaciju. Izuzetak od ovoga predstavljaju situacije kada govornik citira neku sopstvenu raniju izjavu – tada sadržaj citata treba tretirati ravnopravno sa ostalim elementima teksta. Vrednosni stav citiranog teksta takođe treba uzeti u obzir ako citiranje služi samo kao pojašnjenje ranije iznetog stava sa kojim se autor komentara slaže ili ne slaže, shodno kriterijumu K7 o slaganju i podršci / neslaganju.

K11 – Poređenja

Kriterijum: Prilikom poređenja, polarnost iskaza treba odrediti na osnovu odnosa između dela diskursa na kome leži glavni fokus i drugog elementa poređenja. Ukoliko se npr. u komentaru razmatra film A tj. na njemu leži glavni fokus, i ukoliko se naiđe na poređenje tipa *A je bolji od B* ili *B je gori od A*, gde je B neki drugi film, onda takav iskaz treba smatrati pozitivnim, jer je cilj poređenja da iskaže superiornost elementa diskursa koji je u fokusu nad drugim elementom. Slično tome, iskaze tipa *A je gori od B* i *B je bolji od A* treba smatrati negativnim jer je njihov cilj iskazivanje inferiornosti elementa diskursa koji je u fokusu u odnosu na drugi element. Naravno, ovo rezonovanje treba podjednako primenjivati ne samo na filmove već na sve teme/aspekte. Ilustracija ovakvog rezonovanja je data u sledećim primerima (relevantni delovi komentara su podvučeni):

- **+1** Po meni bolji od filma Warrior....
- **-1** pogledao sam ovaj rimejk filma iz devedesete sa Švarcenegerom meni se iskreno više dopada sa Švarcenegerom nego ovaj sa Kolinom Farelom
- **+M** dobar film ali ubedljivo najbolji džejms bond je Šon koneri

- **-M** Pogledao sam oba jedan za drugim… Bolji je onaj iz 1982… ovom novom jednostavno nešto fali… ali može da prođe za ubijanje vremena :D

Međutim, neretko se dešava da se u iskazu izađe iz okvira uskog poređenja posmatranog elementa sa nekim drugim i da se nezavisno od posmatranog elementa navode neki kvaliteti ili nedostaci tog drugog elementa. Takođe, dešava se i da poređenje posmatranog elementa sa nekim drugim nije direktno, već samo implicirano uvođenjem tog drugog elementa u diskurs. Tada ne treba više smatrati da je taj drugi element pomenut jedino u funkciji poređenja sa glavnim fokusom diskursa, te se i njegove dodatno navedene odlike moraju uzeti u razmatranje pri donošenju finalne ocene, na primer:

- **+M** OK je Contagion, ali realno je pomalo dosadan. Preporuka za Carriers, sličan film sa daleko manjim budžetom koji je na mene ostavio dosta jači utisak.
- **-M** Precenjen apsolutno. Lepa sela su mu najbolji film :) Posle njega ne znam šta je gore.

K12 – Upiti

Kriterijum: Ako ne sadrže neki vrednosni stav, pitanja i upite treba tretirati kao objektivne iskaze i označiti ocenom **NS**. Upite koji impliciraju neki stepen interesovanja treba označiti sa **+NS**, dok one koji impliciraju blagi stepen začuđenosti, neverice ili razočaranja treba označiti sa **-NS**:

- **+NS** a gde bi film mogao da se pogleda?
- **+NS** je l' ima još nekih tekstova o filmovima Almodovara?
- **-NS** Kako to da ti pišeš recenzije?
- **-NS** Sve te knjige ste pročitali?
- **-NS** Kažeš da je prvi deo veoma loš a dali ste mu ocenu 8.
- **-NS** Pa kako onda nije zaslužio 10/10?

K13 – Kurtoazni iskazi

Kriterijum: Kurtoazni iskazi predstavljaju formalnost u izražavanju i ne iskazuju bilo pozitivan bilo negativan stav. Stoga ih ako nisu propraćeni nekim jasnim vrednosnim stavovima treba označavati sa **+NS**, ako se radi o izrazima zahvalnosti, odnosno **-NS**, ako se radi o izvinjenjima:

- **+NS** Hvala na lepim rečima! Čitamo se!
- **+NS** Molim i drugi put, javi utiske.
- **-NS** 93. Izvini, ispravio sam

K14 – Najave / zainteresovanost / molbe / želje / zahtevi / sugestije / predlozi

Kriterijum: Najave govornika da će uraditi nešto, izražavanja zainteresovanosti za nešto, molbe, želje, zahteve, sugestije ili predloge da se nešto uradi treba sve anotirati kao objektivne iskaze, ocenom **+NS**, ukoliko nisu propraćeni nekim jasnim vrednosnim stavovima. Na primer:

- **+NS** e ovo sutra idem da gledam na spensu
- **+NS** Interesuje me gde može da se nađe ova knjiga, da se kupi kod nas? Nema veze ako je na engleskom… Hvala unapred!
- **+NS** Da li možeš da uradiš recenziju za film Bad Lieutenant? Unapred zahvalan, stalni čitač :)
- **+NS** voleo bih da se uradi recenzija za film za šaku dolara takođe od ovog reditelja sa Klint Istvudom u glavnoj ulozi
- **+NS** stavite u Kategorije - žanrovi i domaći filmovi

- **+NS** ne bi bilo loše kada bi na naslovnoj strani gde su svi filmovi stavio i ocene, da ne bi moralo da se ulazi u stranicu svakog filma kako bi se videla ista.
- **+NS** Ništa o filmu, samo da kažem da recenzije ostalih Hari Potera ne bi bile na odmet

Obrazloženje: Iako na prvi pogled može delovati da zainteresovanost ili želja da se nešto uradi nose pozitivnu konotaciju, malo dubljim razmatranjem se može zaključiti da je takav pristup problematičan. Naime, sama namera/želja da se obavi neka akcija ne znači nužno da autor u trenutku komentarisanja ima formiran vrednosni stav prema objektu akcije niti da će njegov stav prema objektu akcije biti nužno pozitivan nakon izvršenja akcije. Na primer, bilo bi pogrešno na osnovu želje da se odgleda film ili najave te akcije implicirati da je stav govornika o filmu nužno bilo pozitivan bilo negativan, ili da će on takav biti u budućnosti. Motivacija za interesovanje ili želju da se nešto uradi ne mora nužno biti pozitivna, već može, na primer, biti uzrokovana inatom:

- **+NS** da… ovo ti je najoštrija kritika do sada.. pored avatara.. :) baš ću da pogledam ovaj film jer me je motivisala recenzija.. :) inat je čudo

Slično važi i za generalnu zainteresovanost za neku temu – neko može biti npr. zainteresovan za rat kao istorijsku ili sociološku pojavu, ali to ne znači da ta osoba odobrava ratovanje. Kako molbe, zahtevi, sugestije i predlozi predstavljaju poruke upućene drugom licu u cilju ostvarenja neke sopstvene želje, za njih važe ista pravila kao i za najave, želje i iskaze zainteresovanosti. Ovo ne znači da autor komentara ne može da ima vrednosni stav prema objektu neke akcije i pre izvršenja same akcije, ali u tom slučaju taj vrednosni stav mora da bude jasno izražen, na primer:

- **+1** hm, ovo baš deluje interesantno. vredi gledati, rekla bih.

K15 – Žaljenje

Kriterijum: Iskazi žaljenja najčešće izražavaju jasan negativan sentiment, na primer:

- **-1** a šteta, baš sam se nadala da će biti kul…
- **-1** uh, Goran i ja gledali, nažalost… :)

Međutim, ponekad iskazi žaljenja mogu označavati i pozitivan sentiment. To se dešava kod iskaza u kojima je autoru krivo što neko pozitivno iskustvo nije moguće ponoviti ili kada govornik izražava da bi mu bilo žao ili da bi bila šteta ako se nešto ne bi uradilo ili bi se propustilo. Ovakvi iskazi su ponekad praćeni manjim ili većim prekorom upućenim onima koji ignorišu datu preporuku:

- **+1** uvek kad neko od drugara predloži ovaj film da se gleda, meni je krivo šta sam ga već gledao :D
- **+1** Nema komentara? Bilo bi stvarno šteta da ovaj film ostane nezapažen. Odlična zabava!
- **+M** Kreće film u redovno prikazivanje 17. Novembra. Greota je skinuti ovako dobar film. 300 din. je smešna cena za dobijeni doživljaj. Toliko košta kafa u Beogradu.
- **-M** Iskreno, velika je greota gledati bedni DVDScreener ovakvog filma, ako već postoji bioskop u gradu. Džast sejing.

K16 – Činjenični iskazi

Kriterijum: Prirodna ocena za komentare kojima se saopštavaju faktografske informacije jeste **NS**. Polarnost te ocene zavisi od procene toga kakav sentiment potencijalno leži iza konkretnog činjeničnog iskaza. Na primer, na osnovu prvog od navedenih komentara se može pretpostaviti potencijalno blago pozitivan sentiment prema filmu, dok se na osnovu drugog komentara može pretpostaviti blago nezadovoljstvo ili negativnost:

- **+NS** Baš to, film je slojevit, mora se pažljivo gledati, inače se lako propusti poruka koju šalje.
- **-NS** Taj im je još i najviše uspeo na box office-u... inače Koeni retko kada dobiju pohvale kod "šire" publike...

K17 – Redosled iskaza

Kriterijum: Ako su elementi pozitivne i negativne polarnosti podjednako brojni i jaki, tada redosled njihovog navođenja određuje koja polarnost preovlađuje. Kod naporednih rečenica sa suprotnim odnosom, iskazi koji se javljaju u drugoj klauzi imaju veću težinu od onih koji su prvi navedeni, na primer:

- **-M** Pa to, zabavan jeste, ali je već viđen… u potpunosti se slažemo :D
- **+M** mislim da je film nedovoljno opisao knjigu... ali ipak nije loš

Ako bi se redosled iskaza u ovim komentarima promenio, to bi dovelo i do promene globalne polarnosti komentara:

- **+M** Pa to, već je viđen, ali jeste zabavan… u potpunosti se slažemo :D
- **-M** mislim da film nije loš… ali je ipak nedovoljno opisao knjigu

K18 – Ton komentara

Kriterijum: Sentiment se ponekad može dedukovati na osnovu delova komentara koji ukazuju na ton govornika, i to obično kada je ton negativan. Tada i za komentar predominantno objektivnog sadržaja može da bude adekvatna ocena **-1**, na primer (relevantni delovi komentara su podvučeni):

- **-1** Glavni lik u filmu se zove Adrijan Kronauer, a ne Aron. Lepo piše na posteru.
- **-1** Još jedna stvar - nije imao nominaciju za Oscar. Ne znam otkud ti takva dezinformacija.
- **-1** Pa ne, tačno sam znala. Naši novinari su o tome pisali kao da je sporedna ženska uloga i nosilac zapleta. :)
- **-1** Pisao sam samo jedan od tih tekstova. "Body of Lies" je čist akcioni film, ne pretenduje da bude ništa više od toga. Ti čitaš uopšte ko je koji tekst pisao?
- **-1** Da li si uopšte pogledao film? Ako jesi, znaš odakle im hrana. Naravno da je sve ovo neizvodljivo, zato je i naučna fantastika. Podvučeno fantastika.

Kod nekih komentara, poštapalice, određene rečce/partikule, interpunkcija i ponavljanje karaktera odaju emotivni naboj i ton komentara, a ne sam sadržaj, dok je kod drugih ton komentara istaknut naglašavanjem određenih reči, na primer:

- **-1** Pa zašto se onda javljaš?
- **-1** 8/10 ?????????????? ako se ovaj film kotira ovako kako bi onda trebalo oceniti the dark knight rises, ili bilo koji iz te trilogije ?
- **-1** Ne, jer ne znam zašto je bitno ko ga je snimio. Valjda je bitno ZAŠTO je snimljen.

Ton komentara može ponekad nadjačati druga pravila anotiranja. Na primer, po kriterijumu K14 molbe treba ocenjivati sa **+NS**, ali ta ocena ipak može da varira u zavisnosti od tona komentara:

- **+NS** Može neko samo da mi objasni kraj filma? :)

- **-NS** da li neko može da mi objasni kakav ugovor je imao Louis Cyphre sa Johnny Favorite-om (angel heart-om)? kakvo prodavanje duše i šta je poenta svega?
- **-1** Ja ništa ne kapiram! Da li bi bio neko ljubazan da mi objasni šta se ovde dešava i ko je ko i šta je šta?!

K19 – Smeh

Kada se oznake za smeh, poput *Hahaha* ili *Hehehe*, javljaju same za sebe, bez ikakvog propratnog sadržaja, treba ih obeležiti sa **+NS**, jer nije moguće znati da li u takvom kontekstu uopšte imaju ikakvu konotaciju vezanu za sentiment teksta. Ukoliko je pak na osnovu konteksta jasno da je smeh izraz radosti ili uzbuđenja, treba ga označiti kao pozitivan iskaz, a ako je jasno da se radi o podsmevanju, treba ga označiti kao negativan iskaz.

K20 – Emotikoni

Kriterijum: U odsustvu drugih indikatora sentimenta, emotikone treba koristiti kao reper za označavanje sentimenta tekstova. Na primer, značenje narednih komentara bi u odsustvu emotikona bilo teško utvrditi, ali njihovo prisustvo jasno govori o sentimentu koji tekst izražava:

- **-1** tako sam i mislio :/
- **+1** ♥ Znao sam da će ovaj biti prvi! :)

K21 – Sarkazam

Kriterijum: Pod sarkazmom se za potrebe anotacije podrazumevaju kako sarkazam tako i ironija tj. stilske figure u kojima se značenje poruke, a naročito njena oznaka sentimenta, razlikuje u odnosu na ono što bi se dobilo bukvalnim razumevanjem napisanih reči. Sarkastični komentari po prirodi ne mogu da budu objektivni jer se njima izražava neki vrednosni stav.

Moguće je izdvojiti nekoliko karakterističnih primera javljanja sarkazma. Najpre, kod određenih komentara sarkazam je evidentan iz samog sadržaja koji bi u slučaju bukvalnog razumevanja poruke bio nelogičan ili besmislen, na primer:

- **-1s** Kolega, kako Vas nije sramota da unosite logiku i razum u ovu diskusiju! Znate vrlo dobro da tome ovde nije mesto.
- **-1s** Hvala što si mi pojasnio stvari koje moj prosečan um ne može da shvati. Što se tiče vizuelnog, bilo mi je nelogično da film smešten u dvadesete godine prošlog veka liči na epizodu "MTV Cribs", ali šta ja znam.

Neki sarkastični komentari su formulisani u obliku retoričkih pitanja:

- **-1s** Vau. Najbolja ekranizacija romana ikada snimljena? Znači, to govoriš iz svog iskustva kao neko ko je pogledao svaku snimljenu ekranizaciju romana ikada, u istoriji filma? Stvarno? Toliko je dobro?
- **-1s** Je li sa par piksela više film odjednom dobar?
- **-1s** Nisi gledao a sudiš?! Bravo…

Određene sarkastične komentare je moguće identifikovati preko tona govornika koji je ostao evidentan i u pisanom obliku, pri čemu ponekad na sarkastični ton ne ukazuje toliko značenje poruke koliko rečenična struktura, tj. neuobičajen redosled korišćenih reči:

- **-1s** To sigurno rečiti i pismeni čoveče. :)

- **-1s** Ako si mu presudio. Bravo za tebe. :)
- **-1s** Neće više glupih komentara biti. Sve ću ih banovati.

Oznakom za sarkazam treba označavati i one komentare kod kojih je samo deo teksta sarkastičan, na primer (sarkastični delovi su podvučeni):

- **-1s** "„Red Dragon“ pokriva period pre „The Silence of the Lambs“" ne pokriva on ništa on je samo rimejk Manhunter (1986) i ne može da prismrdi originalu odradi istraživanje kad si već toliki filmski stručnjak
- **-1s** @nenead222 ne poznaješ me, ne možeš izlaziti ovde sa takvim tvrdnjama, ako nemam isti ukus kao ti ne znači da sam kako kažeš filmski neobrazovan... bilo bi neinteligentno da ulazim u dalju polemiku sa tobom. Uživaj u mediokritetskim filmovima i dalje, pozdrav!

Iako ogromna većina sarkastičnih komentara nosi negativan sentiment, moguća je i upotreba ironije u pozitivnom kontekstu:

- **+1s** Moram da priznam da nisam pogledao prva dva dela, ali treći je fenomenalan. Poruka Pixaru: WALL-E, Up, Toy Story 3, pa dokle više??? Dajte, snimite neki osrednji film, da i drugi studiji imaju šansu..... :)

Ocenom **Ms** treba označiti ne samo tekstove kod kojih pored sarkazma, po kriterijumu kompozitnog ocenjivanja (K2), postoji i negativna i pozitivna komponenta, već i primere kod kojih nije moguće sa sigurnošću utvrditi da su sarkastični, ali za koje postoji snažna mogućnost da jesu:

- **-Ms** volim kad ste kolegijalni :)
- **-Ms** hvala bogu da je neko shvatio poentu filma…

Obrazloženje: Prepoznavanje sarkazma u tekstu, pogotovo bez uvida u okolni kontekst diskusije, ume da bude prilično težak problem čak i za ljude. Naime, u principu gotovo svaki napisani iskaz je potencijalno sarkastičan jer u govoru na sarkazam najčešće upućuje ton kojim je iskaz izgovoren. Informacija o tonu govornika se najčešće gubi prilikom prenosa iskaza iz govornog u pisani oblik, što čini identifikaciju sarkazma u pisanom obliku teškom. Stoga su uputstva za anotaciju usmerena na detekciju bar nekoliko karakterističnih primera javljanja sarkazma.

4.2.2 Analiza anotacije sentimenta kratkih tekstova iz SentiComments.SR korpusa

U ovom odeljku najpre je izložena analiza konzistentnosti anotacije sentimenta kratkih tekstova iz *SentiComments.SR* korpusa, a zatim je dat statistički prikaz kreiranih skupova podataka. Na kraju je analizirana efikasnost i ekonomičnost sprovedene anotacije pomoću nove metrike za merenje ekonomičnosti upotrebe uputstava za anotaciju na posmatranom zadatku.

4.2.2.1 Analiza konzistentnosti anotacije sentimenta kratkih tekstova iz SentiComments.SR korpusa

Pošto je glavni *SentiComments.SR* korpus zajednički anotiralo dvoje anotatora, za merenje njihove međusobne saglasnosti bilo je neophodno koristiti dodatne verifikacione skupove podataka, *SentiComments.SR.verif.movies* i *SentiComments.SR.verif.books*, čija je izrada opisana u poglavlju 3.2. Oba anotatora su samostalno označila sentiment tekstova iz ovih skupova. Pored toga, radi objektivne procene kvaliteta i razumljivosti novog sistema označavanja sentimenta, angažovano je još četvoro anotatora koji su samostalno anotirali verifikacione skupove podataka. Niko od dodatnih

anotatora nije imao ranijeg iskustva sa ovim tipom anotacije. Dvoje dodatnih anotatora su dobili celokupna uputstva (navedena u odeljku 4.2.1.1) i definicije svih oznaka sentimenta u novom sistemu označavanja (navedene na početku odeljka 4.2.1). Preostalo dvoje su dobili samo definicije oznaka, ali bez ikakvih smernica za tretiranje različitih jezičkih pojava i problematičnih situacija, što je značilo da pri anotiranju moraju da se oslone samo na sopstveno intuitivno razumevanje zadatka. Ovakva podela je sprovedena radi utvrđivanja inherentne razumljivosti korišćenog sistema oznaka, kao i korisnosti i ekonomičnosti korišćenja sastavljenih uputstava za anotaciju. Stoga su se izdvojile tri grupe od po dvoje anotatora:

1. Inicijalna grupa (IG) – anotatori koji su označili glavni *SentiComments.SR* korpus i koji su stoga bili dobro upoznati sa sistemom označavanja sentimenta i uputstvima za anotaciju;
2. Eksperimentalna grupa (EG) – dodatni anotatori koji su dobili celokupna uputstva za anotaciju i definicije oznaka sentimenta;
3. Kontrolna grupa (KG) – dodatni anotatori koji su dobili samo definicije oznaka sentimenta.

Sve tri grupe su najpre anotirale verifikacioni skup iz domena filmova, a zatim verifikacioni skup iz domena književnosti. Pre početka anotacije, anotatori iz eksperimentalne grupe su utrošili nekoliko sati na upoznavanje sa uputstvima. Na osnovu dobijenih anotacija, na oba korpusa je izračunata međusobna binarna saglasnost anotatora unutar svake grupe, izražena u vidu procentualnog slaganja i Kripendorfovog α koeficijenta. Preko α koeficijenta je izračunata i globalna saglasnost anotatora između svakog para grupa. Saglasnosti su izražene za više različitih interpretacija oznaka sentimenta:

1. Za označavanje polarnosti, gde se gleda samo +/**-** znak u oznaci;
2. Za označavanje subjektivnosti, gde se pravi razlika između objektivnih tekstova (oznake +/-**NS**) i subjektivnih tekstova (sve preostale oznake);
3. Za četvoroklasno označavanje sentimenta, gde se pravi razlika između pozitivnih tekstova (oznaka **+1**), negativnih tekstova (oznaka **-1**), tekstova koji su dvosmisleni ili izražavaju mešavinu sentimenata (oznake +/-**M**) i objektivnih tekstova tj. tekstova koji ne izražavaju bilo kakav sentiment (oznake +/-**NS**);
4. Za puno šestoklasno označavanje sentimenta, u kome se kao odvojene kategorije koriste sve oznake iz kreiranog sistema označavanja;
5. Za označavanje sarkazma, gde se pravi razlika između sarkastičnih tekstova (koji imaju nastavak **s** u oznaci) i nesarkastičnih (koji nemaju taj nastavak u oznaci).

Tabela 6 sadrži međusobne saglasnosti anotatora za sve ove interpretacije oznaka na verifikacionom skupu iz filmskog domena, dok tabela 7 prikazuje međusobne saglasnosti na verifikacionom skupu iz književnog domena. Za razliku od procentualnih poklapanja, Kripendorfov α koeficijent ne uračunava slučajne saglasnosti, te će stoga diskusija vezana za konzistentnost anotacije sentimenta biti fokusirana na ovu metriku.

Razmatrajući najpre verifikacioni skup iz filmskog domena, može se videti da se najveći stepeni saglasnosti postižu na najjednostavnijem zadatku označavanja polarnosti, što nije iznenađujuće. Na ovom zadatku razlike između grupa anotatora su relativno male, i anotatori postižu odlične međusobne saglasnosti unutar svih grupa ($\alpha \geq 0{,}8$). Kako se složenost interpretacija oznaka povećava, tako i stepeni saglasnosti počinju da opadaju, a odstupanja između kontrolne grupe i preostale dve grupe anotatora postaju sve izraženija. Ove razlike počinju da bivaju jasne na označavanju subjektivnosti, a kulminiraju na označavanju sarkazma. Anotatori iz inicijalne grupe konzistentno dostižu najviše međusobne saglasnosti, koje za sve interpretacije oznaka spadaju u odlične ($\alpha \geq 0{,}8$), što je za očekivati zbog njihovog obimnijeg prethodnog iskustva u anotaciji ovog tipa. Rezultati eksperimentalne grupe su nešto niži – saglasnosti su odlične na dve interpretacije oznaka, prihvatljive ($\alpha \geq 0{,}667$) na još dve, i neprihvatljive na označavanju sarkazma. Kontrolna

grupa na većini zadataka ima najniže nivoe saglasnosti anotatora, pri čemu je označavanje polarnosti jedini izuzetak, i jedini slučaj gde su saglasnosti unutar ove grupe odlične. Na ostalim zadacima saglasnosti spadaju u prihvatljive, osim na označavanju sarkazma, gde su saglasnosti izuzetno loše.

Što se saglasnosti između grupa tiče, uočava se da su globalne saglasnosti četvoro anotatora koji pripadaju inicijalnoj i eksperimentalnoj grupi konzistentno najviše, što je prirodna posledica oslanjanja na isti skup detaljnih uputstava. Zapravo, za sve interpretacije oznaka, ova kombinacija grupa postiže odlične stepene međusobne saglasnosti ($\alpha \geq 0{,}8$), osim označavanja sarkazma, gde je saglasnost na granici prihvatljivosti. Za preostale dve kombinacije grupa globalne saglasnosti anotatora su primetno niže (sa izuzetkom označavanja polarnosti) i međusobno dosta slične, što ukazuje na odskakanje oznaka koje su dodelili anotatori iz kontrolne grupe.

Rezultati na verifikacionom skupu iz književnog domena prate iste obrasce, ali u poređenju sa prethodnim rezultatima dodatno demonstriraju važnost dva efekta – promene domena i sticanja iskustva u anotaciji. Naime, saglasnosti anotatora u inicijalnoj grupi su odlične i dosta slične na oba skupa podataka. Nasuprot tome, kontrolna grupa pokazuje primetno lošije rezultate na književnom domenu, sa neprihvatljivim nivoima saglasnosti za sve interpretacije oznaka osim označavanja polarnosti, što je indikator negativnog uticaja promene domena na sprovođenje anotacije bez sveobuhvatnih smernica za anotaciju. Sa druge strane, eksperimentalna grupa postiže značajno bolje saglasnosti na skupu iz književnog domena, gde ponekad dostiže ili čak prestiže inicijalnu grupu, i postiže odlične stepene saglasnosti za sve interpretacije oznaka. Ovo pokazuje da novi sistem označavanja nije vezan samo za jedan domen, već je podjednako primenjiv na različite domene. To takođe demonstrira da je sistem dovoljno intuitivan i jasan da i novi anotatori u okviru njega mogu brzo, već posle par stotina anotiranih primera, dostići nivoe saglasnosti koji su u kvalitativnom smislu bliski onima između znatno iskusnijih anotatora, i to uprkos promenama domena podataka.

Tabela 6. Međusobne saglasnosti anotatora u označavanju sentimenta kratkih tekstova na srpskom jeziku iz *SentiComments.SR.verif.movies* korpusa, izražene u vidu procentualnih vrednosti i Kripendorfovog α koeficijenta

Interpretacija oznaka	Saglasnost unutar grupa						Saglasnost između grupa		
	IG		EG		KG		IG i EG	IG i KG	EG i KG
	%	α	%	α	%	α	α	α	α
Označavanje polarnosti	0,966	0,929	0,933	0,861	0,948	0,887	0,895	0,874	0,857
Označavanje subjektivnosti	0,989	0,896	0,976	0,795	0,970	0,725	0,823	0,754	0,748
Četvoroklasno označavanje sentimenta	0,955	0,934	0,873	0,814	0,815	0,697	0,853	0,724	0,721
Šestoklasno označavanje sentimenta	0,922	0,892	0,821	0,750	0,802	0,679	0,801	0,687	0,678
Označavanje sarkazma	0,991	0,829	0,983	0,628	0,974	0,131	0,658	0,391	0,396

Tabela 7. Međusobne saglasnosti anotatora u označavanju sentimenta kratkih tekstova na srpskom jeziku iz *SentiComments.SR.verif.books* korpusa, izražene u vidu procentualnih vrednosti i Kripendorfovog α koeficijenta

Interpretacija oznaka	Saglasnost unutar grupa						Saglasnost između grupa		
	IG		EG		KG		IG i EG	IG i KG	EG i KG
	%	α	%	α	%	α	α	α	α
Označavanje polarnosti	0,977	0,935	0,977	0,935	0,908	0,731	0,935	0,802	0,807
Označavanje subjektivnosti	0,971	0,929	0,954	0,889	0,838	0,520	0,880	0,661	0,625
Četvoroklasno označavanje sentimenta	0,965	0,948	0,902	0,852	0,751	0,570	0,869	0,700	0,657
Šestoklasno označavanje sentimenta	0,948	0,924	0,884	0,832	0,711	0,517	0,848	0,664	0,623
Označavanje sarkazma	0,994	0,931	1,000	1,000	0,977	0,324	0,859	0,559	0,544

U pogledu saglasnosti između grupa primećuju se isti obrasci kao i na verifikacionom korpusu iz filmskog domena. Pri tome, na podacima iz književnog domena globalne saglasnosti anotatora koji pripadaju inicijalnoj i eksperimentalnoj grupi su primetno više u odnosu na ostale dve kombinacije grupa i na zadatku označavanja polarnosti. Zapravo, na književnom domenu ova kombinacija grupa postiže odlične stepene međusobne saglasnosti anotatora na svim interpretacijama oznaka.

4.2.2.2 Statistički prikaz anotiranih korpusa kratkih tekstova SentiComments.SR

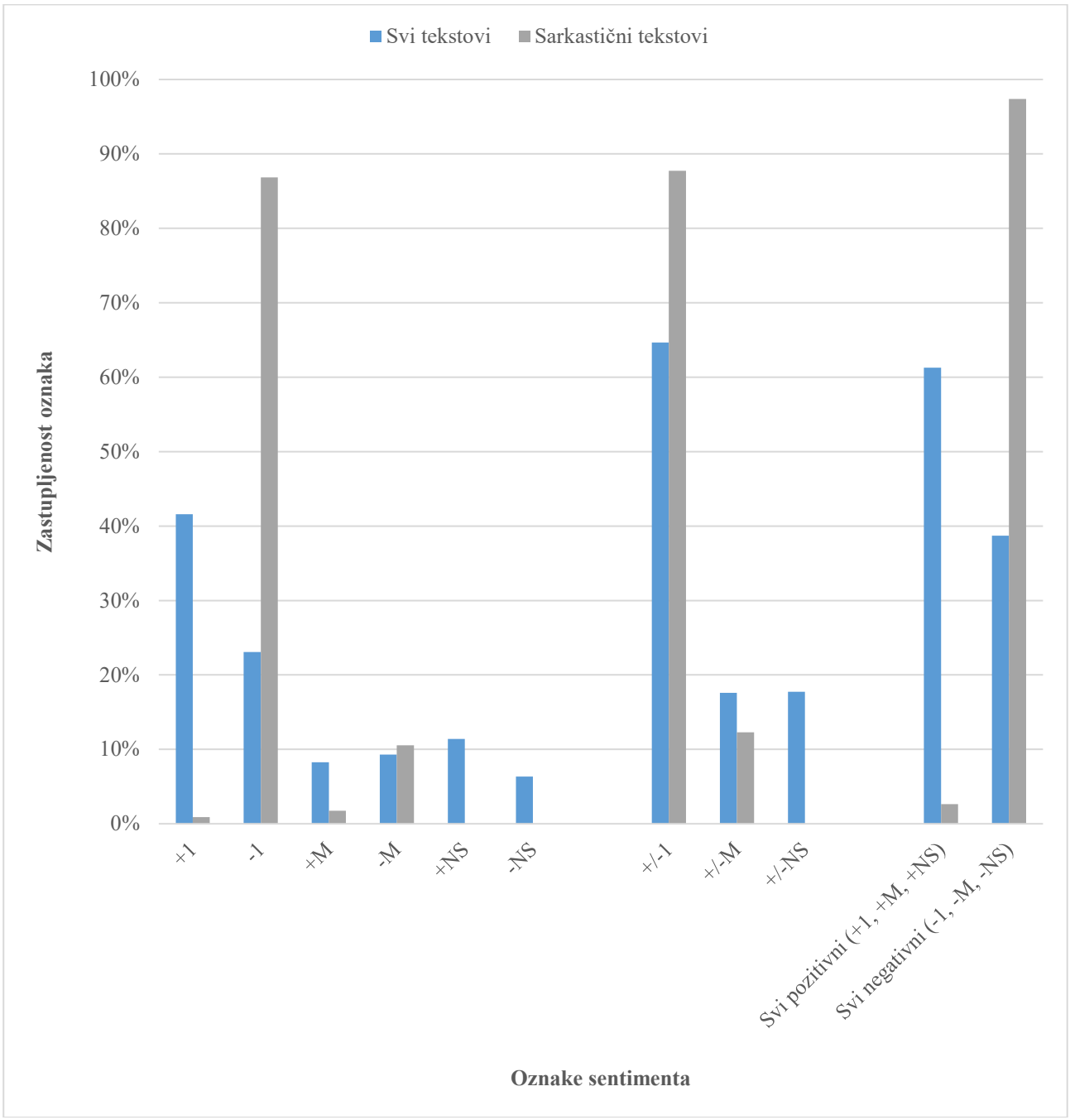
Kao što je već pomenuto u poglavlju 3.2, glavni skup podataka *SentiComments.SR* sadrži 3490 kratkih komentara. Njihova raspodela po oznakama sentimenta je prikazana na slici 8, kako za celokupan korpus, tako i za podskup sarkastičnih komentara. Detaljni numerički podaci o ovoj raspodeli su navedeni u tabeli P3 u prilozima ove disertacije, gde su pored procentualnih zastupljenosti za svaku oznaku ili skup oznaka date i apsolutne numeričke vrednosti.

Kao što se vidi sa slike, *SentiComments.SR* korpus je neizbalansiran, pri čemu su brojniji tekstovi jasne polarnosti, naročito pozitivni (oznaka **+1**). Veća zastupljenost pozitivnih tekstova je takođe evidentna i u okviru **NS** podgrupe, dok su tekstovi sa oznakama **M** prilično ravnomerno raspodeljeni između dve polarnosti. Korpus sadrži 114 tekstova koji su označeni kao sarkastični, odnosno oko 3,27% od ukupnog broja. Primetno je da je velika većina sarkastičnih tekstova negativna, što je i za očekivati. Slično tome, većina takvih tekstova je jasne polarnosti, dok je oko 10% dvosmisleno ili izražava mešavinu sentimenata.

Slika 9 ilustruje raspodelu tekstova iz verifikacionog skupa iz filmskog domena po oznakama sentimenta, dok slika 10 prikazuje tu raspodelu za tekstove iz verifikacionog skupa iz književnog domena. Pošto su ove korpuse odvojeno anotirali svi anotatori, kod njih su oznake sentimenta

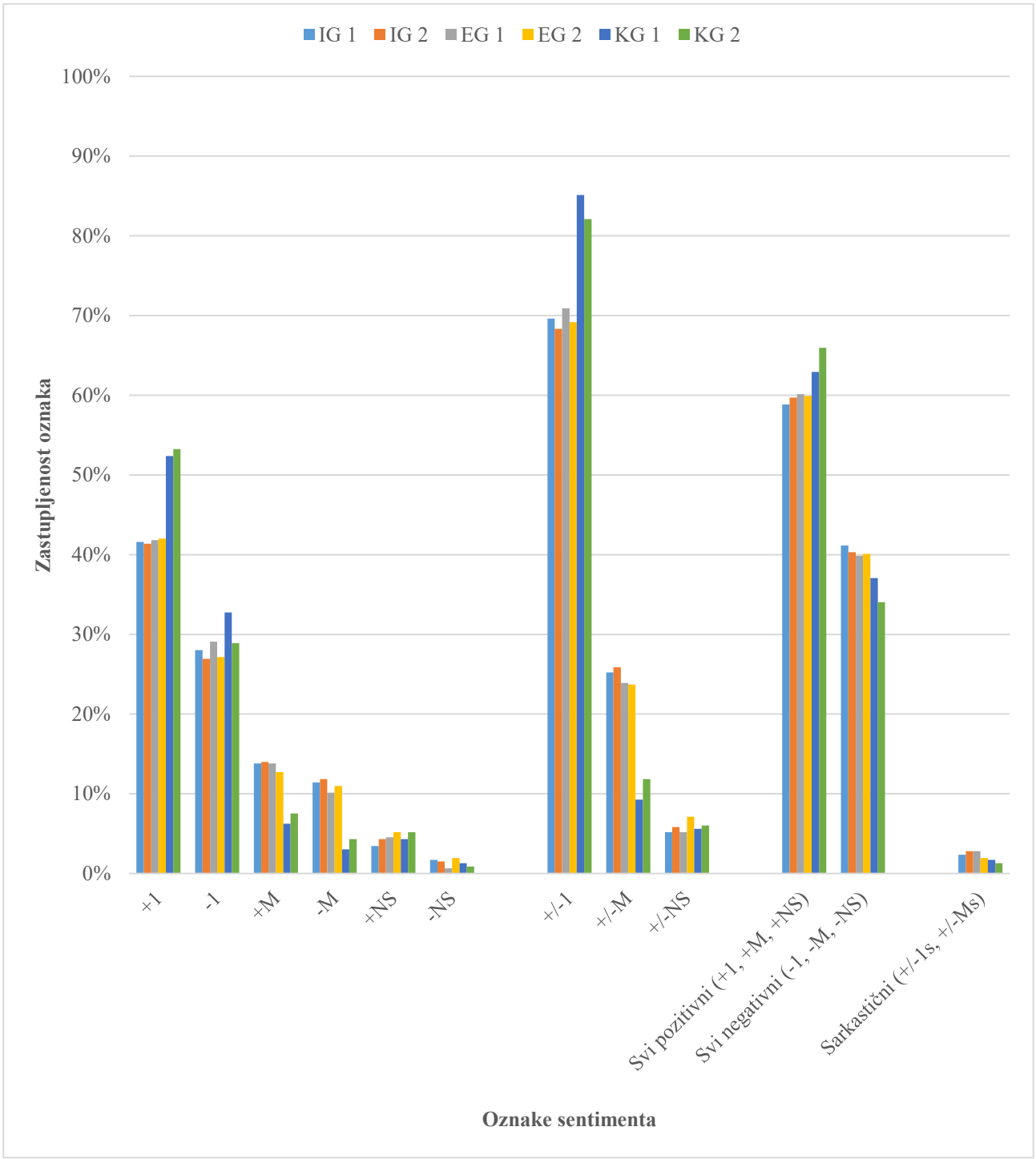
prikazane za svakog anotatora individualno. Detaljni numerički podaci vezani za ove raspodele su prikazani u prilozima disertaciji, u tabelama P4 i P5.

Verifikacioni skup iz filmskog domena se odlikuje sličnim odnosom između pozitivnih i negativnih oznaka kao i glavni *SentiComments.SR* korpus. Sa druge strane, **NS** oznake su ređe nego u glavnom korpusu, dok su **M** oznake češće, sa izuzetkom anotatora iz kontrolne grupe. Oznake **-1** su takođe blago frekventnije u verifikacionom skupu, dok je zastupljenost sarkastičnih komentara nešto niža nego u glavnom korpusu. Što se razlika između grupa anotatora tiče, uočljivo je da su anotatori iz kontrolne grupe bili skloniji da koriste **+1** oznaku, a da su se ređe odlučivali za **M** oznake nego ostali anotatori. Između inicijalne i eksperimentalne grupe anotatora nema primetnih razlika, što je očekivani ishod upotrebe istih uputstava za anotaciju.

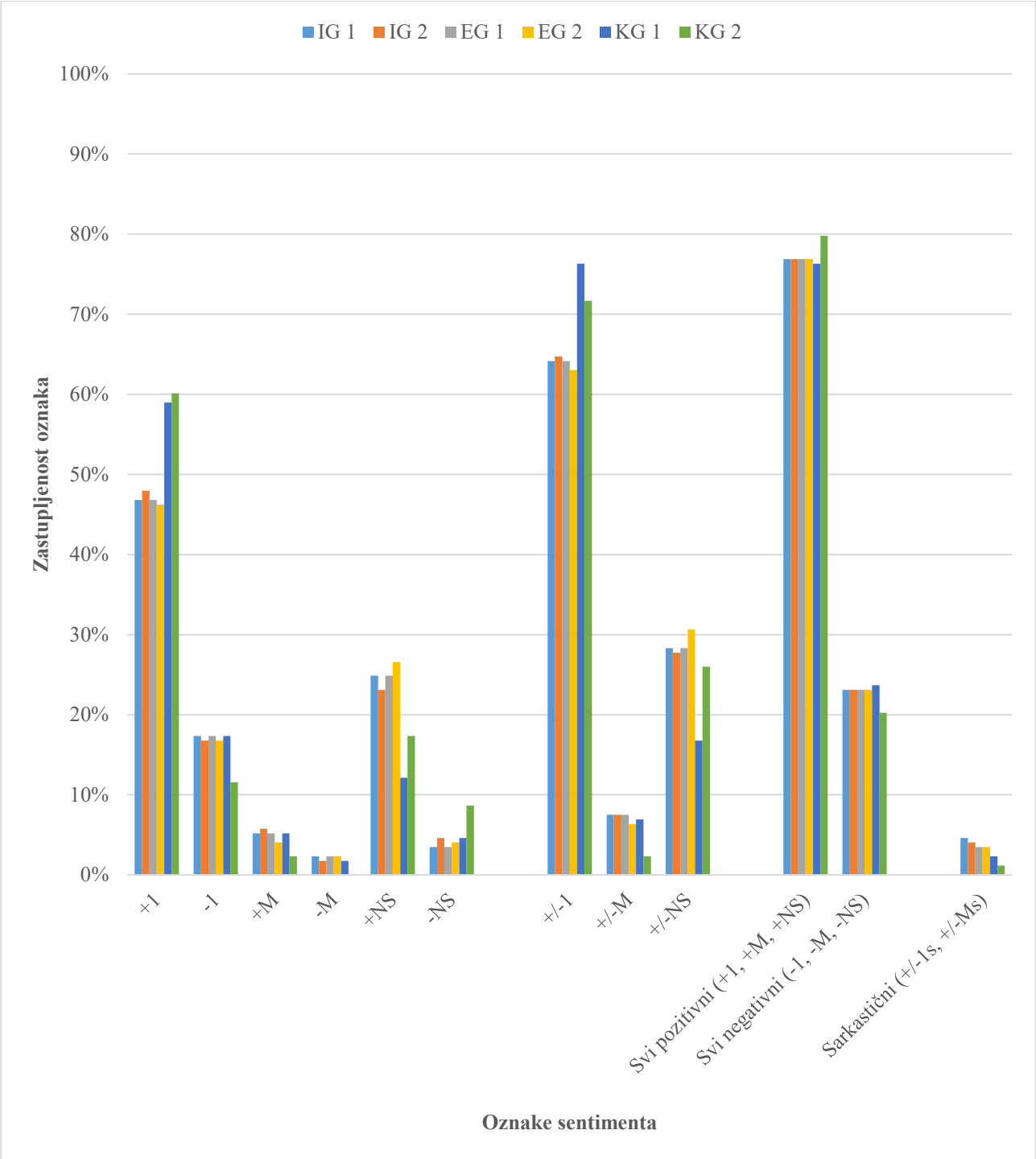


Slika 8. Raspodela tekstova iz glavnog *SentiComments.SR* korpusa po oznakama sentimenta

Verifikacioni skup iz književnog domena ima veći ukupni procenat pozitivnih oznaka nego prethodna dva korpusa, prvenstveno zbog većeg udela oznaka **+1** i **+NS**. Kao posledica toga, udeo preostale četiri oznake sentimenta je manji nego u prethodnim korpusima. Procenat sarkastičnih tekstova je sličan onome u glavnom *SentiComments.SR* korpusu. U pogledu razlika između grupa anotatora, anotatori iz kontrolne grupe su bili skloniji da upotrebljavaju oznaku **+1**, dok su oznaku **+NS** koristili ređe od preostala četiri anotatora. Takođe su bili manje skloni uočavanju sarkastičnih tekstova. Kao i kod verifikacionog skupa iz filmskog domena, ni ovde se ne javljaju primetnije razlike između anotatora u inicijalnoj i u eksperimentalnoj grupi.



Slika 9. Raspodela tekstova iz *SentiComments.SR.verif.movies* korpusa po oznakama sentimenta



Slika 10. Raspodela tekstova iz *SentiComments.SR.verif.books* korpusa po oznakama sentimenta

4.2.2.3 Analiza efikasnosti i ekonomičnosti anotacije sentimenta kratkih tekstova iz SentiComments.SR korpusa

Da bi se utvrdila ekonomičnost upotrebe kreiranog sistema označavanja, mereno je vreme potrebno za anotaciju oba verifikaciona korpusa za sve grupe anotatora. U tabeli 8 su prikazana prosečna trajanja anotacije i prosečne brzine anotatora iz svake grupe.

Tabela 8. Prosečna trajanja anotacije i prosečne brzine anotatora u anotaciji sentimenta kratkih tekstova na srpskom jeziku iz verifikacionih korpusa

Grupa anotatora	*SentiComments.SR.verif.movies* (464 komentara)		*SentiComments.SR.verif.books* (173 komentara)	
	Prosečno trajanje anotacije	Prosečna brzina anotacije	Prosečno trajanje anotacije	Prosečna brzina anotacije
IG	~6 h	~77 tekstova/h	~2 h	~87 tekstova/h
EG	~9 h	~52 tekstova/h	~3 h	~58 tekstova/h
KG	~3,5 h	~133 tekstova/h	~1,25 h	~138 tekstova/h

Anotatori iz eksperimentalne grupe su u proseku bili oko dva i po puta sporiji od onih iz kontrolne grupe. Inicijalna grupa je bila oko 50% brža od eksperimentalne, što pokazuje da se, uz dovoljno prethodnog iskustva, smernice za anotaciju mogu uspešno internalizovati, čime se povećava brzina anotacije.

Prethodni radovi iz oblasti analize i anotacije sentimenta ne pružaju podatke o efikasnosti anotacije podataka, zbog čega nije moguće u ovom pogledu uporediti novi sistem označavanja sa ranijim pristupima. Međutim, efikasnost anotacije je vrlo važna u jezicima sa ograničenim resursima, jer odluka o izboru sistema označavanja zavisi dobrim delom i od raspoloživih ljudskih i materijalnih resursa koje je neophodno utrošiti na sprovođenje anotacije. Naravno, pored efikasnosti anotacije važan je i njen kvalitet. Odgovarajuća metrika za ekonomičnost upotrebe određenog sistema označavanja, koja bi uzela u obzir i efikasnost i kvalitet anotacije, bi stoga bila od velike koristi, ali takva metrika nije do sada predstavljena.

Iz ovog razloga, u ovoj disertaciji je heuristički definisana metrika za određivanje ekonomičnosti anotacije. Ova metrika, označena sa *ACE* (engl. *Annotation Cost-Effectiveness*), definiše se kao odnos između relativnog smanjenja nesaglasnosti anotatora i relativnog smanjenja efikasnosti anotacije, do kojih dolazi upotrebom određenih uputstava za anotaciju. Ukoliko je ovaj odnos veći od jedan, tj. ako su povećanja međusobne saglasnosti anotatora značajnija od usporenja anotacije, upotreba posmatranih uputstava se može smatrati ekonomičnom. Kao polazna tačka u računanju ovakve metrike bi se koristila anotacija sprovedena bez uputstava i smernica.

Za računanje relativnog smanjenja nesaglasnosti anotatora, može se koristiti Kripendorfov α koeficijent na sledeći način:

$$\Delta A_1 = \frac{(1-\alpha_{KG}) - (1-\alpha_{EG})}{(1-\alpha_{KG})} = \frac{\alpha_{EG} - \alpha_{KG}}{1-\alpha_{KG}}$$

U navedenom izrazu, α_{KG} je međusobna saglasnost anotatora u kontrolnoj grupi, tj. kod anotatora kojima nisu data uputstva za anotaciju, izražena u vidu Kripendorfovog α koeficijenta. Slično tome, α_{EG} je međusobna saglasnost anotatora u eksperimentalnoj grupi, tj. kod anotatora kojima jesu data uputstva. $1-\alpha$ predstavlja nivo međusobne nesaglasnosti anotatora unutar određene grupe, pošto je jedan maksimalna vrednost za α. Konačno, ΔA_1 predstavlja relativno smanjenje nesaglasnosti anotatora. Maksimalna vrednost relativnog smanjenja nesaglasnosti je takođe jedan, i dobija se u slučaju kada je $\alpha_{EG} = 1$.

Navedeno relativno smanjenje nesaglasnosti anotatora uzima kao ciljnu gornju vrednost jedino maksimalnu teoretsku vrednost saglasnosti, tj. jedan. Međutim, Kripendorf sugeriše i dva praga u tumačenju vrednosti α koeficijenta (Krippendorff 2004). Po njemu, ako je $\alpha < 0{,}667$, tada saglasnost nije prihvatljiva. Ako je pak $0{,}667 \leq \alpha < 0{,}8$, tada se saglasnost može smatrati prihvatljivom, dok ako je $\alpha \geq 0{,}8$, tada je saglasnost pouzdana. Stoga povećanje saglasnosti koje dovodi do prelaska ovih pragova, a time i do drugačijeg tumačenja kvaliteta saglasnosti, treba dodatno nagraditi pri formulisanju sveukupnog relativnog smanjenja nesaglasnosti anotatora. To je učinjeno tako što se za svaku vrednost praga računa i relativno smanjenje nesaglasnosti u odnosu na njega, ukoliko je početna vrednost saglasnosti, tj. saglasnost kontrolne grupe, ispod nivoa datog praga. Pritom, ukoliko je nova vrednost saglasnosti, tj. saglasnost eksperimentalne grupe, iznad nivoa određenog praga, tada pri računanju relativnog smanjenja nesaglasnosti u odnosu na taj prag ne treba uzimati u obzir saglasnost eksperimentalne grupe, već vrednost praga, da bi se obezbedilo da relativno smanjenje ne može imati vrednost preko 1:

$$\Delta A_{0,8} = \frac{\min\{\alpha_{EG},\ 0{,}8\} - \alpha_{KG}}{0{,}8 - \alpha_{KG}}$$

$$\Delta A_{0,667} = \frac{\min\{\alpha_{EG},\ 0{,}667\} - \alpha_{KG}}{0{,}667 - \alpha_{KG}}$$

Ukupno relativno smanjenje nesaglasnosti anotatora se dobija kao prosek relativnih smanjenja u odnosu na sve ciljne vrednosti koje su iznad početne vrednosti saglasnosti, tj. iznad saglasnosti kontrolne grupe, dokle god je saglasnost eksperimentalne grupe veća od saglasnosti kontrolne grupe. U suprotnom, metrika se računa samo u odnosu na teoretski maksimalnu saglasnost, pošto su navedeni pragovi definisani u odnosu na pretpostavljeno nižu vrednost saglasnosti kontrolne grupe:

$$\Delta A = \begin{cases} \frac{1}{3}\left(\Delta A_1 + \Delta A_{0,8} + \Delta A_{0,667}\right), & \alpha_{KG} \leq \alpha_{EG}, \quad \alpha_{KG} < 0{,}667 \\ \frac{1}{2}\left(\Delta A_1 + \Delta A_{0,8}\right), & \alpha_{KG} \leq \alpha_{EG}, \quad 0{,}667 \leq \alpha_{KG} < 0{,}8 \\ \Delta A_1, & \alpha_{KG} \leq \alpha_{EG}, \quad \alpha_{KG} \geq 0{,}8 \\ \Delta A_1, & \alpha_{KG} > \alpha_{EG} \end{cases}$$

Relativno smanjenje efikasnosti anotacije se može izraziti na sličan način:

$$\Delta E = \frac{S_{KG} - S_{EG}}{S_{KG}}$$

U navedenom izrazu, S_{KG} predstavlja brzinu anotacije u kontrolnoj grupi, S_{EG} je brzina anotacije u eksperimentalnoj grupi, a ΔE je relativno smanjenje efikasnosti anotacije. Za vrednost brzine anotacije koristi se prosečan broj tekstova anotiranih na sat. Naravno, ovo pretpostavlja da su tekstovi koji se anotiraju slične dužine, ali to i jeste najčešće slučaj pri anotaciji.

Konačno, metrika za određivanje ekonomičnosti se matematički definiše kao:

$$ACE = \frac{\Delta \mathrm{A}}{\Delta E}$$

Ako je $ACE \geq 1$, upotreba uputstava za anotaciju jeste ekonomična, dok za $ACE < 1$ korišćena uputstva i smernice nisu ekonomični. Ako je $ACE < 0$, tada su najverovatnije međusobne saglasnosti anotatora niže kada se uputstva primenjuju, nego kada se ne primenjuju ($\Delta A < 0 \Rightarrow \alpha_{EG} < \alpha_{KG}$).

Alternativni uzrok ovakvih vrednosti *ACE* metrike bi teoretski moglo da bude povećanje ne samo stepena saglasnosti anotatora, već i brzine anotacije ($\Delta E < 0 \Rightarrow S_{EG} > S_{KG}$), do kojih je došlo upotrebom uputstava za anotaciju. Međutim, ovakav ishod je vrlo malo verovatan u praksi. Matematički je moguće i da obe ove pojave budu tačne u isto vreme ($\Delta A < 0,\ \Delta E < 0$), u kom slučaju bi *ACE* bilo veće od nule, ali je ovakva situacija takođe vrlo malo verovatna, jer bi značila da korišćenje uputstava snižava saglasnost anotatora, ali povećava brzinu anotacije.

ACE metrika se može koristiti i za poređenje dva skupa uputstava za anotaciju. Radi poređenja potrebno je samo da se kao kontrolna grupa koriste anotatori koji su upotrebljavali ona uputstva koja dovode do nižeg stepena međusobne saglasnosti. Pored toga, zahvaljujući fleksibilnosti Kripendorfovog α koeficijenta, *ACE* metrika se može koristiti ne samo za kategoričke tipove oznaka, već i za redne, numeričke, itd, kao i za proizvoljno velike grupe anotatora.

Na osnovu opisanih matematičkih izraza, međusobnih saglasnosti anotatora na verifikacionim korpusima koje su prikazane u tabelama 6 i 7, kao i prosečnih brzina anotacije predstavljenih u tabeli 8, izračunate su vrednosti *ACE* metrike za sve interpretacije oznaka sentimenta u kreiranom sistemu označavanja. Vrednosti ove metrike su prikazane u tabeli 9.

Tabela 9. Vrednosti predložene *ACE* metrike za određivanje ekonomičnosti anotacije, izračunate za anotaciju sentimenta kratkih tekstova na srpskom jeziku iz verifikacionih korpusa

Interpretacija oznaka	*SentiComments.SR.verif.movies*		*SentiComments.SR.verif.books*	
	IG vs KG	EG vs KG	IG vs KG	EG vs KG
Označavanje polarnosti	0,883	-0,378	2,379	1,517
Označavanje subjektivnosti	1,926	0,975	2,572	1,592
Četvoroklasno označavanje sentimenta	2,116	1,138	2,597	1,527
Šestoklasno označavanje sentimenta	1,975	0,663	2,564	1,525
Označavanje sarkazma	2,219	1,227	2,614	1,725

ACE metrika pokazuje da jedini slučaj kada je upotreba uputstava za anotaciju bila kontraproduktivna ($ACE < 0,\ \Delta A < 0$) jeste označavanje polarnosti koje je eksperimentalna grupa sprovela nad verifikacionim korpusom iz filmskog domena. Uzrok ovoga verovatno delom leži u ograničenom iskustvu koje je eksperimentalna grupa imala u tom trenutku, ali i u većoj jednostavnosti zadatka označavanja polarnosti, koja je učinila upotrebu uputstava neekonomičnom čak i za inicijalnu grupu ($ACE < 1$). Inicijalna grupa postiže vrednosti *ACE* metrike veće od jedan za sve ostale interpretacije oznaka na verifikacionom korpusu iz filmskog domena. Nasuprot tome, za eksperimentalnu grupu korišćenje sastavljenih uputstava za anotaciju se ne pokazuje ekonomičnim ni pri šestoklasnom označavanju sentimenta na istom korpusu, dok je označavanje subjektivnosti na samoj granici ekonomičnosti. Verovatan uzrok ovakvih rezultata jeste kompleksnost korišćenog sistema oznaka i vreme potrebno da anotatori steknu dovoljno iskustva u radu sa njim. Naime, na verifikacionom korpusu iz književnog domena, koji je kasnije anotiran, ekonomičnost anotacije je postignuta i za

inicijalnu i za eksperimentalnu grupu, i to na svim interpretacijama oznaka. Pored toga, konzistentan rast vrednosti *ACE* metrike za obe grupe pri prelasku sa verifikacionog korpusa iz filmskog domena na onaj iz književnog domena pokazuje da promena domena podataka čini korišćenje uputstava za anotaciju još ekonomičnijim, i to na svim interpretacijama oznaka sentimenta, čak i na označavanju polarnosti. Najveći stepen ekonomičnosti se konzistentno dobija na zadatku označavanja sarkazma, što odgovara kompleksnosti tog zadatka i važnosti davanja jasnih uputstava anotatorima kako da mu pristupe.

4.2.3 Poređenje kreiranog sistema označavanja sentimenta kratkih tekstova sa ranijim rešenjima

Iako je kreirani sistem označavanja sentimenta kratkih tekstova osmišljen tako da bude pogodan za upotrebu u jezicima sa ograničenim resursima, skup korišćenih oznaka nužno ima nekih sličnosti sa ranije razvijenim rešenjima. Stoga će u ovom odeljku biti detaljno upoređen novi sistem označavanja sentimenta sa prethodnim poznatijim pristupima za višeklasno označavanje sentimenta.

Glavna osobina po kojoj se sistem označavanja predstavljen u ovoj disertaciji razlikuje od prethodnih jeste fleksibilnost koju nudi u pogledu interpretacije oznaka sentimenta. Neki od ranijih pristupa pružaju mogućnost da se pojedine klase sentimenta spoje nakon anotiranja (što je npr. urađeno u (Mohammad et al. 2017) sa sistemom oznaka predstavljenim u (Mohammad 2016)), ali takve pojave su retke. Pritom, uputstva za anotaciju korišćena u ovim radovima nisu osmišljena sa naknadnim kombinovanjima i spajanjima klasa na umu, što dovodi u pitanje validnost takvih postupaka u smislu ispravnosti finalnih ocena sentimenta za sve označene tekstove.

Sistem oznaka koji je predložio (Mohammad 2016) omogućava anotatorima da lako označe neke problematične iskaze, ali im ne pruža smernice za druge, poput npr. dvosmislenih iskaza, za koje postoji namenska oznaka (**M**) u novom sistemu označavanja. Pored toga, Mohammad se fokusira na vokabular prisutan u datom tekstu, što može biti prepreka kod kratkih tekstova, jer njihov sentiment često ne zavisi samo od korišćenog vokabulara, već i od drugih faktora poput tona iskaza. Iako usmeravanje na vokabular u principu olakšava anotaciju sentimenta, ono može navesti i na pogrešne zaključke. Na primer, Mohammad predlaže da se iskazi iz sportskog domena poput *A je pobedio B* označavaju kao pozitivni, a iskazi poput *A je izgubio od B* kao negativni, zbog korišćenja glagola koji se mogu smatrati primerima vokabulara sa pozitivnim, odnosno negativnim sentimentom. Međutim, pri ovakvom rezonovanju postaje nejasno kako bi trebalo označiti iskaze poput *A je porazio B*, pošto su takvi iskazi semantički identični sa *A je pobedio B*, ali bi se zbog upotrebe glagola *poraziti* (ili, alternativno, imenice *poraz*) mogli legitimno smatrati primerima vokabulara sa negativnim sentimentom. Drugim rečima, u ovakvom sistemu postoji opasnost da anotatori dodele drugačije oznake sentimenta iskazima koji su semantički ekvivalentni. Nasuprot tome, u sistemu označavanja izrađenom u ovoj disertaciji oba navedena primera bi bila označena na isti način, na osnovu kriterijuma K11 o poređenju, predstavljenog u odeljku 4.2.1.1. Sa druge strane, treba istaći da su ovde predstavljena uputstva značajno duža od onih koja je izložio Mohammad, što podrazumeva duži period obučavanja anotatora i čini ih neadekvatnim za *crowdsourcing* metode anotacije.

Sistem označavanja koji su predstavili (Abdul-Mageed & Diab 2011) je napravljen za potrebe označavanja sentimenta tekstova iz novinskog domena, koji su po prirodi pretežno objektivni. Stoga je njihov sistem u više pogleda primetno drugačiji od onog predstavljenog u ovoj disertaciji, koji je pogodniji za tekstove iz domena koji nemaju inherentnih ograničenja u pogledu izražavanja sentimenta. Najistaknutija razlika između ova dva pristupa je u tome što Abdul-Mageed i Diab ne uzimaju u obzir samo stav govornika pri određivanju sentimenta. Umesto toga, oni se oslanjaju na koncept privatnih stanja (engl. *private states*) (Quirk et al. 1985) – stanja koja se ne mogu direktno

verifikovati – te stoga razmatraju i sentimente i stavove osoba koje nisu autori posmatranog teksta. Drugim rečima, tekst koji sadrži pozitivna privatna stanja bi bio svrstan u njihovu subjektivno-pozitivnu kategoriju, bez obzira na to o čijim privatnim stanjima se radi. Na primer, u njihovom sistemu označavanja, rečenica *Nade za oslobađanje talaca oživele u poslednja 24h zbog intervencije Libije* (primer preveden iz (Abdul-Mageed & Diab 2011)) bi bila primer subjektivno-pozitivne klase. Nasuprot tome, u sistemu označavanja predstavljenom u ovoj disertaciji, ovakva rečenica bi dobila oznaku **+NS**, jer ona samo prenosi sentiment nekih drugih ljudi, dok stav govornika ostaje nepoznat (kriterijum K6). Sličan primer koji bi dobio oznaku **-NS**, odnosno koji bi Abdul-Mageed i Diab kategorizovali u svoju klasu subjektivno-negativnih iskaza, jeste sledeća rečenica: *Tursko ministarstvo spoljnih poslova je izjavilo da Turska sa velikom zabrinutošću prati terorističke napade koji su se prethodnih dana desili u Uzbekistanu i Kirgiziji.* Abdul-Mageed i Diab su spekulacije o budućnosti svrstavali u subjektivno-neutralnu kategoriju. Takvi iskazi bi u novom sistemu označavanja najčešće bili anotirani ocenom **NS**, kao na primer sledeća rečenica: *Svi su pokazatelji da se situacija neće promeniti nakon izbora*. Abdul-Mageed i Diab su na isti način obeležavali tvrdnje, kao što je sledeća rečenica: *Već sam rekao i ponavljam, problem nije u sirovoj nafti već u naftnim derivatima.* Uzevši u obzir jasan negativan stav koji govornik izražava prema naftnim derivatima, ovakav iskaz bi bio obeležen ocenom **-1** u novom sistemu označavanja. Abdul-Mageed i Diab takođe pominju ulogu ilokucionih govornih činova (engl. *illocutionary speech acts*) (Searle 1975), poput ekspresiva, u izražavanju sentimenta. U sistemu označavanja razvijenom u ovoj disertaciji se takođe razmatra njihov tretman, sa posebnim smernicama za različite vrste izraza (za izraze zahvalnosti – kriterijum K13, za sugestije – kriterijum K14, za žaljenja – kriterijum K15, itd.).

U svom kasnijem radu (Abdul-Mageed & Diab 2012), isti autori su proširili svoja uputstva za anotaciju smernicama vezanim za slaganje i neslaganje, kao i smernicama zasnovanim na teoriji učtivosti (Brown & Levinson 1987). Njihov pristup je da se iskazi slaganja automatski označavaju kao pozitivni, a iskazi neslaganja ili neodobravanja kao negativni, u slučaju direktnog neslaganja, ili neutralni, u slučaju indirektnog/ublaženog neslaganja. Pristup predstavljen u ovoj disertaciji je fundamentalno drugačiji, jer se tu izrazi slaganja/neslaganja u odsustvu drugih indikatora vrednosnog stava obeležavaju sa **+/-M**, pošto se tada ne može znati sa kojim sentimentom se govornik slaže (kriterijum K7). Naravno, ako se govornik slaže/ne slaže sa jasno izraženim vrednosnim stavom, onda se takvom iskazu dodeljuje ocena **+/-1**. Konačno, u slučajevima kada je jasno da se slaganje/neslaganje odnose samo na neke objektivne informacije i ocena **+/-NS** može biti adekvatna. Dodatna razlika u odnosu na pomenuti raniji sistem označavanja se tiče obeležavanja zahteva, koji se, sami za sebe, u novom sistemu obeležavaju sa **+NS** (kriterijum K14). Sa druge strane, zbog oslanjanja na teoriju učtivosti, Abdul-Mageed i Diab dele zahteve na direktne, koji u njihovom sistemu spadaju u negativnu klasu, i indirektne/ublažene zahteve, koji u njihovom sistemu spadaju u pozitivnu klasu.

Za razliku od pristupa u ovoj disertaciji, gde se dvosmisleni iskazi obeležavaju sa **M**, (Al-Twairesh et al. 2017) takve iskaze svrstavaju u posebnu kategoriju za neodređene iskaze, koja je odvojena od kategorije za mešavinu sentimenata. Pored toga, njihova uputstva nalažu da se, u slučaju javljanja emotikona čija je polarnost suprotna polarnosti ostatka teksta, tekst označi kao da ispoljava mešavinu sentimenata. U novom sistemu označavanja to ne bi bila automatska odluka, jer se često dešava da takve diskrepance ukazuju na sarkazam, koji bi onda trebalo adekvatno obeležiti. Al-Twairesh et al. takođe razmatraju neke karakteristične pojave koje su njihovim anotatorima zadavale teškoće u radu, poput citata i obraćanja Bogu (koja takođe pominje i (Mohammad 2016)). Citati su obrađeni posebnim smernicama i u sistemu kreiranom u ovoj disertaciji (kriterijum K10), dok obraćanja Bogu nisu, jer su se u razmatranim podacima na srpskom jeziku takva obraćanja konzistentno koristila za jasno izražavanje negativnog sentimenta (npr. *O bože pa dokle više s tim Titanikom. Ovo samo na brzo premotavanje može da se gleda*).

5 Formulisanje, obučavanje i evaluacija modela

U ovoj glavi prikazana je faza formulisanja, obučavanja i evaluacije modela, kao i rezultati evaluacije svih modela. Ponovo je najpre razmotren problem određivanja semantičke sličnosti kratkih tekstova, a zatim problem analize sentimenta. Za oba problema, prvo su formulisana i evaluirana osnovna rešenja, koja se uglavnom odlikuju većim stepenom jednostavnosti i služe kao polazišna referentna tačka u ispitivanjima složenijih pristupa. Ova rešenja se takođe odlikuju širokom primenjivošću, jer su sintaksno agnostička, te stoga ne zahtevaju posebne sintaktičke alate vezane za određeni jezik. Nakon toga su razmotrena moguća unapređenja osnovnih modela, korišćenjem različitih tehnika pretprocesiranja tekstualnih podataka, kao i bazičnih jezičkih alata i korpusa koji su neretko raspoloživi i u jezicima sa ograničenim resursima.

Kao jedna vrsta unapređenja, evaluirana je upotreba statičkih vektora značenja reči, kao vida transfernog učenja tj. prenošenja šireg znanja o značenju reči na konkretan NLP problem koji se razmatra. Vektori značenja reči su konstruisani na osnovu najvećeg javno dostupnog korpusa tekstova na srpskom – srpskog veb korpusa *srWaC* (Ljubešić & Klubička 2014), koji sadrži oko 555 miliona tokena uključujući interpunkciju. Nakon uklanjanja interpunkcije i normalizacije na mala slova dobijen je korpus od oko 470 miliona reči i vokabular od oko 3,8 miliona odrednica.

Pored toga, razmotreno je poboljšanje performansi modela putem morfološke normalizacije tekstova. Naime, srpski jezik spada u jezike sa bogatom morfologijom, zbog čega se jedna ista reč može javiti u velikom broju oblika, čime se povećava proređenost već ograničenog skupa anotiranih podataka, a time i otežava sposobnost generalizacije modela. U cilju minimizacije ovog efekta, tipično se koriste alati za morfološku normalizaciju, poput stemera, koji normalizaciju vrše odsecanjem krajeva reči, i lematizatora, koji normalizaciju vrše zamenom različitih oblika reči njenim osnovnim, rečničkim oblikom, tj. lemom. Pronađeno je više javno dostupnih rešenja koja se mogu primeniti za stemovanje i lematizaciju tekstova na srpskom jeziku, i to:

- Optimalni i pohlepni algoritmi stemovanja predstavljeni u (Kešelj & Šipka 2008);
- Unapređenje pohlepnog algoritma Kešelja i Šipke (Milošević 2012);
- Stemer za hrvatski jezik koji su razvili Ljubešić i Pandžić[21] i koji je unapređenje pristupa predstavljenog u (Ljubešić et al. 2007);
- *BTagger* lematizator za srpski, koji je dostupan u dve varijante – jednoj u kojoj se vrši normalizacija samo sufiksa reči (Gesmundo & Samardžić 2012a), i drugoj u kojoj se normalizuju i prefiksi, tj. vrši se potpuna lematizacija (Gesmundo & Samardžić 2012b);
- Lematizacioni model za hrvatski jezik (Agić et al. 2013) razvijen za *CST* lematizator (Jongejan & Dalianis 2009);
- *ReLDI* lematizator za srpski jezik, koji koristi veliki flektivni leksikon (Ljubešić et al. 2016).

Komparativna evaluacija ovih rešenja je ranije sprovedena samo u kontekstu dužih dokumenata, i to na problemu određivanja polarnosti teksta (Batanović & Nikolić 2016, 2017). Stoga je odlučeno da se svi navedeni alati razmotre u kontekstu pretprocesiranja podataka za probleme određivanja semantičke sličnosti i analize sentimenta kratkih tekstova. Radi lakšeg sprovođenja ove komparacije, svi navedeni stemeri su reimplementirani u programskom jeziku Java u vidu zajedničkog paketa *SCStemmers* (Batanović et al. 2016), koji je objavljen na posebnom *GitHub* repozitorijumu[22]. Naime, stemeri Kešelja i Šipke su izvorno bili predstavljeni u vidu skripti za programski jezik Perl,

[21] http://nlp.ffzg.hr/resources/tools/stemmer-for-croatian/
[22] https://vukbatanovic.github.io/SCStemmers/

Miloševićev stemer je implementiran u jeziku PHP[23], dok je stemer Ljubešića i Pandžića napisan u Python-u, što je otežavalo postupak njihovog poređenja. Što se lematizatora tiče, zbog njihove složenosti i oslanjanja na druge biblioteke, izrada objedinjenog paketa lematizatora nije bila praktična, te su u evaluaciji korišćene izvorne implementacije.

Kao tehnika pretprocesiranja tekstova takođe je razmotreno i jednostavno obeležavanje negiranih reči koje je prvobitno predloženo u (Pang et al. 2002), radi njihovog razlikovanja od nenegiranih oblika. Ono se zasniva na dodavanju jedinstvenog negacionog prefiksa na tokene koji se u tekstu javljaju nakon negacija, te stoga ne zahteva nikakvu sintaktičku analizu teksta, što je pogodno za jezike sa ograničenim resursima. Navedena tehnika je već poboljšala rezultate analize sentimenta na srpskom u domenu filmskih recenzija (Batanović et al. 2016) i tvitova (Ljajić & Marovac 2019), te je bilo logično ispitati njenu učinkovitost i nad novim skupom podataka za ovu problematiku, kao i na problemu određivanja semantičke sličnosti.

Pored unapređenja osnovnih modela, za problem određivanja semantičke sličnosti kratkih tekstova je formulisano i nekoliko namenskih novih modela. Kao što je prikazano u odeljku 5.1.4, iako su predloženi novi modeli bolji od osnovnih pristupa, oni ipak ne dostižu aktuelne najbolje neuralne modele, te stoga novi pristupi ovog tipa nisu razvijeni za problem analize sentimenta.

Na kraju, za rešavanje oba razmatrana semantička problema evaluirani su i najnoviji neuralni pristupi u vidu jezičkih modela zasnovanih na *Transformer* arhitekturama, koji su fino podešeni korišćenjem kreiranih anotiranih skupova podataka na srpskom jeziku. Među njima, pronađena su četiri koja su prethodno obučena na višejezičnim podacima i koja podržavaju i srpski jezik:

- *Multilingual BERT* (Devlin et al. 2019) – u ovoj disertaciji je korišćena novija varijanta ovog modela koja uzima u obzir kapitalizaciju slova i ima 12 slojeva (*BERT Base Multilingual Cased*), dok vektori imaju 768 dimenzija. Model je prethodno obučen na objedinjenom sadržaju Vikipedija na prvih 104 jezika sa najvećim Vikipedijama, uključujući srpski jezik.
- *Multilingual distilBERT* (Sanh et al. 2019) – model dobijen kroz proces kompresije (engl. *distillation*) *multilingual BERT* modela. Kompresija modela (Bucilă et al. 2006; Hinton et al. 2015) predstavlja postupak putem koga je moguće napraviti kompaktniji model koji je obučen da reprodukuje ponašanje polaznog većeg modela ili polaznog ansambla modela. Konkretno, *multilingual distilBERT* model je putem kompresije redukovan na 6 slojeva, uz istu dimenzionalnost vektora, i podržava istih 104 jezika.
- *XLM* (Conneau & Lample 2019) – u ovoj disertaciji je korišćena *MLM* (*Masked Language Modeling*) varijanta *XLM* (*Cross-lingual Language Model*) modela, sa 16 slojeva i vektorima od 1280 dimenzija, koja je prethodno obučena na 100 jezika, uključujući srpski.
- *XLM-RoBERTa* (Conneau et al. 2020a) – novija i naprednija varijanta *XLM* modela koja je obučena na znatno većem korpusu od 2,5 TB pročišćenih *Common Crawl* tekstova na 100 jezika (Wenzek et al. 2019), uključujući srpski. Dostupna u *Base* varijanti sa 12 slojeva i vektorima od 768 dimenzija i *Large* varijanti sa 24 sloja i vektorima od 1024 dimenzije.

Međutim, *XLM-RoBERTa* model ipak nije mogao da bude uključen u evaluaciju zbog stalnih problema sa konvergencijom nad podacima iz kreiranih skupova podataka na srpskom jeziku. Stoga su u ovoj disertaciji prva tri navedena neuralna modela evaluirana i upoređena sa ostalim pristupima. Korišćene su implementacije modela iz *HuggingFace Transformers* biblioteke (Wolf et al. 2019), kojima je pristupano putem *Simple Transformers* interfejsa[24].

[23] Autor ovog stemera je naknadno predstavio i verziju na programskom jeziku Python.
[24] https://github.com/ThilinaRajapakse/simpletransformers

5.1 Formulisanje, obučavanje i evaluacija modela za problem određivanja semantičke sličnosti kratkih tekstova

Obučavanje i evaluacija modela za problem određivanja semantičke sličnosti kratkih tekstova na srpskom jeziku sprovedeni su korišćenjem anotiranih podataka iz korpusa *STS.news.sr*, predstavljenog u poglavlju 4.1. Kao metrika za merenje performansi upotrebljavan je Pirsonov koeficijent korelacije između ocena sličnosti koje su izlaz modela i uprosečenih ocena sličnosti dobijenih putem anotacije, jer je to ustaljena metrika za ovu problematiku (Agirre et al. 2012, 2013, 2014, 2015, 2016; Cer et al. 2017).

U ostatku ovog poglavlja, najpre je prikazana evaluacija osnovnih modela za određivanje semantičke sličnosti kratkih tekstova i njihovih unapređenja, a zatim su opisani novi namenski modeli za rešavanje ovog problema koji su razvijeni u okviru ove disertacije. Svi modeli su implementirani u zajedničkom paketu *STSFineGrain* (Batanović et al. 2018a), koji je napisan na programskom jeziku Java i sadrži i zajednički okvir za evaluaciju algoritama za određivanje semantičke sličnosti kratkih tekstova. Pomenuti paket je javno dostupan na posebnom *GitHub* repozitorijumu[25]. Na kraju ovog poglavlja razmotreno je fino podešavanje neuralnih jezičkih modela i izvršeno je poređenje performansi različitih grupa modela.

5.1.1 Osnovni modeli za određivanje semantičke sličnosti kratkih tekstova i njihova unapređenja

U ovom odeljku su predstavljeni osnovni modeli za određivanje semantičke sličnosti kratkih tekstova, njihova unapređenja i rezultati. Kao polazni model upotrebljen je nenadgledan pristup koji je u tom svojstvu korišćen i u okviru *SemEval* takmičenja u rešavanju ovog problema (Agirre et al. 2012, 2013, 2014, 2015, 2016; Cer et al. 2017). Radi se o jednostavnoj tehnici traženja podudarnih reči između zadatog para tekstova, gde se tekstovi prvo tokenizuju korišćenjem blanko znakova, a onda prevode u vektorski oblik po principu vreće reči (engl. *bag-of-words*). U ovom modelu, tekst se predstavlja u vidu *n*-dimenzionalnog vektora, gde je *n* veličina vokabulara u okviru datog para tekstova, a svaka dimenzija vektora odgovara jednoj od reči iz vokabulara. Binarna vrednost svake dimenzije vektora predstavlja indikaciju prisustva ili odsustva odgovarajuće reči u određenom tekstu, a sličnost tekstova se izražava kao kosinusna sličnost njihovih vektora. Ovaj pristup je zatim unapređen normalizacijom tekstova na mala slova, uklanjanjem interpunkcije i upotrebom tokenizatora za srpski jezik[26] koji je razvijen u okviru *Regional Linguistic Data Initiative* (ReLDI) projekta (Samardžić et al. 2015). Pošto se taj tokenizator pokazao kao dobro rešenje, korišćen je u pripremi podataka i za sve naredne modele.

Drugi, nešto složeniji osnovni pristup koji je razmotren jeste model u kome se tekstovi takođe predstavljaju u vektorskom obliku, ali gde se ti vektori dobijaju uprosečavanjem statičkih vektora značenja reči prisutnih u tekstu. Vektori značenja reči su napravljeni pomoću *word2vec* algoritma (Mikolov et al. 2013a, b), i to njegove implementacije u *gensim* biblioteci (Řehůřek & Sojka 2010). *Word2vec* algoritam je odabran za izradu vektora značenja reči jer je pokazano da je ovaj model, tačnije njegova *skip-gram* varijanta, bolji od alternativnih, kada se kreirani vektori koriste na opisani način u okviru problematike određivanja semantičke sličnosti (Cer et al. 2017). Kao što je već napomenuto, vektori značenja su izrađeni na osnovu srpskog veb korpusa *srWaC* (Ljubešić & Klubička 2014), pri čemu je dimenzionalnost vektora postavljena na 100, dok je kontekstni opseg širine 10 reči. Svi ostali parametri su zadržani na podrazumevanim (engl. *default*) vrednostima.

[25] https://vukbatanovic.github.io/STSFineGrain/
[26] https://reldi.spur.uzh.ch/hr-sr/tokenizator-hrvatski-srpski/

Pored ova dva osnovna pristupa, evaluirana je i njihova kombinacija, u kojoj se tekstovi predstavljaju u vektorskom obliku kao konkatenacija obe opisane vektorske reprezentacije. Za sve navedene modele, istražena je i korisnost tehnike obeležavanja negiranih reči, po kojoj se prvoj reči nakon svake negacije dodaje poseban prefiks *NE_*. Kod modela koji koriste vektore značenja reči, ova tehnika je primenjena i na sadržaj *srWaC* korpusa pri izradi vektora.

Kako opisani pristupi predstavljaju nenadgledane modele, njihova evaluacija je sprovedena na celokupnom *STS.news.sr* korpusu. Rezultati evaluacije su predstavljeni u tabeli 10.

Tabela 10. Rezultati osnovnih modela za određivanje semantičke sličnosti kratkih tekstova na *STS.news.sr* korpusu, izraženi u vidu Pirsonovog koeficijenta korelacije r

Model	r
Podudarnost reči (tokenizacija pomoću blanko znakova)	0,6461
Podudarnost reči (ReLDI tokenizator)	0,6869
Podudarnost reči (ReLDI tokenizator + obeležavanje negiranih reči)	0,6862
Prosek vektora značenja reči (ReLDI tokenizator)	0,6211
Prosek vektora značenja reči (ReLDI tokenizator + obeležavanje negiranih reči)	0,6257
Podudarnost reči + prosek vektora značenja reči (ReLDI tokenizator)	**0,6949**
Podudarnost reči + prosek vektora značenja reči (ReLDI tokenizator + obeležavanje negiranih reči)	0,6943

Rezultati pokazuju da model podudarnosti reči dovodi do boljih performansi u odnosu na upotrebu proseka vektora značenja reči. Ovo je indikator višeg stepena leksičkog poklapanja između rečenica u *STS.news.sr* korpusu, što je očekivano zbog prirode postupka prikupljanja izvornih parova tekstova, opisanog u poglavlju 3.1. Ipak, kombinacija dve vrste vektorskog predstavljanja rečenica nadmašuje obe varijante pojedinačno. Tehnika obeležavanja negiranih reči se pokazuje korisnom samo pri korišćenju isključivo proseka vektora značenja reči kao vektorske reprezentacije teksta, dok u modelu podudarnosti reči i u kombinovanom modelu dovodi do lošijih rezultata. Stoga ova tehnika nije primenjivana u daljim eksperimentima.

Nakon ovoga, razmotren je efekat različitih alata za morfološku normalizaciju, predstavljenih u uvodu glave 5, na ponašanje osnovnih modela, a rezultati su prikazani u tabeli 11. Oznaka (S) u tabelama znači da se radi o stemeru, dok oznaka (L) predstavlja lematizatore. Što se vektora značenja reči tiče, ispitivanje efekata morfološke normalizacije je sprovedeno konstruisanjem odvojenih skupova vektora za svaku morfološki normalizovanu varijantu *srWaC* korpusa. Na osnovu rezultata se vidi da primena morfološke normalizacije ima konzistentan pozitivan efekat na model podudarnosti reči, kao i na kombinovani model, i konzistentno negativan efekat na pristup zasnovan samo na vektorima značenja reči. U proseku, stemeri bolje utiču na performanse modela od lematizatora. Kao najbolje rešenje se izdvaja stemer za hrvatski jezik Ljubešića i Pandžića, ali je razlika između njega i optimalnog stemera Kešelja i Šipke mala. Lematizator koji su predstavili Ljubešić et al. se u ovom kontekstu izdvaja kao najbolje rešenje za lematizaciju, ali ipak ne dovodi do tako dobrih rezultata kao najbolja dva stemera.

Tabela 11. Efekti metoda morfološke normalizacije na rezultate osnovnih modela za određivanje semantičke sličnosti kratkih tekstova na *STS.news.sr* korpusu, izražene u vidu Pirsonovog koeficijenta korelacije r

Metod morfološke normalizacije	Model		
	Podudarnost reči	Prosek vektora značenja reči	Podudarnost reči + prosek vektora značenja reči
Bez morfološke normalizacije	0,6869	0,6211	0,6949
(S) Kešelj & Šipka – optimalni	0,7291	0,5971	0,7338
(S) Kešelj & Šipka – pohlepni	0,7218	0,5966	0,7271
(S) Milošević	0,7210	0,5986	0,7266
(S) Ljubešić i Pandžić	0,7287	0,6077	**0,7339**
(L) BTagger – sufiks	0,7031	0,5936	0,7126
(L) BTagger – prefiks + sufiks	0,7019	0,5921	0,7112
(L) Agić et al.	0,7064	0,5915	0,7143
(L) Ljubešić et al.	0,7225	0,5937	**0,7283**

Preliminarni eksperimenti sprovedeni sa kombinovanim modelom, uz primenu stemera Ljubešića i Pandžića, pokazali su da dimenzionalnost vektora značenja reči i širina kontekstnog opsega korišćenog pri njihovoj izradi imaju mali uticaj na performanse modela. Na primer, povećanje dimenzionalnosti vektora sa 100 na 1000 dovodi do poboljšanja koeficijenta korelacije r od samo 0,001, i to po cenu znatno dužeg i hardverski zahtevnijeg obučavanja *word2vec* modela. Iz tog razloga, i u narednim eksperimentima je svuda korišćena ista dimenzionalnost vektora od 100 i širina kontekstnog opsega od 10 reči.

5.1.2 Novi namenski modeli za određivanje semantičke sličnosti kratkih tekstova

U ovom odeljku su predstavljeni novi namenski modeli za određivanje semantičke sličnosti kratkih tekstova i njihovi rezultati. Kao početna osnova za razvoj novih modela korišćen je pristup predstavljen u (Islam & Inkpen 2008), pošto on ne zavisi od napredne sintaktičke analize teksta, što ga čini primenjivim i u jezicima sa ograničenim resursima. Ovaj algoritam se zasniva na ekskluzivnom uparivanju svake reči iz kraćeg teksta sa njoj najsličnijom reči iz dužeg teksta, pri čemu se kao mera sličnosti reči koristi kombinacija sličnosti stringova, tj. nizova karaktera, i semantičke sličnosti. Oslanjanje i na sličnost stringova je bilo dodatan razlog za korišćenje ovog algoritma kao polazne tačke za razvoj novih modela, pošto njena upotreba može da pomogne modelu da se bolje nosi sa sličnošću različitih flektivnih oblika retkih vlastitih imenica (Islam & Inkpen 2008). Sličnost stringova se u ovom modelu dobija kao prosek tri normalizovane varijante LCS (*Longest Common*

Subsequence) mere (Bergroth et al. 2000), koja služi za određivanje najduže zajedničke podsekvence karaktera između dve reči:

- NLCS (*Normalized Longest Common Subsequence*) – LCS mera u kojoj se traži najduža zajednička sekvenca nekonsekutivnih karaktera, normalizovana u odnosu na dužinu reči koje se porede;
- $NMCLCS_1$ (*Normalized Maximal Consecutive Longest Common Subsequence starting at character 1*) – normalizovana varijanta LCS mere u kojoj se traži najduža zajednička sekvenca konsekutivnih karaktera, pri čemu početni karakter sekvence u obe reči mora da bude prvi karakter reči;
- $NMCLCS_N$ (*Normalized Maximal Consecutive Longest Common Subsequence starting at character N*) – normalizovana varijanta LCS mere u kojoj se traži najduža zajednička sekvenca konsekutivnih karaktera, pri čemu početni karakter sekvence u obe reči može da bude proizvoljni karakter reči.

Normalizacija LCS mere omogućava da njena vrednost bude u opsegu od 0 do 1, gde 0 predstavlja nepostojanje bilo kakve sličnosti stringova, a 1 njihovo potpuno podudaranje. Za izražavanje semantičke sličnosti reči, Islam i Inkpen su predložili korišćenje mere zasnovane na distribucionoj semantici, te je stoga u ovoj disertaciji za to upotrebljena kosinusna sličnost statičkih vektora značenja reči kreiranih pomoću *word2vec* algoritma. Zbog prirode kosinusne sličnosti, i ova mera se kreće u opsegu od 0 do 1. Ukupna sličnost između para reči (i, j) se dobija kao težinski prosek sličnosti stringova i semantičke sličnosti tih reči, pri čemu je optimalan odnos između težina tih vrsta sličnosti neophodno utvrditi tokom obučavanja modela. Pošto se uzima da je zbir pondera za obe vrste sličnosti jednak jedinici, u praksi se tokom obučavanja optimizuje samo jedna od te dve vrednosti:

$$\begin{aligned} Similarity(i,j) &= w_{string}String(i,j) + w_{semantic}Semantic(i,j) \\ &= w_{string}String(i,j) + \big(1 - w_{string}\big)Semantic(i,j) \end{aligned}$$

Postupak uparivanja reči iz kraćeg teksta sa po jednom reči iz dužeg teksta počinje tako što se najpre tekstovi normalizuju na mala slova, a zatim se identifikuju identične reči u oba teksta. Njihova sličnost je, prirodno, maksimalna, i iznosi 1. Broj reči iz kraćeg teksta za koje postoje identični parnjaci u dužem tekstu se zapisuje u promenljivu S_{same}, a takve reči se uklanjaju iz oba teksta. Nakon toga se prelazi na računanje ukupnih sličnosti između svih mogućih parova preostalih reči, i inicijalizuje se promenljiva za čuvanje sume sličnosti odabranih parova $S_{different} = 0$. Zatim se bira onaj par reči sa najvećim stepenom sličnosti, vrednost te sličnosti se dodaje u promenljivu $S_{different}$, a posmatrane reči se uklanjaju iz oba teksta, čime se sprečava njihovo dalje uparivanje. Ovaj proces se iterativno ponavlja dokle god se ne iscrpe sve reči iz kraćeg teksta. Konačno, finalna sličnost tekstova se normalizuje recipročnom harmonijskom sredinom dužina oba teksta, tj. brojeva reči u njima, označenim sa *m* i *n*, da bi se dobila vrednost sličnosti između 0 i 1:

$$S = \big(S_{same} + S_{different}\big)\frac{m+n}{2mn}$$

Obučavanje opisanog modela se svodi na odabir optimalnog balansa između težina za sličnost stringova i za semantičku sličnost, tj. na odabir težinskog parametra za sličnost stringova. Pošto ovaj model spada u nadgledane, kao i svi novi razvijeni modeli zasnovani na njemu i prikazani u ovom odeljku, njihova evaluacija je sprovedena upotrebom desetoslojne unakrsne validacije sa sortiranom stratifikacijom. U ovoj disertaciji je za ovaj model, kao i za sve naredne nove modele zasnovane na njemu, za težinski parametar sličnosti stringova razmatran opseg vrednosti [0,3 , 0,7]. Pri tome, radi izbegavanja preterane prilagođenosti podacima korišćenim za obučavanje, upotrebljavan je relativno veliki korak od 0,1 za promenu vrednosti težinskog parametra. U nastavku ovog poglavlja su

prikazana tri modela razvijena na osnovu pristupa iz (Islam & Inkpen 2008), nakon čega je izvršena njihova evaluacija i poređenje rezultata.

5.1.2.1 LInSTSS model

Prvi nov model zasnovan na algoritmu koji su izložili Islam i Inkpen, pod nazivom *LInSTSS* (*Language Independent Short-Text Semantic Similarity*) (Furlan et al. 2013), poboljšava ovaj osnovni pristup tako što se ukupna sličnost svakog para reči ponderiše težinskim faktorom zasnovanim na učestalosti tih reči (engl. *term frequency – TF*). Osnovna ideja ovog modela, po uzoru na pristup predstavljen u (Mihalcea et al. 2006), jeste da sličnostima vrlo frekventnih reči treba dati manju težinu nego sličnostimaređih reči. Naime, prisustvo semantički bliskih retkih reči u oba zadata teksta je mnogo jači indikator njihove semantičke povezanosti nego prisustvo semantički bliskih ali frekventnih reči, zato što će oba teksta često sadržati iste ili slične funkcionalne reči, poput predloga ili pomoćnih glagola, koje su prirodno vrlo frekventne u jeziku.

Frekventnost reči je potrebno utvrditi uvidom u neki veliki korpus tekstova. Za svaku reč *i* računa se logaritamska vrednost njene relativne frekventnosti u korpusu kao:

$$TF_{log}(i) = -\log\frac{Count(i)}{\sum_{j\in V} Count(j)}$$

gde je $Count(i)$ broj javljanja reči *i* u korpusu, a V je vokabular posmatranog korpusa. Tako dobijene logaritamske vrednosti se zatim smeštaju u opseg od 0 do 1 preko min-max normalizacije:

$$TF_{min-max}(i) = \frac{TF_{log}(i)}{\max\{TF_{log}\}}$$

gde je $\max\{TF_{log}\}$ maksimalna uočena logaritamska vrednost relativne frekventnosti za neku reč iz korpusa. Konačno, za ponderisanje sličnosti para reči (i,j) upotrebljava se sledeći izraz:

$$TF(i,j) = 2^{(TF_{min-max}(i)\cdot TF_{min-max}(j))-1}$$

$$\begin{aligned} Similarity(i,j) &= \left(w_{string}String(i,j) + w_{semantic}Semantic(i,j)\right)\cdot TF(i,j) \\ &= \left(w_{string}String(i,j) + \left(1-w_{string}\right)Semantic(i,j)\right)\cdot TF(i,j) \end{aligned}$$

Upotrebom navedene min-max normalizacije na ovaj način postiže se da se vrednost ponderacionog faktora $TF(i,j)$ kreće u opsegu [0,5 , 1]. Time se sličnosti parova frekventnih reči penalizuju, i to smanjenjem do čak 50% od njihove inicijalne vrednosti, dok sličnosti parova retkih reči zadržavaju svoju punu vrednost. Naime, retke reči će imati visoke apsolutne vrednosti logaritama relativne frekvencije, pa će time i njihove min-max normalizovane vrednosti biti bliske jedinici, što znači da će ponderacioni faktor $TF(i,j)$ takođe biti blizak jedinici. Sa druge strane, frekventne reči će imati male apsolutne vrednosti logaritama relativne frekvencije, te će njihove min-max normalizovane vrednosti u krajnosti težiti nuli, što dovodi do toga da ponderacioni faktor $TF(i,j)$ bude blizak vrednosti 0,5. Pošto navedeni sistem ponderisanja uzima u obzir frekventnost obe reči u paru, on u određenoj meri smanjuje sličnost i onih parova gde je samo jedna reč frekventna.

Pri uparivanju identičnih reči iz oba teksta takođe se koristi navedeno ponderisanje, s tim što su tada min-max normalizovane vrednosti logaritama relativne frekventnosti identične za obe reči, čime se ponderacioni izraz svodi na:

$$TF(i,i) = 2^{(TF_{min-max}(i)^2)-1}$$

Stoga ukupna vrednost sličnosti takvih parova S_{same} nije više jednaka broju parova identičnih reči, već sumi njihovih $TF(i,i)$ ponderacionih faktora. Dakle, parovi identičnih reči nemaju više maksimalnu vrednost sličnosti od 1, već im sličnost zavisi od relativne frekventnosti posmatrane reči.

Ostatak *LInSTSS* modela je identičan već opisanom pristupu iz (Islam & Inkpen 2008). Za utvrđivanje relativne frekventnosti reči u ovoj disertaciji upotrebljavan je veb korpus srpskog jezika *srWaC*. Obučavanje ovog modela na određenom anotiranom skupu parova kratkih tekstova se takođe svodi na odabir optimalnog balansa između težina za sličnost stringova i za semantičku sličnost, koje se sprovodi na način već opisan u odeljku 5.1.2.

5.1.2.2 POST STSS model

Drugi nov model zasnovan na algoritmu koji su izložili Islam i Inkpen jeste *POST STSS* (*Part-of-Speech Tag-supported Short-Text Semantic Similarity*) (Batanović & Bojić 2015; Batanović et al. 2018a). Kao i *LInSTSS*, i ovaj model se oslanja na ideju težinskog ponderisanja sličnosti parova reči, ali se u njemu težinski faktori ne dobijaju na osnovu frekventnosti reči, već na osnovu vrsta reči. Naime, intuitivno je jasno da ne nose sve vrste reči u tekstu istu količinu semantički relevantnog sadržaja – reči koje spadaju u glagole ili imenice su važnije za ukupnu semantiku teksta od onih koje spadaju u veznike ili predloge. Pored toga, *POST STSS* model uzima u obzir i smislenost uparivanja različitih vrsta reči, tako što se određene kombinacije dozvoljavaju, a druge zabranjuju.

Za primenu *POST STSS* modela na određenom jeziku je stoga neophodno postojanje modula za označavanje vrsta reči, što donekle sužava opštost njegove primenjivosti. Ipak, takvi moduli su najosnovniji alati za sintaktičku analizu i kao takvi su neretko dostupni i u jezicima sa ograničenim resursima. U ovoj disertaciji je za dobijanje oznaka vrsta reči u tekstovima na srpskom upotrebljavan alat za srpski jezik predstavljen u (Ljubešić et al. 2016), koji proizvodi morfosintaktičke oznake po MULTEXT-East v5 (Erjavec 2017) standardu za srpskohrvatski makrojezik[27]. U ovom standardu, POS (*Part-of-Speech*) oznaka, tj. oznaka vrste reči, predstavlja prvo slovo obimnije morfosintaktičke oznake.

Pored već opisanih osnovnih elemenata modela koji su zajednički sa pristupom koji su predložili Islam i Inkpen, *POST STSS* model koristi skup POS težinskih parametara, po jedan za svaku vrstu reči, kao i matricu interakcija između različitih vrsta reči koja je simetrična i sadrži binarne/Bulove vrednosti putem kojih se uparivanje određenih vrsta reči dozvoljava ili zabranjuje. Stoga *POST STSS* model ima tri tipa parametara čije se vrednosti optimizuju tokom obučavanja modela:

- POS težinske parametre za svaku od vrsta reči;
- Matricu interakcija vrsta reči;
- Balans između sličnosti stringova i semantičke sličnosti, tj. težinski parametar za sličnost stringova, koji se koristi i u *LInSTSS* i osnovnom (Islam & Inkpen 2008) modelu.

Finalna ponderaciona vrednost $POS(i,j)$ za sličnost određenog para reči (i,j) se određuje na osnovu vrednosti težinskih parametara vezanih za njihovu vrstu. Ako obe reči u određenom paru pripadaju istoj vrsti reči, onda je logično da za ponderisanje njihove sličnosti treba koristiti upravo vrednost POS težinskog parametra za tu vrstu reči. Međutim, ako reči u okviru para pripadaju različitim vrstama reči, tada se najpre proverava u matrici interakcija da li je takva kombinacija vrsta reči dozvoljena ili nije. Ako nije dozvoljena, ponderaciona vrednost $POS(i,j)$ se za taj par reči postavlja

[27] http://nl.ijs.si/ME/V5/msd/html/msd-hbs.html

na nulu, čime se efektivno uklanja mogućnost takvog uparivanja reči. Ako pak uparivanje jeste dozvoljeno, pri određivanju ponderacione vrednosti u obzir treba uzeti dva različita POS težinska parametra, te je moguće koristiti više različitih pristupa, od kojih su u inicijalnim eksperimentima razmotrene sledeće:

- Odabir vrednosti onog POS težinskog parametra sa većom vrednošću, odnosno:

$$POS(i,j) = \max\left\{POS_{weight}(i), POS_{weight}(j)\right\}$$

- Odabir vrednosti onog POS težinskog parametra sa manjom vrednošću, odnosno:

$$POS(i,j) = \min\left\{POS_{weight}(i), POS_{weight}(j)\right\}$$

- Računanje aritmetičke sredine vrednosti oba POS težinska parametra, odnosno:

$$POS(i,j) = \frac{POS_{weight}(i) + POS_{weight}(j)}{2}$$

- Računanje geometrijske sredine vrednosti oba POS težinska parametra, odnosno:

$$POS(i,j) = \sqrt{POS_{weight}(i) \cdot POS_{weight}(j)}$$

- Računanje harmonijske sredine vrednosti oba POS težinska parametra, odnosno:

$$POS(i,j) = \frac{2 \cdot POS_{weight}(i) \cdot POS_{weight}(j)}{POS_{weight}(i) + POS_{weight}(j)}$$

Kao najbolje rešenje u inicijalnim eksperimentima na srpskom jeziku, kao i u primeni *POST STSS* modela na engleskom jeziku (Batanović & Bojić 2015), pokazala se aritmetička sredina vrednosti oba POS težinska parametra. Stoga je ova funkcija korišćena u ovoj disertaciji u nastavku istraživanja.

Konačno, ukupna sličnost para reči (i,j) se računa kao:

$$\begin{aligned} Similarity(i,j) &= \left(w_{string}String(i,j) + w_{semantic}Semantic(i,j)\right) \cdot POS(i,j) \\ &= \left(w_{string}String(i,j) + \left(1 - w_{string}\right)Semantic(i,j)\right) \cdot POS(i,j) \end{aligned}$$

Ukoliko se u paru nalaze identične reči, tada se vrednost njihove sličnosti svodi na vrednost ponderacione funkcije $POS(i,j)$, koja se zatim dodaje u sumu S_{same}. Ukupna sličnost tekstova se dobija na isti način kao i u pristupu iz (Islam & Inkpen 2008), s tim što se sada maksimalna vrednost sličnosti mora eksplicitno ograničiti na jedinicu:

$$S = min\left\{\left(S_{same} + S_{different}\right)\frac{m+n}{2mn}, 1\right\}$$

Naime, kod *POST STSS* modela teoretski je moguće da ukupna sličnost tekstova bude viša od 1, ako se u tekstovima koji se porede javljaju samo reči čiji su POS težinski parametri veći od 1. Ipak, takva situacija je u praksi malo verovatna i nije se dešavala na anotiranim tekstovima iz *STS.news.sr* skupa.

POS težinski parametri se u ovom modelu optimizuju u opsegu [0,7 , 1,3], koji je centriran na neutralnoj vrednosti 1, i simetričan u odnosu na nju. Drugim rečima, ponderisanje ocena sličnosti

para reči u *POST STSS* modelu može da uveća ili umanji vrednost sličnosti za najviše 30%. Ovaj opseg vrednosti za POS težinske parametre je odabran kao kompromis između maksimizacije mogućnosti težinskog ponderisanja i minimizacije dužine obučavanja modela. Takvim opsegom se zadržava i da prosečne vrednosti ukupne sličnosti, koje se kreću između 0 i 1, gravitiraju ka poželjnoj srednjoj vrednosti od 0,5, čime se sprečava veća neizbalansiranost ukupnih ocena sličnosti. Pri tome, u toku obučavanja upotrebljava se relativno veliki korak od 0,1 za promene vrednosti svih težinskih parametara, da bi se izbegla njihova preterana prilagođenost podacima korišćenim za obučavanje.

Glavni problem u obučavanju *POST STSS* modela jeste veliki broj parametara koje treba optimizovati. Taj broj je direktno zavisan od broja različitih vrsta reči koje se koriste u modelu. Ako se razmatra *k* vrsta reči, tada je broj parametara koje treba optimizovati, pod pretpostavkom korišćenja navedenih opsega vrednosti parametara i navedenog koraka za promene njihovih vrednosti, jednak:

$$N_{parameter\_space} = 7^k \cdot 2^{\frac{k(k-1)}{2}} \cdot 5$$

Naime, uz korak od 0,1 svaki POS težinski parametar može imati jednu od sedam vrednosti u opsegu [0,7 , 1,3], svako moguće uparivanje različitih vrsta reči, kojih ima $\frac{k(k-1)}{2}$, moguće je dozvoliti ili zabraniti, a težinski parametar vezan za važnost sličnosti stringova može imati jednu od 5 vrednosti u opsegu [0,3 , 0,7]. U MULTEXT-East v5 standardu definisano je 13 različitih glavnih oznaka za vrste reči, odnosno 12 ne računajući interpunkciju. Pri tome, mnoge vrste reči imaju svoje podtipove – npr. imenice se dele na zajedničke i vlastite, glagoli se dele na glavne i pomoćne, itd. Potpuna pretraga kombinatornog prostora vrednosti svih parametara bi čak i samo za glavne vrste reči bila izuzetno računski skupa operacija, jer bi podrazumevala razmatranje ogromnog broja kombinacija:

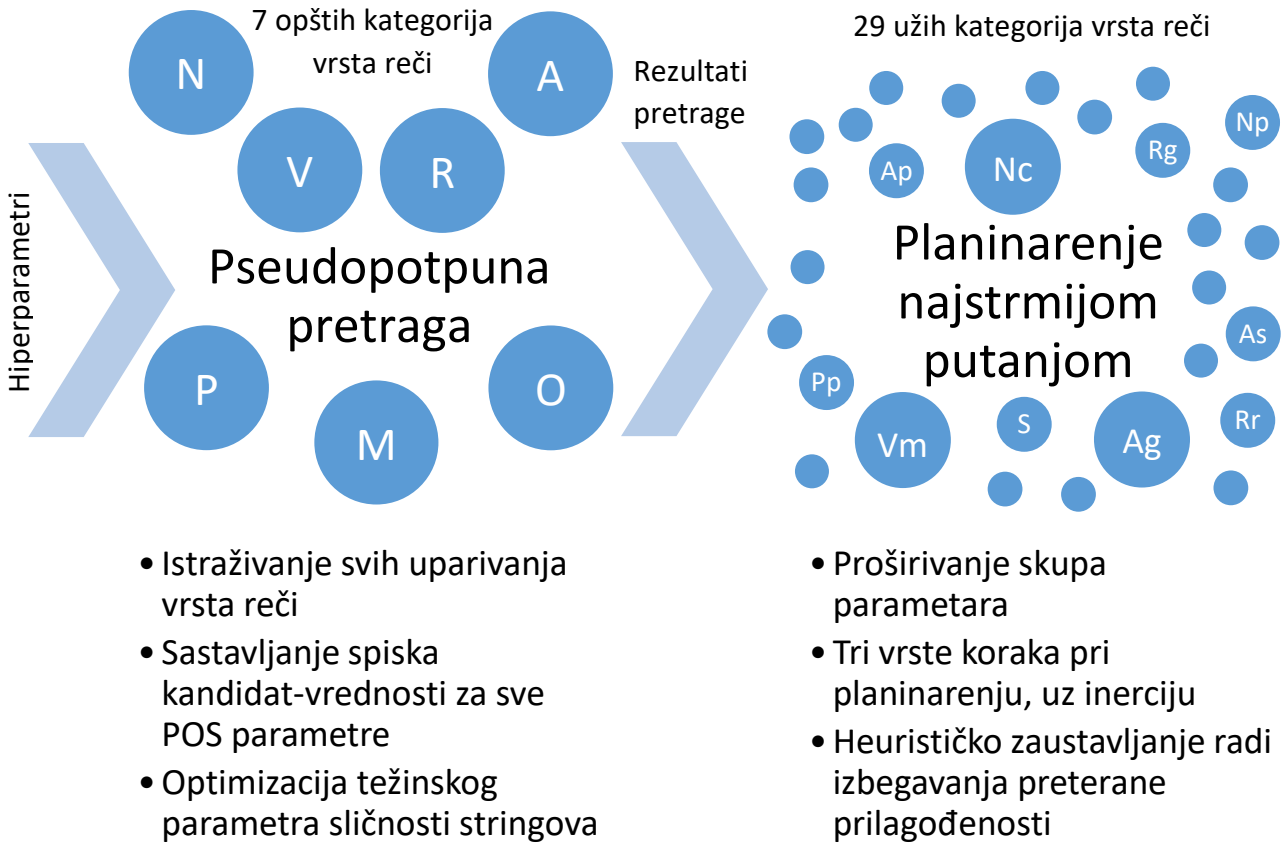
$$N_{parameter\_space\_exhaustive} = 7^{12} \cdot 2^{\frac{12(12-1)}{2}} \cdot 5 = 7^{12} \cdot 2^{66} \cdot 5 \approx 5{,}1 \cdot 10^{30}$$

Međutim, jasno je da ograničavanje samo na glavne vrste reči nije primereno, jer npr. nema smisla dodeljivati istu težinu glavnim i pomoćnim glagolima u rečenici, pošto su glavni glagoli ti koji nose najveći deo semantički bitnih informacija. Radi rešavanja problema ove kombinatorne eksplozije, usvojen je princip dvostepenog obučavanja *POST STSS* modela, gde se parametri najpre delimično optimizuju u nižedimenzionalnom prostoru vrednosti parametara, a zatim se, počevši od tako dobijene delimično optimizovane konfiguracije, dodatno podešavaju u višedimenzionalnom prostoru. Osnovna ideja ovakvog pristupa jeste da model najpre treba da nauči relativne važnosti i interakcije između opštih kategorija vrsta reči, a da se tek nakon toga usredsredi na specifičnosti pojedinih podvrsta. Stoga se obučavanje *POST STSS* modela može podeliti na dve faze, čije su glavne odlike i koraci prikazani na slici 11:

1. Pseudopotpunu pretragu u nižedimenzionalnom prostoru vrednosti parametara;
2. Planinarenje najstrmijom putanjom (engl. *steepest ascent hill climbing*) u višedimenzionalnom prostoru vrednosti parametara.

U fazi pseudopotpune pretrage, sve reči se svrstavaju u jednu od narednih sedam opštih kategorija:

- Imenice – morfosintaktička oznaka počinje slovom *N*;
- Glagoli – morfosintaktička oznaka počinje slovom *V*;
- Prilozi – morfosintaktička oznaka počinje slovom *R*;
- Pridevi – morfosintaktička oznaka počinje slovom *A*;
- Zamenice – morfosintaktička oznaka počinje slovom *P*;
- Brojevi – morfosintaktička oznaka počinje slovom *M*;
- Ostalo – sve preostale morfosintaktičke oznake.



Slika 11. Glavne odlike i koraci faza obučavanja *POST STSS* i *POS-TF STSS* modela za određivanje semantičke sličnosti kratkih tekstova

Korišćenje sedam kategorija vrsta reči značajno smanjuje kombinatornu eksploziju, te redukovani broj kombinacija vrednosti parametara koje treba razmotriti postaje:

$$N_{parameter\_space\_exhaustive\_reduced} = 7^7 \cdot 2^{\frac{7(7-1)}{2}} \cdot 5 = 7^7 \cdot 2^{21} \cdot 5 \approx 8{,}6 \cdot 10^{12}$$

Ipak, i ovo je preveliki broj kombinacija za sprovođenje prave potpune pretrage prostora vrednosti parametara. Umesto toga, u prvom koraku obučavanja sprovodi se pseudopotpuna pretraga, gde je ideja da se pretraga sprovede samo među onim kombinacijama vrednosti parametara koje imaju neku šansu da budu optimalne.

Pseudopotpuna pretraga počinje tako što se svim POS težinskim parametrima dodeljuje zajednička inicijalna vrednost. Svi elementi matrice interakcija vrsta reči se takođe inicijalizuju jednom istom binarnom vrednošću. Ove inicijalne vrednosti predstavljaju hiperparametre modela. Konačno, usvaja se i inicijalna neutralna vrednost od 0,5 za težinski parametar koji određuje važnost sličnosti stringova. Sa takvim inicijalnim podešavanjima sprovodi se evaluacija modela nad podacima za obučavanje i beleži inicijalni nivo performansi.

Zatim se odvojeno istražuju sva moguća uparivanja različitih vrsta reči. Za svako moguće uparivanje, iterira se kroz sve moguće kombinacije vrednosti POS težinskih parametara vezanih za vrste reči u trenutno posmatranom paru, i za svaku kombinaciju se beleže performanse modela na podacima za obučavanje. Tokom ovog iteriranja, svi ostali POS težinski parametri se drže fiksiranim na svojim

inicijalnim vrednostima, što važi i za binarne vrednosti u matrici interakcije, kao i za težinski faktor sličnosti stringova. Zatim se element matrice interakcije koji se odnosi na trenutno posmatrani par vrsta reči invertuje, a ceo proces iteriranja ponavlja. Drugim rečima, parametri modela se zamrzavaju na svojim inicijalnim vrednostima, a jedino se razmatraju promene POS težinskih parametara i elementa matrice interakcije koji se odnose na trenutno posmatrani par vrsta reči. Nakon što se završi razmatranje određenog para vrsta reči, model se resetuje na inicijalno stanje i prelazi se na sledeći par. Ova procedura se ponavlja za sva moguća uparivanja različitih vrsta reči. Pri tome se za svaki par identifikuje ona kombinacija vrednosti POS težinskih parametara i odgovarajućeg elementa matrice interakcije koja dovodi do maksimalnog nivoa performansi modela na podacima za obučavanje. Pošto u niždimenzionalnom prostoru postoji 21 različit par vrsta reči, veličina kombinatornog prostora koji treba istražiti u ovom prvom koraku pretrage je značajno manja:

$$N_{parameter\_space\_step\_one} = 7^2 \cdot 2 \cdot 21 = 2058$$

Opisana procedura se oslanja na pretpostavku da je moguće zasebno razmotriti međusobni odnos između svakog para vrsta reči i relativan odnos važnosti između njih. Drugim rečima, ovaj deo obučavanja modela podrazumeva da se izbor optimalnih vrednosti parametara može podeliti na veći broj odvojenih odluka za svaki par vrsta reči. Iako je potpuna validnost navedene pretpostavke upitna, ona omogućava drastičnu redukciju veličine kombinatornog prostora.

U skladu sa navedenom pretpostavkom, ako neka vrednost POS težinskog parametra za određenu vrstu reči nije detektovana kao optimalna ni u jednom od parova, onda je malo verovatno da bi ta vrednost na kraju dovela do optimalnih performansi modela. Stoga se takva vrednost odbacuje iz daljeg razmatranja, čime se smanjuje broj kombinacija vrednosti parametara koje treba istražiti u nastavku obučavanja. Slično tome, ako jedna binarna vrednost elementa matrice interakcija vezanog za posmatrani par vrsta reči dovodi do lošijih performansi nego druga, ona se odbacuje. Ipak, za mnoge parove je moguće da postoji više kombinacija vrednosti parametara koje su podjednako dobre.

Sledeći korak pseudopotpune pretrage podrazumeva sastavljanje spiska kandidat-vrednosti za svaki POS parametar. Za svaku vrstu reči prolazi se kroz sva njena uparivanja obrađena u prethodnom koraku i prikupljaju se sve vrednosti njenog POS težinskog parametra koje su u bilo kom uparivanju bile optimalne. Kandidat-vrednosti za svaki element matrice interakcija se dobijaju kao optimalne vrednosti tih elemenata detektovane u prethodnom koraku pri obradi odgovarajućeg para vrsta reči. Nakon što su sve kandidat-vrednosti za sve parametre prikupljene, vrši se potpuna pretraga svih njihovih mogućih kombinacija, pri čemu se za svaku kombinaciju beleže performanse modela na podacima za obučavanje. I u ovom koraku se vrednost težinskog parametra sličnosti stringova drži na inicijalnoj vrednosti. Na kraju ove pretrage usvaja se ona kombinacija vrednosti parametara koja dovodi do najboljih performansi modela. Ako postoji više kombinacija koje daju podjednako dobre rezultate, sve one se čuvaju za naredni korak pretrage. Tačnu veličinu kombinatornog prostora vrednosti parametara u ovom koraku nije moguće definisati unapred, jer zavisi od rezultata prethodnog koraka. Ipak, ona se u praksi kreće od nekoliko desetina do nekoliko hiljada kombinacija, što je i dalje primetno manje od broja kombinacija koje bi trebalo razmotriti u potpunoj pretrazi celog kombinatornog prostora.

U poslednjem koraku pseudopotpune pretrage vrši se iteriranje kroz sve moguće vrednosti težinskog parametra sličnosti stringova, i to za svaku konfiguraciju preostalih parametara koja je usvojena kao optimalna u prethodnom koraku. Pri tome se, kao i do sada, beleže performanse modela na podacima za obučavanje. Pošto pomenutih konfiguracija obično ima samo jedna ili dve, ovaj korak pretrage je vrlo efikasan. Onaj skup vrednosti svih parametara koji u ovom koraku dovodi do najboljih rezultata modela predstavlja izlaz faze pseudopotpune pretrage i koristi se kao početna tačka za dalje obučavanje modela u narednoj fazi.

Druga faza obučavanja – planinarenje najstrmijom putanjom – sprovodi se u višedimenzionalnom prostoru vrednosti parametara. U ovoj fazi se broj različitih kategorija vrsta reči proširuje na 29, tako što se kao posebna kategorija razmatra svaki podtip morfosintaktičkih grupa navedenih u MULTEXT-East v5 standardu. Na primer, morfosintaktička grupa *Imenice* je podeljena na dva podtipa – zajedničke imenice (morfosintaktička oznaka počinje sa *Nc*) i vlastite imenice (morfosintaktička oznaka počinje sa *Np*). Naravno, postoje i vrste reči koje nisu podeljene na podtipove, poput predloga (engl. *adpositions* – morfosintaktička oznaka *S*), kod kojih opšta vrsta reči predstavlja jednu kategoriju. Jedini izuzetak od ovog sistema kategorija predstavlja interpunkcija (oznaka *Z*), koja se ignoriše, pošto se interpunkcija izbacuje iz teksta prilikom tokenizacije, kao i reziduali (oznaka počinje sa *X*), koji se skupa tretiraju kao jedna kategorija, pošto se samo jedan njihov podtip – strani reziduali – pojavljuje u podacima iz *STS.news.sr* korpusa.

Ipak, opisani sistem kategorija se primenjuje samo na one podtipove vrsta reči za koje u MULTEXT-East v5 standardu zaista postoje specifikovane pune vrednosti morfosintaktičkih oznaka. Na primer, standard dozvoljava poseban podtip kopulativnih glagola (morfosintaktička oznaka počinje sa *Vc*), ali za taj podtip nisu navedene pune vrednosti morfosintaktičkih oznaka, te ih ni korišćeni modul za morfosintaktičko označavanje ne upotrebljava. Ista situacija je i sa morfosintaktičkim oznakama za upitno-odnosne zamenice, koje počinju sa *Pr*. Stoga takvi podtipovi nisu tretirani kao posebna kategorija u *POST STSS* modelu.

U drugoj fazi obučavanja cilj je da se konfiguracija parametara koja je dobijena kao rezultat pseudopotpune pretrage dodatno podesi u višedimenzionalnom prostoru vrednosti parametara. Da bi se to uradilo, najpre se vrši proširivanje skupa parametara iz nižedimenzionalnog u višedimenzionalni prostor na sledeći način. Svakoj kategoriji u višedimenzionalnom prostoru dodeljuje se POS težinski parametar čija je vrednost jednaka težinskom parametru odgovarajuće opšte vrste reči iz nižedimenzionalnog prostora. Na primer, vrednost nižedimenzionalnog POS težinskog parametra za glagole (oznaka *V*) se prenosi na višedimenzionalne POS težinske parametre za glavne glagole (oznaka *Vm*) i pomoćne glagole (oznaka *Va*). Pored toga, matrica interakcija se takođe proširuje, ali tako da svi utvrđeni odnosi između vrsta reči ostanu nepromenjeni. Na primer, ako je u nižedimenzionalnom prostoru bilo dozvoljeno uparivanje prideva i priloga, u višedimenzionalnom prostoru će ostati dozvoljeno uparivanje između svih njihovih podtipova. Takođe se u višedimenzionalnom prostoru dozvoljavaju međusobne interakcije podtipova koji su u nižedimenzionalnom prostoru spadali u jednu opštu kategoriju. Na primer, dozvoljavaju se uparivanja veznika i predloga, jer su ove vrste reči u prvoj fazi obučavanja bile grupisane u kategoriju *Ostalo*. Konačno, kao i u nižedimenzionalnom prostoru, ponovo se automatski dozvoljavaju uparivanja dve reči koje pripadaju istoj kategoriji, samo što se sada to odnosi na pripadnost podtipovima vrsta reči.

Pošto je višedimenzionalni prostor vrednosti parametara izuzetno veliki, njegova bilo kakva sistematičnija pretraga nije izvodljiva. Stoga se u drugoj fazi obučavanja primenjuje algoritam planinarenja najstrmijom putanjom, pri čemu se koriste tri vrste koraka:

1. Povećanje ili smanjenje vrednosti određenog POS težinskog parametra;
2. Inverzija određenog elementa matrice interakcija;
3. Povećanje ili smanjenje vrednosti težinskog parametra sličnosti stringova.

Planinarenje se obavlja evaluiranjem efekta svakog mogućeg koraka na performanse modela na podacima za obučavanje. Standardan metod planinarenja najstrmijom putanjom bi uvek birao onaj korak koji dovodi do najvećeg poboljšanja performansi modela. Međutim, u obučavanju *POST STSS* modela svaka izmena nekog težinskog parametra ima svoju inerciju (engl. *momentum*). To znači da ako u određenom trenutku promena nekog težinskog parametra ima najpovoljniji efekat na performanse modela, tada će ta promena biti ponovljena dokle god performanse modela nastavljaju da rastu, čak i ako se u nekom trenutku pojavi neki drugi potencijalni korak koji bi doveo do većeg

rasta performansi. Ovakva inercija omogućava modelu da izbegne neke lokalne maksimume u planinarenju i pokazuje se da povećava ne samo finalne performanse modela nego i efikasnost uspona. Ako inercijalno ponavljanje promene nekog težinskog parametra više ne poboljšava performanse modela, planinarenje se nastavlja upotrebom najstrmije putanje.

Radi ubrzavanja uspona, u obučavanju se dozvoljava da korak koji menja vrednost nekog POS težinskog parametra simultano invertuje i neki element matrice interakcija koji se odnosi na istu kategoriju kao i dati POS težinski parametar. Pored toga, u usponu se dozvoljavaju i skokovi – koraci u kojima se vrednost težinskog parametra ne menja za standardnu vrednost od 0,1, već za dvostruko toliko. Kod skokova je takođe dozvoljena simultana izmena nekog od odgovarajućih elemenata matrice interakcija.

Planinarenje bi moglo da se izvršava dokle god postoje koraci koji poboljšavaju performanse modela na podacima za obučavanje, ali bi to dovelo do preterane prilagođenosti modela. Da bi se to izbeglo, korišćena je jednostavna heuristika – uspon se prekida u trenutku kada više ne postoji nijedan korak koji bi na podacima za obučavanje doveo do smanjenja greške koje iznosi bar 5% od smanjenja greške koje je ostvareno u prvom koraku uspona.

Kao što je već pomenuto, u hiperparametre *POST STSS* modela, koje je neophodno odabrati pre samog obučavanja, spadaju:

- Zajedničke inicijalne vrednosti za sve POS težinske parametre;
- Zajedničke inicijalne vrednosti za elemente matrice interakcija vrsta reči.

Radi optimizacije hiperparametara korišćena je ugnježđena unakrsna validacija sa tri unutrašnja i deset spoljašnjih slojeva, uz primenu sortirane stratifikacije. Manji broj unutrašnjih slojeva je odabran radi ubrzavanja celokupne procedure optimizacije modela.

Opisana procedura obučavanja *POST STSS* modela dozvoljava da se veći deo koraka u njenom sprovođenju paralelizuje, što povećava efikasnost obučavanja i značajno smanjuje njegovo trajanje. Međutim, njen nedostatak jeste to što ne garantuje pronalaženje globalnog optimuma vrednosti parametara. Ograničavajući faktor u primeni *POST STSS* modela na određenom jeziku jeste i tačnost korišćenog modula za označavanje vrsta reči, koja mora da bude dovoljno visoka da bi primena ovog modela imala smisla.

5.1.2.3 POS-TF STSS model

Treći i poslednji nov model zasnovan na pristupu koji su izložili Islam i Inkpen jeste *POS-TF STSS* (*Part-of-Speech and Term Frequency weighted Short-Text Semantic Similarity*) (Batanović et al. 2018a). *POS-TF STSS* model predstavlja direktnu kombinaciju *LInSTSS* i *POST STSS* modela, tj. proširenje *POST STSS* modela tako da se sličnosti parova reči ne ponderišu samo na osnovu njihove vrste, već i na osnovu njihove učestalosti:

$$\begin{aligned} Similarity(i,j) &= \Big(w_{string} String(i,j) + w_{semantic} Semantic(i,j)\Big) \cdot TF(i,j) \cdot POS(i,j) \\ &= \Big(w_{string} String(i,j) + \big(1 - w_{string}\big) Semantic(i,j)\Big) \cdot TF(i,j) \cdot POS(i,j) \end{aligned}$$

U ovom modelu računanje $TF(i,j)$ težinskih vrednosti se sprovodi na identičan način kao i u *LInSTSS* modelu, pri čemu je opet korišćen *srWaC* korpus za određivanje relativne frekventnosti reči. Ostatak *POS-TF STSS* modela, uključujući njegovo obučavanje, identičan je *POST STSS* modelu.

5.1.2.4 *Rezultati novih namenskih modela*

U ovom odeljku su prikazani rezultati evaluacije novih namenskih modela za određivanje semantičke sličnosti kratkih tekstova i diskutovane su optimalne vrednosti njihovih parametara. Korišćenjem desetoslojne unakrsne validacije sa sortiranom stratifikacijom takođe su iznova evaluirani i nenadgledani modeli, radi mogućnosti upoređivanja svih rezultata. Svi modeli su evaluirani kako u varijanti gde morfološka normalizacija tekstova nije primenjena, tako i u varijantama gde je primenjen stemer Ljubešića i Pandžića, koji se najbolje pokazao pri ranijoj evaluaciji nenadgledanih modela. Takođe je opet razmotren i lematizator predstavljen u (Ljubešić et al. 2016), jer je on deo istog paketa alata pomoću kojih je vršeno i označavanje vrsta reči. Rezultati evaluacije su prikazani u tabeli 12.

Tabela 12. Rezultati novih namenskih modela za određivanje semantičke sličnosti kratkih tekstova na *STS.news.sr* korpusu, izraženi u vidu Pirsonovog koeficijenta korelacije r

Model	Metod morfološke normalizacije		
	Bez morfološke normalizacije	(S) Ljubešić i Pandžić	(L) Ljubešić et al.
Nenadgledani modeli			
Podudarnost reči	0,6970	0,7367	0,7278
Prosek vektora značenja reči	0,6405	0,6295	0,6136
Podudarnost reči + prosek vektora značenja reči	0,7050	0,7417	0,7335
Nadgledani modeli			
Islam i Inkpen	0,7387	0,7444	0,7350
LInSTSS	0,7534	0,7573	0,7494
POST STSS	0,7538	0,7593	0,7491
POS-TF STSS	**0,7599**	**0,7665**	**0,7606**

Kao što se moglo i očekivati, nadgledani modeli ostvaruju primetno bolje performanse u odnosu na nenadgledane pristupe, ali je razlika između njih znatno primetnija kada se koriste morfološki nenormalizovani tekstovi. Stemovanje ima jasan pozitivan efekat na performanse skoro svih modela, dok lematizacija popravlja rezultate nenadgledanih modela koji se oslanjaju na podudarnost reči, ali su njeni efekti na nadgledane modele uglavnom negativni. Svi predloženi novi namenski modeli konzistentno ostvaruju bolje rezultate od osnovnog pristupa koji su predložili Islam i Inkpen. *LInSTSS* i *POST STSS* modeli ostvaruju prilično uporedive performanse, ali ih kombinovani *POS-TF STSS* model nadmašuje. Stoga se *POS-TF STSS* model, u kombinaciji sa stemerom Ljubešića i Pandžića, izdvaja kao najbolje rešenje za određivanje semantičke sličnosti kratkih tekstova od svih do sada razmotrenih.

Optimalne vrednosti parametara nadgledanih modela se prirodno donekle razlikuju u zavisnosti od toga da li se primenjuje neka forma morfološke normalizacije i koja. Ipak, moguće je uočiti neke zajedničke obrasce. U osnovnom modelu koji su predstavili Islam i Inkpen, vrednost težinskog parametra sličnosti stringova teži ka gornjoj granici od 0,7, što znači da težina semantičke sličnosti teži ka vrednosti 0,3. Ovo je očekivana posledica nešto većeg stepena leksičkog poklapanja između parova rečenica u *STS.news.sr* korpusu. Međutim, u *LInSTSS* modelu optimalna vrednost težinskog parametra sličnosti stringova je blago niža i iznosi 0,6, što znači da dodavanje frekvencijskog ponderisanja ističe važnost kompleksnijih načina izražavanja sličnosti.

Između *POST STSS* i *POS-TF STSS* modela postoje manje razlike u optimalnim konfiguracijama parametara, ali nisu uočena bilo kakva sistematska odstupanja. Optimizacija hiperparametara najčešće dovodi do toga da se POS težinski parametri inicijalizuju na neutralnu vrednost od 1,0, dok je optimalna inicijalizacija matrice interakcija obično takva da se dozvoljavaju uparivanja između svih vrsta reči.

Zajedničke imenice najčešće zadržavaju neutralnu POS težinsku vrednost od 1,0, po čemu se ističu kao značajnije od vlastitih imenica, čije se težine kreću oko 0,8 – 0,9. Težinske vrednosti glavnih glagola se gotovo uvek optimizuju ka maksimalnoj vrednosti od 1,3, što ističe ključnu ulogu glagola u prenošenju značenja rečenice. Ovaj efekat je ranije uočen i pri primeni *POST STSS* modela na engleskom jeziku (Batanović & Bojić 2015), kao i od strane drugih istraživača (Wiemer-Hastings 2004). Sa druge strane, pomoćni glagoli nose znatno manju količinu semantički bitnog sadržaja i stoga im se dodeljuju niže težinske vrednosti, obično u opsegu 0,7 – 0,9. U skladu sa tim, glagolski pridevi konzistentno dobijaju maksimalne težinske vrednosti. Težine prisvojnih prideva su takođe nešto veće, ali ne u toj meri, dok su optimalne vrednosti težinskih parametara za ostale vrste prideva najčešće 0,9. Slično tome, glagolski prilozi dobijaju veće težinske vrednosti, oko 1,2, dok se težine pravih priloga kreću oko 1,0 – 1,1. I brojevi, naročito redni, imaju visoke težinske vrednosti, često blizu maksimalnih. Visoke vrednosti težinskih parametara za brojeve, priloge i veći deo prideva su verovatno indikator njihove važnosti u pravilnom merenju stepena semantičke sličnosti tekstova čiji su glagoli isti. POS težinski parametri zamenica su u principu niži i kreću se u opsegu 0,7 – 1,0, dok se skraćenicama dodeljuju težine u opsegu 0,8 – 1,0. Preostale vrste reči uglavnom spadaju u funkcionalne/gramatičke reči, kao što su veznici, predlozi, rečce/partikule, i sl, koje ne sadrže bitan semantički sadržaj, te im se stoga dodeljuju niske težinske vrednosti u opsegu 0,7 – 0,8.

Optimizovana konfiguracija matrice interakcija u većini slučajeva dozvoljava parove različitih vrsta reči, ali su neka nelogična uparivanja ipak zabranjena, poput uparivanja zamenica sa funkcionalnim rečima kao što su veznici. Ipak, činjenica da većina parova ostaje principijelno moguća pokazuje da su stroge zabrane korisne tek u manjem broju situacija i da, tokom obučavanja, *POST STSS* i *POS-TF STSS* modeli pre svega vrše modifikaciju vrednosti POS težinskih parametara. U prilog tome govori i to što u ovim modelima optimalna težinska vrednost sličnosti stringova obično ostaje na početnoj vrednosti od 0,5, čime se ostvaruje jednaka težina sličnosti stringova i semantičke sličnosti.

5.1.3 Fino podešavanje neuralnih jezičkih modela radi određivanja semantičke sličnosti kratkih tekstova

U ovom odeljku je opisana evaluacija i rezultati neuralnih jezičkih modela zasnovanih na *Transformer* arhitekturama na problemu određivanja semantičke sličnosti kratkih tekstova iz *STS.news.sr* korpusa. U eksperimentima su korišćena tri prethodno obučena višejezična modela opisana u uvodu glave 5.

Navedeni neuralni modeli su najpre podvrgnuti finom podešavanju u trajanju od jedne epohe, pri čemu je razmotrena i primena do sada konzistentno najbolje metode morfološke normalizacije – stemera Ljubešića i Pandžića. Za fino podešavanje korišćene su podrazumevane vrednosti iz *Simple Transformers* biblioteke (*batch size* = 8, *learning rate* = 4e-5), pri čemu je očuvana originalna kapitalizacija slova, jer su navedeni modeli obučeni da i nju uzimaju u obzir. Rezultati ovako dobijenih modela, uprosečeni sa pet pokretanja sa različitim nasumičnim inicijalizacionim vrednostima (engl. *random seed*), prikazani su u tabeli 13. Može se videti da neuralni modeli daju bolje rezultate na morfološki nenormalizovanim tekstovima, što je očekivano jer je to vrsta tekstova na kojima su ovi modeli inicijalno obučavani.

Dužina procedure finog podešavanja je zatim podignuta na tri epohe, kao što je preporučeno za ovakve modele u (Devlin et al. 2019). Rezultati pokazuju da je *multilingual BERT* model najbolji u određivanju semantičke sličnosti kratkih tekstova, ali je *XLM* model tik uz njega. Iako je pokušano da se rezultati neuralnih modela dodatno poprave podizanjem broja epoha na pet, to nije dovelo do konzistentnih poboljšanja.

Tabela 13. Rezultati neuralnih jezičkih modela zasnovanih na *Transformer* arhitekturama u određivanju semantičke sličnosti kratkih tekstova na *STS.news.sr* korpusu, izraženi u vidu Pirsonovog koeficijenta korelacije r

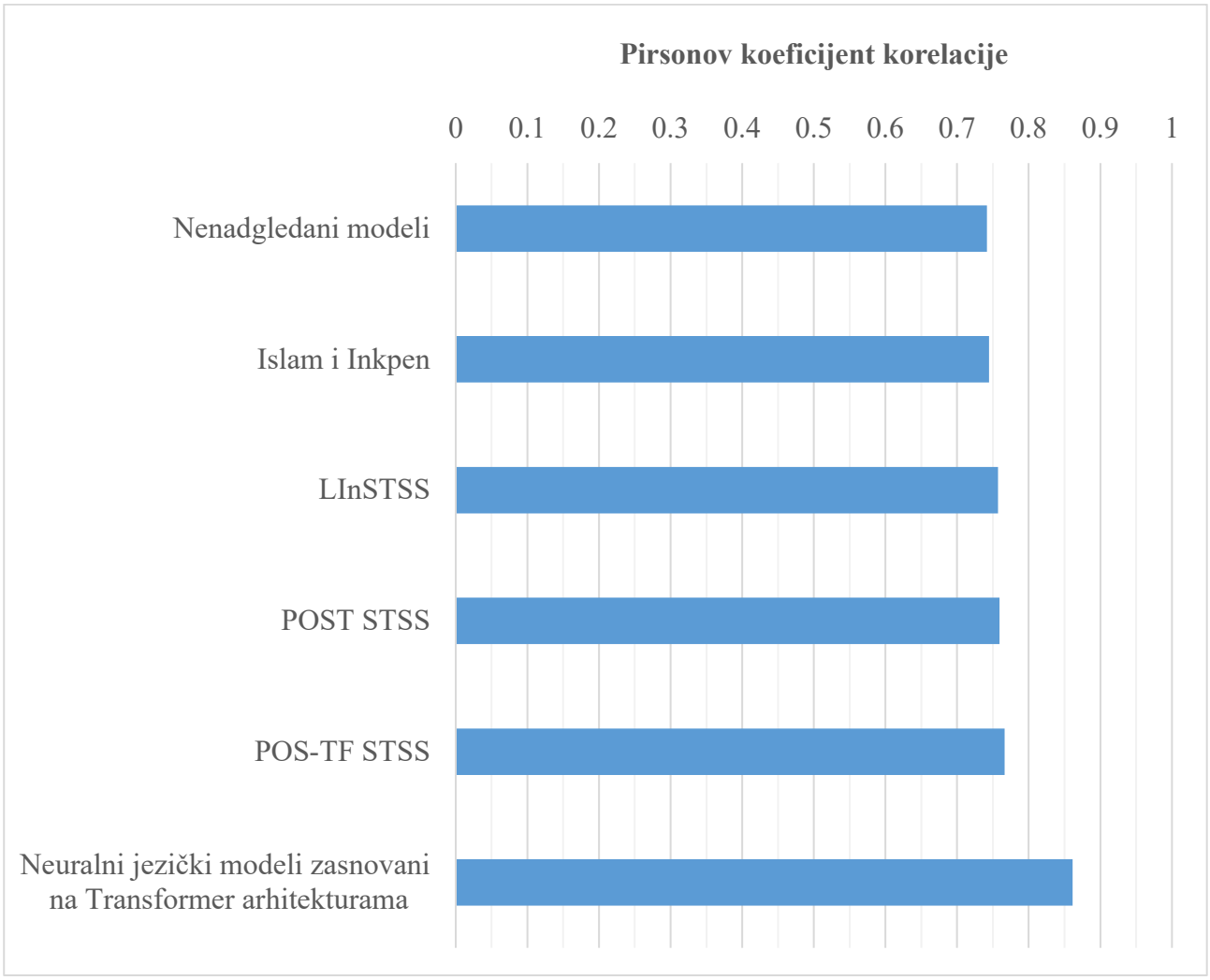
Model	Metod morfološke normalizacije	
	Bez morfološke normalizacije	(S) Ljubešić i Pandžić
Broj epoha = 1		
BERT Base Multilingual Cased	0,8327	0,8109
DistilBERT Base Multilingual Cased	0,7641	0,7630
XLM MLM	0,8277	0,8115
Broj epoha = 3		
BERT Base Multilingual Cased	**0,8612**	0,8441
DistilBERT Base Multilingual Cased	0,8189	0,8096
XLM MLM	0,8609	0,8454

5.1.4 Poređenje i diskusija rezultata modela za određivanje semantičke sličnosti kratkih tekstova na *STS.news.sr* korpusu

Na slici 12 grafički je predstavljeno poređenje najboljih rezultata evaluacije različitih vrsta modela za određivanje semantičke sličnosti kratkih tekstova na *STS.news.sr* korpusu. Na osnovu njega može se uočiti da fino podešavanje neuralnih jezičkih modela zasnovanih na *Transformer* arhitekturama dovodi do primetno boljih rezultata od svih ostalih razmotrenih pristupa. Pored toga, iako su svi novi namenski modeli, opisani u odeljku 5.1.2, bolji od osnovnih nenadgledanih i nadgledanih pristupa, razlika između najboljeg od njih, *POS-TF STSS* modela, i osnovnih referentnih rešenja je znatno

manja od razlike između *POS-TF STSS* i neuralnih modela. Stoga se može zaključiti da je, u slučaju dostupnosti neuralnih jezičkih modela za posmatrani jezik sa ograničenim resursima, njihova upotreba jasno bolja opcija od klasičnijih pristupa. Njihova dodatna prednost jeste to što daju jako dobre rezultate bez potrebe za morfološkom normalizacijom ili sintaktičkom analizom teksta.

Pošto je za anotaciju *STS.news.sr* korpusa korišćena standardizovana metodologija koja je primenjivana i za izradu anotiranih skupova podataka ovog tipa na engleskom jeziku, moguće je na osnovu podataka iz literature razmotriti ponašanje modela između srpskog i engleskog jezika. Kao standard za evaluaciju modela za određivanje semantičke sličnosti kratkih tekstova na engleskom jeziku koristi se skup *STS Benchmark* (Cer et al. 2017) koji sadrži oko 7600 parova rečenica. Ovaj skup je domenski raznovrstan jer predstavlja amalgam većeg broja manjih skupova podataka anotiranih tokom *SemEval* takmičenja (Agirre et al. 2012, 2013, 2014, 2015, 2016; Cer et al. 2017), koji se međusobno razlikuju po domenu i veličini. Stoga nije primereno direktno porediti rezultate modela na *STS.news.sr* i *STS Benchmark* skupovima, ali je moguće proceniti da su dobijeni nivoi Pirsonovog koeficijenta korelacije r uporedivi između jezika, kao što je pokazano u tabeli 14 za sve razmotrene modele za koje su dostupni podaci iz literature. Na osnovu ovog pregleda se može zaključiti da fino podešavanje neuralnih jezičkih modela dovodi do sličnog nivoa performansi čak i na jezicima sa ograničenim resursima.



Slika 12. Poređenje najboljih rezultata različitih pristupa na problemu određivanja semantičke sličnosti kratkih tekstova iz *STS.news.sr* korpusa

Tabela 14. Pregled rezultata modela za određivanje semantičke sličnosti kratkih tekstova na srpskom i na engleskom jeziku, izraženih u vidu Pirsonovog koeficijenta korelacije r

Model	Performanse na *STS.news.sr* skupu podataka (SR)	Performanse na *STS Benchmark* skupu podataka (EN)
Prosek *word2vec* vektora značenja reči	0,621	0,565 (Cer et al. 2017)
BERT	0,861	0,871[28] (Devlin et al. 2019)
DistilBERT	0,819	0,812 (Sanh et al. 2019)
XLM	0,861	0,888[29] (Conneau & Lample 2019)

5.2 Formulisanje, obučavanje i evaluacija modela za problem analize sentimenta kratkih tekstova

Za obučavanje i evaluaciju modela za problem analize sentimenta upotrebljen je glavni anotirani *SentiComments.SR* skup podataka. Naime, kao što je opisano u odeljku 4.2.2, verifikacioni korpusi *SentiComments.SR.verif.movies* i *SentiComments.SR.verif.books* su napravljeni sa ciljem merenja kvaliteta, efikasnosti i ekonomičnosti anotacije sentimenta u okviru novog sistema označavanja, te za njih nije sprovedeno usaglašavanje ocena sentimenata između anotatora. Zbog toga, kao i zbog male količine podataka u navedenim korpusima, oni nisu korišćeni u obučavanju i evaluaciji modela.

Evaluacija modela za analizu sentimenta je, zahvaljujući fleksibilnosti razvijenog sistema označavanja sentimenta, sprovedena za više interpretacija oznaka sentimenta pomenutih u odeljku 4.2.2.1, tj. za sledećih pet problema klasifikacije:

- Određivanje polarnosti teksta;
- Određivanje subjektivnosti teksta;
- Četvoroklasnu klasifikaciju sentimenta teksta;
- Šestoklasnu klasifikaciju sentimenta teksta;
- Detekciju sarkastičnih tekstova.

U svim eksperimentima, tekstovi komentara su preslovljeni sa ćiriličnog na latinično pismo, a zatim tokenizovani pomoću ReLDI tokenizatora za srpski jezik, primenjenog i u pripremi podataka za problem određivanja semantičke sličnosti. Nakon toga, u okviru svakog od navedenih klasifikacionih problema ispitani su efekti različitih tehnika pretprocesiranja tekstualnih podataka. Najpre je razmotrena korisnost korekcije slovnih grešaka i nedostajućih dijakritičkih znakova, koja je ručno sprovedena u okviru anotacije podataka. Zatim su evaluirana dva postupka normalizacije teksta. Prvi od njih se odnosi na emotikone, koji su normalizovani u tri klase – pozitivne, negativne i dvosmislene – korišćenjem rečnika emotikona. Drugi postupak normalizacije se tiče ponavljanja pojedinačnih karaktera ili nizova karaktera, koji su normalizovani na jedno javljanje, uz ubacivanje posebnog *CHAR_REP* tokena u normalizovani tekst. Ovakvu proceduru je motivisalo to što ponavljanja karaktera često služe da izraze postojanje emotivnog naboja u tekstu (Thelwall et al. 2010).

[28] https://gluebenchmark.com/leaderboard
[29] https://github.com/facebookresearch/XLM

Kao i kod problema određivanja semantičke sličnosti, i pri evaluaciji modela za analizu sentimenta izvršeno je poređenje efekata različitih rešenja za morfološku normalizaciju tekstova na srpskom jeziku. Razmatran je isti skup stemera i lematizatora koji je već predstavljen u uvodu glave 5.

Svi modeli su evaluirani korišćenjem desetoslojne stratifikovane unakrsne validacije, dok je težinski uprosečena F-mera služila za merenje performansi klasifikatora. Jedini izuzetak od ovoga predstavlja problem detekcije sarkastičnih tekstova gde je, zbog ogromne neizbalansiranosti podataka po klasama, F-mera sarkastične klase upotrebljavana za merenje performansi. U ostatku ovog poglavlja, najpre je prikazana evaluacija osnovnih modela za analizu sentimenta kratkih tekstova i njihovih unapređenja, a zatim je razmotreno fino podešavanje neuralnih jezičkih modela.

5.2.1 Osnovni modeli za analizu sentimenta kratkih tekstova i njihova unapređenja

U ovom odeljku su predstavljeni osnovni modeli za analizu sentimenta, njihova unapređenja i rezultati. U sklopu toga, najpre je razmotren skup linearnih klasifikatora, implementiranih u *Scikit-learn* biblioteci (Pedregosa et al. 2011), koji koriste odlike dobijene po principu vreće reči (engl. *bag-of-words* – BOW), odnosno vreće *n*-grama (engl. *bag-of-n-grams*), gde se svaki tekst predstavlja u vidu neuređenog skupa reči/*n*-grama u njima, a svaka reč/*n*-gram predstavlja jednu odliku. Nakon toga su ispitani linearni modeli koji koriste odlike dobijene po principu vreće vektora značenja reči (engl. *bag-of-embeddings* – BOE), gde se svaki tekst predstavlja u vidu proseka vektora značenja reči u njemu, a svaka dimenzija takvog uprosečenog vektora se tretira kao odlika u klasifikaciji.

U okviru evaluacije ovih modela razmotrene su i različite mogućnosti njihovog unapređenja upotrebom šireg spektra tehnika za pretprocesiranje teksta, pomenutih u uvodu glave 5. Pošto je ranije opisana tehnika obeležavanja negiranih reči dala obećavajuće rezultate na problemima analize sentimenta, razmotreni su različiti opsezi primene te tehnike, od minimalnog obeležavanja samo jedne reči iza negacije do maksimalnog obeležavanja svih reči između negacije i prvog narednog znaka interpunkcije. U ovim eksperimentima praćena su pravila i obrasci negiranja koji su predloženi u (Ljajić 2019; Ljajić & Marovac 2019).

Zbog velikog broja opcija koje je trebalo razmotriti za sve pomenute modele, one su u evaluaciji grupisane u određene logičke celine – npr. jedna grupa opcija se odnosi na osnovne tehnike pretprocesiranja teksta, druga na metode morfološke normalizacije, itd. Razmatranju različitih grupa podešavanja se pristupalo iterativno – optimalne opcije iz prethodno evaluiranih grupa su korišćene u eksperimentima za naredne grupe. Pri odabiru optimalne opcije unutar svake grupe podešavanja, uziman je u obzir njihov efekat na sve klasifikatore, odnosno na sve klasifikacione probleme. Da bi se izbegao rizik od preterane prilagođenosti odabira, kao optimalne opcije nisu usvajane one koje imaju pozitivan efekat samo u nekim situacijama, a negativan u drugim. U slučaju da takvih opcija nije ni bilo u okviru posmatrane grupe, nije usvajano nijedno od razmatranih podešavanja. U nastavku ovog odeljka najpre su prikazani rezultati evaluacije *bag-of-words* modela, a zatim su izloženi rezultati *bag-of-embeddings* modela.

5.2.1.1 Bag-of-words modeli

U ovom odeljku prikazani su rezultati linearnih klasifikatora koji koriste odlike dobijene po principu vreće reči. Razmotreni su sledeći klasifikatori: multinomijalni (engl. *Multinomial Naïve Bayes* – MNB) i komplementni naivni Bajesov klasifikator (engl. *Complement Naïve Bayes* – CNB) (Rennie et al. 2003), kao i logistička regresija (LR) i metoda potpornih vektora (engl. *Support Vector Machine* – SVM) bez kernela. Za poslednja dva navedena algoritma korišćena je numerička implementacija

na osnovu LIBLINEAR biblioteke (Fan et al. 2008). Za probleme binarne klasifikacije, u obzir je uzet i NBSVM algoritam, koji predstavlja kombinaciju naivnog Bajesovog klasifikatora i metode potpornih vektora i za koji je pokazano da ih u binarnoj klasifikaciji tekstova nadmašuje (Wang & Manning 2012). Kao što je predloženo u (Wang & Manning 2012), u logističkoj regresiji, metodi potpornih vektora i NBSVM-u upotrebljavana je L_2 funkcija gubitka i L_2 regularizacija.

Za optimizaciju hiperparametra $C \in [10^{-2}, 10^{2}]$ koji se koristi u logističkoj regresiji, metodi potpornih vektora i NBSVM algoritmu, kao i za optimizaciju $\beta \in \{0{,}25\,, 0{,}5\}$ hiperparametra NBSVM algoritma, upotrebljavana je petoslojna ugnježđena stratifikovana unakrsna validacija. Svi ostali hiperparametri modela su postavljeni na podrazumevane vrednosti. Pri klasifikaciji svi tekstovi su normalizovani na mala slova.

U tabelama 15 – 19 su prikazani rezultati evaluacije navedenih *bag-of-words* klasifikatora na problemima određivanja polarnosti teksta, određivanja subjektivnosti teksta, četvoroklasne klasifikacije sentimenta teksta, šestoklasne klasifikacije sentimenta teksta i detekcije sarkastičnih tekstova. Evaluacija je za svaki od problema sprovedena istim redosledom. Najpre su ispitani efekti osnovnog pretprocesiranja tekstualnih podataka, a zatim su međusobno upoređene različite tehnike morfološke normalizacije. Nakon toga je razmatrano obeležavanje negiranih reči, a na kraju je evaluirano težinsko ponderisanje na osnovu TF-IDF (engl. *Term Frequency – Inverse Document Frequency*) vrednosti, kao i dodavanje bigramskih i trigramskih odlika.

Ime svake grupe podešavanja je u tabelama navedeno masnim slovima, dok su ispod imena prikazana ostala podešavanja korišćena u evaluaciji te grupe. Najbolji rezultati za svaki klasifikator su obeleženi masnim slovima, ako su bolji od početnih vrednosti za tu grupu i klasifikator, dok je usvojena optimalna opcija unutar svake grupe zasenčena. Za pomenutu normalizaciju emotikona i ponavljanja karaktera, u tabelama je korišćena skraćenica NEPK. Oznaka (S) u tabelama znači da se radi o stemeru, dok oznaka (L) predstavlja lematizatore.

Tabela 15 sadrži rezultate *bag-of-words* klasifikatora na problemu određivanja polarnosti teksta, gde automatski odabir najfrekventnije klase dovodi do početnog referentnog nivoa F-mere od 0,466. Performanse svih klasifikatora pokazuju da korekcija slovnih i dijakritičkih grešaka ima konzistentno pozitivne efekte, kao i normalizacija emotikona i ponavljanja karaktera. Morfološka normalizacija je takođe korisna, pri čemu je stemer Ljubešića i Pandžića najčešće najbolji izbor. Pored toga, i obeležavanje negiranih reči je od koristi, ali se različiti opsezi negiranja pokazuju dobro kod različitih klasifikatora. Iz tog razloga, kao optimalan opseg odabrano je obeležavanje dve reči nakon negacije, pošto je ovo jedino podešavanje koje nije snižavalo performanse nijednog modela. Efekti TF-IDF ponderisanja su nekonzistentni – pozitivan uticaj postoji kod logističke regresije i metode potpornih vektora, uz znatno negativniji uticaj na ostala tri algoritma. Slično tome, dodavanje bigramskih i trigramskih odlika ima zanemarljiv pozitivan uticaj na logističku regresiju i metodu potpornih vektora, ali primetno negativan na preostale modele.

Tabela 16 sadrži rezultate *bag-of-words* klasifikatora na problemu određivanja subjektivnosti teksta, gde automatski odabir najfrekventnije klase dovodi do početnog referentnog nivoa F-mere od 0,743. Kao i na problemu određivanja polarnosti teksta, i ovde se ispravke slovnih grešaka i predložene metode normalizacije pokazuju korisnim. Morfološka normalizacija takođe ponovo ima pozitivne efekte, pri čemu se kao najbolja rešenja izdvajaju stemer Ljubešića i Pandžića i lematizator koji su predstavili Ljubešić et al. Radi konzistentnosti dalje evaluacije na svim problemima, i ovde je kao optimalno rešenje usvojen pomenuti stemer. Nasuprot rezultatima na problemu određivanja polarnosti teksta, pri određivanju subjektivnosti utvrđeno je da obeležavanje negiranih reči ne pomaže gotovo nijednom klasifikatoru. Uticaj TF-IDF ponderisanja zavisi od modela, kao i na prethodnom problemu, dok je dodavanje bigramskih i trigramskih odlika konzistentno negativno za sve modele.

Tabela 15. Rezultati *bag-of-words* klasifikatora na problemu određivanja polarnosti kratkih tekstova iz glavnog *SentiComments.SR* korpusa, izraženi u vidu težinski uprosečene F-mere

Podešavanje	MNB	CNB	LR	SVM	NBSVM
Osnovna pretprocesiranja teksta Unigramske odlike					
Originalni tekstovi	0,688	0,710	0,720	0,708	0,717
Korigovani tekstovi	0,716	0,731	0,731	0,731	0,739
Korigovani tekstovi + NEPK	**0,746**	**0,755**	**0,755**	**0,748**	**0,760**
Morfološka normalizacija Korigovani tekstovi + NEPK, unigramske odlike					
(S) Kešelj & Šipka – optimalni	0,760	**0,775**	0,757	0,757	**0,772**
(S) Kešelj & Šipka – pohlepni	0,754	0,767	0,756	**0,760**	**0,772**
(S) Milošević	0,761	0,773	0,756	0,755	**0,772**
(S) Ljubešić i Pandžić	**0,765**	0,773	**0,758**	0,756	**0,772**
(L) BTagger – sufiks	0,746	0,764	0,752	0,748	0,758
(L) BTagger – prefiks + sufiks	0,749	0,767	0,751	0,753	0,760
(L) Agić et al.	0,749	0,764	0,745	0,743	0,762
(L) Ljubešić et al.	0,761	**0,775**	0,756	0,751	0,766
Obeležavanje negiranih reči Korigovani tekstovi + NEPK, (S) Ljubešić i Pandžić, unigramske odlike					
Opseg negacija = 1 token	0,762	0,776	0,762	0,759	0,774
Opseg negacija = 2 tokena	0,767	0,776	0,762	0,758	**0,782**
Opseg negacija = 3 tokena	**0,769**	**0,779**	0,759	0,753	**0,782**
Opseg negacija = 5 tokena	0,764	0,775	0,761	0,761	**0,782**
Opseg negacija = do interpunkcije	0,754	0,771	**0,767**	**0,766**	0,778
Ponderisanje i n-grami Korigovani tekstovi + NEPK, (S) Ljubešić i Pandžić, opseg negacija = 2 tokena					
TFIDF ponderisanje	0,746	0,749	**0,772**	**0,770**	0,765
Unigramske + bigramske odlike	0,735	0,751	0,766	0,762	0,769
Unigramske + bigramske + trigramske odlike	0,720	0,732	0,760	0,759	0,768

Tabela 16. Rezultati *bag-of-words* klasifikatora na problemu određivanja subjektivnosti kratkih tekstova iz glavnog *SentiComments.SR* korpusa, izraženi u vidu težinski uprosečene F-mere

Podešavanje	MNB	CNB	LR	SVM	NBSVM
Osnovna pretprocesiranja teksta Unigramske odlike					
Originalni tekstovi	0,784	0,820	0,841	0,837	0,830
Korigovani tekstovi	0,787	0,823	0,845	0,843	0,836
Korigovani tekstovi + NEPK	**0,793**	**0,833**	**0,865**	**0,861**	**0,849**
Morfološka normalizacija Korigovani tekstovi + NEPK, unigramske odlike					
(S) Kešelj & Šipka – optimalni	0,807	0,853	0,870	0,865	0,858
(S) Kešelj & Šipka – pohlepni	0,805	0,849	0,866	0,861	0,858
(S) Milošević	**0,808**	0,850	0,868	0,868	0,861
(S) Ljubešić i Pandžić	0,807	**0,855**	0,871	0,863	**0,864**
(L) BTagger – sufiks	0,795	0,847	0,876	0,869	0,859
(L) BTagger – prefiks + sufiks	0,796	0,846	0,873	0,873	0,858
(L) Agić et al.	0,794	0,841	0,875	0,868	0,846
(L) Ljubešić et al.	0,804	0,854	**0,879**	**0,875**	0,862
Obeležavanje negiranih reči Korigovani tekstovi + NEPK, (S) Ljubešić i Pandžić, unigramske odlike					
Opseg negacija = 1 token	0,806	0,851	0,867	0,863	0,856
Opseg negacija = 2 tokena	0,803	0,847	0,869	**0,864**	0,861
Opseg negacija = 3 tokena	0,800	0,847	0,865	0,861	0,862
Opseg negacija = 5 tokena	0,799	0,844	0,861	0,857	0,857
Opseg negacija = do interpunkcije	0,797	0,841	0,864	0,858	0,859
Ponderisanje i n-grami Korigovani tekstovi + NEPK, (S) Ljubešić i Pandžić, bez obeležavanja negiranih reči					
TFIDF ponderisanje	**0,861**	**0,864**	0,869	**0,869**	0,846
Unigramske + bigramske odlike	0,797	0,819	0,867	0,862	0,856
Unigramske + bigramske + trigramske odlike	0,791	0,811	0,859	0,856	0,842

Tabela 17. Rezultati *bag-of-words* klasifikatora na problemu četvoroklasne klasifikacije sentimenta kratkih tekstova iz glavnog *SentiComments.SR* korpusa, izraženi u vidu težinski uprosečene F-mere

Podešavanje	MNB	CNB	LR	SVM
Osnovna pretprocesiranja teksta Unigramske odlike				
Originalni tekstovi	0,460	0,544	0,567	0,566
Korigovani tekstovi	0,498	0,562	0,579	0,584
Korigovani tekstovi + NEPK	**0,523**	**0,602**	**0,616**	**0,617**
Morfološka normalizacija Korigovani tekstovi + NEPK, unigramske odlike				
(S) Kešelj & Šipka – optimalni	0,558	0,600	**0,629**	0,626
(S) Kešelj & Šipka – pohlepni	0,556	0,599	0,626	0,624
(S) Milošević	**0,563**	0,603	0,610	0,613
(S) Ljubešić i Pandžić	0,560	**0,613**	**0,629**	**0,627**
(L) BTagger – sufiks	0,545	0,599	0,621	0,620
(L) BTagger – prefiks + sufiks	0,542	0,600	0,618	0,615
(L) Agić et al.	0,526	0,590	0,615	0,610
(L) Ljubešić et al.	0,558	0,603	0,622	0,624
Obeležavanje negiranih reči Korigovani tekstovi + NEPK, (S) Ljubešić i Pandžić, unigramske odlike				
Opseg negacija = 1 token	0,560	**0,616**	**0,640**	**0,631**
Opseg negacija = 2 tokena	**0,562**	0,613	0,632	0,628
Opseg negacija = 3 tokena	0,555	0,609	0,626	0,628
Opseg negacija = 5 tokena	0,550	0,609	0,630	0,625
Opseg negacija = do interpunkcije	0,542	0,603	0,625	0,623
Ponderisanje i n-grami Korigovani tekstovi + NEPK, (S) Ljubešić i Pandžić, opseg negacija = 1 token				
TFIDF ponderisanje	**0,595**	0,563	0,630	**0,633**
Unigramske + bigramske odlike	0,504	0,597	0,623	0,618
Unigramske + bigramske + trigramske odlike	0,477	0,576	0,615	0,617

Tabela 18. Rezultati *bag-of-words* klasifikatora na problemu šestoklasne klasifikacije sentimenta kratkih tekstova iz glavnog *SentiComments.SR* korpusa, izraženi u vidu težinski uprosečene F-mere

Podešavanje	MNB	CNB	LR	SVM
Osnovna pretprocesiranja teksta Unigramske odlike				
Originalni tekstovi	0,384	0,484	0,506	0,498
Korigovani tekstovi	0,403	0,501	0,522	0,515
Korigovani tekstovi + NEPK	**0,417**	**0,532**	**0,555**	**0,547**
Morfološka normalizacija Korigovani tekstovi + NEPK, unigramske odlike				
(S) Košelj & Šipka – optimalni	0,441	0,536	**0,557**	0,554
(S) Košelj & Šipka – pohlepni	0,436	0,533	0,551	0,550
(S) Milošević	0,443	0,539	0,556	0,553
(S) Ljubešić i Pandžić	**0,444**	0,540	**0,557**	0,547
(L) BTagger – sufiks	0,435	0,538	0,549	0,555
(L) BTagger – prefiks + sufiks	0,437	0,537	0,554	**0,556**
(L) Agić et al.	0,421	0,526	0,548	0,541
(L) Ljubešić et al.	0,442	**0,548**	**0,557**	0,555
Obeležavanje negiranih reči Korigovani tekstovi + NEPK, (S) Ljubešić i Pandžić, unigramske odlike				
Opseg negacija = 1 token	0,443	**0,547**	0,562	0,560
Opseg negacija = 2 tokena	**0,445**	0,546	**0,566**	0,561
Opseg negacija = 3 tokena	0,442	0,539	0,563	0,561
Opseg negacija = 5 tokena	0,439	0,541	0,563	**0,562**
Opseg negacija = do interpunkcije	0,436	0,534	0,565	0,560
Ponderisanje i n-grami Korigovani tekstovi + NEPK, (S) Ljubešić i Pandžić, opseg negacija = 2 tokena				
TFIDF ponderisanje	**0,531**	0,510	**0,569**	**0,569**
Unigramske + bigramske odlike	0,409	0,516	0,556	0,554
Unigramske + bigramske + trigramske odlike	0,399	0,504	0,552	0,544

Tabela 19. Rezultati *bag-of-words* klasifikatora na problemu detekcije sarkastičnih kratkih tekstova iz glavnog *SentiComments.SR* korpusa, izraženi u vidu F-mere klase sarkastičnih tekstova

Podešavanje	MNB	CNB	LR	SVM	NBSVM
Osnovna pretprocesiranja teksta Unigramske odlike					
Originalni tekstovi	0	0	0,040	0,059	0,038
Korigovani tekstovi	0	0	0,030	0,043	0,055
Korigovani tekstovi + NEPK	0	0	**0,096**	**0,124**	**0,070**
Morfološka normalizacija Korigovani tekstovi + NEPK, unigramske odlike					
(S) Kešelj & Šipka – optimalni	0	**0,014**	**0,167**	0,175	0,052
(S) Kešelj & Šipka – pohlepni	0	0	0,104	0,137	0,040
(S) Milošević	0	0	0,121	0,141	0,060
(S) Ljubešić i Pandžić	0	0	0,112	**0,186**	0,054
(L) BTagger – sufiks	0	0	0,095	0,115	0,038
(L) BTagger – prefiks + sufiks	0	0	0,107	0,129	0,053
(L) Agić et al.	0	**0,014**	0,082	0,098	**0,079**
(L) Ljubešić et al.	0	0	0,100	0,115	0,072
Obeležavanje negiranih reči Korigovani tekstovi + NEPK, bez morfološke normalizacije, unigramske odlike					
Opseg negacija = 1 token	0	0	0,029	0,095	0,070
Opseg negacija = 2 tokena	0	0	0,014	0,079	**0,072**
Opseg negacija = 3 tokena	0	0	0,031	0,081	0,045
Opseg negacija = 5 tokena	0	0	0,014	0,105	0,057
Opseg negacija = do interpunkcije	0	0	0,046	0,108	0,058
Ponderisanje i n-grami Korigovani tekstovi + NEPK, bez morfološke normalizacije, bez obeležavanja negiranih reči					
TFIDF ponderisanje	**0,156**	**0,149**	0,071	0,081	0,017
Unigramske + bigramske odlike	0	0,015	0,015	0,030	0
Unigramske + bigramske + trigramske odlike	0	0	0	0	0

Tabela 17 sadrži rezultate *bag-of-words* klasifikatora na problemu četvoroklasne klasifikacije sentimenta teksta, gde automatski odabir najfrekventnije klase dovodi do početnog referentnog nivoa F-mere od 0,244. I ovde se uočavaju isti obrasci kao i na prethodna dva klasifikaciona problema – korekcije slovnih grešaka, predložene metode normalizacije teksta i morfološka normalizacija, pogotovo u vidu stemera Ljubešića i Pandžića, kod svih modela dovode do boljih rezultata. Obeležavanje negiranih reči se ponovo pokazuje korisnim, najverovatnije zbog prisustva polarnih oznaka sentimenta među klasama. Pri tome se kao optimalni opseg obeležavanja ističe jedna reč iza negacija. TF-IDF ponderisanje opet dovodi do mešovitih efekata, dok je dodavanje bigramskih i trigramskih odlika loše po sve razmotrene modele.

Tabela 18 sadrži rezultate *bag-of-words* klasifikatora na problemu šestoklasne klasifikacije sentimenta teksta, gde automatski odabir najfrekventnije klase dovodi do početnog referentnog nivoa F-mere od 0,244. Kao i na prethodno razmotrenim problemima, i ovde predložene osnovne tehnike pretprocesiranja teksta i morfološka normalizacija pomažu svim klasifikatorima. Ponovo se kao optimalna rešenja za morfološku normalizaciju izdvajaju stemer Ljubešića i Pandžića i lematizator od Ljubešića et al, pri čemu je opet radi konzistentnosti odabrana primena stemera u narednim eksperimentima. Obeležavanje negiranih reči sa opsegom od dve reči nakon negacije se pokazuje kao najbolja opcija na ovom problemu, dok su efekti TF-IDF ponderisanja i dodavanja *n*-gramskih odlika višeg reda isti kao i kod četvoroklasne klasifikacije sentimenta tekstova.

Tabela 19 sadrži rezultate *bag-of-words* klasifikatora na problemu detekcije sarkastičnih tekstova, gde automatski odabir najfrekventnije klase dovodi do početnog referentnog nivoa binarne F-mere od 0. Detekcija sarkastičnih tekstova je jedan od najtežih problema u analizi sentimenta, te ne čudi što su čak i opcije koje su na prethodnim problemima uvek bile pozitivne ovde od male ili upitne koristi. Predložene osnovne tehnike normalizacije teksta dovode do blagih poboljšanja rezultata klasifikatora, ali su efekti morfološke normalizacije, obeležavanja negiranih reči, TF-IDF ponderisanja i dodavanja bigramskih i trigramskih odlika veoma nekonzistentni i vrlo često negativni. Ovo je najverovatnije posledica vrlo malog broja primera sarkastičnih tekstova u *SentiComments.SR* korpusu, što dovodi do veoma neizbalansiranog klasifikacionog problema i povećava varijansu svih rezultata na ovom problemu. Stoga je zaključeno da je za izvlačenje pouzdanih zaključaka o vrednosti razmatranih opcija na problemu detekcije sarkastičnih tekstova neophodan veći skup podataka, sa bogatijim podskupom primera sarkastičnih tekstova, te na *SentiComments.SR* korpusu nisu sprovođeni dalji eksperimenti vezani za ovaj problem.

Sveukupno, NBSVM algoritam se pokazuje kao najbolji izbor za određivanje polarnosti teksta, dok se logistička regresija i metoda potpornih vektora ističu u višeklasnoj klasifikaciji. Ispravljanje slovnih grešaka, normalizacija emotikona i ponavljanja karaktera i morfološka normalizacija su od koristi na svim problemima iz analize sentimenta, kada se primenjuju uz odlike dobijene po principu vreće reči, dok se uticaj označavanja negiranih reči razlikuje od problema do problema. Ipak, tamo gde ima pozitivan efekat, obeležavanje negiranih reči je najefektivnije uz vrlo ograničen opseg označavanja tj. uz obeležavanje samo jedne do dve reči iza negacije.

5.2.1.2 Bag-of-embeddings modeli

U ovom odeljku predstavljeni su rezultati evaluacije linearnih SVM modela sa odlikama zasnovanim na uprosečavanju vektora značenja reči iz teksta, tj. odlikama dobijenim po principu vreće vektora značenja reči. Vektori značenja reči su, kao i u poglavlju 5.1, dobijeni primenom *word2vec* algoritma (Mikolov et al. 2013a, b), implementiranog u *gensim* biblioteci (Řehůřek & Sojka 2010). Pri tome je ponovo korišćena njegova *skip-gram* varijanta, koja je u preliminarnim eksperimentima pokazala konzistentno bolje rezultate od alternativnog CBOW (engl. *continuous bag-of-words*) pristupa. Takođe su sprovedeni preliminarni eksperimenti i sa novijim *fastText* algoritmom (Bojanowski et al.

2017), ali se pokazalo da vektori značenja reči dobijeni ovim putem dovode do vrlo sličnih performansi kao i oni generisani putem *word2vec* modela. Međutim, obučavanje *fastText* modela je primetno sporije od *word2vec* algoritma, zbog čega *fastText* nije upotrebljavan u daljoj evaluaciji. Kao i pri izradi modela za određivanje semantičke sličnosti, i ovde je za pravljenje vektora značenja reči korišćen srpski veb korpus *srWaC* (Ljubešić & Klubička 2014).

Za SVM model je ponovo korišćena L_2 funkcija gubitka i L_2 oblik regularizacije, a optimizacija hiperparametra $C \in [10^{-2}, 10^{2}]$ opet je sprovedena kroz petoslojnu ugnježdenu stratifikovanu unakrsnu validaciju. Svi ostali hiperparametri modela su postavljeni na podrazumevane vrednosti.

Kao početno podešavanje korišćeni su vektori značenja reči sa 100 dimenzija, kreirani uz kontekstni opseg širine 10 reči, dok su ostali *word2vec* parametri ostali na podrazumevanim vrednostima za *skip-gram* arhitekturu. Tekstovi su normalizovani na mala slova, kako oni iz *srWaC* korpusa, korišćeni pri izradi vektora značenja reči, tako i oni iz *SentiComments.SR* korpusa, korišćeni pri obučavanju i evaluaciji klasifikatora. Za sve klasifikacione probleme prvo su razmotrene varijante osnovnog pretprocesiranja teksta, a onda one vezane za morfološku normalizaciju. Zatim su ispitane opcije koje se tiču dimenzionalnosti vektora značenja reči i širine kontekstnog opsega koji se koristi pri njihovoj izradi, a na kraju su evaluirane ostale opcije koje ne spadaju u prethodne grupe. Najbolji rezultati i usvojene optimalne opcije su obeležavani na isti način kao u prethodnom odeljku.

Rezultati iz tabele 20 pokazuju da su, i u ovom pristupu, tipografska i dijakritička ispravnost teksta, kao i normalizacija emotikona i ponavljanja karaktera, od koristi za rešavanje svih problema analize sentimenta. Ispitivanje efekata morfološke normalizacije je sprovedeno konstruisanjem odvojenih skupova vektora značenja reči za svaku morfološki normalizovanu varijantu *srWaC* korpusa. Kao i kod *bag-of-words* modela, i ovde se morfološka normalizacija konzistentno pokazuje korisnom, i ponovo se stemer Ljubešića i Pandžića izdvaja kao najbolje rešenje.

Uticaj dimenzionalnosti vektora značenja reči i širine kontekstnog opsega korišćenog pri njihovoj izradi je ispitan razmatranjem opsega dimenzionalnosti od 100 do 1000 i dve vrednosti širine opsega – 5 i 10. Ove vrednosti parametara su upotrebljene za obučavanje *word2vec* modela nad varijantom *srWaC* korpusa koja je prethodno morfološki normalizovana stemerom Ljubešića i Pandžića. Rezultati pokazuju da performanse klasifikatora rastu kako se povećava dimenzionalnost vektora i širina kontekstnog opsega. Ovaj efekat je najmanje izražen na problemu određivanja subjektivnosti teksta, a najočigledniji je na problemima višeklasne klasifikacije sentimenta. Najbolji rezultati se dobijaju za najveću vrednost dimenzionalnosti i širine opsega – 1000, odnosno 10.

Korišćenjem navedenih vrednosti parametara, ispitane su i dve dodatne opcije. Prvo je razmotrena prethodno predstavljena tehnika obeležavanja negiranih reči, tako što je u pretprocesiranom i stemovanom *srWaC* korpusu obeležena po jedna reč iza svake negacije, a zatim su pomoću tako označenog korpusa kreirani vektori značenja reči. Da bi se takvi vektori primenili za klasifikaciju, bilo je neophodno obeležiti negirane reči i u tekstovima iz *SentiComments.SR* korpusa, što je urađeno odvojeno za svaki klasifikacioni zadatak, u optimalnim opsezima utvrđenim u prethodnom odeljku. Međutim, ovaj pristup je doveo samo do blagog rasta performansi na problemu određivanja polarnosti teksta, po cenu blagog pada performansi na problemu četvoroklasne klasifikacije sentimenta.

Nakon toga, razmotreno je kombinovanje odlika dobijenih uprosečavanjem vektora značenja reči sa odlikama dobijenim po principu vreće reči, uz već utvrđena optimalna podešavanja za svaki klasifikacioni problem. Ova kombinacija, pogotovo u varijanti gde se kod *bag-of-words* odlika koristi i obeležavanje negiranih reči, skoro uvek pokazuje primetno bolje rezultate od upotrebe samo odlika dobijenih uprosečavanjem vektora značenja reči. Jedini izuzetak je problem određivanja polarnosti, gde se ne uočava razlika. Konačno, obeležavanje negiranih reči pri izradi vektora značenja reči dovodi do blago lošijih rezultata od običnih vektora značenja pri ovakvom kombinovanju odlika.

Tabela 20. Rezultati *bag-of-embeddings* SVM klasifikatora u analizi sentimenta kratkih tekstova iz glavnog *SentiComments.SR* korpusa, izraženi u vidu težinski uprosečene F-mere

Podešavanje	Problem			
	Određivanje polarnosti	Određivanje subjektivnosti	Četvoroklasna klasifikacija sentimenta	Šestoklasna klasifikacija sentimenta
Osnovna pretprocesiranja teksta Uprosečeni *word2vec* vektori, Dim = 100, WS = 10				
Originalni tekstovi	0,710	0,836	0,516	0,462
Korigovani tekstovi	0,720	0,842	0,536	0,474
Korigovani tekstovi + NEPK	**0,745**	**0,858**	**0,559**	**0,506**
Morfološka normalizacija Korigovani tekstovi + NEPK, uprosečeni *word2vec* vektori, Dim = 100, WS = 10				
(S) Kešelj & Šipka – optimalni	0,744	0,859	0,559	0,509
(S) Kešelj & Šipka – pohlepni	0,745	0,864	0,548	0,494
(S) Milošević	0,740	0,858	0,562	0,499
(S) Ljubešić i Pandžić	**0,761**	**0,866**	**0,566**	**0,511**
(L) BTagger – sufiks	0,748	0,858	0,562	0,510
(L) BTagger – prefiks + sufiks	0,746	0,861	0,557	0,502
(L) Agić et al.	0,747	0,851	0,553	0,498
(L) Ljubešić et al.	0,741	0,857	0,564	0,506
Dimenzionalnost vektora (Dim) i veličina kontekstnog opsega (WS) Korigovani tekstovi + NEPK, (S) Ljubešić i Pandžić, uprosečeni *word2vec* vektori				
Dim = 100, WS = 5	0,757	0,861	0,564	0,512
Dim = 100, WS = 10	0,761	0,866	0,566	0,511
Dim = 300, WS = 5	0,767	0,865	0,604	0,533
Dim = 300, WS = 10	0,770	0,869	0,599	0,535
Dim = 500, WS = 5	0,775	0,872	0,614	0,543
Dim = 500, WS = 10	0,777	0,870	0,616	0,546
Dim = 1000, WS = 5	**0,783**	0,871	0,626	**0,558**
Dim = 1000, WS = 10	**0,783**	**0,873**	**0,628**	0,557

Ostala podešavanja Korigovani tekstovi + NEPK, (S) Ljubešić i Pandžić, uprosečeni *word2vec* vektori, Dim = 1000, WS = 10				
Obeležavanje negiranih reči pri kreiranju vektora značenja reči	**0,788**	/	0,622	0,558
Dodavanje BOW odlika bez obeležavanja negiranih reči	0,780	**0,885**	**0,655**	0,576
Dodavanje BOW odlika sa optimalnim obeležavanjem negiranih reči za svaki problem	0,783	**0,885**	**0,655**	**0,586**
Obeležavanje negiranih reči pri kreiranju vektora značenja reči + dodavanje BOW odlika sa optimalnim obeležavanjem negiranih reči za svaki problem	0,778	/	0,652	0,585

5.2.2 Fino podešavanje neuralnih jezičkih modela radi analize sentimenta kratkih tekstova

U ovom odeljku je opisana evaluacija i rezultati neuralnih jezičkih modela zasnovanih na *Transformer* arhitekturama na problemima analize sentimenta kratkih tekstova iz *SentiComments.SR* korpusa. Korišćena su ista tri prethodno obučena višejezična modela primenjena i na problemu određivanja semantičke sličnosti kratkih tekstova i opisana u uvodu glave 5. Navedeni neuralni modeli su odvojeno podvrgnuti finom podešavanju u trajanju od jedne epohe na svakom od klasifikacionih problema iz analize sentimenta. Pri tome, razmotrene su različite tehnike pretprocesiranja teksta kroz četiri varijante tekstova iz *SentiComments.SR* korpusa:

- Originalni tekstovi, preslovljeni sa ćirilice na latinicu, gde je to bilo potrebno;
- Ručno korigovani tekstovi u kojima su uklonjene slovne i dijakritičke greške;
- Ručno korigovani tekstovi u kojima je pored tipografskih ispravki primenjena i normalizacija emotikona i ponavljanja karaktera;
- Stemovani ručno korigovani i normalizovani tekstovi, dobijeni pomoću stemera Ljubešića i Pandžića, koji se pokazao kao optimalan izbor za morfološku normalizaciju u analizi sentimenta na prethodno razmotrenim modelima.

Za fino podešavanje modela korišćena su podrazumevane vrednosti iz *Simple Transformers* biblioteke (*batch size* = 8, *learning rate* = 4e-5), pri čemu je očuvana originalna kapitalizacija slova, jer su navedeni modeli obučeni da i nju uzimaju u obzir. Rezultati ovako dobijenih modela, uprosečeni sa pet pokretanja sa različitim nasumičnim inicijalizacionim vrednostima, prikazani su u tabeli 21. Može se videti da neuralni modeli najčešće daju najbolje rezultate na korigovanim ali nenormalizovanim tekstovima, što je za očekivati jer je to vrsta tekstova na kojima su ovi modeli inicijalno i obučavani. Razlike između performansi na korigovanim i na normalizovanim verzijama korpusa su male, ali je primena stemovanja u ovoj situaciji očito kontraproduktivna. Slično tome, performanse na originalnim tekstovima su najvećim delom tek nešto ispod onih dobijenih na korigovanim tekstovima, ali razlike između ove dve varijante jesu nešto izraženije u određenim kombinacijama modela i klasifikacionih problema.

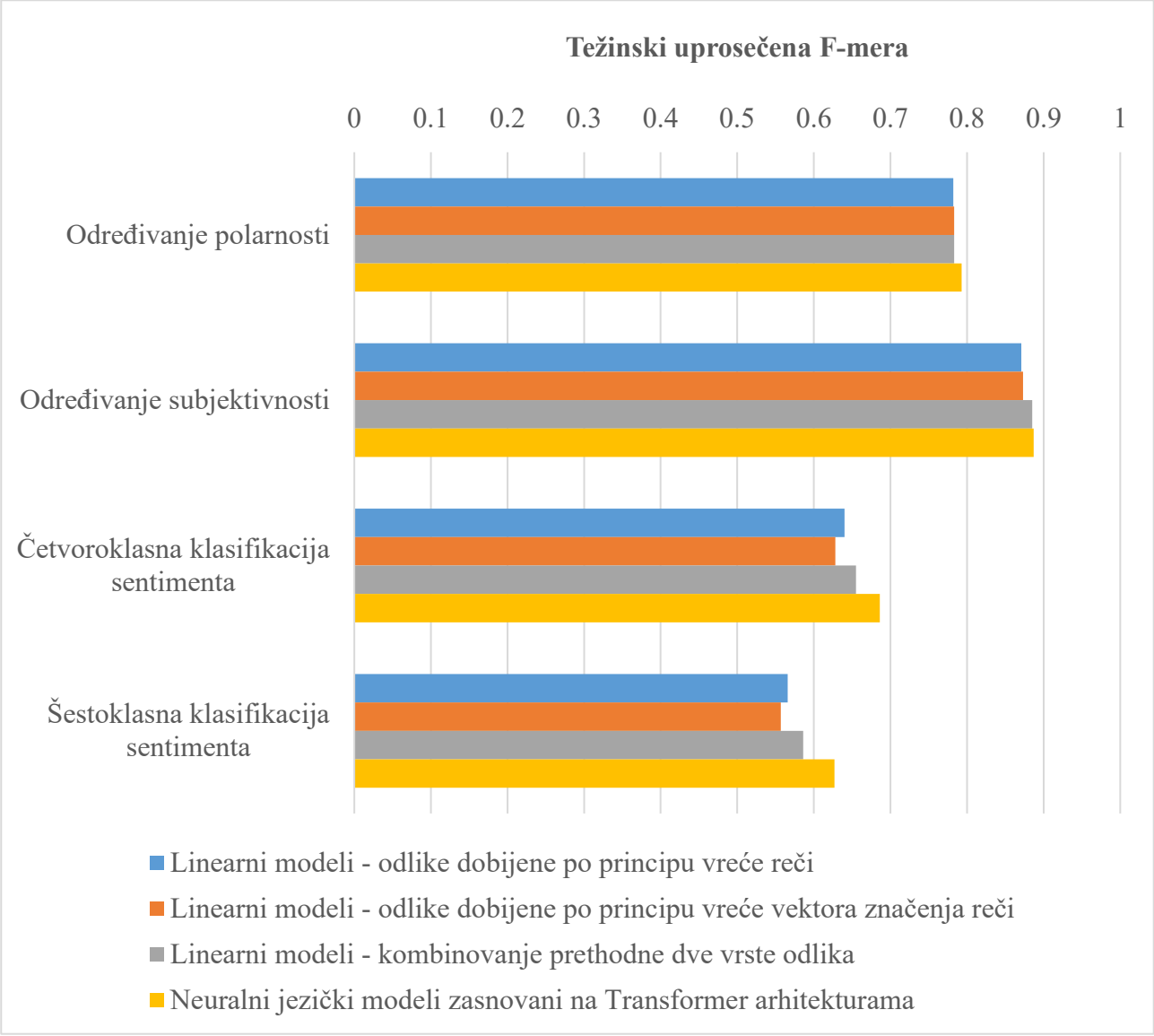
Tabela 21. Rezultati neuralnih jezičkih modela zasnovanih na *Transformer* arhitekturama u analizi sentimenta kratkih tekstova iz glavnog *SentiComments.SR* korpusa, izraženi u vidu težinski uprosečene F-mere

Model / Ulazni podaci		Problem			
		Određivanje polarnosti	Određivanje subjektivnosti	Četvoroklasna klasifikacija sentimenta	Šestoklasna klasifikacija sentimenta
Broj epoha = 1					
BERT Base Multilingual Cased	Originalni tekstovi	0,725	0,862	0,538	0,493
	Korigovani tekstovi	0,735	**0,872**	**0,578**	0,497
	Korigovani tekstovi + NEPK	**0,739**	0,867	0,573	**0,502**
	Stemovani tekstovi + NEPK	0,715	0,864	0,574	0,478
DistilBERT Base Multilingual Cased	Originalni tekstovi	0,720	0,864	**0,548**	0,455
	Korigovani tekstovi	**0,725**	**0,869**	0,545	**0,465**
	Korigovani tekstovi + NEPK	0,715	0,867	0,542	0,459
	Stemovani tekstovi + NEPK	0,713	0,857	0,538	0,451
XLM MLM	Originalni tekstovi	0,739	0,873	0,634	0,553
	Korigovani tekstovi	**0,788**	0,873	**0,647**	**0,571**
	Korigovani tekstovi + NEPK	0,779	**0,879**	0,646	0,547
	Stemovani tekstovi + NEPK	0,760	0,870	0,618	0,532
Broj epoha = 3 Korigovani tekstovi					
BERT Base Multilingual Cased		0,785	0,879	0,635	0,604
DistilBERT Base Multilingual Cased		0,772	0,883	0,634	0,576
XLM MLM		**0,793**	**0,887**	**0,686**	**0,627**

Nakon ovoga, dužina procedure finog podešavanja je podignuta na tri epohe, kao što je preporučeno za ovakve modele u (Devlin et al. 2019), pri čemu su korišćeni korigovani tekstovi kao ulaz. Rezultati jasno pokazuju da je *XLM* model najbolji na svim klasifikacionim problemima, dok *DistilBERT* model postiže performanse blizu onih koje ima puni *multilingual BERT*. Iako je pokušano da se rezultati neuralnih modela dodatno poprave podizanjem broja epoha na pet, to nije dovelo do konzistentnih poboljšanja.

5.2.3 Poređenje i diskusija rezultata modela za analizu sentimenta kratkih tekstova na *SentiComments.SR* korpusu

Na slici 13 grafički je predstavljeno poređenje najboljih rezultata evaluacije različitih vrsta modela za analizu sentimenta kratkih tekstova na *SentiComments.SR* korpusu. Na osnovu njega može se uočiti nekoliko jasnih trendova.



Slika 13. Poređenje najboljih rezultata različitih pristupa na problemima analize sentimenta tekstova iz *SentiComments.SR* korpusa

Najpre, jasno je da linearni modeli, čak i oni koji koriste samo odlike dobijene po principu vreće reči, predstavljaju veoma jaka osnovna rešenja kada se primenjuju uz jednostavne tehnike pretprocesiranja tekstova. Samostalno korišćenje odlika dobijenih uprosečavanjem statičkih vektora značenja reči je slično ili blago lošije od prethodne varijante, ali kombinovanje ove dve vrste odlika se pokazuje boljim od bilo kog individualnog rešenja. Ipak, finim podešavanjem neuralnih jezičkih modela zasnovanih na *Transformer* arhitekturama postižu se iste ili bolje performanse od linearnih modela. Odstupanja u rezultatima između različitih pristupa su mala na jednostavnijim zadacima binarne klasifikacije, u šta spada određivanje polarnosti i subjektivnosti teksta, ali su izraženija na komplikovanijim problemima višeklasne klasifikacije sentimenta tekstova.

Dodatna prednost neuralnih modela jeste to što oni daju jako dobre rezultate i bez razmotrenih tehnika pretprocesiranja teksta, što ih čini lako primenjivim. Međutim, ovakvi višejezični modeli su i dalje dostupni za samo stotinak jezika. Stoga je upotreba jednostavnijih modela nužnost za mnoge jezike sa ograničenim resursima. Ipak, kao što rezultati evaluacije pokazuju, jednostavnost modela ne mora da dovede do uočljivo lošijih rezultata na manje zahtevnim problemima binarne klasifikacije sentimenta.

Kao što je opisano u poglavlju 4.2, *SentiComments.SR* korpus je anotiran korišćenjem potpuno novog sistema označavanja, koji je razvijen u okviru ove disertacije. Stoga pre izrade anotiranih skupova podataka na drugim jezicima po novoj metodologiji nije moguće na adekvatan način uporediti rezultate razmatranih modela na srpskom jeziku sa performansama na nekom drugom jeziku. Naime, kao što je ilustrovano u odeljku 4.2.3, različiti sistemi anotacije sentimenta neretko koriste različite oznake za ista izražavanja sentimenta, te stoga poređenje performansi sistema na skupovima podataka anotiranih po različitim metodologijama ne bi bilo primereno.

6 Pregled metodologije razvoja statističkih rešenja semantičkih problema u prirodnim jezicima sa ograničenim resursima

Na osnovu opisanih iskustava i dobijenih rezultata u izradi rešenja za probleme određivanja semantičke sličnosti i analize sentimenta kratkih tekstova, u ovoj glavi je iznet pregled metodologije razvoja statističkih rešenja semantičkih problema u prirodnim jezicima sa ograničenim resursima. Proces razvoja po ovoj metodologiji je izložen po fazama i ilustrovan u vidu dijagrama na slikama 14 i 15.

Opšta odlika celokupne metodologije jeste da je u uslovima rada sa ograničenim resursima neophodno mudro i pažljivo konceptualno isplanirati svaku fazu razvoja, oslanjajući se na ranije predstavljena rešenja i standarde tamo gde je to moguće. Najpre, u fazi prikupljanja tekstualnog sadržaja za određeni NLP problem, prvo treba razmotriti da li na datom jeziku već postoje kvalitetni javno dostupni skupovi podataka dovoljnog obima koji bi se mogli prilagoditi za potrebe konkretnog NLP problema. Ukoliko to jeste slučaj, onda je preporučljivo da se prikupljanje tekstualnog sadržaja ne obavlja potpuno iz početka, već da se postojeći resursi prilagode i iskoriste za rešavanje novog problema. Na ovaj način se ne ostvaruje samo ušteda vremena u ovoj fazi razvoja, već se i omogućuje postojanje paralelnih oznaka različitog tipa nad istim podacima, što u nekim situacijama može biti od značaja za dalja istraživanja. U ovoj varijanti prikupljanja tekstualnog sadržaja često je potrebno isfiltrirati elemente skupa podataka radi kontrole njihovog kvaliteta i adekvatnosti u kontekstu aktuelnog NLP problema. U ovoj disertaciji, prikupljanje tekstualnog sadržaja na opisani način je ilustrovano na problemu određivanja semantičke sličnosti kratkih tekstova, u poglavlju 3.1, gde je kao polazni materijal za izradu novog anotiranog skupa podataka upotrebljen sadržaj postojećeg korpusa parafraza.

Ukoliko pak nije moguće pronaći javno dostupne skupove podataka koji bi bili adekvatni za upotrebu u okviru rešavanja novog NLP problema, ili ako je kvalitet ili obim postojećih skupova podataka upitan, onda je neophodno prikupiti sasvim nov tekstualni sadržaj. To najčešće podrazumeva prikupljanje podataka sa interneta, pri čemu je potrebno razmotriti dostupne izvore adekvatnih podataka za posmatrani problem, odabrati jedan ili više pogodnih izvora, i zatim sprovesti sakupljanje tekstualnog sadržaja i njegovo filtriranje. U ovoj disertaciji ova varijanta prikupljanja tekstualnog sadržaja je ilustrovana na problemu analize sentimenta, u poglavlju 3.2, izgradnjom novog skupa kratkih komentara iz filmskog i književnog domena.

Faza anotacije podataka je izuzetno važna u izradi svih statističkih rešenja, ali je zbog znatnog vremenskog i materijalnog utroška potrebnog za njeno sprovođenje verovatno ključna u kreiranju rešenja u jezicima sa ograničenim resursima. Naime, u ovakvim jezicima prirodna dinamika razvoja novih anotiranih skupova podataka ne dozvoljava laku i brzu zamenu loše osmišljenih ili nedovoljno kvalitetno anotiranih skupova, te stoga greške u ovoj fazi mogu biti dugoročno pogubne po formulisanje i evaluaciju statističkih modela.

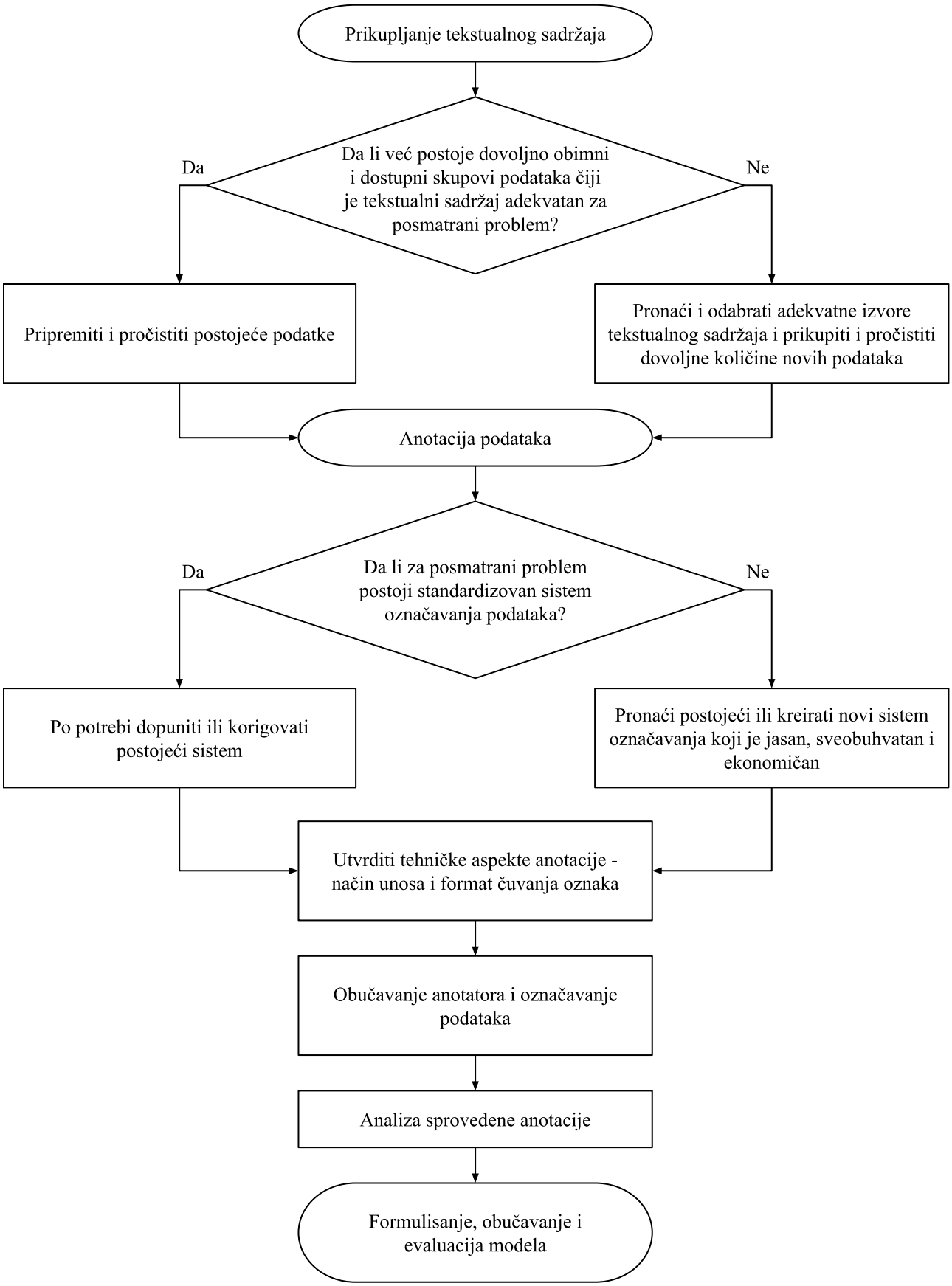
Kao što je već rečeno u uvodnoj glavi, u većini situacija za kvalitetnu anotaciju podataka u jezicima sa ograničenim resursima neophodno je direktno angažovanje tima anotatora. Pošto je time pitanje načina organizovanja anotacije rešeno, najvažnija preostala konceptualna odluka u ovoj fazi vezana je za izbor sistema označavanja podataka, što obuhvata skup oznaka i uputstava za anotaciju. Ako za razmatrani NLP problem već postoji prihvaćen međunarodno standardizovan sistem označavanja, tada je uglavnom najbolje držati se tog sistema. Takav pristup omogućava laku uporedivost novog anotiranog skupa podataka sa ranijim skupovima na drugim jezicima. On takođe pruža i određenu uporedivost rezultata istih modela između jezika, a doprinosom široj porodici resursa koji su svi

anotirani upotrebom standardizovane metodologije pospešuje se razvoj višejezičnih NLP modela. Od ovakvog standarda ima smisla odstupiti samo u slučaju uočavanja njegovih jasnih nedostataka. Ukoliko anotacija treba da obuhvati i neke aspekte NLP problema o kojima se u formulisanju standarda označavanja nije vodila pažnja, ili ako se uoče određene manjkavosti u standardu, preporučljivije je korigovati i dopuniti standardizovan sistem nego izrađivati potpuno nov. U ovoj disertaciji ovaj pristup anotaciji podataka je upotrebljen za označavanje semantičke sličnosti kratkih tekstova, kao što je opisano u poglavlju 4.1.

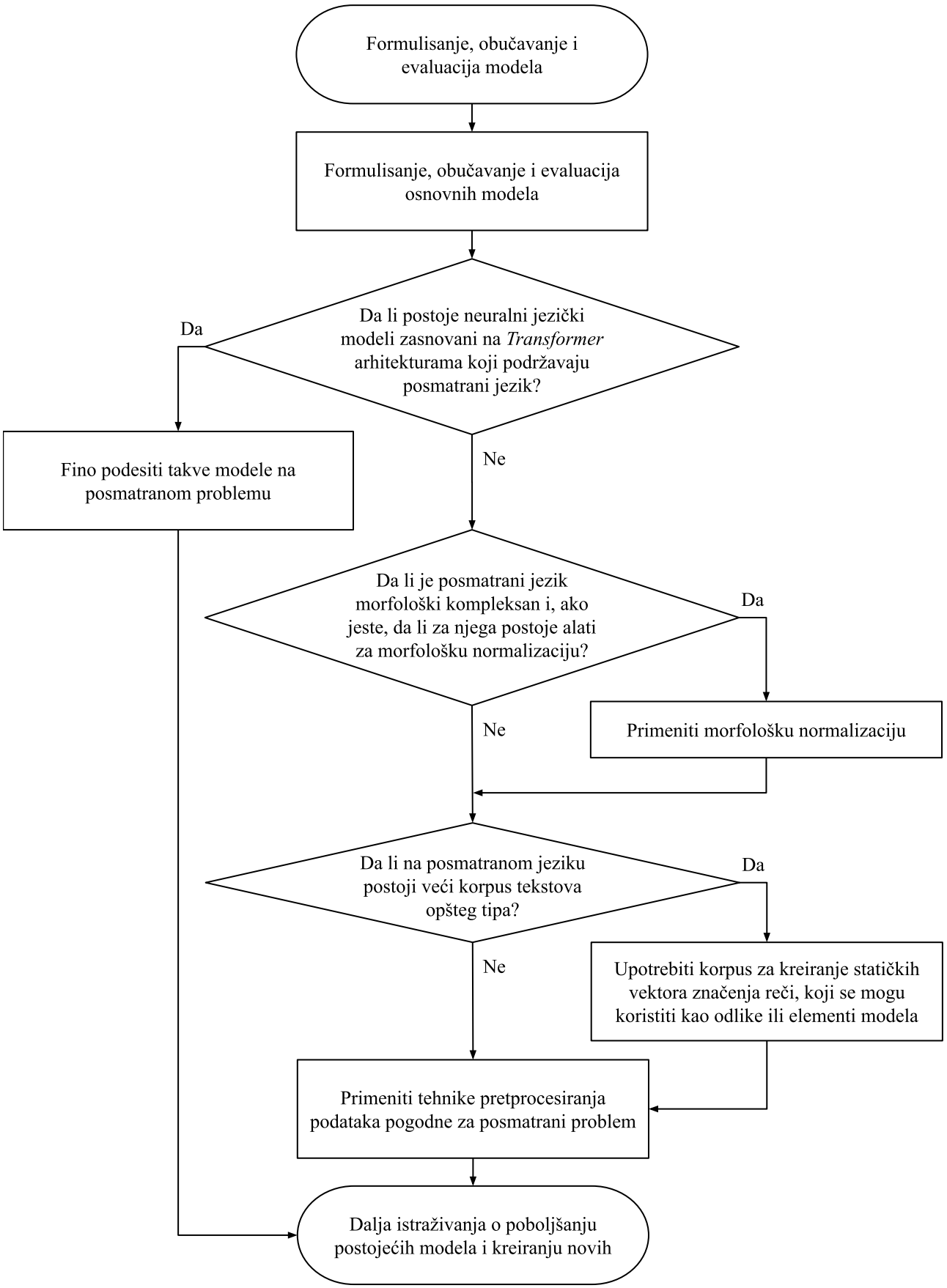
Međutim, ako za razmatrani NLP problem ne postoji jasan standard označavanja podataka, tada je izbor optimalnog pristupa anotaciji teži. Preporučljivo je da se najpre izvrši temeljna analiza postojećih srodnih sistema označavanja i utvrdi da li su oni dovoljno jasni i sveobuhvatni da bi bili pogodni za primenu u konkretnoj situaciji. Pored toga, poželjno je voditi računa i o ekonomičnosti upotrebe određenog sistema anotacije, ali je taj faktor gotovo u potpunosti zapostavljen u NLP literaturi, te je stoga teško doći do pouzdanih zaključaka po ovom pitanju bez sprovođenja nekih preliminarnih anotacionih eksperimenata. Ukoliko se među postojećim pristupima pronađe neki koji bi bio adekvatan po navedenim parametrima, onda se on može koristiti u sprovođenju anotacije, uz eventualne dorade i poboljšanja. U suprotnom, neophodno je osmisliti nov sistem označavanja podataka, pri čemu je takođe neophodno imati u vidu kriterijume jasnoće, sveobuhvatnosti i ekonomičnosti. Za procenu ekonomičnosti upotrebe sistema označavanja može se koristiti *ACE* metrika predložena u odeljku 4.2.2.3. U ovoj disertaciji izrada novog sistema označavanja je demonstrirana na primeru problema analize sentimenta, u poglavlju 4.2.

Pored izbora sistema označavanja, u ovoj fazi razvoja neophodno je utvrditi i tehničke aspekte sprovođenja anotacije. U to spada način unošenja anotacija, koji može biti direktno u izlazne fajlove ili pomoću nekog anotacionog alata, i format čuvanja anotiranih podataka, koji jako zavisi od konkretne vrste anotacije. Za neke NLP probleme, naročito one kod kojih je potrebno anotirati sekvencijalnu ili strukturnu povezanost podataka, primena namenskih alata za anotaciju je gotovo neminovna, pri čemu izbor konkretnog alata treba da zavisi od njegovih mogućnosti, željene vrste alata (desktop ili veb aplikacija), dizajna i intuitivnosti interfejsa i lakoći snimanja anotiranih podataka u željenom izlaznom formatu. Ponekad se dešava da ne postoje ranije izrađeni alati pogodni za tip anotacije koji je potrebno sprovesti. Tada je preporučljivo izraditi nov namenski alat samo ako se proceni da njegovo korišćenje primetno olakšava proces anotacije. Primer ovakvog novog alata izrađenog u ovoj disertaciji jeste *STSAnno*, predstavljen u odeljku 4.1.2, koji je korišćen u anotaciji semantičke sličnosti kratkih tekstova. Nasuprot tome, u anotaciji sentimenta je procenjeno da se upotreba posebnog alata ne bi značajno razlikovala od upotrebe bilo kog generičkog programa za unos i izmenu tekstualnih podataka, zbog čega poseban alat nije korišćen na tom problemu. Naravno, u graničnim slučajevima u proceni potrebe za namenskim anotacionim alatom, poželjno je odluku doneti uz konsultacije sa anotatorima koji će sprovoditi označavanje podataka.

Nakon rešavanja tehničkih pitanja može se preći na obučavanje anotatora, što uključuje i eventualnu kalibraciju uputstava za anotaciju, a zatim i na sprovođenje anotacije. Po završetku označavanja podataka važno je analizirati konzistentnost anotacije i statističku raspodelu dobijenih oznaka, što je u ovoj disertaciji urađeno za oba razmatrana problema i predstavljeno u odeljcima 4.1.3 i 4.2.2. Ukoliko je upotrebljavan standardizovan sistem označavanja dobra praksa je sprovođenje uporedne analize novog skupa podataka sa ranijim skupovima anotiranim po istom sistemu. U ovoj disertaciji to je urađeno za *STS.news.sr* korpus, vezan za problem određivanja semantičke sličnosti kratkih tekstova, što je opisano u odeljku 4.1.3.3. U suprotnom, ako je korišćen potpuno nov sistem označavanja, dobro je da se na kraju anotacije evaluira njegova ukupna ekonomičnost i da se uporedi sa postojećim pristupima sličnog tipa, što je u ovoj disertaciji ilustrovano na problemu analize sentimenta, u odeljcima 4.2.2.3 i 4.2.3.



Slika 14. Dijagram faza prikupljanja tekstualnog sadržaja i anotacije podataka pri razvoju statističkih rešenja semantičkih problema u jezicima sa ograničenim resursima po predloženoj metodologiji



Slika 15. Dijagram faze formulisanja, obučavanja i evaluacije modela pri razvoju statističkih rešenja semantičkih problema u jezicima sa ograničenim resursima po predloženoj metodologiji

U fazi formulisanja, obučavanja i evaluacije modela, dobra praksa je, kao i u svakoj drugoj primeni mašinskog učenja, najpre razmotriti jednostavne, osnovne modele, da bi se dobila referentna tačka za procenu performansi složenijih pristupa. Pri tome, zbog ograničene količine anotiranih podataka, pri obučavanju i evaluaciji svih vrsta modela neophodno je koristiti stratifikovanu unakrsnu validaciju, čime se smanjuje varijansa rezultata.

Ipak, glavni cilj u ovoj fazi jeste postizanje što boljih performansi na posmatranom NLP problemu. U zavisnosti od dostupnosti različitih NLP resursa zavisi i koji je pristup najbolji za postizanje ovog cilja. Naime, kao što su eksperimenti na problemima određivanja semantičke sličnosti i analize sentimenta pokazali (odeljci 5.1.3 i 5.2.2), fino podešavanje neuralnih jezičkih modela zasnovanih na *Transformer* arhitekturama je rešenje koje u ovom trenutku konzistentno donosi najbolje performanse, čak i u kontekstu ograničene dostupnosti anotiranih podataka. Dakle, ukoliko postoje modeli ovog tipa koji podržavaju konkretan jezik sa ograničenim resursima za koji se pokušava razvoj statističkih rešenja, jasno je da oni treba da predstavljaju prvi izbor. Ipak, i najveći višejezični modeli ovog tipa trenutno podržavaju tek stotinak jezika, te se stoga ne mogu smatrati opšte primenjivim.

Ako rešenja tog tipa nisu dostupna, neophodno je okrenuti se jednostavnijim postojećim modelima, uz razmatranje različitih mogućnosti njihovog unapređenja, korišćenjem jezičkih alata i korpusa dostupnih za dati jezik. Ako je jezik morfološki kompleksan, primena alata za morfološku normalizaciju može znatno podići performanse sistema, što je na primeru srpskog jezika u ovoj disertaciji pokazano kod svih pristupa osim neuralnih jezičkih modela. Ukoliko na datom jeziku postoji korpus opšteg tipa i adekvatne veličine, preporučljivo je iskoristiti ga za izradu statičkih vektora značenja reči, koji se potom mogu upotrebiti kao elementi ili odlike u modelima za rešavanje konkretnog NLP problema. Rezultati na NLP problemima razmotrenim u ovoj disertaciji (odeljci 5.1.1, 5.1.2 i 5.2.1.2) pokazuju da kombinovanje statičkih vektora značenja reči sa drugim vrstama elemenata ili odlika u okviru modela dovodi do boljih performansi od samostalne upotrebe vektora značenja reči. Za određene probleme postoje i tehnike pretprocesiranja tekstova koje su opšte primenjive, a koje mogu takođe unaprediti performanse modela, poput razmotrenih tehnika obeležavanja negacija ili normalizacije ponavljanja karaktera i emotikona, koje imaju efekta na problemima analize sentimenta.

Naravno, i sa najboljim modelima uvek postoji prostor za dodatna istraživanja u pogledu njihovog poboljšanja ili razvijanja još boljih pristupa. U ovoj disertaciji kreiranje novih namenskih modela je demonstrirano na problemu određivanja semantičke sličnosti kratkih tekstova, u odeljku 5.1.2. Međutim, razvoj višejezičnih varijanti neuralnih jezičkih modela i njihova fleksibilnost u smislu mogućnosti finog podešavanja za raznovrsne NLP probleme učinili su da izrada namenskih modela za određeni problem bude adekvatna samo u ograničenom broju situacija, i to pre svega u slučaju da neuralni jezički modeli ne podržavaju posmatrani jezik. Time je u izradi statističkih rešenja semantičkih problema u jezicima sa ograničenim resursima još veći akcenat stavljen na faze prikupljanja tekstualnog sadržaja i anotacije podataka i na kvalitet njihovog sprovođenja.

7 Zaključak

U ovoj disertaciji predstavljena je metodologija razvoja statističkih rešenja za semantičke probleme u obradi kratkih tekstova napisanih na jezicima sa ograničenim resursima. Ona je ilustrovana kroz rešavanje problema određivanja semantičke sličnosti i analize sentimenta kratkih tekstova na srpskom jeziku, i to kroz sve faze razvoja, počevši od prikupljanja tekstualnog sadržaja, preko anotacije podataka, do formulisanja, obučavanja i evaluacije modela mašinskog učenja.

Tokom izrade disertacije, potvrđene su sve polazne hipoteze, prvobitno navedene u poglavlju 1.3:

- Moguće je pribaviti dovoljne količine neophodnih tekstualnih resursa na srpskom jeziku za razvoj i evaluaciju statističkih rešenja za posmatrane semantičke probleme – potvrđeno rezultatima faze prikupljanja tekstualnog sadržaja i opisano u glavi 3.
- Osnovna referentna rešenja (engl. *baselines*) za posmatrane probleme mogu se unaprediti korišćenjem dostupnih korpusa opšteg tipa i osnovnih NLP alata – potvrđeno uspešnom upotrebom statičkih vektora značenja reči obučenih na osnovu veb korpusa srpskog jezika *srWaC* (Ljubešić et al. 2016), kao i primenom alata za morfološku normalizaciju, i to na oba razmotrena problema (odeljci 5.1.1 i 5.2.1).
- Problem proređenosti podataka koji je prouzrokovan morfološkom kompleksnošću jezika poput srpskog može se minimizovati primenom alata za morfološku normalizaciju – potvrđeno na osnovnim modelima iz oba razmotrena problema (odeljci 5.1.1 i 5.2.1) kao i na novim namenskim modelima za određivanje semantičke sličnosti kratkih tekstova razvijenim u ovoj disertaciji (odeljak 5.1.2). Jedina vrsta modela na kojima morfološka normalizacija nije dovela do poboljšanja performansi jesu neuralni jezički modeli, ali neuspeh primene morfološke normalizacije je kod njih povezan sa činjenicom da su oni prethodno obučeni na znatnim količinama standardnog, morfološki nenormalizovanog teksta.
- Određivanje semantičke sličnosti i analiza sentimenta kratkih tekstova se mogu realizovati i bez potpune sintaktičke analize teksta, čak i u jezicima sa složenijim sintaktičkim pravilima kao što je srpski – potvrđeno uspešnim razvojem i evaluacijom šireg spektra modela na oba problema, prikazanim u glavi 5.
- Iako su neuralni modeli tipično veoma zahtevni u pogledu količine podataka potrebnih za njihovo obučavanje, moguće je primeniti određene varijante takvih modela i na rešavanje semantičkih problema u jezicima sa ograničenim resursima – potvrđeno razmatranjem aktuelne literature u poglavlju 2.2 i uspešnom primenom neuralnih jezičkih modela zasnovanih na *Transformer* arhitekturama na rešavanje posmatranih semantičkih problema (odeljci 5.1.3 i 5.2.2).

Pored toga, uspešno su ostvareni svi planirani ciljevi i doprinosi disertacije, takođe prvobitno navedeni u poglavlju 1.3:

1. Nakon što je u glavama 3, 4 i 5 na problemima određivanja semantičke sličnosti i analize sentimenta detaljno ilustrovana svaka faza razvoja statističkih rešenja semantičkih problema u obradi kratkih tekstova na jezicima sa ograničenim resursima, u glavi 6 je iznet pregled metodološkog doprinosa u pogledu budućeg lakšeg i bolje planiranog razvoja takvih rešenja.
2. Identifikacija postojećih pristupa za rešavanje semantičkih problema u obradi kratkih tekstova koji su pogodni za primenu u jezicima sa ograničenim resursima – analiza literature data u poglavlju 2.2 je pokazala da je transferno učenje načelni pristup koji je najadekvatniji za jezike sa ograničenim resursima. U okviru toga, višejezične varijante prethodno obučenih neuralnih jezičkih modela predstavljaju najbolji izbor, jer se ovi modeli ne obučavaju iznova na

konkretnom NLP problemu, već samo fino podešavaju, za šta su dovoljne i manje količine anotiranih podataka. U slučaju da modeli ovog tipa ne podržavaju dati jezik, transferno učenje se može primeniti i preko vektora značenja reči, izrađenih na osnovu velikih neanotiranih korpusa tekstova opšteg tipa, koji omogućavaju transfer znanja o semantici reči na proizvoljne druge NLP probleme.

3. Kreiranje metrike za merenje ekonomičnosti anotacije, tj. ekonomičnosti upotrebe uputstava za anotaciju u okviru određenog sistema označavanja podataka – kako do sada nijedna lako primenjiva metrika ovog tipa nije postojala, za potrebe utvrđivanja ekonomičnosti anotacije sentimenta razvijena je *ACE* metrika, prikazana u odeljku 4.2.2.3. Ova metrika se može upotrebljavati ne samo za anotaciju sentimenta, već i za druge vrste semantičkih anotacija.
4. Izrada novog, fleksibilnog i ekonomičnog sistema označavanja sentimenta tekstova, koji omogućava više nivoa interpretacije za oznake sentimenta i posebno je pogodan za primenu u jezicima sa ograničenim resursima – u nedostatku standardizovanog ili fleksibilnog i sveobuhvatnog sistema označavanja sentimenta, razvijen je nov sistem označavanja, uključujući skup oznaka i smernica za anotaciju, što je opisano u poglavlju 4.2.
5. Stvaranje prvih referentnih javno dostupnih anotiranih skupova podataka za probleme određivanja semantičke sličnosti i analize sentimenta kratkih tekstova na srpskom jeziku, izloženo u glavi 4, će olakšati dalji razvoj ovih oblasti u računarskoj obradi srpskog jezika i omogućiti kasniju uporednu evaluaciju rešenja razvijenih u ovoj disertaciji sa nekim novijim.
6. Pomoću anotiranih skupova podataka kreirana su i evaluirana osnovna referentna rešenja za određivanje semantičke sličnosti i analizu sentimenta kratkih tekstova na srpskom jeziku, kao i njihove varijante i unapređenja, što je prikazano u odeljcima 5.1.1 i 5.2.1.
7. U glavi 5 izložena je, između ostalog, i prva komparativna evaluacija efekata morfološke normalizacije kratkih tekstova na srpskom jeziku i različitih rešenja razvijenih za ove potrebe, u kontekstu problema određivanja semantičke sličnosti i analize sentimenta kratkih tekstova. Rezultati su pokazali da je stemovanje obično bolji izbor od lematizacije, pri čemu se stemer Ljubešića i Pandžića za hrvatski jezik pokazao kao optimalno rešenje.
8. Korišćenjem ideje težinskog ponderisanja sličnosti reči na osnovu njihove učestalosti i/ili vrste reči, razvijena su i evaluirana tri nova namenska modela za određivanje semantičke sličnosti kratkih tekstova, koji su adekvatni za upotrebu u jezicima sa ograničenim resursima i koji postižu bolje performanse od osnovnih modela i njihovih unapređenja, što je pokazano u odeljku 5.1.2.
9. Pomoću kreiranih anotiranih skupova podataka uspešno je sprovedeno fino podešavanje i evaluacija neuralnih jezičkih modela zasnovanih na *Transformer* arhitekturama na posmatranim problemima na srpskom jeziku, što je opisano u odeljcima 5.1.3 i 5.2.2.

Shodno preporukama za razvoj NLP tehnologija za manje jezike (Maxwell & Hughes 2006; Streiter et al. 2006), svi NLP resursi izrađeni u okviru ove disertacije su javno dostupni na sledećim repozitorijumima:

- *STS.news.sr* – anotirani korpus za problem određivanja semantičke sličnosti kratkih tekstova na srpskom jeziku: https://vukbatanovic.github.io/STS.news.sr/
- *SentiComments.SR* – anotirani korpus za problem analize sentimenta kratkih tekstova na srpskom jeziku: https://vukbatanovic.github.io/SentiComments.SR/
- *STSAnno* alat za anotaciju semantičke sličnosti kratkih tekstova: https://vukbatanovic.github.io/STSAnno/
- *SCStemmers* – skup stemera za srpski i hrvatski jezik: https://vukbatanovic.github.io/SCStemmers/
- *STSFineGrain* – skup algoritama za određivanje semantičke sličnosti kratkih tekstova i okvir za njihovo obučavanje i evaluaciju: https://vukbatanovic.github.io/STSFineGrain/

Prirodan nastavak naučnih istraživanja vezanih za temu disertacije bila bi primena predložene metodologije u izradi statističkih rešenja semantičkih problema na još nekom jeziku sa ograničenim resursima. Pored toga, dalji pravci istraživanja se mogu okvirno podeliti na one vezane za anotaciju podataka i izradu jezičkih resursa, i one koji se odnose na dalji razvoj računarskih modela.

U pogledu razvoja resursa za određivanje semantičke sličnosti kratkih tekstova, naredna istraživanja bi mogla da se usredsrede na izgradnju anotiranih skupova podataka na srpskom jeziku i iz drugih domena, a ne samo novinskog. Takođe, kao što je pokazano u odeljku 4.1.3.3, velika većina anotiranih korpusa ovog tipa iz novinskog domena se odlikuje neizbalansiranim raspodelama prosečnih ocena semantičke sličnosti. Stoga bi vredan pravac ispitivanja bilo razmatranje uticaja ove neizbalansiranosti na ponašanje NLP modela. Konačno, izazovan pravac daljih istraživanja bi bilo i razmatranje srodnog problema određivanja semantičke sličnosti između tekstova različitih dužina (engl. *cross-level semantic similarity*), za koji su do sada sprovedena vrlo ograničena istraživanja čak i na engleskom jeziku (Jurgens et al. 2014, 2016).

Izloženu metodologiju anotacije sentimenta bi u budućnosti bilo korisno primeniti za izradu uporedivih resursa i iz drugih domena ili na drugim jezicima. Pored toga, dalja istraživanja vezana za predloženu *ACE* metriku bi bila usmerena na njenu primenu ili eventualnu doradu u kontekstu projekata anotacije drugačijeg tipa.

Što se razvoja NLP modela tiče, najveće mogućnosti za dalja istraživanja leže u razmatranju različitih neuralnih pristupa. U tom pogledu, izdvaja se nekoliko pravaca potencijalnih daljih istraživanja:

- Fino podešavanje višejezičnih varijanti neuralnih jezičkih modela na konkretnom NLP problemu uz kombinovano korišćenje anotiranih podataka na više jezika. Naime, za skoro sve NLP probleme količine dostupnih anotiranih podataka su primetno veće na engleskom jeziku. Stoga bi se performanse ovakvih modela mogle potencijalno poboljšati kombinovanom upotrebom i relevantnih skupova podataka na engleskom, kao i dostupnih anotiranih podataka na jeziku sa ograničenim resursima.
- Kombinovano fino podešavanje neuralnih modela na više problema (engl. *multi-task*) – ovaj pristup je već pokazao dobre performanse na engleskom na većem skupu NLP problema, uključujući određivanje semantičke sličnosti i analizu sentimenta (Liu et al. 2019). U jezicima sa ograničenim resursima bi se na ovaj način potencijalno mogli sinergijski upotrebiti manji skupovi anotiranih podataka za različite NLP probleme.
- Razmatranje povezanosti između sličnosti osobina određenih jezika i performansi višejezičnih varijanti neuralnih jezičkih modela. Već postoje indikacije da uspeh ovakvih modela zasnovanih na *Transformer* arhitekturama u transferu znanja između jezika u primetnoj meri zavisi od strukturnih sličnosti između jezika (Conneau et al. 2020b; Karthikeyan et al. 2020), ali nije jasno ustanovljeno koje tačno lingvističke odlike su od kolike važnosti u ovom smislu.

U zavisnosti od raspoloživosti hardverskih resursa u budućnosti i od razmatranja vezanih za neophodne veličine korpusa neanotiranih podataka, vredelo bi istražiti i kreiranje posebnog neuralnog kontekstno osetljivog jezičkog modela za srpski jezik. Na primeru engleskog, francuskog i drugih jezika je već pokazano da jednojezični modeli ovog tipa ostvaruju generalno bolje performanse od višejezičnih (Devlin et al. 2019; Le et al. 2020; Martin et al. 2020; Nozza et al. 2020), te bi se na ovaj način verovatno ostvarilo dodatno poboljšanje rezultata na mnogim problemima i na srpskom jeziku.

Prilozi

Tabela P1. Raspodela parova rečenica iz *STS.news.sr* korpusa po finalnim, uprosečenim ocenama semantičke sličnosti

Prosečna ocena	Parovi	
	#	%
0	57	*4,78%*
0,2	27	*2,27%*
0,4	24	*2,01%*
0,6	23	*1,93%*
0,8	33	*2,77%*
1,0	41	*3,44%*
1,2	51	*4,28%*
1,4	30	*2,52%*
1,6	34	*2,85%*
1,8	39	*3,27%*
2,0	46	*3,86%*
2,2	35	*2,94%*
2,4	38	*3,19%*
2,6	42	*3,52%*
2,8	70	*5,87%*
3,0	226	*18,96%*
3,2	72	*6,04%*
3,4	60	*5,03%*
3,6	52	*4,36%*
3,8	64	*5,37%*
4,0	39	*3,27%*
4,2	26	*2,18%*
4,4	16	*1,34%*
4,6	12	*1,01%*
4,8	10	*0,84%*
5,0	25	*2,10%*
Ukupno	1192	*100,00%*

Tabela P2. Raspodela parova rečenica iz *STS.news.sr* korpusa po individualnim ocenama semantičke sličnosti za svakog anotatora

Ocena semantičke sličnosti	Anotator 1		Anotator 2		Anotator 3		Anotator 4		Anotator 5	
	#	%	#	%	#	%	#	%	#	%
0	129	*10,82%*	69	*5,79%*	114	*9,56%*	134	*11,24%*	148	*12,42%*
1	180	*15,10%*	151	*12,67%*	212	*17,79%*	190	*15,94%*	143	*12,00%*
2	178	*14,93%*	208	*17,45%*	207	*17,37%*	158	*13,26%*	147	*12,33%*
3	424	*35,57%*	527	*44,21%*	327	*27,43%*	468	*39,26%*	582	*48,83%*
4	216	*18,12%*	174	*14,60%*	292	*24,50%*	186	*15,60%*	128	*10,74%*
5	65	*5,45%*	63	*5,29%*	40	*3,36%*	56	*4,70%*	44	*3,69%*
Ukupno					1192	*100%*				

Tabela P3. Raspodela tekstova iz glavnog *SentiComments.SR* korpusa po oznakama sentimenta

Oznaka sentimenta	Svi tekstovi		Sarkastični tekstovi	
	#	%	#	%
+1	1452	*41,60%*	1	*0,88%*
-1	805	*23,07%*	99	*86,84%*
+M	289	*8,28%*	2	*1,75%*
-M	325	*9,31%*	12	*10,53%*
+NS	398	*11,40%*	0	*0%*
-NS	221	*6,33%*	0	*0%*
+/-1	2257	*64,67%*	100	*87,72%*
+/-M	614	*17,59%*	14	*12,28%*
+/-NS	619	*17,74%*	0	*0%*
Svi pozitivni (**+1, +M, +NS**)	2139	*61,29%*	3	*2,63%*
Svi negativni (**-1, -M, -NS**)	1351	*38,71%*	111	*97,37%*
Ukupno	3490	*100,00%*	114	*100,00%*

Tabela P4. Raspodela tekstova iz *SentiComments.SR.verif.movies* korpusa po oznakama sentimenta

Oznaka sentimenta	IG anotatori		EG anotatori		KG anotatori	
	IG 1	IG 2	EG 1	EG 2	KG 1	KG 2
+1	193 (*41,59%*)	192 (*41,38%*)	194 (*41,81%*)	195 (*42,03%*)	243 (*52,37%*)	247 (*53,23%*)
-1	130 (*28,02%*)	125 (*26,94%*)	135 (*29,09%*)	126 (*27,16%*)	152 (*32,76%*)	134 (*28,88%*)
+M	64 (*13,79%*)	65 (*14,01%*)	64 (*13,79%*)	59 (*12,72%*)	29 (*6,25%*)	35 (*7,54%*)
-M	53 (*11,42%*)	55 (*11,85%*)	47 (*10,13%*)	51 (*10,99%*)	14 (*3,02%*)	20 (*4,31%*)
+NS	16 (*3,45%*)	20 (*4,31%*)	21 (*4,53%*)	24 (*5,17%*)	20 (*4,31%*)	24 (*5,17%*)
-NS	8 (*1,72%*)	7 (*1,51%*)	3 (*0,65%*)	9 (*1,94%*)	6 (*1,29%*)	4 (*0,86%*)
+/-1	323 (*69,61%*)	317 (*68,32%*)	329 (*70,91%*)	321 (*69,18%*)	395 (*85,13%*)	381 (*82,11%*)
+/-M	117 (*25,22%*)	120 (*25,86%*)	111 (*23,92%*)	110 (*23,71%*)	43 (*9,27%*)	55 (*11,85%*)
+/-NS	24 (*5,17%*)	27 (*5,82%*)	24 (*5,17%*)	33 (*7,11%*)	26 (*5,60%*)	28 (*6,03%*)
Svi pozitivni (**+1**, **+M**, **+NS**)	273 (*58,84%*)	277 (*59,70%*)	279 (*60,13%*)	278 (*59,91%*)	292 (*62,93%*)	306 (*65,95%*)
Svi negativni (**-1**, **-M**, **-NS**)	191 (*41,16%*)	187 (*40,30%*)	185 (*39,87%*)	186 (*40,09%*)	172 (*37,07%*)	158 (*34,05%*)
Sarkastični (**+/-1s**, **+/-Ms**)	11 (*2,37%*)	13 (*2,80%*)	13 (*2,80%*)	9 (*1,94%*)	8 (*1,72%*)	6 (*1,29%*)
Ukupno	464 (*100%*)					

Tabela P5. Raspodela tekstova iz *SentiComments.SR.verif.books* korpusa po oznakama sentimenta

Oznaka sentimenta	IG anotatori		EG anotatori		KG anotatori	
	IG 1	IG 2	EG 1	EG 2	KG 1	KG 2
+1	81 (*46,82%*)	83 (*47,98%*)	81 (*46,82%*)	80 (*46,24%*)	102 (*58,96%*)	104 (*60,12%*)
-1	30 (*17,34%*)	29 (*16,76%*)	30 (*17,34%*)	29 (*16,76%*)	30 (*17,34%*)	20 (*11,56%*)
+M	9 (*5,20%*)	10 (*5,78%*)	9 (*5,20%*)	7 (*4,05%*)	9 (*5,20%*)	4 (*2,31%*)
-M	4 (*2,31%*)	3 (*1,73%*)	4 (*2,31%*)	4 (*2,31%*)	3 (*1,73%*)	0 (*0%*)
+NS	43 (*24,86%*)	40 (*23,12%*)	43 (*24,86%*)	46 (*26,59%*)	21 (*12,14%*)	30 (*17,34%*)
-NS	6 (*3,47%*)	8 (*4,62%*)	6 (*3,47%*)	7 (*4,05%*)	8 (*4,62%*)	15 (*8,67%*)
+/-1	111 (*64,16%*)	112 (*64,74%*)	111 (*64,16%*)	109 (*63,01%*)	132 (*76,30%*)	124 (*71,68%*)
+/-M	13 (*7,51%*)	13 (*7,51%*)	13 (*7,51%*)	11 (*6,36%*)	12 (*6,94%*)	4 (*2,31%*)
+/-NS	49 (*28,32%*)	48 (*27,75%*)	49 (*28,32%*)	53 (*30,64%*)	29 (*16,76%*)	45 (*26,01%*)
Svi pozitivni (**+1**, **+M**, **+NS**)	133 (*76,88%*)	133 (*76,88%*)	133 (*76,88%*)	133 (*76,88%*)	132 (*76,30%*)	138 (*79,77%*)
Svi negativni (**-1**, **-M**, **-NS**)	40 (*23,12%*)	40 (*23,12%*)	40 (*23,12%*)	40 (*23,12%*)	41 (*23,70%*)	35 (*20,23%*)
Sarkastični (+/-**1s**, +/-**Ms**)	8 (*4,62%*)	7 (*4,05%*)	6 (*3,47%*)	6 (*3,47%*)	4 (*2,31%*)	2 (*1,16%*)
Ukupno	173 (*100%*)					

Literatura

Abdul-Mageed M, Diab M (2012) AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*. European Language Resources Association (ELRA), Istanbul, Turkey, pp. 3907–3914

Abdul-Mageed M, Diab MT (2011) Subjectivity and Sentiment Analysis of Modern Standard Arabic Newswire. In: *Proceedings of the Fifth Language Annotation Workshop (LAW V)*. Association for Computational Linguistics, Portland, Oregon, USA, pp. 110–118

Agarwal D, Mujadia V, Sharma DM, Mamidi R (2017) A Modified Annotation Scheme for Semantic Textual Similarity. In: *Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2017)*. Budapest, Hungary

Agić Ž, Ljubešić N, Merkler D (2013) Lemmatization and Morphosyntactic Tagging of Croatian and Serbian. In: *Proceedings of the Fourth Biennial International Workshop on Balto-Slavic Natural Language Processing (BSNLP 2013)*. Association for Computational Linguistics, Sofia, Bulgaria, pp. 48–57

Agirre E, Banea C, Cardie C, et al (2014) SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. In: *Proceedings of the Eighth International Workshop on Semantic Evaluation (SemEval 2014)*. Association for Computational Linguistics, Dublin, Ireland, pp. 81–91

Agirre E, Banea C, Cardie C, et al (2015) SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In: *Proceedings of the Ninth International Workshop on Semantic Evaluation (SemEval 2015)*. Association for Computational Linguistics, Denver, Colorado, USA, pp. 252–263

Agirre E, Banea C, Cer D, et al (2016) SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*. Association for Computational Linguistics, San Diego, California, USA, pp. 497–511

Agirre E, Cer D, Diab M, et al (2013) *SEM 2013 shared task: Semantic Textual Similarity. In: *Second Joint Conference on Lexical and Computational Semantics (*SEM)*. Association for Computational Linguistics, Atlanta, Georgia, USA, pp. 32–43

Agirre E, Cer D, Diab M, Gonzalez-Agirre A (2012) SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In: *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*. Association for Computational Linguistics, Montreal, Canada, pp. 385–393

Al-Twairesh N, Al-Khalifa H, Al-Salman A, Al-Ohali Y (2017) AraSenTi-Tweet: A Corpus for Arabic Sentiment Analysis of Saudi Tweets. *Procedia Computer Science* 117, pp. 63–72. https://doi.org/10.1016/j.procs.2017.10.094

Artetxe M, Schwenk H (2019) Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics* 7, pp. 597–610. https://doi.org/10.1162/tacl_a_00288

Artstein R, Poesio M (2008) Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics* 34(4), pp. 555–596. https://doi.org/10.1162/coli.07-034-R2

Balamurali AR, Joshi A, Bhattacharyya P (2012) Cost and Benefit of Using WordNet Senses for Sentiment Analysis. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*. European Language Resources Association (ELRA), Istanbul, Turkey, pp. 3090–3097

Batanović V (2011) Ekspertski sistem za određivanje semantičke sličnosti kratkih tekstova na srpskom jeziku. Univerzitet u Beogradu

Batanović V, Bojić D (2015) Using Part-of-Speech Tags as Deep-Syntax Indicators in Determining Short-Text Semantic Similarity. *Computer Science and Information Systems* 12(1), pp. 1–31. https://doi.org/10.2298/CSIS131127082B

Batanović V, Cvetanović M, Nikolić B (2018a) Fine-grained Semantic Textual Similarity for Serbian. In: *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan, pp. 1370–1378

Batanović V, Furlan B, Nikolić B (2011) Softverski sistem za određivanje semantičke sličnosti kratkih tekstova na srpskom jeziku. *Zbornik radova sa 19. telekomunikacionog foruma (TELFOR 2011)*. IEEE, Beograd, Srbija, pp. 1249–1252

Batanović V, Ljubešić N, Samardžić T (2018b) SETimes.SR – A Reference Training Corpus of Serbian. In: *Proceedings of the Conference on Language Technologies & Digital Humanities 2018 (JT-DH 2018)*. Ljubljana University Press, Faculty of Arts, Ljubljana, Slovenia, pp. 11–17

Batanović V, Nikolić B (2016) Sentiment Classification of Documents in Serbian: The Effects of Morphological Normalization. In: *Proceedings of the 24th Telecommunications Forum (TELFOR 2016)*. IEEE, Belgrade, Serbia, pp. 889–892

Batanović V, Nikolić B (2017) Sentiment Classification of Documents in Serbian: The Effects of Morphological Normalization and Word Embeddings. *Telfor Journal* 9(2), pp. 104–109. https://doi.org/10.5937/telfor1702104B

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)*. European Language Resources Association (ELRA), Portorož, Slovenia, pp. 2688–2696

Bender EM (2011) On Achieving and Evaluating Language-Independence in NLP. *Linguistic Issues in Language Technology* 6(3)

Bender EM (2009) Linguistically Naïve != Language Independent: Why NLP Needs Linguistic Typology. In: *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?* Association for Computational Linguistics, Athens, Greece, pp. 26–32

Benjamin M (2018) Hard Numbers: Language Exclusion in Computational Linguistics and Natural Language Processing. In: *Proceedings of the LREC 2018 Workshop "CCURL2018 – Sustaining*

Knowledge Diversity in the Digital Age." Miyazaki, Japan, pp. 13–18

Bergroth L, Hakonen H, Raita T (2000) A Survey of Longest Common Subsequence Algorithms. In: *Proceedings of the Seventh International Symposium on String Processing and Information Retrieval (SPIRE 2000)*. IEEE, A Coruña, Spain, pp. 39–48

Berment V (2002) Several directions for minority languages computerization. In: *Proceedings of the 19th International Conference on Computational Linguistics: Project Notes (COLING 2002)*. Association for Computational Linguistics, Taipei, Taiwan

Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5, pp. 135–146

Bowman SR, Angeli G, Potts C, Manning CD (2015) A large annotated corpus for learning natural language inference. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*. Association for Computational Linguistics, Lisbon, Portugal, pp. 632–642

Brown P, Levinson SC (1987) Politeness: Some universals in language usage (Studies in Interactional Sociolinguistics 4). Cambridge University Press, New York, New York, USA

Bucilă C, Caruana R, Niculescu-Mizil A (2006) Model Compression. In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)*. Philadelphia, Pennsylvania, USA

Callison-Burch C, Ungar L, Pavlick E (2015) Crowdsourcing for NLP. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorial Abstracts (NAACL-HLT 2015)*. Association for Computational Linguistics, Denver, Colorado, USA, pp. 2–3

Caruana RA (1993) Multitask Learning: A Knowledge-Based Source of Inductive Bias. In: *Proceedings of the Tenth International Conference on Machine Learning (ICML 1993)*. Amherst, Massachusetts, USA, pp. 41–48

Cer D, Diab M, Agirre E, et al (2017) SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Cross-lingual Focused Evaluation. In: *Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval 2017)*. Association for Computational Linguistics, Vancouver, Canada, pp. 1–14

Cer D, Yang Y, Kong S, et al (2018) Universal Sentence Encoder for English. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP 2018)*. Association for Computational Linguistics, Brussels, Belgium, pp. 169–174

Chen Z, Badrinarayanan V, Lee C-Y, Rabinovich A (2018) GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In: *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*. Stockholm, Sweden, pp. 794–803

Chidambaram M, Yang Y, Cer D, et al (2019) Learning Cross-Lingual Sentence Representations via a Multi-task Dual-Encoder Model. In: *Proceedings of the Fourth Workshop on Representation Learning for NLP (RepL4NLP-2019)*. Association for Computational Linguistics, Florence, Italy, pp. 250–259

Cohen J (1960) A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20(1), pp. 37–46. https://doi.org/10.1177/001316446002000104

Collobert R, Weston J (2008) A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In: *Proceedings of the 25th International Conference on Machine Learning (ICML 2008)*. ACM, Helsinki, Finland, pp. 160–167

Conneau A, Khandelwal K, Goyal N, et al (2020a) Unsupervised Cross-lingual Representation Learning at Scale. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*. Association for Computational Linguistics, pp. 8440–8451

Conneau A, Kiela D, Schwenk H, et al (2017) Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*. Association for Computational Linguistics, Copenhagen, Denmark, pp. 670–680

Conneau A, Lample G (2019) Cross-lingual Language Model Pretraining. In: *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*. Vancouver, Canada

Conneau A, Wu S, Li H, et al (2020b) Emerging Cross-lingual Structure in Pretrained Language Models. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*. Association for Computational Linguistics, pp. 6022–6034

Corley C, Mihalcea R (2005) Measuring the Semantic Similarity of Texts. In: *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment (EMSEE 2005)*. Association for Computational Linguistics, Ann Arbor, Michigan, USA, pp. 13–18

Deerwester S, Dumais ST, Furnas GW, et al (1990) Indexing by Latent Semantic Analysis. *Journal Of The American Society For Information Science* 41(6), pp. 391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Deng L, Wiebe J (2015) MPQA 3.0: An Entity/Event-Level Sentiment Corpus. In: *Proceedings of Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL (NAACL-HLT 2015)*. Association for Computational Linguistics, Denver, Colorado, USA, pp. 1323–1328

Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*. Association for Computational Linguistics, Minneapolis, Minnesota, USA, pp. 4171–4186

Dolan B, Quirk C, Brockett C (2004) Unsupervised construction of large paraphrase corpora. In: *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*. Association for Computational Linguistics, Geneva, Switzerland, pp. 350–356

Dolan WB, Brockett C (2005) Automatically Constructing a Corpus of Sentential Paraphrases. In: *Proceedings of the Third International Workshop on Paraphrasing (IWP 2005)*. Asia Federation of Natural Language Processing, Jeju Island, South Korea, pp. 9–16

Duong LT (2017) Natural Language Processing for Resource-Poor Languages. University of

Melbourne

El-Haj M, Kruschwitz U, Fox C (2015) Creating language resources for under-resourced languages: methodologies, and experiments with Arabic. *Language Resources and Evaluation* 49(3), pp. 549–580. https://doi.org/10.1007/s10579-014-9274-3

Elsevier (2018) Artificial Intelligence: How knowledge is created, transferred, and used

Erjavec T (2017) MULTEXT-East. In: Ide N, Pustejovsky J (eds) *Handbook of Linguistic Annotation*. Springer, Dordrecht, pp. 441–462

Ethayarajh K (2019) How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the Ninth International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*. Association for Computational Linguistics, Hong Kong, China, pp. 55–65

Fan R-E, Chang K-W, Hsieh C-J, et al (2008) LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research* 9(2008), pp. 1871–1874. https://doi.org/10.1038/oby.2011.351

Finlayson MA, Erjavec T (2017) Overview of Annotation Creation: Processes and Tools. In: Pustejovsky J, Ide N (eds) *Handbook of Linguistic Annotation*. Springer, Dordrecht, pp. 167–191

Firth JR (1957) A synopsis of linguistic theory 1930-1955. *Studies in Linguistic Analysis*, pp. 1–32

Fleiss JL (1971) Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5), pp. 378–382. https://doi.org/10.1037/h0031619

Fonseca ER, Santos LB dos, Criscuolo M, Aluísio SM (2016) Visão Geral da Avaliação de Similaridade Semântica e Inferência Textual. *Linguamática* 8(2), pp. 3–13

Furlan B, Batanović V, Nikolić B (2013) Semantic similarity of short texts in languages with a deficient natural language processing support. *Decision Support Systems* 55(3), pp. 710–719. https://doi.org/10.1016/j.dss.2013.02.002

Gage P (1994) A New Algorithm for Data Compression. The C Users Journal

Ganitkevitch J, Durme B Van, Callison-Burch C (2013) PPDB: The Paraphrase Database. In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*. Association for Computational Linguistics, Atlanta, Georgia, USA, pp. 758–764

Gesmundo A, Samardžić T (2012a) Lemmatising Serbian as Category Tagging with Bidirectional Sequence Classification. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*. European Language Resources Association (ELRA), Istanbul, Turkey, pp. 2103–2106

Gesmundo A, Samardžić T (2012b) Lemmatisation as a tagging task. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*. Association for Computational Linguistics, Jeju Island, Korea, pp. 368–372

Grljević O (2016) Sentiment u sadržajima sa društvenih mreža kao instrument unapređenja poslovanja visokoškolskih institucija. Univerzitet u Novom Sadu

Harris Z (1954) Distributional Structure. *Word* 10, pp. 146–162

Hayashi Y, Luo W (2016) Extending Monolingual Semantic Textual Similarity Task to Multiple Cross-lingual Settings. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Portorož, Slovenia, pp. 1233–1239

Hill F, Cho K, Korhonen A (2016) Learning Distributed Representations of Sentences from Unlabelled Data. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2016)*. Association for Computational Linguistics, San Diego, California, USA, pp. 1367–1377

Hinton G, Vinyals O, Dean J (2015) Distilling the Knowledge in a Neural Network. In: *Proceedings of the NIPS Deep Learning and Representation Learning Workshop*. Montreal, Canada

Hirschberg J, Manning CD (2015) Advances in natural language processing. *Science* 349(6245), pp. 261–266. https://doi.org/10.1126/science.aaa8685

Hovy E, Lavid J (2010) Towards a "Science" of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics. *International Journal of Translation* 22(1), pp. 13–36

Howard J, Ruder S (2018) Universal Language Model Fine-tuning for Text Classification. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers) (ACL 2018)*. Association for Computational Linguistics, Melbourne, Australia, pp. 328–339

Injac-Malbaša V (2012) Crowdsourcing - određenje pojma, tipologija i srodni termini. Glasnik Narodne biblioteke Srbije

Islam A, Inkpen D (2008) Semantic Text Similarity Using Corpus-Based Word Similarity and String Similarity. *ACM Transactions on Knowledge Discovery from Data* 2(2), Article No. 10. https://doi.org/10.1145/1376815.1376819

Jongejan B, Dalianis H (2009) Automatic training of lemmatization rules that handle morphological changes in pre- , in- and suffixes alike. In: *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the Fourth International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-AFNLP 2009)*. ACL and AFNLP, Suntec, Singapore, pp. 145–153

Joshi A, Mishra A, Senthamilselvan N, Bhattacharyya P (2014) Measuring Sentiment Annotation Complexity of Text. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers) (ACL 2014)*. Association for Computational Linguistics, Baltimore, Maryland, USA, pp. 36–41

Jurgens D, Pilehvar MT, Navigli R (2014) SemEval-2014 Task 3: Cross-Level Semantic Similarity. In: *Proceedings of the Eighth International Workshop on Semantic Evaluation (SemEval 2014)*. Association for Computational Linguistics, Dublin, Ireland, pp. 17–26

Jurgens D, Pilehvar MT, Navigli R (2016) Cross Level Semantic Similarity: An Evaluation Framework for Universal Measures of Similarity. *Language Resources and Evaluation* 50(1),

pp. 5–33. https://doi.org/10.1007/s10579-015-9318-3

Kale M, Siddhant A, Nag S, et al (2019) Supervised Contextual Embeddings for Transfer Learning in Natural Language Processing Tasks. In: *Proceedings of the Second Learning from Limited Labeled Data (LLD) Workshop*. New Orleans, Louisiana, USA

Kann K, Cho K, Bowman SR (2019) Towards Realistic Practices In Low-Resource Natural Language Processing: The Development Set. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the Ninth International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*. Association for Computational Linguistics, Hong Kong, China, pp. 3342–3349

Karthikeyan K, Wang Z, Mayhew S, Roth D (2020) Cross-Lingual Ability of Multilingual BERT: An Empirical Study. In: *Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020)*

Kešelj V, Šipka D (2008) Pristup izgradnji stemera i lematizora za jezike s bogatom fleksijom i oskudnim resursima zasnovan na obuhvatanju sufiksa. *INFOTEKA - časopis za bibliotekarstvo i informatiku* 9(1–2), pp. 21–31

Kessler JS, Eckert M, Clark L, Nicolov N (2010) The ICWSM 2010 JDPA Sentiment Corpus for the Automotive Domain. In: *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media Data Challenge Workshop (ICWSM-DCW 2010)*. Washington DC, USA

King BP (2015) Practical Natural Language Processing for Low-Resource Languages. University of Michigan

Kiros R, Zhu Y, Salakhutdinov R, et al (2015) Skip-Thought Vectors. In: *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS 2015)*. pp. 3294–3302

Koppel M, Schler J (2006) The Importance of Neutral Examples for Learning Sentiment. *Computational Intelligence* 22(2), pp. 100–109. https://doi.org/10.1111/j.1467-8640.2006.00276.x

Kornai A (2013) Digital Language Death. *PLoS ONE* 8(10), pp. e77056. https://doi.org/10.1371/journal.pone.0077056

Krauwer S (2003) The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap. In: *Proceedings of the 2003 International Workshop Speech and Computer (SPECOM 2003)*. Moscow, Russia

Krippendorff K (2004) Content Analysis: An Introduction to Its Methodology, 2nd edn. Sage, Beverly Hills, California, USA

Le H, Vial L, Frej J, et al (2020) FlauBERT: Unsupervised Language Model Pre-training for French. In: *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*. European Language Resources Association (ELRA), Marseille, France, pp. 2479–2490

Levy O, Goldberg Y, Dagan I (2015) Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics* 3, pp. 211–225. https://doi.org/10.1162/tacl_a_00134

Li Y, McLean D, Bandar ZA, et al (2006) Sentence Similarity Based on Semantic Nets and Corpus Statistics. *IEEE Transactions on Knowledge and Data Engineering* 18(8), pp. 1138–1150. https://doi.org/10.1109/TKDE.2006.130

Liu B, Zhang L (2012) A Survey of Opinion Mining and Sentiment Analysis. In: Aggarwal CC, Zhai C (eds) *Mining Text Data*. Springer, Boston, Massachusetts, USA, pp. 415–463

Liu X, He P, Chen W, Gao J (2019) Multi-Task Deep Neural Networks for Natural Language Understanding. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*. Association for Computational Linguistics, Florence, Italy, pp. 4487–4496

Ljajić A (2019) Obrada negacije u kratkim neformalnim tekstovima u cilju poboljšanja klasifikacije sentimenta. Univerzitet u Nišu

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), pp. 289–311. https://doi.org/10.2298/CSIS180122013L

Ljubešić N, Boras D, Kubelka O (2007) Retrieving Information in Croatian: Building a Simple and Efficient Rule-Based Stemmer. In: *INFuture2007: Digital Information and Heritage*. Department for Information Sciences, Faculty of Humanities and Social Sciences, Zagreb, Croatia, pp. 313–320

Ljubešić N, Klubička F (2014) {bs,hr,sr}WaC –Web corpora of Bosnian, Croatian and Serbian. In: *Proceedings of the Ninth Web as Corpus Workshop (WaC-9)*. Association for Computational Linguistics, Gothenburg, Sweden, pp. 29–35

Ljubešić N, Klubička F, Agić Ž, Jazbec I-P (2016) New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Portorož, Slovenia, pp. 4264–4270

Logeswaran L, Lee H (2018) An efficient framework for learning sentence representations. In: *Proceedings of the Sixth International Conference on Learning Representations (ICLR 2018)*. Vancouver, Canada

Lund K, Burgess C (1996) Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers* 28(2), pp. 203–208. https://doi.org/10.3758/BF03204766

Maas AL, Daly RE, Pham PT, et al (2011) Learning Word Vectors for Sentiment Analysis. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*. Association for Computational Linguistics, Portland, Oregon, USA, pp. 142–150

Mariani J, Francopoulo G, Paroubek P (2019) The NLP4NLP Corpus (I): 50 Years of Publication, Collaboration and Citation in Speech and Language Processing. *Frontiers in Research Metrics and Analytics* 3(February), Article 36. https://doi.org/10.3389/frma.2018.00036

Martin L, Muller B, Suárez PJO, et al (2020) CamemBERT: a Tasty French Language Model. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*. Association for Computational Linguistics, pp. 7203–7219

Maxwell M, Hughes B (2006) Frontiers in linguistic annotation for lower-density languages. In: *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006*. Association for Computational Linguistics, Sydney, Australia, pp. 29–37

Maynard D, Greenwood MA (2014) Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*. European Language Resources Association (ELRA), Reykjavik, Iceland, pp. 4238–4243

Mihalcea R, Corley C, Strapparava C (2006) Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In: *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI 2006)*. AAAI Press, Boston, Massachusetts, USA, pp. 775–780

Mikolov T, Chen K, Corrado G, Dean J (2013a) Efficient Estimation of Word Representations in Vector Space. In: *Proceedings of the International Conference on Learning Representations Workshop (ICLR 2013)*. Scottsdale, Arizona, USA

Mikolov T, Sutskever I, Chen K, et al (2013b) Distributed Representations of Words and Phrases and their Compositionality. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS 2013)*. Curran Associates, Inc., Lake Tahoe, Nevada, USA, pp. 3111–3119

Mikolov T, Yih W, Zweig G (2013c) Linguistic Regularities in Continuous Space Word Representations. In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*. Association for Computational Linguistics, pp. 746–751

Miller GA (1995) WordNet: A Lexical Database for English. *Communications of the ACM* 38(11), pp. 39–41. https://doi.org/10.1145/219717.219748

Miller GA, Beckwith R, Fellbaum C, et al (1990) Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography* 3(4), pp. 235–244

Milošević N (2012) Stemmer for Serbian language. arXiv:1209.4471

Mladenović M (2016) Informatički modeli u analizi osećanja zasnovani na jezičkim resursima. Univerzitet u Beogradu

Mladenović M, Mitrović J, Krstev C, Vitas D (2015) Hybrid sentiment analysis framework for a morphologically rich language. Journal of Intelligent Information Systems. https://doi.org/10.1007/s10844-015-0372-5

Mohammad SM (2020) NLP Scholar: A Dataset for Examining the State of NLP Research. In: *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*. European Language Resources Association (ELRA), Marseille, France, pp. 868–877

Mohammad SM (2016) A Practical Guide to Sentiment Annotation: Challenges and Solutions. In: *Proceedings of the Seventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, San Diego, California, USA, pp. 174–179

Mohammad SM, Salameh M, Kiritchenko S (2016) How Translation Alters Sentiment. *Journal of Artificial Intelligence Research* 55, pp. 95–130. https://doi.org/10.1613/jair.4787

Mohammad SM, Sobhani P, Kiritchenko S (2017) Stance and Sentiment in Tweets. *Special Section of the ACM Transactions on Internet Technology on Argumentation in Social Media* 17(3), Article No. 26. https://doi.org/10.1145/3003433

Mozetič I, Grčar M, Smailović J (2016) Multilingual Twitter Sentiment Classification: The Role of Human Annotators. *PLoS ONE* 11(5). https://doi.org/10.1371/journal.pone.0155036

Nabil M, Aly M, Atiya AF (2015) ASTD: Arabic Sentiment Tweets Dataset. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*. Association for Computational Linguistics, Lisbon, Portugal, pp. 2515–2519

Nozza D, Bianchi F, Hovy D (2020) What the [MASK]? Making Sense of Language-Specific BERT Models. arXiv:2003.02912

O'Shea J, Bandar Z, Crockett K (2013) A New Benchmark Dataset with Production Methodology for Short Text Semantic Similarity Algorithms. *ACM Transactions on Speech and Language Processing* 10(4), Article No. 19

Pagliardini M, Gupta P, Jaggi M (2018) Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018)*. Association for Computational Linguistics, New Orleans, Louisiana, USA, pp. 528–540

Pan SJ, Yang Q (2010) A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* 22(10), pp. 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Pang B, Lee L (2008) Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval* 2(1–2), pp. 1–135. https://doi.org/10.1561/1500000001

Pang B, Lee L (2005) Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*. Association for Computational Linguistics, Ann Arbor, Michigan, USA, pp. 115–124

Pang B, Lee L (2004) A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In: *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*. Association for Computational Linguistics, Morristown, New Jersey, USA, Article No. 271

Pang B, Lee L, Vaithyanathan S (2002) Thumbs up? Sentiment Classification using Machine Learning Techniques. In: *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp. 79–86

Pedregosa F, Varoquaux G, Gramfort A, et al (2011) Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, pp. 2825–2830

Pennington J, Socher R, Manning CD (2014) GloVe: Global Vectors for Word Representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*. Association for Computational Linguistics, Doha, Qatar, pp. 1532–1543

Peters ME, Neumann M, Iyyer M, et al (2018) Deep contextualized word representations. In:

Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (NAACL-HLT 2018). Association for Computational Linguistics, New Orleans, Louisiana, USA, pp. 2227–2237

Peters ME, Ruder S, Smith NA (2019) To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks. In: *Proceedings of the Fourth Workshop on Representation Learning for NLP (RepL4NLP-2019)*. Association for Computational Linguistics, Florence, Italy, pp. 7–14

Pham NT, Kruszewski G, Lazaridou A, Baroni M (2015) Jointly optimizing word representations for lexical and sentential tasks with the C-PHRASE model. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the Seventh International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2015)*. Beijing, China, pp. 971–981

Pires T, Schlinger E, Garrette D (2019) How Multilingual is Multilingual BERT? In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*. Association for Computational Linguistics, Florence, Italy, pp. 4996–5001

Pontiki M, Galanis D, Papageorgiou H, et al (2015) SemEval-2015 Task 12: Aspect Based Sentiment Analysis. In: *Proceedings of the Ninth International Workshop on Semantic Evaluation (SemEval 2015)*. Association for Computational Linguistics, Denver, Colorado, USA, pp. 486–495

Pontiki M, Galanis D, Papageorgiou H, et al (2016) SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*. Association for Computational Linguistics, San Diego, California, USA, pp. 19–30

Pontiki M, Galanis D, Pavlopoulos J, et al (2014) SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In: *Proceedings of the Eighth International Workshop on Semantic Evaluation (SemEval 2014)*. Association for Computational Linguistics, Dublin, Ireland, pp. 27–35

Ptáček T, Habernal I, Hong J (2014) Sarcasm Detection on Czech and English Twitter. In: *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers (COLING 2014)*. Dublin, Ireland, pp. 213–223

Pustejovsky J, Bunt H, Zaenen A (2017) Designing Annotation Schemes: From Theory to Model. In: *Handbook of Linguistic Annotation*. Springer, Dordrecht, pp. 21–72

Pustejovsky J, Stubbs A (2012) Natural Language Annotation for Machine Learning: A Guide to Corpus-Building for Applications, First Edit. O'Reilly Media

Quirk R, Greenbaum S, Leech G, Svartvik J (1985) A Comprehensive Grammar of the English Language. Longman, New York, New York, USA

Radford A, Narasimhan K, Salimans T, Sutskever I (2018) Improving Language Understanding by Generative Pre-Training

Radford A, Wu J, Child R, et al (2019) Language Models are Unsupervised Multitask Learners

Řehůřek R, Sojka P (2010) Software Framework for Topic Modelling with Large Corpora. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. European Language Resources Association (ELRA), Valletta, Malta, pp. 45–50

Rei M (2017) Semi-supervised Multitask Learning for Sequence Labeling. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*. Association for Computational Linguistics, Vancouver, Canada, pp. 2121–2130

Reimers N, Gurevych I (2020) Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. arXiv:2004.09813

Rennie JDM, Shih L, Teevan J, Karger D (2003) Tackling the Poor Assumptions of Naive Bayes Text Classifiers. In: *Proceedings of the 20th International Conference on Machine Learning (ICML 2003)*. Washington DC, USA

Rohde DLT, Gonnerman LM, Plaut DC (2005) An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence

Ruder S (2019) Neural Transfer Learning for Natural Language Processing. National University of Ireland, Galway

Ruder S, Vulić I, Søgaard A (2019) A Survey of Cross-lingual Word Embedding Models. *Journal of Artificial Intelligence Research* 65, pp. 569–631. https://doi.org/10.1613/jair.1.11640

Sabou M, Bontcheva K, Derczynski L, Scharl A (2014) Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*. European Language Resources Association (ELRA), Reykjavik, Iceland, pp. 859–866

Samardžić T, Ljubešić N, Miličević M (2015) Regional Linguistic Data Initiative (ReLDI). In: *Proceedings of the Fifth Workshop on Balto-Slavic Natural Language Processing (BSNLP 2015)*. Hissar, Bulgaria, pp. 40–42

Samardžić T, Starović M, Agić Ž, Ljubešić N (2017) Universal Dependencies for Serbian in Comparison with Croatian and Other Slavic Languages. In: *Proceedings of the Sixth Workshop on Balto-Slavic Natural Language Processing (BSNLP 2017)*. Association for Computational Linguistics, Valencia, Spain, pp. 39–44

Sanh V, Debut L, Chaumond J, Wolf T (2019) DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In: *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing (EMC^2)*. Vancouver, Canada

Scannell KP (2007) The Crúbadán Project: Corpus building for under-resourced languages. *Cahiers du Cental* 5(1)

Schuster M, Nakajima K (2012) Japanese and Korean voice search. In: *Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012)*. Kyoto, Japan, pp. 5149–5152

Searle JR (1975) A taxonomy of speech acts. In: *Language, mind, and knowledge*. University of Minnesota Press, Minneapolis, Minnesota, USA, pp. 344–369

Sennrich R, Haddow B, Birch A (2016) Neural Machine Translation of Rare Words with Subword Units. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*. Association for Computational Linguistics, Berlin, Germany, pp. 1715–1725

Siegert I, Böck R, Wendemuth A (2014) Inter-rater reliability for emotion annotation in human-computer interaction: Comparison and methodological improvements. *Journal on Multimodal User Interfaces* 8(1), pp. 17–28. https://doi.org/10.1007/s12193-013-0129-9

Singla K, Can D, Narayanan S (2018) A Multi-task Approach to Learning Multilingual Representations. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*. Association for Computational Linguistics, Melbourne, Australia, pp. 214–220

Socher R, Perelygin A, Wu JY, et al (2013) Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*. Association for Computational Linguistics, Seattle, Washington, USA, pp. 1631–1642

Søgaard A (2013) Semi-Supervised Learning and Domain Adaptation in Natural Language Processing, 1st edn. Morgan & Claypool Publishers

Streiter O, Scannell KP, Stuflesser M (2006) Implementing NLP projects for noncentral languages: Instructions for funding bodies, strategies for developers. *Machine Translation* 20(4), pp. 267–289. https://doi.org/10.1007/s10590-007-9026-x

Subramanian S, Trischler A, Bengio Y, Pal CJ (2018) Learning General Purpose Distributed Sentence Representations via Large Scale Multi-task Learning. In: *Proceedings of the Sixth International Conference on Learning Representations (ICLR 2018)*. Vancouver, Canada

Täckström O, Mcdonald R (2011) Semi-supervised latent variable models for sentence-level sentiment analysis. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*. Association for Computational Linguistics, Portland, Oregon, USA, pp. 569–574

Thelwall M, Buckley K, Paltoglou G, et al (2010) Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology* 61(12), pp. 2544–2558. https://doi.org/10.1002/asi.21416

Turney PD (2002) Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp. 417–424

Van de Kauter M, Desmet B, Hoste V (2015) The good, the bad and the implicit: a comprehensive approach to annotating explicit and implicit sentiment. *Language Resources and Evaluation* 49(3), pp. 685–720. https://doi.org/10.1007/s10579-015-9297-4

Vaswani A, Shazeer N, Parmar N, et al (2017) Attention Is All You Need. In: *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*. Long Beach, California, USA

Virtanen P, Gommers R, Oliphant TE, et al (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* 17, pp. 261–272. https://doi.org/10.1038/s41592-019-0686-2

Wang S, Manning CD (2012) Baselines and Bigrams: Simple, Good Sentiment and Topic

Classification. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*. Association for Computational Linguistics, Jeju Island, South Korea, pp. 90–94

Wenzek G, Lachaux M-A, Conneau A, et al (2019) CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data

Widdows D (2004) Geometry and Meaning. CSLI Publications

Wiebe J, Wilson T, Cardie C (2005) Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation* 39(2–3), pp. 165–210. https://doi.org/10.1007/s10579-005-7880-9

Wiemer-Hastings P (2004) All parts are not created equal: SIAM-LSA. In: *Proceedings of the 26th Annual Conference of the Cognitive Science Society*. Erlbaum, Chicago, Illinois, USA

Wieting J, Bansal M, Gimpel K, Livescu K (2016) Towards Universal Paraphrastic Sentence Embeddings. In: *Proceedings of the Fourth International Conference on Learning Representations (ICLR 2016)*. San Juan, Puerto Rico

Wolf T, Debut L, Sanh V, et al (2019) HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv:1910.03771

Wróblewska A, Krasnowska-Kieraś K (2017) Polish evaluation dataset for compositional distributional semantics models. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*. Association for Computational Linguistics, Vancouver, Canada, pp. 784–792

Wu S, Dredze M (2019) Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the Ninth International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*. Association for Computational Linguistics, Hong Kong, China, pp. 833–844

Xu W, Callison-Burch C, Dolan WB (2015) SemEval-2015 Task 1: Paraphrase and Semantic Similarity in Twitter (PIT). In: *Proceedings of the Ninth International Workshop on Semantic Evaluation (SemEval 2015)*. Association for Computational Linguistics, Denver, Colorado, USA, pp. 1–11

Zhu Y, Kiros R, Zemel R, et al (2015) Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In: *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV 2015)*. IEEE, Santiago, Chile, pp. 19–27

Spisak slika

Spisak tabela

Spisak skraćenica

ACE	Annotation Cost-Effectiveness
ACL	Association for Computational Linguistics
BERT	Bidirectional Encoder Representations from Transformers
BOE	Bag-of-Embeddings – Vreća vektora značenja reči
BOW	Bag-of-Words – Vreća reči
BPE	Byte-Pair Encoding
CBOW	Continuous Bag-of-Words
CNB	Complement Naïve Bayes – Komplementni naivni Bajesov klasifikator
Dim	Dimenzionalnost vektora značenja reči
EG	Eksperimentalna grupa anotatora
GPU	Graphics Processing Unit – Grafička procesorska jedinica
HTML	Hypertext Markup Language
IG	Inicijalna grupa anotatora
LCS	Longest Common Subsequence
LInSTSS	Language Independent Short-Text Semantic Similarity
KG	Kontrolna grupa anotatora
LR	Logistička regresija
MLM	Masked Language Modeling
MNB	Multinomial Naïve Bayes – Multinomijalni naivni Bajesov klasifikator
MULTEXT-East	Multilingual Text Tools and Corpora for Central and Eastern European Languages
NBSVM	Naïve Bayes Support Vector Machine
NEPK	Normalizacija emotikona i ponavljanja karaktera
NLCS	Normalized Longest Common Subsequence
$NMCLCS_1$	Normalized Maximal Consecutive Longest Common Subsequence starting at character 1
$NMCLCS_N$	Normalized Maximal Consecutive Longest Common Subsequence starting at character N
NLP	Natural Language Processing – Obrada prirodnih jezika
POS	Part-of-Speech – Vrsta reči
POST STSS	Part-of-Speech Tag-supported Short-Text Semantic Similarity
POS-TF STSS	Part-of-Speech and Term Frequency weighted Short-Text Semantic Similarity
ReLDI	Regional Linguistic Data Initiative
STS	Semantic Textual Similarity – Semantička sličnost kratkih tekstova
STSS	Short-Text Semantic Similarity – Semantička sličnost kratkih tekstova
SVD	Singular Value Decomposition
SVM	Support Vector Machine – Metoda potpornih vektora
TF	Term Frequency
TF-IDF	Term Frequency – Inverse Document Frequency
TPU	Tensor Processing Unit – Tenzorska procesorska jedinica
WS	Window size – Širina kontekstnog opsega
XLM	Cross-lingual Language Model

Biografija autora

Vuk Batanović je rođen 02.06.1987. u Beogradu. Osnovne akademske studije na Elektrotehničkom fakultetu Univerziteta u Beogradu upisao je 2006. godine. Diplomirao je na modulu Računarska tehnika i informatika 2010. godine, sa prosečnom ocenom 9,56. Master akademske studije na istom modulu Elektrotehničkog fakulteta je upisao 2010. i završio 2011. godine, sa prosečnom ocenom 10. U prolećnom semestru školske 2011/2012. godine upisao je doktorske akademske studije na modulu Softversko inženjerstvo istog fakulteta. Ispite na doktorskim studijama je položio sa prosečnom ocenom 10. U toku studija usavršavao se prisustvovanjem većem broju letnjih škola iz obrade prirodnih jezika i mašinskog učenja.

Sarađivao je na međunarodnom projektu *Regional Linguistic Data Initiative* (ReLDI) i slovenačkom projektu CLARIN.SI (*Common Language Resources and Technology Infrastructure*). Od 2018. godine radi kao istraživač-saradnik u Inovacionom centru Elektrotehničkog fakulteta u Beogradu. Bio je angažovan na projektu Programa Ujedinjenih nacija za razvoj (UNDP) o automatizaciji semantičke pretrage pravnih dokumenata na srpskom jeziku. Od školske 2017/2018. godine angažovan je u nastavi na master akademskim studijama Elektrotehničkog fakulteta, na predmetima Obrada prirodnih jezika i Pronalaženje skrivenog znanja.

Naučna interesovanja su mu usmerena na rešavanje semantičkih problema iz obrade prirodnih jezika, pre svega primenom metoda mašinskog učenja. Naročito ga interesuju specifičnosti obrade kratkih tekstova i rešenja primenjiva i u jezicima sa ograničenim resursima. Objavio je 16 naučnih publikacija, od čega 2 rada u međunarodnim časopisima (M20), 2 rada u časopisima nacionalnog značaja (M50), 7 radova na međunarodnim (M30) i 3 na domaćim konferencijama (M60) i 2 tehnička rešenja (M80).

Изјава о ауторству

Име и презиме аутора Вук Батановић

Број индекса 5045/2011

Изјављујем

да је докторска дисертација под насловом

Методологија решавања семантичких проблема у обради кратких текстова написаних на природним језицима са ограниченим ресурсима

- резултат сопственог истраживачког рада;
- да дисертација у целини ни у деловима није била предложена за стицање друге дипломе према студијским програмима других високошколских установа;
- да су резултати коректно наведени и
- да нисам кршио/ла ауторска права и користио/ла интелектуалну својину других лица.

Потпис аутора

У Београду, 20.08.2020

Вук Батановић

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора Вук Батановић

Број индекса 5045/2011

Студијски програм Софтверско инжењерство

Наслов рада Методологија решавања семантичких проблема у обради кратких текстова написаних на природним језицима са ограниченим ресурсима

Ментори проф. др Бошко Николић, проф. др Милош Цветановић

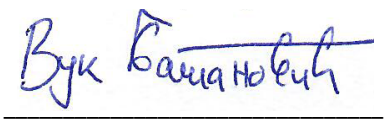
Изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла ради похрањивања у **Дигиталном репозиторијуму Универзитета у Београду.**

Дозвољавам да се објаве моји лични подаци везани за добијање академског назива доктора наука, као што су име и презиме, година и место рођења и датум одбране рада.

Ови лични подаци могу се објавити на мрежним страницама дигиталне библиотеке, у електронском каталогу и у публикацијама Универзитета у Београду.

Потпис аутора

У Београду, 20.08.2020.



Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

Методологија решавања семантичких проблема у обради кратких текстова написаних на природним језицима са ограниченим ресурсима

која је моје ауторско дело.

Дисертацију са свим прилозима предао/ла сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију похрањену у Дигиталном репозиторијуму Универзитета у Београду и доступну у отвореном приступу могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучио/ла.

1. Ауторство (CC BY)

2. Ауторство – некомерцијално (CC BY-NC)

3. Ауторство – некомерцијално – без прерада (CC BY-NC-ND)

4. Ауторство – некомерцијално – делити под истим условима (CC BY-NC-SA)

5. Ауторство – без прерада (CC BY-ND)

6. Ауторство – делити под истим условима (CC BY-SA)

(Молимо да заокружите само једну од шест понуђених лиценци.
Кратак опис лиценци је саставни део ове изјаве).

Потпис аутора

У Београду, 20.08.2020.

Вук Батањевић

1. **Ауторство**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце, чак и у комерцијалне сврхе. Ово је најслободнија од свих лиценци.

2. **Ауторство – некомерцијално**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела.

3. **Ауторство – некомерцијално – без прерада**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела. У односу на све остале лиценце, овом лиценцом се ограничава највећи обим права коришћења дела.

4. **Ауторство – некомерцијално – делити под истим условима**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца не дозвољава комерцијалну употребу дела и прерада.

5. **Ауторство – без прерада**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца дозвољава комерцијалну употребу дела.

6. **Ауторство – делити под истим условима**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца дозвољава комерцијалну употребу дела и прерада. Слична је софтверским лиценцама, односно лиценцама отвореног кода.