SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 118

ODREĐIVANJE SEMANTIČKE ORIJENTACIJE SUBJEKTIVNIH RIJEČI I FRAZA

Ivan Šolta

Zagreb, lipanj 2010.

Zahvaljujem kolegama i prijateljima na uloženom trudu i vremenu za označavanje uzoraka:

Doris Alajbegović
Ana Brebrić
Ognjen Lajšić
Marko Martinis
dr. sc. Jan Šnajder

Sadržaj

1 Uvod

Velika količina teksta s kojim se svakodnevno susrećemo u dnevnom tisku i na Internetu prenosi ili izražava nečije subjektivno mišljenje, stav ili doživljaj. Kolumne u novinama, rasprave na forumima, recenzije filmova, klubova, restorana ili novih elektroničkih naprava imaju kao glavnu funkciju prenijeti čitatelju subjektivne stavove autora. Kažemo da se radi o subjektivnim tekstovima, budući da su objektivne činjenice navedene samo kao argumenti ili u nekim slučajevima čak i potpuno izostavljene. Klasični pristup dubinskoj analizi teksta fokusira se na činjenice u tekstu, a informaciju o subjektivnosti stavlja u drugi plan ili potpuno zanemaruje. Kako je često upravo subjektivni dio teksta zanimljiv korisniku potrebno ga je na odgovarajući način obraditi. Prepoznavanje subjektivnih tekstova i njihova strojna obrada naziva se analiza subjektivnosti (eng. *subjectivity analysis*).

Subjektivni tekstovi u kojima autor opisuje neki kompleksni objekt komentirajući njegove brojne značajke nisu nikada u potpunosti pozitivni ili negativni. U recenziji restorana autor može pohvaliti kvalitetu jela, ugodnu atmosferu i glazbu, a kritizirati poslugu, siromašan izbor na jelovniku i slično. Zbog toga je potrebno analizu subjektivnosti provesti na razini rečenice, pa i niže, na razini subjektivne fraze. Za frazu kažemo da je subjektivna ako ima neku semantičku orijentaciju, tj. ako možemo reći da je njome izraženo neko pozitivno ili negativno mišljenje.

Pronalaženje subjektivnih fraza i određivanje njihove semantičke orijentacije ima široku primjenu. Pretraživanje Interneta bi automatizacijom analize subjektivnosti dobilo sasvim novu dimenziju. Rezultati koje bi tražilica dohvatila korisniku bi pružili pregledan uvid u stavove drugih ljudi o pojmu koji traži. Sortiranje na pozitivne i negativne komentare o traženom pojmu, navođenje najčešćih zamjerki i značajki koje su dobile najviše komentara i slične mogućnosti drastično bi smanjile vrijeme potrebno za procesiranje rezultata. Čitanje recenzija i dugih forumskih rasprava s ciljem stvaranje slike o nekom proizvodu bilo bi nepotrebno. S druge strane, kompanije bi za analizu tržišta mogle iskoristiti sve postojeće dokumente na Internetu kako bi saznale kakvu sliku o njihovim proizvodima imaju ciljne skupine potrošača. Analiza komentara na političke teme dala bi zanimljive informacije o stavu birača u vrijeme izbora.

Strojna analiza subjektivnosti pretpostavlja mogućnost identifikacije subjektivnih fraza i određivanje semantičke orijentacije. Oba problema su izuzetno zahtjevna. U sklopu rada obrađen je problem određivanja semantičke orijentacije subjektivnih fraza. Iduće poglavlje daje opsežan pregled postojećih rješenja. U trećem poglavlju opisana je komponenta za automatsku izgradnju leksikona u kojem su pohranjene a priori orijentacije riječi. Leksikon je korišten prilikom određivanja orijentacije subjektivne fraze. Postupak je opisan u četvrtom poglavlju. Provedena je evaluacija leksikona i analiza značajki koje bi mogle biti upotrjebljene za određivanje orijentacije fraze. Uzorci korišteni za evaluaciju opisani su u poglavlju 5, a rezultati su prikazani u poglavlju 6.

2 Pregled područja

Strojna analiza subjektivnog teksta je relativno novo i zanimljivo područje istraživanja sa širokim područjem primjene. Pristup problemu na razini dokumenta, u smislu klasifikacije na pozitivne i negativne tekstove u određenoj mjeri sliči na klasifikaciju teksta po temama. U odjeljku 2.1 prikazan je jedan takav pristup koji je pokazao da se radi o daleko složenijem problemu te da klasične metode ne daju dovoljno dobre rezultate. Promatramo li problem analize subjektivnog teksta na razini subjektivne fraze, primjećuje se trend u istraživanjima koji se svodi na određivanje *a priori* orijentacije riječi i kompozicije sentimenta do razine fraze. U odjeljku 2.2 prikazan je postupak za automatsko generiranje leksikona sa semantičkim orijentacijama riječi. Također je dan pregled ručno označenog leksikona koji je često korišten u domeni engleskog jezika. Istraživanja kompozicije sentimenta u svrhu određivanja orijentacije fraze opisana su u odjeljku 2.3.

2.1 Klasifikacija na razini dokumenta

Reprezentativan problem klasične analize teksta jest klasifikacija tekstova po temama. Unatoč tome što teme koje je potrebno razlikovati mogu biti vrlo slične i tvoriti složenu taksonomiju, klasificiranje stava iznesenog u subjektivnom tekstu na samo dvije klase (pozitivan i negativan stav), pokazalo se kao daleko teži problem. Ljudima nije teško prepoznati stavove u tekstu, niti odrediti jesu li oni pozitivni ili negativni. No taj je postupak teško automatizirati jer se velikim dijelom oslanja na razumijevanje konteksta u kojem je stav iznesen. Kada u književnoj kritici naiđete na frazu *pročitajte knjigu* jasno je da se radi o pozitivnom stavu autora o komentiranoj knjizi. Međutim, nađe li se ista fraza u filmskoj kritici radi se o izrazito negativnom stavu. Postupcima strojne analize ovakve je slučajeve iznimno teško razlikovati.

Pang et al. (2002) pokušali su problem klasifikacije filmskih kritika riješiti metodom koja se pokazala učinkovitom u klasifikaciji tekstova po temama. Metoda je objašnjena u ostatku odjeljka. Rezultati istraživanja i njihovi zaključci dobro ilustriraju prethodnu tvrdnju.

Kako bismo klasificirali tekst, potrebno ga je pretvoriti u oblik na koji možemo primijeniti postojeće algoritme. Neka je $\{f_1, f_2, ..., f_m\}$ m-člani skup riječi i dvočlanih izraza. Broj $n_i(d)$ označava koliko se puta izraz f_i javio u tekstu d. Svaki od tekstova koje želimo klasificirati pretvaramo u vektor $\boldsymbol{x} := (n_1(d), n_2(d), ..., n_m(d))$.

Bayesov klasifikator na temelju vjerojatnosti pojave pojedine klase i vjerojatnosti pojave teksta unutar određene klase računa vjerojatnost da dani tekst pripada klasi *c*.

$$p(c|d) = \frac{p(c)p(d|c)}{p(d)}$$

Problem klasifikacije svodi se na problem maksimizacije gornjeg izraza. Za procjenu uvjetne vjerojatnosti $p(d|c)$ pretpostavljamo da su pojave izraza f_i nezavisne. Tada gornji izraz postaje

$$p(c|d) = \frac{p(c)\prod_{i=1}^{m} p(f_i|c)^{n_i(d)}}{p(d)}$$

Unatoč jednostavnosti i pretpostavci koja u realnim slučajevima nije opravdana Bayesov klasifikator daje iznenađujuće dobre rezultate.

Stroj s potpornim vektorima (eng. *support vector machine*) traži hiperravninu predstavljenu vektorom $\boldsymbol{w}$ koja razdvaja klase uzoraka. Udaljenost hiperravnine i najbližeg uzorka nazivamo *margina razdvajanja*. SVM hiperravninu u prostor smješta tako da je margina razdvajanja maksimalna. Opisani postupak se svodi na problem optimizacije izraza

$$\boldsymbol{w} = \sum_i \alpha_i c_i \boldsymbol{x}_i$$

Klasa *c* predstavljena je sa vrijednosti iz skupa {1, -1}. Težine α_i dobivamo rješavanjem dualnog problema. Oni uzorci $\boldsymbol{x_i}$ za koje su težine α_i nenegativne nazivaju se potporni vektori budući da su to jedini uzorci koji utječu na $\boldsymbol{w}$. Klasifikacija se obavlja tako da se određuje strana hiperravnine na kojoj se nalazi uzorak. U klasifikaciji teksta SVM općenito daje bolje rezultate od Bayesovog klasifikatora. Detaljan opis metode je dan u 4.1.2.

Rezultati ispitivanja prikazani su u tablici 2.1. Stupac NB prikazuje rezultate Bayesovog klasifikatora, a stupac SVM rezultate dobivene uporabom stroja s potpornim vektorima.

Pang i Lee (2002) pokazali su nekoliko zanimljivih činjenica (tablica 2.1). Prvo, za razliku od klasifikacije teksta po temama, učestalost pojave nekog izraza nije presudna. Pokazalo se da su rezultati znatno bolji ukoliko prilikom preslikavanja teksta u vektor $\boldsymbol{x}$ umjesto frekvencije pojave izraza f_i promatramo samo je li se izraz pojavio ili nije. Neočekivano, dvočlani izrazi nisu toliko dobre značajke koliko jednočlani izrazi. Konačno, pokazalo se da se uporabom samo nekoliko najčešćih izraza postižu jednako dobri rezultati.

Tablica 2.1 Rezultati klasifikacije na razini dokumenta

Značajke	Broj značajki	Frekv. ili prisustvo	NB	ME	SVM
unigrami	16165	frekv.	**78.7**	NA	72.8
unigrami	16165	pris.	81.0	80.4	**82.9**
unigrami + bigrami	32330	pris.	80.6	80.8	**82.7**
bigrami	16165	pris.	77.3	**77.4**	77.1
unigrami + POS	16695	pris.	81.5	80.4	**81.9**
pridjevi	2633	pris.	77.0	**77.7**	75.1
2633 unigrama	2633	pris.	80.3	81.0	**81.4**
unigrami + položaj	22430	pris.	81.0	80.1	**81.6**

Na prvi pogled učinkovitost od preko 80% izgleda kao jako dobar rezultat. Unatoč tome što se u ovom slučaju radi o klasifikaciji na samo dva razreda, primjena iste metode na problem klasifikacije teksta na nekoliko tema daje znatno bolje rezultate. Već tu je jasno da se radi o složenijem problemu. Valja napomenuti i to da označavanje cijelog teksta kao pozitivnog ili negativnog ne izvlači onoliko informacija koliko bi korisnik možda želio dobiti. Gotovo svaki tekst se sastoji od pozitivnih i negativnih dijelova. Ako se u tekstu komentira neki složeniji objekt ili događaj vrlo je vjerojatno da će autor teksta o nekim komponentama imati pozitivno mišljenje, a o nekima negativno. Prethodna metoda promatra ukupnost teksta i zanemaruje informacije na nižoj razini.

2.2 Izgradnja leksikona

2.2.1 Leksikon temeljen na korelaciji

Turney i Littman (2003) predlažu da se semantička orijentacija svake riječi određuje na temelju semantičke veze dotične riječi s riječima iz malog, unaprijed definiranog skupa čija je semantička orijentacija poznata. Heuristično pravilo na kojem se postupak temelji pretpostavlja da se riječi koje su jednake semantičke orijentacije upotrebljavaju zajedno. Prema tome njihova će udaljenost u tekstu biti mala, tj. semantička veza između njih je velika. Kao referentni skup predlažu:

pWords = {dobar, lijep, odličan, pozitivan, sretan, ispravan, superioran}
pWords = {loš, ružan, jadan, negativan, nesretan,pogrešan, inferioran}.

Podskup *pWords* je skup riječi koje su neovisno o kontekstu pozitivne semantičke orijentacije, dok su riječi iz podskupa *nWords* uvijek negativne orijentacije. Skupovi *pWords* i *nWords* nisu skupovi za treniranje, već skupovi koji definiraju pozitivnu, tj. negativnu orijentaciju te je postupak učenja koji predlažu u biti nenadzirani postupak.

Definiramo mjeru PMI (eng. *pointwise mutual information*) (Turney i Littman, 2003) između dvije riječi *r* i *q* na sljedeći način:

$$PMI(r,q) = \log_2 \frac{p(rBLIZUq)}{p(r)p(q)}$$

Operator *BLIZU* označava da se riječ *r* pojavila blizu riječi *q*. Vjerojatnost *p(r BLIZU q)* označava vjerojatnost da se dvije riječi *r* i *q* pojave blizu jedna druge. Vjerojatnosti *p(r)* i *p(q)* su vjerojatnosti pojave pojedinih riječi u jeziku. Jednostavan i učinkovit način za odrediti potrebne vjerojatnosti je upotreba Internetske tražilice pri čemu je pretragu potrebno ograničiti na jezik kojim se bavimo. Semantičku orijentaciju riječi temeljenu na PMI (kratko SO-PMI) određujemo prema sljedećem izrazu:

$$SOPMI\ (r) = \sum_{k\ \epsilon\ Prije\check{c}i} PMI(r,k) - \sum_{l\ \epsilon\ Nrije\check{c}i} PMI(r,l)$$

Riječi s pozitivnom vrijednosti SO-PMI klasificirati ćemo kao riječi s pozitivnom semantičkom orijentacijom, dok će negativna vrijednost SO-PMI označavati da se radi o riječi s negativnom semantičkom orijentacijom. Apsolutna vrijednost SO-PMI označava u kojoj je mjeri riječ pozitivna, tj. negativna.

Na primjer, semantičku orijentaciju riječi *dugotrajan* odredili bismo dakle tako da izračunamo PMI sa svakom od riječi iz oba referentna skupa te se odlučimo za onaj kod kojeg je suma pojedinih vrijednosti PMI veća. Ako bi to bio skup *pWords* to bi značilo da se riječ više koristi za izražavanje pozitivne semantičke orijentacije.

Ispitivanje je izvršeno nad listom od 3,596 ručno označenih riječi. Korištena je tražilica AltaVista. Operator BLIZU u obzir uzima okolinu od 10 riječi. Korpus AV-ENG čine dokumenti na engleskom jeziku koje je tražilica indeksirala. Korpus AV-CA su dokumenti na engleskom jeziku koji su dohvaćeni sa stranica koje se nalaze u domeni *.ca*. Nakon izračunavanja SO-PMI vrijednosti riječi su sortirane prema apsolutnom iznosu izračunate vrijednosti. Klasificirano je prvih *n* riječi. Rezultati su prikazani tablicom 2.2.

Tablica 2.2 Točnost leksikona izgrađenog na temelju korelacije

Postotak skupa za testiranje	Veličina skupa za testiranje	AV-ENG	AV-CA
100%	3596	82.84%	76.06%
75%	2697	90.66%	81.76%
50%	1798	95.49%	87.26%
25%	899	97.11%	89.88%
Broj riječi u korpusu		1×10^{11}	2×10^{9}

2.2.2 Izgradnja leksikona pomoću pojmovnika

Esuli (2005) predlaže drugačiji pristup problemu klasifikacije na razini riječi. Riječi koje su identične semantičke orijentacije često u rječniku imaju slično tumačenje. To možemo iskoristiti tako da kao uzorke promatramo tumačenja riječi. Nad tumačenjima ručno označenih riječi obavimo treniranje klasifikatora. Za neku danu riječ koju je potrebno

klasificirati potražimo tumačenje te ga klasificiramo. Ispitivanja su pokazala da je učinkovitost ove metode otprilike jednaka prethodno opisanoj metodi. Ovaj princip korišten je u izgradnji SentiWordNeta (Esuli, 2006). Svakoj riječi iz WordNeta (Fellbaum, 1998) na temelju opisa dodijeljena je ocjena pozitivnosti, negativnosti i objektivnosti.

2.2.3 Ručno izgrađeni leksikon

DAL (eng. *Dictionary of Affect in Language*) (Whissel, 1989) je leksikon koji se sastoji od 8742 ručno označene riječi. Uzorci koji su korišteni u izradi leksikona prikupljeni su iz raznih izvora kao što su intervjui, adolescentski opisi emotivnih stanja i razni studentski radovi i eseji. Time se izbjegla pristranost prema određenoj domeni ili izvoru. Svakoj riječi u leksikonu pridružene su tri ocjene u intervalu od 1 (niska ocjena) do 3 (visoka ocjena). Ugodnost (eng. *plesantness*), označena sa *ee*, je mjera semantičke orijentacije riječi. Aktivnost (eng. *activeness, aa*) označava dinamiku osjećaja vezanih uz riječ. U tablici 2.3 redak tri i četiri dobro ilustriraju ovu mjeru. Slikovitost (eng. *imagery, ii*) kazuje u kojoj je mjeri jednostavno ili teško stvoriti mentalnu sliku vezanu uz riječ. Prvi i zadnji redak u tablici dobro ilustriraju ovu mjeru.

Tablica 2.3 Primjeri riječi s pripadajućim ocjenama iz leksikona DAL (Agarwal et al., 2009)

Riječ	*ee*	*aa*	*ii*
Affect	1.75	1.85	1.60
Affection	2.77	2.25	2.00
Slug	1.00	1.18	2.40
Energetic	2.25	3.00	3.00
Flower	2.75	1.07	3.00

Agarwal et al. (2009) u svom istraživanju koriste DAL proširen pomoću WordNeta (Fellbaum, 1998). Pretpostavka koju koriste je da sinonimi neke riječi imaju iste ocjene kao i početna riječ, dok antonimi imaju recipročne ocjene. Ukoliko se prilikom obrade naiđe na riječ koja se ne nalazi u leksikonu DAL, pomoću WordNeta se stvara lista antonima i sinonima nepoznate riječi. Listom se prolazi sekvencijalno dok se ne naiđe na riječ za koju

su poznate ocjene. Ako se niti jedna riječ iz liste ne nalazi u leksikonu, početnoj riječi se ne dodjeljuju nikakve ocjene.

Tri ocjene iz leksikona su nekorelirane (Cowie et al., 2001) pa zato ima smisla definirati normu koja će ih objediniti (Agarwal et al., 2009):

$$norm = \frac{\sqrt{ee^2 + aa^2}}{ii}$$

Nazivnik ovisi o ocjeni ugodnosti i aktivnosti riječi i predstavlja AE-reprezentaciju (eng. *Activation-Evaluation*) (Cowine et al., 2001). Riječi mogu pripadati u jednu od četiri kategorije:

1. Visok AE, visok ii: Riječi koje su jako polarizirane i u maloj mjeri subjektivne (npr. *anđeo*)
2. Nizak AE, nizak ii: Neutralne riječi podložne polaritetu konteksta
3. Visok AE, nizak ii: Polarne i subjektivne riječi (npr. *uspjeti*, *dobar*)
4. Nizak AE, visok ii: Neutralne riječi koje je lako zamisliti (npr. *vrata*, *auto*)

Glavna uloga ove ocjene je razlikovanje između jako subjektivnih riječi koje u kontekstu mogu lako promijeniti orijentaciju i manje subjektivnih riječi koje rijetko mijenjaju orijentaciju. Rezultati istraživanja su pokazali da je norma jedna od najutjecajnijih značajki prilikom kompozicije sentimenta.

2.3 Određivanje semantičke orijentacije fraze

Kada su poznate orijentacije pojedinih riječi potrebno je na temelju njih izračunati semantičku orijentaciju cijele fraze. Orijentacija fraze može biti drugačija od *a priori* orijentacija riječi koji se u njoj javljaju. Pozitivno orijentirane riječi koriste se za izražavanje negativnih stavova, i obratno. Neke riječi koje imaju orijentaciju izvan konteksta fraze ne nose sentiment unutar fraze. Neutralne riječi često mogu poprimiti orijentaciju ovisno o kontekstu. Negacije poprilično jednoznačno mijenjaju orijentaciju riječi na koje se odnose. Utvrditi koje su to riječi i koliki je doseg negacije nije uvijek jednostavan problem.

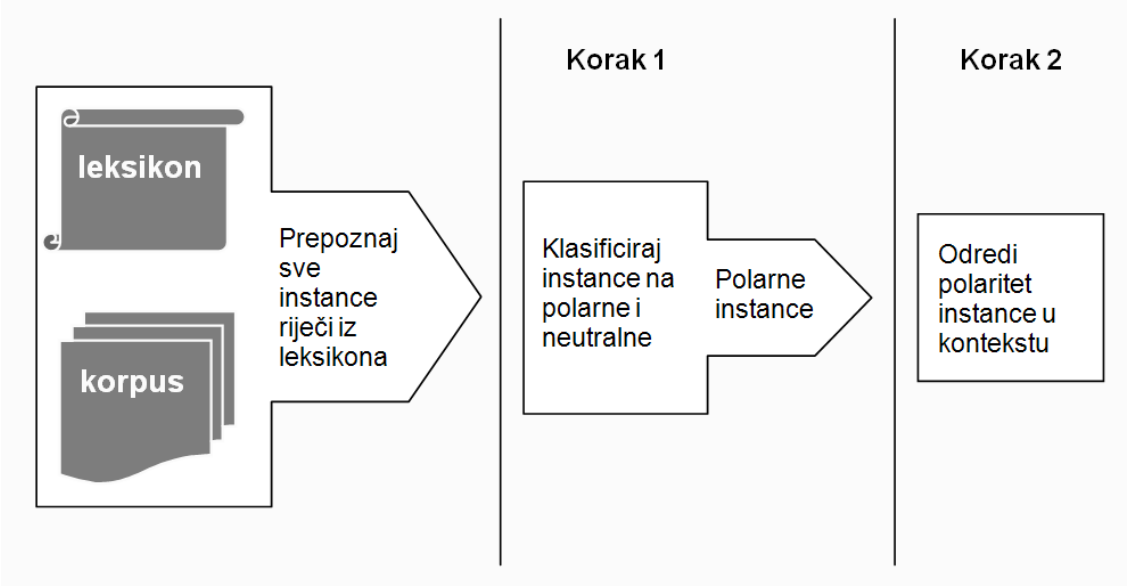
Osim negacija postoji mnoštvo riječi koje može promijeniti ili na drugi način utjecati na orijentaciju riječi na koje se odnose (npr. *jako*, *eliminirati*...).

U primjeru vidimo da označena fraza ima pozitivnu orijentaciju. Riječ *uništen* mijenja orijentaciju riječi *trud*, no kako se radi o *pljačkašima* cijela fraza u kontekstu izražava pozitivnu činjenicu.

[Trud(+) pljačkaša(-) je uništen(-)] pravovremenom policijskom akcijom.

Domena i tema također utječu na orijentaciju fraze i riječi. Tako na primjer, riječ *nepredvidiv* u recenziji automobila ima negativnu orijentaciju, dok u filmskoj kritici izražava pozitivnu karakteristiku.

2.3.1 Semantička orijentacija riječi u kontekstu



Slika 2.1 Postupak određivanja semantičke orijentacije instance u dva koraka (Wilson et al. 2009)

Na skupu od 3761 instanci pokazalo se da samo 48% riječi zadržava svoju *a priori* orijentaciju u kontekstu fraze (Wilson et al. 2009). Čak 76% greške nastaje jer su pozitivne ili negativne riječi u kontekstu fraze neutralne. Samo 11% greške je uzrokovano okretanjem polariteta riječi unutar fraze. To pokazuje da je opravdano u prvom koraku odrediti je li instanca unutar fraze neutralna ili nosi polaritet.

Wilson et al. (2009) koriste ručno izgrađen leksikon, a problemu određivanja orijentacije fraze pristupaju u dva koraka (slika 2.1). Leksikon se sastoji od 8000 riječi koje mogu izražavati subjektivnost. To uključuje i riječi koje mogu imati i objektivno značenje. Leksikon je izgrađen od liste subjektivnih riječi koju su sastavili Riloff i Wiebe (2003), proširene pomoću rječnika i tezaurusa. Također su dodane pozitivne i negativne riječi iz General Inquirera (Stone et al. 1966). Svaka riječ ima oznaku: *pozitivna*, *negativna*, *oboje* ili *neutralna*. Riječi je dodijeljena oznaka *oboje* ukoliko istovremeno pobuđuje pozitivne i negativne reakcije. Primjer takve riječi je *ludnica* koja u značenju mentalne ustanove ima negativnu orijentaciju, a u žargonu pozitivnu orijentaciju. Osim navedenih oznaka svakoj je riječi dodijeljena kategorija *slaba subjektivna riječ* ili *jaka subjektivna riječ*. Ove kategorije koreliraju s normom koju su definirali Agarwal et al (2009), a označuju u kojem je opsegu riječ subjektivna. U leksikonu većina riječi (92.8%) ima pozitivan (33.1%) ili negativan (59.7%) polaritet. Samo je 6.9% riječi označeno kao neutralno.

U prvom koraku sve instance riječi klasificiraju na neutralne i polarne. Instance su riječi iz leksikona upotrijebljene u frazi. Ideja ovog koraka je prepoznati slučajeve u kojima je polarna riječ u kontekstu fraze neutralna, tj. nema ulogu izražavanja subjektivnog stava. Svaka instanca predstavlja se skupom značajki. Značajke su podijeljene u šest skupina: značajke riječi, opće modifikacijske značajke, modifikacijske značajke polariteta, strukturalne značajke, značajke rečenice i značajka dokumenta. Značajke riječi uključuju samu riječ, vrstu riječi instance te prethodne i iduće riječi, *a priori* orijentaciju i kategoriju iz leksikona. Opće modifikacijske značajke su binarne značajke koje opisuju stablo ovisnosti (eng. *dependency tree*) dobiveno parsiranjem rečenice u kojoj se instanca nalazi (Collins 1997). Također označuju prethodi li instanci vrsta riječi koja može imati ulogu modifikatora, te je li instanca modifikator. Modifikacijske značajke polariteta opisuju vezu između instance i ostalih instanci u rečenici. Strukturalne značajke označuju koju funkciju ima instanca u rečenici (subjekt, predikat...). Značajke rečenice broje pojave pridjeva, instanci itd. u okolnim rečenicama, a značajka dokumenta označava domenu ili temu dokumenta.

U drugom koraku se određuje polaritet instance na temelju značajki podijeljenih u četiri skupine: značajke riječi, modifikacijske značajke polariteta, značajke negacije, promjene polariteta. Prve dvije skupine značajki jednake su kao u prvom koraku postupka. Značajke negacije su binarne značajke koje su postavljene ako je instanca unutar dosega negacije. Pri tome je napravljena razlika u udaljenosti negacije od instance. Zadnja skupina značajki

generira se na temelju prozora od četiri riječi koje prethode instanci. One označavaju javlja li se u tom prozoru neka riječ koja može promijeniti polaritet instance. Skupovi riječi koje se traže uključuju one koje mijenjaju polaritet riječi (npr. *loš uzor*, *loš napad*), okreću ga na pozitivno (npr. *izbjeći štetu*) ili na negativno (npr. *nedostatak razumijevanja*).

Tablica 2.4 Usporedba učinkovitosti klasifikacije u jednom i dva koraka za različite klasifikator

	Točnost	Poz F	Neg F	Oboje F	Neut F
BoosTexter					
dva koraka	**74.5**	47.1	57.5	12.9	**83.4**
jedan korak	74.3	**49.1**	**59.8**	**14.1**	82.9
TiMBL					
dva koraka	**74.1**	47.6	56.4	13.8	**83.2**
jedan korak	73.9	**49.6**	**59.3**	**15.2**	82.6
Ripper					
dva koraka	68.9	26.6	49.0	00.0	**80.1**
jedan korak	**69.5**	**30.2**	**52.8**	**14.0**	79.4
SVM					
dva koraka	**73.1**	**46.6**	**58.0**	13.0	**82.1**
jedan korak	71.6	43.4	51.7	**17.0**	81.6

Tablica 2.4 prikazuje rezultate za četiri korištena algoritma. U eksperimentima je korišten Multi-perspective Question Answering korpus (Wiebe et al., 2005). Korpus sadrži označene fraze koje izražavaju privatna stanja (eng. *private states*) (Quirk et al., 1985). Privatna stanja su mentalna i emocionalna stanja, a uključuju vjerovanja, nagađanja, namjere i osjećaje. Kao podskup privatnih stanja dodatno su označene subjektivne fraze. Taj podskup se sastoji od indirektnih i direktnih izražavanja subjektivnog mišljenja (npr. upravni govor). Prvi redak svakog algoritma prikazuje rezultate dobivene opisanim postupkom. Drugi redak prikazuje rezultate dobivene u jednom koraku, koristeći sve značajke istovremeno. Zanimljivo je da rezultati koje taj pristup daje nisu značajno lošiji od klasifikacije u dva koraka.

2.3.2 Glasanje i kompozicijska semantika

Kao najjednostavnija heuristika u postupku određivanja semantičke orijentacije fraze može se koristiti glasanje (eng. *vote*) (Choi i Cardie, 2008). Subjektivnoj frazi dodjeljuje se ona orijentacija koju ima većina subjektivnih riječi od kojih se fraza sastoji. To znači da se prvo svim riječima koje se nalaze u frazi i koje su sadržane u leksikonu dodjeljuje *a priori* orijentacija. Nakon toga se broji koliko se u frazi javilo negativnih riječi, a koliko pozitivnih. Ako je rezultat glasanja izjednačen frazi se dodjeljuje ona orijentacija koja prevladava u ostalim uzorcima.

Naprednija varijanta glasanja uključuje u obzir i negacije (eng. *function-word negator*). Pretpostavlja se da negacija okreće polaritet fraze. Choi i Cardie (2008) ispituju ulogu subjektivnih riječi koje mogu imati isti utjecaj kao negacije (eng. *content-word negators*). Primjer takve riječi je *eliminirati*. Semantička orijentacija riječi *sumnja* je negativna, a riječi *prednost* pozitivna.

[[*eliminirati*]- [*sumnju*]-]+
[[*eliminirati*-] [*prednost*]+]-

Prema opisanom postupku glasanja obje fraze bi bile klasificirane kao negativne. Prva fraza bi bila negativna kao rezultat glasanja, a druga kao posljedica većeg broja negativnih fraza u uzorcima. Kako *eliminirati* ima ulogu negacije, orijentacija fraza se određuje okretanjem polariteta ostalih riječi. U frazama koje se sastoje od više riječi moguća je pojava dvostrukih negacija. Česta je i pojava negacija u kombinaciji sa riječima koje okreću polaritet:

[*nije* [*eliminiralo*]- [*sumnju*]-]+

U skladu sa iznesenim primjerima Choi i Cardie (2008) definiraju još četiri metode temeljene na glasanju. Metoda Neg(1) ista je kao i osnovna metoda glasanja, osim što na kraju okreće orijentaciju ukoliko fraza sadrži negaciju. Metoda Neg(*n*) je slična metodi Neg(1) osim što se orijentacija okreće *n* puta, gdje je *n* broj negacija koje su se javile u frazi. Ta metoda može prepoznati dvostruku negaciju.

Metode NegEx(1) i NegEx(*n*) definirane su slično kao i metode Neg(1) i Neg(*n*). Razlika je u tome što u obzir uzimaju i subjektivne riječi koje okreću orijentaciju. Metode prilikom glasanja te riječi ne uzimaju u obzir. Lista riječi je sakupljena ručno.

Osim prethodno navedenih heuristika Choi i Cardie (2008) eksperimentiraju i s heuristikama temeljenim na kompozicijskoj semantici (eng. *compositional semantics*). Slika 2.2 prikazuje skup ručno definiranih pravila za engleski jezik i pripadajućih motivacijskih primjera. Prije primjene pravila u frazi su prepoznate imenične (eng. *noun phrase*) i glagolske fraze (eng. *verb phrase*). Funkcija Compose prvo provjerava je li prvi argument negacija. Ako je, okreće orijentaciju drugog argumenta. U slučaju da je drugi argument negacija ne okreće se polaritet prvog, budući da on nije u dosegu negacije. Dvije varijante pravila CompoPR i CompoMC razlikuju se u pretpostavljenoj vrijednosti orijentacije. Varijanta CompoPR koristi vrijednost prvog argumenta, dok CompoMC koristi većinsku orijentaciju uzoraka.

	Rules		Examples
1	Polarity(not_[arg1]) =	¬ Polarity(arg1)	not [bad]$_{arg1}$.
2	Polarity([VP]_[NP]) =	Compose([VP], [NP])	[destroyed]$_{VP}$ [the terrorism]$_{NP}$.
3	Polarity([VP1]_to_[VP2]) =	Compose([VP1], [VP2])	[refused]$_{VP1}$ to [deceive]$_{VP2}$ the man.
4	Polarity([adj]_to_[VP]) =	Compose([adj], [VP])	[unlikely]$_{adj}$ to [destroy]$_{VP}$ the planet.
5	Polarity([NP1]_[IN]_[NP2]) =	Compose([NP1], [NP2])	[lack]$_{NP1}$ [of]$_{IN}$ [crime]$_{NP2}$ in rural areas.
6	Polarity([NP]_[VP]) =	Compose([VP], [NP])	[pollution]$_{NP}$ [has decreased]$_{VP}$.
7	Polarity([NP]_be_[adj]) =	Compose([adj], [NP])	[harm]$_{NP}$ is [minimal]$_{adj}$.

Definition of Compose(arg1, arg2)

For CompoMC: (**COMPO**sition with **M**ajority **C**lass)	Compose(arg1, arg2) = if (arg1 is a negator) then ¬ Polarity(arg2) else if (Polarity(arg1) == Polarity(arg2)) then Polarity(arg1) else the majority polarity of data
For CompoPR: (**COMPO**sition with **PR**iority)	Compose(arg1, arg2) = if (arg1 is a negator) then ¬ Polarity(arg2) else Polarity(arg1)

Slika 2.2 Kompozicijska pravila (Choi i Cardie, 2008)

Rezultati su prikazani u tablici 2.2. Evaluacija je provedena na Multi-Perspective Question Answering (MPQA) korpusu (Wiebe et al., 2005) koji se sastoji od 535 dokumenata ručno označenih s informacijama o subjektivnosti na razini fraza. Korišteni su izrazi koji imaju oznaku intenziteta *srednje* ili više. U obzir su uzimani samo izrazi koji imaju pozitivnu ili

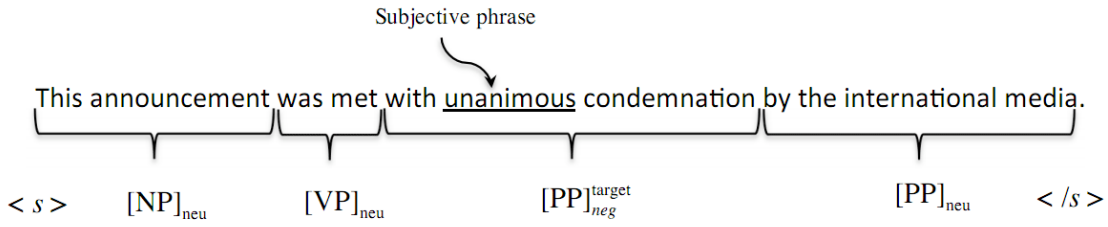
negativnu orijentaciju. Zanimljivo je primijetiti da je preciznost metoda koje u obzir uzimaju negacije znatno lošija od metode Vote koja negacije ignorira. Ukoliko se u obzir uzmu i subjektivne riječi koje okreću polaritet preciznost se povećava. Time je dokazan značajan utjecaj subjektivnih riječi u ulozi negacije. Pravila temeljena na kompozicijskoj semantici daju nešto bolje rezultate.

Tablica 2.2 Točnost metoda glasanja i kompozicijskih pravila (Choi i Cardie, 2008)

Vote	Neg(1)	Neg(n)	NegEx(1)	NegEx(n)	CompoMC	CompoMR
86.5	82.0	82.2	87.7	87.7	89.7	89.4

2.3.3 N-grami rečeničnih odsječaka

Agarwal et al. (2009) kao alternativu glasanju predlažu upotrebu jednostavnog deterministčkog konačnog automata u svrhu obrade negacija. Nakon što su riječima u frazi dodijeljene *a priori* ocjene iz proširenog leksikona DAL, fraza se stavlja na ulaz automata. Automat ima dva stanja: *zadrži* i *invertiraj*. U stanju *invertiraj* automat invertira dodijeljenu ocjenu trenutnoj riječi mijenjajući joj predznak. Iznos ocjene ostaje isti. U stanju *zadrži* automat ne mijenja ocjenu trenutne riječi. Inicijalno, automat se nalazi u stanju *zadrži*. Nailaskom na negaciju automat prelazi u stanje *invertiraj*. Povratak u stanje *zadrži* ostvaruje se nailaskom na suprotni veznik.



Slika 2.3 Oznake rečeničnih komada i pripadajuće oznake orijentacije (Agarwal et al., 2009)

Subjektivne fraze se opisuju skupom značajki i stavljaju na ulaz klasifikatora. Prvi skup značajki broji pojave pojedinih vrsta riječi (eng. *part of speech*). Za izvlačenje ostalih značajki rečenica u kojoj se nalazi fraza se dijeli na odsječke (eng. *chunking)*. Ako subjektivna fraza ne sadrži cijeli odsječak (eng. *chunk*) onda se granica proširuje tako da fraza uključuje

odsječak. Proširena subjektivna fraza naziva se ciljna fraza (eng. *target phrase*). Riječi u svakom odsječku rečenice dobivaju AE ocjenu. Ocjene riječi unutar komada se zbrajaju i normaliziraju brojem riječi. Na temelju pragova određuje se semantička orijentacija odsječka rečenice.

Slika 2.3 ilustrira stanje nakon provedenog postupka. Vitičaste zagrade obuhvaćaju odsječke rečenice. Subjektivna fraza je podcrtana. Ciljna fraza je odsječak u kojem se nalazi subjektivna fraza.

Iz svih rečenica izvučeni su unigrami, bigrami i trigrami komada. U primjeru na slici 2.3 jedan bigram se sastoji od $[VP]_{neu}$ kojeg slijedi komad $[PP]_{neg}^{target}$. Ako rečenica u kojoj se nalazi fraza sadrži neki n-gram odsječka, odgovarajuća binarna vrijednost u vektoru značajki postavljena je na 1. Na ovom skupu značajki provedena je automatska selekcija, mali podskup svih n-grama uključen je u konačni vektor značajki.

Informacije o kontekstu pohranjene su u skup značajki koji opisuje *a priori* orijentaciju komada koji se nalaze s lijeva ili desna ciljnoj frazi.

Tablica 2.3 Analiza utjecaja značajki na rezultate klasifikacije

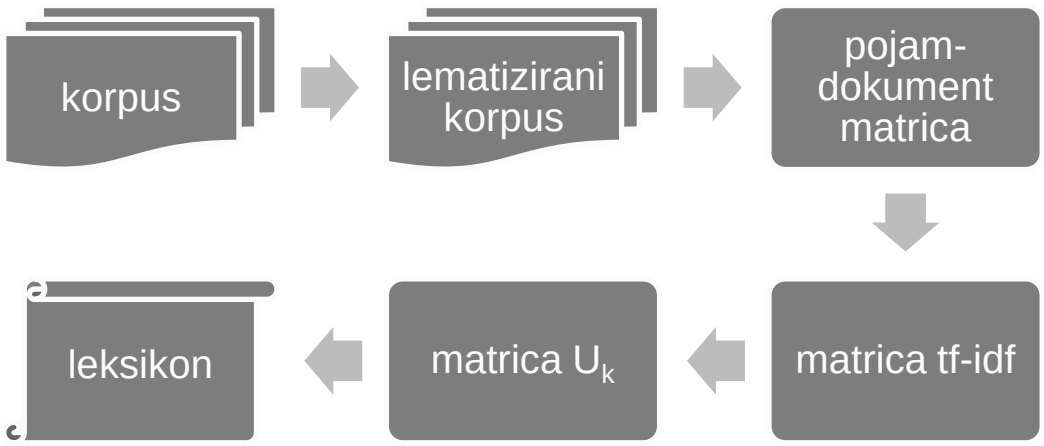
Značajke	Točnost	Poz F	Neg F
DAL ocjena	77.02	0.763	0.728
+ POS	79.02	0.788	0.792
+ odsječci	80.72	**0.807**	0.807
+ n-grami	**82.32**	0.802	**0.823**

Tablica 2.3 prikazuje rezultate. Za ispitivanje je korišten korpus MPQA. Uzorci se sastoje od 2779 instanci pozitivne i negativne klase. Zanimljivo je da je analiza utjecaja značajki pokazala da svaka od kategorija ima predstavnika među 10 najutjecajnijih značajki.

3 Automatski postupak izgradnje leksikona

U prethodnom poglavlju prikazani su neki ručno prikupljeni leksikoni koji se sastoje od skupa riječi i pripadajućih ocjena koje direktno opisuju semantičku orijentaciju ili daju dobru podlogu za računanje semantičke orijentacije. Ručna izgradnja takvog leksikona je zahtjevan posao. Kako je za izgradnju traženog sustava neophodno poznavati *a priori* orijentaciju riječi iz hrvatskog jezika, leksikon je potrebno izgraditi automatskim postupkom. Teorijska osnova i opis pojedinih koraka u postupku izgradnje leksikona dani su u odjeljku 3.1. Opis korištenih resursa i implementacijski detalji opisani su u 3.2.

3.1 Teorijska osnova



Slika 3.1 Koraci u postupku automatske izgradnje leksikona

Turney (2003) je pokazao da postoji veza između semantičke orijentacije i korelacije supojavljivanja između dvije riječi. Na temelju te povezanosti predložio je metodu za automatsko određivanje semantičke orijentacije riječi. Velika prednost metode je ta što ne zahtijeva nikakve resurse osim velike količine teksta, kojeg je lako prikupiti. Za izgradnju leksikona korištena je varijanta metode opisane u 2.2.1 primjenjiva na statičnom korpusu. Za korpus je izabrana skupina članaka iz Vjesnika. Raznolikost tema trebala bi osigurati reprezentativnost.

3.1.1 Normalizacija tf-idf

Prvi korak je predstavljanje korpusa pomoću matrice *X*. Redci matrice predstavljaju riječi, tj. leme koje su se pojavile u korpusu. Radi uklanjanja šuma i smanjivanja dimenzionalnosti matrice u obzir se uzimaju samo one leme koje su se javile u više različitih članaka. Stupci matrice predstavljaju članke. Element matrice *X*[*i, j*] označava koliko puta se lema *i* javila u članku *j*. Idući korak je normalizacija matrice. Normalizacija se radi zamjenom ne-nul elemenata matrice odgovarajućim tf-idf težinama (eng. *term frequency - inverse document frequency*). Vrijedi sljedeće:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

gdje je $n_{i,j}$ broj pojavljivanja leme *i* u članku *j*, a nazivnik je suma svih lema koje su se javile u članku *j*. Dalje vrijedi:

$$idf_i = \log \frac{|D|}{|\{d: t_i \in d\}|}$$

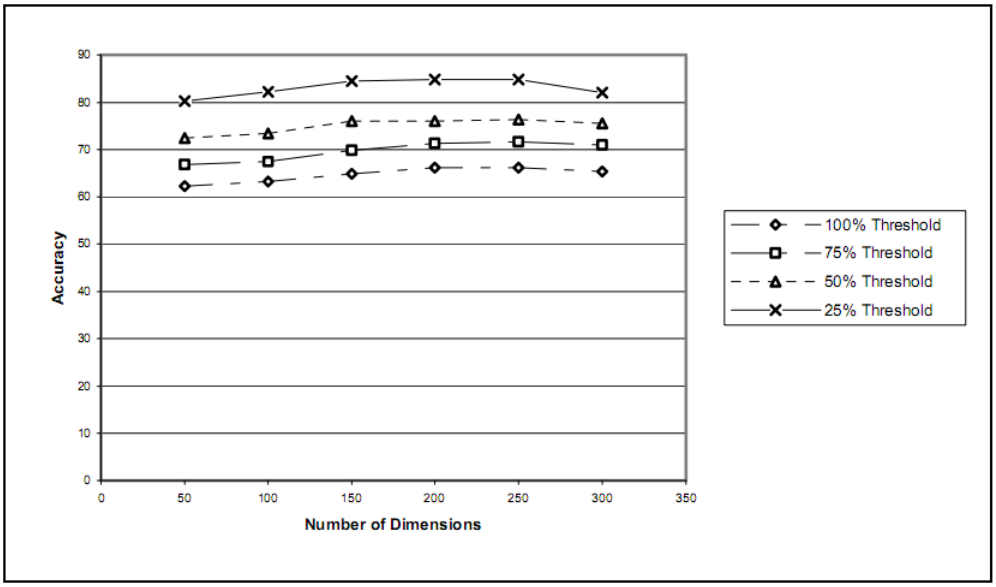
gdje je $|D|$ ukupni broj članaka u korpusu, a nazivnik je jednak broju članaka u kojima se javlja lema *i*. Konačna vrijednost tf-idf težine je produkt prethodne dvije formule:

$$tfidf_{i,j} = tf_{i,j} \times idf_i$$

3.1.2 Latent Semantic Analysis

Idući korak je računanje SVD dekompozicije (eng. *Singular Value Decompozition*) transformirane matrice *X*. SVD dekomponira matricu *X* na produkt tri matrice $U\Sigma V^T$. Matrica Σ je dijagonalna matrica dimenzija $m \times n$ koja sadrži jedinstvene vrijednosti u padajućem redoslijedu. U većini primjena dekompozicije nije potrebno raditi sa svim jedinstvenim vrijednostima, već se uzima samo *k* najvećih. U svojim istraživanjima Turney et al. (2003) su pokazali da je optimalna vrijednost *k* između 150 i 250, te da promjena unutar tog intervala ne utječe znatno na rezultate. Produkt smanjenih dimenzija označavamo sa $U_k\Sigma_k V_k^T$.

LSA metoda (eng. *Latent Semantic Analysis*) mjeri semantičku povezanost dvije riječi koristeći dobiveni rastav matrice $U\Sigma V^T$ (Landauer i Dumais, 1997). Povezanost dvije riječi, *LSA(word,word)*, odgovara kosinusnoj udaljenosti odgovarajućih redaka u krnjoj matrici U_k.



Slika 3.2 Utjecaj parametra *k* na točnost metode LSA (Turney, 2003)

3.1.3 Izgradnja leksikona

Uzmimo dva skupa riječi. Označimo ih sa *nWords* i *pWords*. Oba skupa sadrže jednak broj riječi kojima je lako odrediti semantičku orijentaciju neovisno o kontekstu. Pri tome *nWords* sadrži one riječi koje imaju negativnu semantičku orijentaciju, dok skup *pWords* sadrži one riječi koje imaju pozitivnu semantičku orijentaciju. U implementaciji sustava pretpostavljeni elementi skupa su:

nWords = {dobar, lijep, izvrstan, pozitivan, sretan, točan, pametan }
pWords = {loš, žalostan, siromašan, negativan, nesretan, pogrešan, tragičan}

Semantička orijentacija nove riječi *w* određuje se prema formuli:

$$SO(w) = \sum_{pWord \in pWords} LSA(w, pWord) - \sum_{nWord \in nWords} LSA(w, nWord)$$

pri čemu je *LSA(word,word)* mjera korelacije dobivena LSA metodom.

Dakle, odlučiti ćemo se za onu semantičku orijentaciju koju ima skup riječi kojem riječ *w* u većoj mjeri teži (Turney et al., 2003). Pri tome je bitno uočiti da rezultat *SO(word)* poprima vrijednosti u kontinuiranom intervalu [-1, 1]. Veća apsolutna vrijednost rezultata označava da je riječ u većoj mjeri pozitivna ili negativna.

3.2 Implementacija

3.2.1 Korpus

Za izradu leksikona korištena je arhiva Vjesnika od 31.05.1999. do 01.11.2009. Arhiva sadrži preko 250 000 članaka i pohranjena je u jednoj XML-datoteci. Osnovni elementi datoteke prikazani su u odsječku.[1]

```
<doc name="vjesnik-1999-5-31-tem-2">
    <content language="hr">
    <title>Ruski mirovni plan dijeli Kosovo na ...</title>
    <body>
      <subhead>Dok plan G-8 predviđa povlačenje...</subhead>
      <section>ZAGREB, 30. Svibnja – S pregovora...</section>
    </body>
    </content>
    <categories/>
    <extraInfo/>
</doc>
```

Element <doc> sadrži jedan članak i pripadajuće metapodatke. Sadržaj članka pohranjen je u elementu <content>, a sastoji se od naslova, podnaslova, i svih odsječaka. Oznaka kategorije nalazi se u elementu <category>. Podaci o datumu izdavanja članaka, autoru i lokaciji na Internetu nalaze se unutar elementa <extraInfo>. Za izradu leksikona potreban je samo sadržaj članka unutar <section> elemenata.

3.2.2 Lematizacija

Lematizacija je postupak određivanja osnovnog oblika, tj. leme neke riječi. Prije predstavljanja korpusa matricom potrebno je provesti lematizaciju kako bi se u daljnjem postupku svi oblici iste riječi tretirali jednako. Prilikom određivanja leme problem

[1] XMLSchema nalazi se na http://ktlab.fer.hr/download/documentSet.xsd

predstavljaju riječi koje imaju iste oblike. Na primjer, riječi *zidati* i *zid* imaju oblik *zida*. U prvom slučaju radi se o glagolu u prvom licu jednine, muškom rodu, prezent. Drugi slučaj je imenica, u genitivu jednine. Lematizacija se radi pomoću automatski prikupljenog flektivnog leksikona (Šnajder et al., 2008). Korišteni lematizator prilikom određivanja leme neke riječi ne uzima u obzir kontekst. To znači da će nailaskom na riječ *zida* vratiti sve leme koje imaju taj oblik. U trenutnoj implementaciji u obzir se uzima samo prva lema, a ostale se odbacuju. Ova funkcionalnost omogućena je preko javne metode *getFirstLemma*(). Ista komponenta korištena je i za određivanje vrste riječi. Kako je izvan konteksta nemoguće utvrditi točnu lemu oblika metoda *getMSDs*() vraća sve deskriptore koji odgovaraju obliku. U deskriptorima je, između ostalog, pohranjena i informacija o vrsti riječi. uzimanjem prvog deskriptora unosi se pogreška u slučaju da se radi o obliku neke druge leme. Bez informacije o kontekstu ove pogreške nije moguće izbjeći.

3.2.3 Rijetka matrica

Izgradnja matrice nad cijelim korpusom postupkom opisanim u 3.1 nije moguća. Dimenzije matrice bi bile nekoliko redova veličine veće od raspoložive radne memorije računala. Kako je matrica rijetka, velik dio memorije nepotrebno je utrošen na pohranu elemenata s vrijednosti nula. Klasa *SparseMatrix* ne-nul elemente pohranjuje u tri vektora. Vektori *columns* i *rows* sadrže indekse elemenata, a vektor *values* vrijednosti. Vrijedi $X(rows(k), columns(k)) = values(k)$.

Sučelje klase je svedeno na dvije glavne metode zadužene za dodavanje i dohvaćanje elemenata iz rijetke matrice. Metoda *set*() prima parametar *safe* koji omogućuje efikasnije inicijalno punjenje matrice. Prilikom punjenja sigurno je da je vrijednost elementa koji se dodaje prethodno bila jednaka nuli. To znači da nije potrebno izmijeniti postojeću vrijednost nego samo dodati novu. Time se izbjegava pretraživanje sadržaja matrice i smanjuje vremenska složenost. U svim slučajevima osim inicijalizacije potrebno je postaviti zastavicu *safe* kako bi se očuvao integritet rijetke matrice. Metoda *toFiles*() je praktična metoda za pohranu matrice u binarne datoteke. Parametar *matlabIndex* olakšava povezivanje s programskim alatom Matlab.

Prilikom normalizacije potrebno je proći kroz sve ne-nul elemente matrice. Zato klasa *SparseMatrix* implementira sučelje *Iterable*. U svakom koraku iteracije kroz matricu vraća se

objekt tipa *Elemet* koji sadrži indeks elementa i njegovu vrijednost. Uobičajene metode koje ostvaruju aritmetičke operacije nad matricom nisu implementirane budući da nisu potrebne prilikom normalizacije.

```
public class SparseMatrix implements Iterable<Element> {
	private Vector<Float> values;
	private Vector<Integer> rows;
	private Vector<Integer> columns;
	private int m, n;
	private int nz;

	public void set(int row, int column, float value, boolean safe)
	public float get(int row, int column)
	public double[][] toFull()
	public void toFiles(boolean matlabIndex)

	public Vector<Float> getValues()
	public Vector<Integer> getRows()
	public Vector<Integer> getColumns()
	...
}
```

3.2.4 Dekomponiranje matrice

SVD-dekompozicija matrice računa se u programskom alatu Matlab. Skripta koja se poziva iz Java koda učitava vektore iz datoteka, generira rijetku matricu te poziva funkciju *svds*() za izračun SVD-dekompozicije.

Dekomponiranje matrice se vrši na opisani način zbog toga što ne postoji slobodna Java biblioteka koja implementira učinkovitu metodu za izračun dekompozicije rijetke matrice.

Parametar k koji označava rang krnje matrice U_k, sukladno diskusiji u 3.2, fiksiran je na vrijednost 250.

Nakon izračuna dekompozicije matrica U_k se pohranjuje u tekstualnu datoteku. Ta je matrica puna i dimenzija $n \times k$, gdje je n broj riječi. Zbog relativno malih dimenzija pohranjivanje u tekstovnu datoteku ne stvara probleme, a olakšava učitavanje u objekt tipa *Matrix* iz biblioteke Jama,[2] korištenjem statičke metode *read*().

```
function [  ] = lsa( m, n, folder )
      disp('Loading files...')

      rid = fopen([folder 'r.bin'], 'r');
      r = fread(rid, 'uint32', 'ieee-be');
      cid = fopen([folder 'c.bin'], 'r');
      c = fread(cid, 'uint32', 'ieee-be');
      vid = fopen([folder 'v.bin'], 'r');
      v = fread(vid, 'float32', 'ieee-be');

      disp('Building sparse matrix...')

      sm = sparse(r, c, v, m, n);
      clear r c v m n;

      disp('Decomposing matrix...')

      [U, S, V] = svds(sm, 250);

      dlmwrite('D:\\_temp\\U.txt', U, ' ');
      exit
end
```

3.2.5 Računanje semantičke orijentacije i izgradnja leksikona

Klasa *CorpusMatrix* enkapsulira matricu U_k. Metoda *getCorelation*() prima dvije riječi te vraća kosinus kuta između odgovarajućih vektora. Taj iznos govori u kojoj mjeri dvije riječi koreliraju. Ukoliko jedna ili obje riječi ne postoje u leksikonu vrijednost se ne može izračunati te se riječi dodjeljuje ocjena nula.

[2] http://math.nist.gov/javanumerics/jama/

```
public class CorpusMatrix implements Serializable, Iterable<String>{
	private Matrix matrix;
	private Vector<String> words;

	public CorpusMatrix
	public CorpusMatrix(double[][] matrix, Vector<String> words)
	public CorpusMatrix(Matrix matrix, Vector<String> words)
	public double getCorelation(String l, String m)
	private double cosineDistance(Matrix a, Matrix b)
	public void printWordsToFile(String filepath) throws IOException
	public void printMatrixToFile(String filepath) throws IOException

	@Override
	public Iterator<String> iterator() {
		return words.iterator();
	}
}
```

Za izgradnju leksikona potrebno je proći kroz sve riječi čija je vektorska reprezentacija pohranjena u matrici i izračunati korelaciju sa svakom od referentnih riječi (eng. *seed words*). Riječ i ocjena dobivena izrazom iz 3.1.1 pohranjuje se u tekstualnu datoteku.

4 Semantička orijentacija fraza

Uz poznatu semantičku orijentaciju riječi potrebno je odrediti semantičku orijentaciju subjektivne fraze. Značajke koje opisuju subjektivnu frazu i korišteni klasifikator opisani su u odjeljku 4.1. Implementacija komponente za izvlačenje značajki prikazana je u 4.2.

4.1 Teorijska osnova

Glavni problem prilikom određivanja semantičke orijentacije fraze na hrvatskom jeziku je nedostatak resursa i alata. Osim leksikona generiranog postupkom opisanim u prethodnom poglavlju od jezičnih alata raspoloživ je samo lematizator opisan u odjeljku 3.2.2. Prilikom izvlačenja značajki iz subjektivne fraze lematizator je korišten i za određivanje vrste riječi. Fraze je potrebno klasificirati na temelju poznate orijentacije i vrste riječi od kojih se fraza sastoji.

Većina istraživanja u području analize subjektivnog mišljenja prilikom klasifikacije subjektivnih fraza uzima u obzir i kontekst. Subjektivna fraza koja je u kontekstu pozitivna izvan konteksta može imati negativnu orijentaciju. Primjer takvog slučaja je fraza *odlično snašla*. Izvan konteksta fraza je pozitivna, ali u rečenici:

Skupina kriminalaca se i u novonastaloj situaciji [odlično snašla].

fraza mijenja orijentaciju. Kontekst u kojem se fraza nalazi nosi mnogo informacija čijim se zanemarivanjem unosi greška u postupak klasifikacije. Kako ne postoje alati koji bi izvukli te informacije iz rečenica na hrvatskom jeziku prilikom razvoja sustava postavljeno je ograničenje na klasifikaciju fraza izvan konteksta.

4.1.1 Značajke

Prvi skup značajki predstavlja rezultate različitih metoda glasanja. Metode temeljene na glasanju su jednostavne heurističke metode koje daju dobre rezultate u jednostavnim frazama. Korištene su metode koje su opisali Choi i Cardie (2008). Rezultat pojedine metode pohranjen je u vektor značajki kao binarna vrijednost. Ako je metoda odlučila da je dana subjektivna fraza pozitivna značajka je postavljena na vrijednost 1, u suprotnome je postavljena na vrijednost 0.

Kao sofisticiranija alternativa glasanju korištena je i metoda koja koristi jednostavni deterministički konačni automat (Agarwal et al., 2009). Osim negacija prelazak iz stanja *zadrži* u stanje *invertiraj* ostvaruje se i nailaskom na bilo koju od riječi iz ručno sakupljene liste riječi. Lista sadrži one riječi koje u znatnom broju primjena mijenjaju polaritet riječi koje slijede (npr. *eliminirati*). Metoda koristi leksikon sa kontinuiranim ocjenama. Nakon promjene polariteta prolaskom kroz automat, ocjene riječi u frazi se zbrajaju. Ako je rezultat pozitivan vrijednost značajke se postavlja na vrijednost 1. U suprotnom značajka poprima vrijednost 0.

Slika 4.1 Vektor značajki sa izdvojenim skupinama

U vektoru značajki (slika 4.1) zadnja skupina su binarne značajke koje označavaju prisutnost unigrama, bigrama i trigrama vrste riječi u frazi. N-grami su izvučeni iz cijelog skupa označenih fraza. Svakoj riječi u frazi dodijeljena je oznaka koja označava vrstu riječi i *a priori* orijentaciju iz leksikona.

U prijateljskoj [A+] *i dobrosusjedskoj* [A+] *atmosferi* [N0]

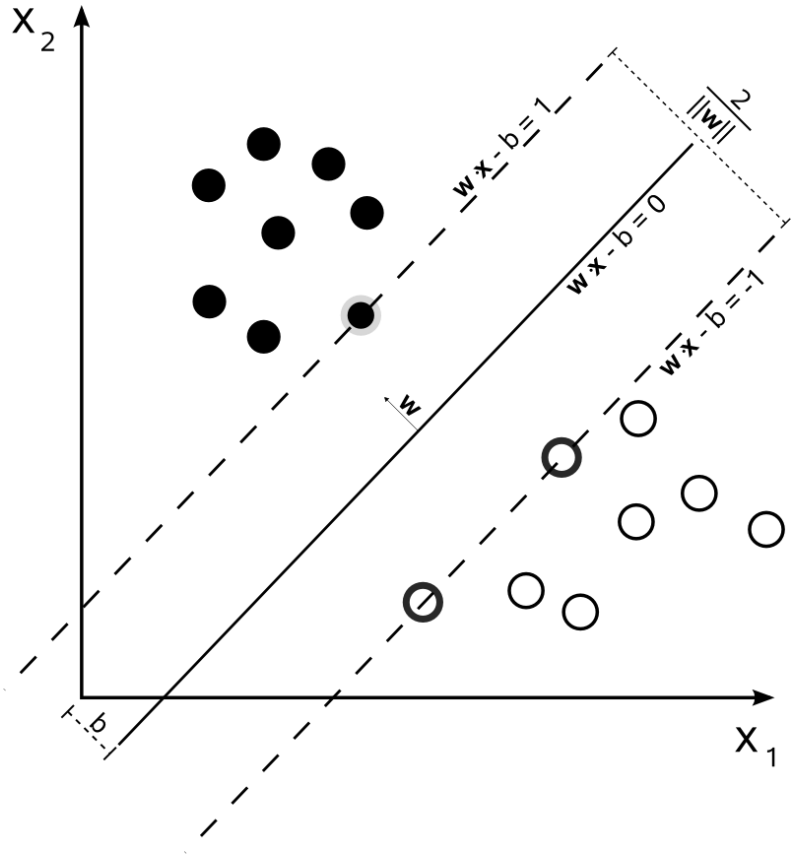
U primjeru oznaka [A+] označava pridjev koji ima pozitivnu orijentaciju, a oznaka [N0] označava neutralnu imenicu. Prilikom prikupljanja n-grama obje oznake bile bi dodane na listu budući da se radi o unigramima. U frazi se nalazi i bigram [A+N0]. Zbog veznika *i* koji se nalazi između pridjeva, to su svi n-grami koji su sadržani u frazi. Prilikom generiranja vektora značajki za frazu u primjeru odgovarajuće tri vrijednosti bi bile postavljene na vrijednost 1.

Da bi neka riječ u frazi dobila oznaku moraju biti zadovoljena dva uvjeta. Prvo, riječ mora biti imenica, pridjev ili glagol da bi joj se dodijelila oznaka N, A, ili V. Drugo, riječ se mora nalaziti u leksikonu da bi joj se dodijelila oznaka +, - ili 0, ako se radi o pozitivnoj,

negativnoj ili neutralnoj riječi. Riječima koje se ne nalaze u leksikonu ne dodjeljuje se nikakva pretpostavljena vrijednost, nego ih se zanemaruje.

Bigrami i trigrami vrsta riječi označavaju uzastopnu pojavu dvije ili tri označene riječi. Ako su riječi odvojene veznikom ili nekom riječi kojoj nije uspješno određena vrsta i orijentacija onda ne čine n-gram.

4.1.2 Stroj s potpornim vektorima



Slika 4.2 Uzorci dvije klase odvojeni decizijskom ravninom uz maksimalnu marginu razdvajanja

Stroj s potpornim vektorima (Vapnik, 1992) je linearni klasifikator. Uzorci koji pripadaju jednoj od dvije klase, predstavljeni su *p*-dimenzionalnim vektorom značajki. Osnovna ideja klasifikatora je je konstrukcija (*p*-1)-dimenzionalne decizijske hiperravnine, uz uvjet da je margina razdvajanja maksimalna (slika 4.2). Pretpostavlja se da su klase linearno odvojive.

Skup uzoraka za učenje *D* sastoji se od *n* vektora oblika

$$D = \{(\boldsymbol{x}_i, c_i) | \boldsymbol{x}_i \in \mathbb{R}^p, c_i \in \{-1, 1\}\}_{i=1}^{n}$$

gdje je uzorak $\boldsymbol{x}_i$ *p*-dimenzionalan vektor realnih vrijednosti, kojem je pridružena oznaka klase c_i. Tražimo hiperravninu koja odvaja uzorke s c_i = 1 od onih koji imaju c_i = -1. Hiperravninu opisuje jednadžba

$$\boldsymbol{w} \cdot x - b = 0$$

gdje je vektor $\boldsymbol{w}$ normala na hiperravninu. Cilj optimizacije je pronaći vrijednosti parametara $\boldsymbol{w}$ i *b* takve da su margine maksimalno udaljene, ali da i dalje razdvajaju uzorke. Margine su opisane jednadžbama

$$\boldsymbol{w} \cdot x - b = 1$$

$$\boldsymbol{w} \cdot x - b = -1$$

Udaljenost između margina je $2/\|\boldsymbol{w}\|$ pa je cilj minimizirati $\|\boldsymbol{w}\|$. Kako bi se spriječilo da se uzorci nađu unutar margina postavljamo uvjet

$$\boldsymbol{w} \cdot \boldsymbol{x}_i - b \geq 1$$

za uzorke iz prve klase i uvjet

$$\boldsymbol{w} \cdot \boldsymbol{x}_i - b \geq 1$$

za uzorke iz druge klase. Ovako postavljen problem je teško riješiti budući da ovisi o normi vektora $\boldsymbol{w}$ koja uključuje korijen. Kvadrat funkcije ima optimum u istoj točki kao i početna funkcija pa problem određivanja optimalne decizijske ravnine možemo definirati kao traženje minimuma kriterijske funkcije

$$J(\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}^{\boldsymbol{T}} \cdot \boldsymbol{w}$$

uz zadovoljavanje prethodnih graničenja.

Pomoću Langrangeovih multiplikatora λ_i postavljamo dualan problem. Ako prvotni problem ima optimum, onda ga ima i dualni problem i odgovarajuća optimalna rješenja su jednaka. Kriterijska funkcija poprima oblik

$$J(\boldsymbol{w}, b, \lambda) = \frac{1}{2}\boldsymbol{w}\boldsymbol{w}^T - \sum_{i=1}^{n} \lambda_i [d_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) - 1]$$

Vrijedi

$$\boldsymbol{w} = \sum_{i=1}^{n} \lambda_i d_i \boldsymbol{x}_i$$

Tražimo maksimum uz uvjete

$$\sum_{i=1}^{n} \lambda_i d_i = 0$$

$$\lambda_i \geq 0, i = 1,2, \dots, n$$

Nakon sređivanja dobivamo konačni oblik dualnog problema uz neizmijenjene uvjete.

$$\max\left(\sum_{i=1}^{n} \lambda_i - \frac{1}{2}\sum_{i,j=1}^{n} \lambda_i \lambda_j d_i d_j \boldsymbol{x}_i^T \boldsymbol{x}_j\right)$$

Optimalna hiperravnina ovisi samo o uzorcima koji leže na margini. Te uzorke nazivamo potpornim vektorima (eng. *support vectors*). Tražena hiperravnina se dobiva sa

$$\boldsymbol{w} = \sum_{i=1}^{n} \lambda_i d_i \boldsymbol{x}_i$$

gdje su λ_i optimalni Lagrangeovi multiplikatori.

4.2 Implementacija

4.2.1 Stvaranje vektora značajki

Prilikom izvlačenja značajki glavnu odgovornost ima klasa Features. Metoda *lemmatizePatterns*() kao parametar prima skup subjektivnih fraza i pripadajućih oznaka klasa. Nakon provedene lematizacije metoda vraća mapu u kojoj je svakoj lematiziranoj frazi pridružena oznaka klase. Ostale metode rade nad lematiziranim frazama.

Metoda *getVoteFeatures*() vraća listu značajki dobivenih na temelju rezultata glasanja. Lista sadrži pet vrijednosti, dobivenih različitim postupcima glasanja. Metoda prolazi kroz lematiziranu frazu jednom te broji pojave pozitivnih i negativnih riječi. Svaki postupak glasanja te pojave tretira na drugačiji način. Metoda *getFSAFeatures*() vraća listu u kojoj se nalazi jedna značajka. Povratna vrijednost je lista kako bi se omogućilo jednako postupanje sa svim metodama za izvlačenje značajki. Obje metode koriste liste riječi *negations* i *negators*. Lista *negations* sadrži negacije, a lista *negators* riječi koje imaju ulogu negacije. Prilikom glasanja postupcima Neg(1) i Neg(*n*) riječi na listi *negations* se ne broje kao riječi negativne orijentacije, već samo utječu na rezultat glasanja. U postupcima NegEx(1) i NegEx(*n*) riječi iz liste *negations* i *negators* se ne broje kao negativne, nego samo utječu na rezultat glasanja. Sadržaj listi prikazan je u nastavku.

negations = {ne, nije, ni, nema, niti, neće}

negators = {eliminirati, uništiti, unatoč, osuditi, nadoknaditi, bez, otklonjen, nekoć}

Metoda *getNgrams*() na ulaz prima lematiziranu frazu u kojoj pronalazi sve unigrame, bigrame i trigrame. Kako niti u jednom trenutku nije bitno koliko se puta neki n-gram javio u frazi, nego samo je li se javio, metoda vraća set pronađenih n-grama.

Metoda *getNgramFeatures*() stvara listu značajki za danu frazu. U prvom koraku se preko metode *getNgrams*() dohvaća set n-grama koji se nalaze u frazi. U drugom koraku se prolazi kroz listu odabranih n-grama te se ispituje je li se n-gram javio u frazi ili nije. U skladu s tim postavlja se vrijednost odgovarajuće binarne značajke. Odabrani n-grami su oni n-grami koji se nalaze u listi koju vraća metoda *selectFeatures*().

```
public class Features {
	private static final List<String> negations;
	private static final List<String> negators;

	private HashMap<String, Float> dalc;
	private HashMap<String, String> dald;
	private Lemmatizer l;

	...
	public HashMap<String, Integer> lemmatizePatterns(...)
	public List<String> selectFeatures()
	public Set<String> getNgrams(String lemmatizedPhrase)

	public List<Float> getVoteFeatures(String lemmatizedPhrase)
	public List<Float> getFSAFeatures(String lemmatizedPhrase)
	public List<Float> getNgramFeatures(String lemmatizedPhrase, ...)
	...
}
```

Metoda *selectFeatures*() radi selekciju n-grama koji se koriste u izgradnji vektora značajki. Kombinacijom tri orijentacije riječi (pozitivna, negativna, neutralna) i tri vrste riječi (pridjev, imenica, glagol) moguće je dobiti 9 različitih unigrama. To daje 81 mogućnost za bigrame i 729 mogućnosti za trigrame. Ako bi vektor značajki sadržavao vrijednosti koje bi indicirale pojavu svakog od 819 n-grama postupak klasifikacije ne bi mogao dati dobre rezultate. Zato je potrebno odrediti koji n-grami svojom pojavom ili odsustvom daju najveću količinu informacije o orijentaciji subjektivne fraze.

Tablica 4.1 Matrica kombinacija za izračun χ^2 mjere

	pozitivno	negativno
n-gram	a	b
!n-gram	c	d

U prvom koraku metoda *selectFeatures*() prikuplja sve n-grame koji se nalaze u skupu označenih fraza. Prolaskom kroz skup za svaku frazu se poziva metoda *getNgrams*().

Rezultati se pohranjuju u set *allNgrams*. U idućem koraku se za svaki n-gram iz seta *allNgrams* računa χ^2 mjera. Ispunjava se tablica dimenzija 2×2 prikazana tablicom 4.1.

Element *a* broji pozitivno orijentirane uzorke u kojima se javio n-gram, a element *c* u kojima se n-gram nije javio. Elementi *b* i *d* broje negativne uzorke u kojima se n-gram javio, tj. nije javio. Mjera χ^2 za n-gram dobiva se uvrštavanjem vrijednosti u formulu:

$$\chi^2 = \frac{(ad - bc)^2(a + b + c + d)}{((a + b) * (c + d) * (a + c) * (b + d))}$$

Metoda kao rezultat vraća one n-grame koji imaju vrijednost χ^2 veću od neke fiksne granice. Tim je postupkom značajno smanjen broj značajki odbacivanjem onih n-grama koji ne nose informaciju korisnu za postupak klasifikacije.

4.2.2 SVM^{light}

Za određivanje semantičke orijentacije vektora značajki koristi se implementacija stroja sa potpornim vektorima SVM^{light} (Joachims, 1999). Program je implementacija algoritma kojeg je opisao Vapnik (1992) u programskom jeziku C uz brojne optimizacije koje omogućavaju rad sa velikim brojem uzoraka. Implementacija se sastoji od dva modula: *svm_learn* i *svm_classify*.

Da bi se uzorci mogli predočiti modulu za učenje potrebno ih je pohraniti u datoteku odgovarajućeg formata. Klasa *PreparePatterns* koristi sučelje klase *Features* kako bi svaki uzorak predstavila vektorom značajki. Prilikom pripreme uzoraka moguće je odabrati koje skupine značajki će biti pohranjene u vektor. Format tekstualne datoteke koji koristi SVM^{light} prikazan je u odsječku.

```
<line> .=. <target> <feature>:<value> <feature>:<value> ... # <info>
<target> .=. +1 | -1 | 0 | <float>
<feature> .=. <integer> | "qid"
<value> .=. <float>
<info> .=. <string>
```

Svaka linija u datoteci predstavlja jedan uzorak. Element <target> označava klasu kojoj uzorak pripada. Vrijednost 1 označava pozitivnu frazu, dok -1 označava negativnu frazu.

Nakon oznake klase slijedi niz značajki. Svaka značajka prikazana je parom <feature>:<value> koji označava redni broj značajke i pripadajuću vrijednost. Značajke koje imaju vrijednost nula moguće je izostaviti (rijetki prikaz vektora). Ta mogućnost se ne koristi s obzirom na mali broj značajki. Vrijednost značajke je kontinuirana vrijednost. Kako su u razvijenom sustavu sve značajke binarne nije potrebno vršiti normalizaciju.

5 Uzorci za evaluaciju

Postupci izgradnje leksikona i klasifikacije subjektivnih fraza su u potpunosti automatizirani u smislu da ne zahtijevaju nikakve ručno označene resurse. Za evaluaciju je bilo potrebno prikupiti određen broj označenih subjektivnih riječi i fraza. Kako se radi o novom području ne postoje odgovarajući skupovi na hrvatskom jeziku. Svi uzorci su izvađeni iz Vjesnikovog korpusa i ručno označeni i klasificirani. U odjeljku 5.1 opisan je postupak klasifikacije subjektivnih riječi, a u odjeljku 5.2 postupak označavanja subjektivnih fraza.

5.1 Označavanje subjektivnih riječi

Za potrebe podešavanja parametara i evaluacije izgrađenog leksikona označeno je ukupno 1500 nasumično odabranih riječi. Svakoj imenici, pridjevu ili glagolu moguće je s obzirom na semantičku orijentaciju dodijeliti jednu od četiri oznake: *pozitivno* (+), *negativno* (-), *oboje* (+-) ili *neutralno* (0). U kategoriju pozitivno spadaju one riječi koje sudcu predstavljaju nešto pozitivno. Većina ljudi bi u tu kategoriju stavila riječi kao što su: *sreća*, *ljeto*, *sunce*, *napredak* i sl. U kategoriju negativno spadaju primjerice: *siromaštvo*, *smrt*, *glad* itd. Oko navedenih riječi ne postoji neslaganje, one su u svakom (i izvan njega) kontekstu pozitivne ili negativne. Postoji međutim mnogo riječi koje ovisno o kontekstu ili sudcu mijenjaju semantičku orijentaciju. Isto tako postoje riječi koje istovremeno predstavljaju i nešto pozitivno i nešto negativno. Primjerice, riječ *ludnica* u kontekstu medicinske ustanove ima negativnu orijentaciju. U žargonu je istu riječ moguće upotrijebiti za izražavanje pozitivnog stava o nekom događaju. Osim u slučaju višeznačnosti u kategoriju *oboje* moguće je svrstati i riječ *osuditi*. U frazi *pljačkaš je osuđen na zatvorsku kaznu* semantička orijentacija je pozitivna iz perspektive društva, ali negativna iz perspektive pljačkaša. Riječ *kulminacija* u filmskoj kritici ima pozitivno značenje, ali u povijesti bolesti negativno. Ako je sudac u trenu kada je pročitao riječ pomislio samo na jednu moguću orijentaciju, označio je riječ kao pozitivnu ili negativnu. Kako bi zadržali subjektivnost sudci su upućeni da označavanje rade u nekoliko etapa.

Svaki od 5 nezavisnih sudaca dobio je skup od 500 riječi. Skupovi sadrže 1250 disjunktnih i 250 zajedničkih riječi, na temelju kojih je izračunato slaganje sudaca.

Koeficijent κ je statistička mjera koja izražava slaganje između dva sudca u slučaju svrstavanja uzoraka u dvije ili više kategorija. U odnosu na jednostavno računanje postotka slaganja κ daje preciznije rezultate jer u obzir uzima i slaganje dobiveno slučajno. Koeficijent je definiran izrazom (Cohen, 1960):

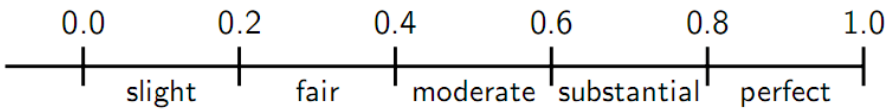
$$\kappa = \frac{A_o - A_e}{1 - A_e}$$

pri čemu je A_o opaženo slaganje, a A_e očekivano slaganje. Za izračun vrijednosti A_o i A_e gradi se matrica dimenzija $C \times C$, gdje je C broj kategorija. Element matrice u retku i i stupcu j predstavlja broj slučaja u kojima je prvi sudac uzorak svrstao u kategoriju i, a drugi sudac u kategoriju j. U slučaju potpunog slaganja matrica ima ne-nul elemente samo na dijagonali. Izrazi za vrijednosti A_o i A_e su:

$$A_o = \sum_{i=1}^{C} p_{ii}$$

$$A_e = \sum_{i=1}^{C} p_{.i} p_{i.}$$

gdje p_{ij} označava vrijednost elementa $[i, j]$ nakon normalizacije matrice dijeljenjem svih elemenata brojem uzoraka. Indeks $[., i]$ označava i-ti element horizontalne margine matrice.



Slika 5.1 Skala koeficijenta κ sa opisnim ocjenama (Landis i Koch, 1977)

Vrijednost koeficijenta nalazi se u intervalu od 0 do 1. Vrijednost 0 označava potpuno neslaganje između dva sudca dok vrijednost 1 označava potpuno slaganje. Potpuno slaganje znači da su za sve uzorke oba sudca dodijelila istu oznaku. Slika 5.1 prikazuje intervale i pripadajuće opisne ocjene (Landis i Koch, 1977). Prikazana podjela je jedna od

mnogih, budući da oko tumačenja vrijednosti koeficijenta postoji mnogo neslaganja. Priroda uzoraka u velikoj mjeri utječe na način na koji će se koeficijent tumačiti.

Tablica 5.1 Težinske mjere neslaganja

	neutralno	pozitivno	negativno	oboje
neutralno	1	1.25	1.25	1.25
pozitivno	1.25	1	1.5	1
negativno	1.25	1.5	1	1
oboje	1.25	1	1	1

Pretpostavimo da su istu riječ tri sudca označila različito. Prvi sudac stavio je oznaku +, drugi -, a treći +-. Jasno je da se niti koja dva od njih ne slažu. Međutim prvi i drugi sudac imaju potpuno suprotno mišljenje o danoj riječi, dok je treći sudac sklon reći da riječ ima i pozitivno i negativno značenje. To znači da se prvi i drugi sudac u većoj mjeri ne slažu oko oznake. Kako neslaganja u primjeru predstavljaju krajnje slučajeve potrebno je neslaganja podijeliti u tri kategorije. Težine dodijeljene svakom mogućem obliku neslaganja prikazane su u tablici 5.1. Prilikom korištenja težina u koeficijent κ se računa na temelju neslaganja. Tada izrazi postaju (Cohen, 1968):

$$\kappa = 1 - \frac{D_o}{D_e}$$

$$D_o = \sum_{i \neq j}^{C} w_{ij} p_{ii}$$

$$D_e = \sum_{i \neq j}^{C} w_{ij} p_{i.} p_{.j}$$

gdje D_o i D_e označavaju opaženo i očekivano neslaganje. Rezultati koeficijenata za svaka dva sudca izračunati gore navedenim izrazima i uzimajući u obzir težine izražene u tablici 5.1 prikazani su u tablici 5.2.

Riječi izvan konteksta mogu poprimiti bilo koju od tri subjektivne kategorije. Svaki sudac je dao svoje subjektivno mišljenje ovisno o tome koje značenje pojedina riječ ima za njega. Kako se radi o potpuno subjektivnom označavanju dobivena vrijednost od 0.365, uz standardnu devijaciju 0.55 nije loš rezultat.

Tablica 5.2 Koeficijent κ za sve parove sudaca

	sudac 1	sudac 2	sudac 3	sudac 4	sudac 5
sudac 1	1	0.396	0.397	0.459	0.399
sudac 2	0.396	1	0.260	0.379	0.288
sudac 3	0.397	0.260	1	0.378	0.336
sudac 4	0.459	0.379	0.378	1	0.354
sudac 5	0.399	0.288	0.336	0.354	1

5.2 Označavanje subjektivnih fraza

Uzorci za evaluaciju komponente za klasifikaciju subjektivnih fraza prikupljeni su iz Vjesnikova korpusa. U nasumično odabranim člancima označene su subjektivne fraze koje imaju pozitivnu ili negativnu semantičku orijentaciju, neovisno o tome je li slaganje oko polariteta univerzalno.

rješenje kosovske krize
originalan zaplet filma

Obje subjektivne fraze su pozitivne. Razlika je u tome što će se svi složiti da je prva fraza pozitivna neovisno o kontekstu, vremenu i bilo čemu drugome. Ta fraza ne izražava nečije mišljenje ili stav nego činjenicu koja je u ovom slučaju pozitivna. Druga fraza izražava subjektivno mišljenje nekog kritičara o nekom filmu. Različiti kritičari ne moraju dijeliti isto mišljenje. Obje fraze zadovoljavaju definiciju subjektivne fraze. Za evaluaciju sustava koji određuje semantičku orijentaciju fraze izvan njenog konteksta ove razlike nisu bitne.

Određivanje granice subjektivne fraze unutar rečenice je poseban problem na kojem se također intenzivno radi. Razvijeni sustav pretpostavlja strogo definirane granice fraze. Uzorci su morali biti označeni u skladu s tom pretpostavkom. Ispitivanje slaganja sudaca

oko granice izlazi iz opsega rada. S ciljem prikupljanja čim većeg broja uzoraka nezavisni sudci su označavali disjunktne skupove članaka.

U napadu na Vranje jedna [osoba je poginula], a jedna je [teško ranjena]...

Primjer pokazuje odsječak rečenice u kojem su označene dvije subjektivne fraze. Kako sustav klasificira fraze izvan konteksta bitno je da budu označene na način da je značenje i smisao fraze sačuvan. U drugoj frazi se objekt (osoba) ne spominje eksplicitno, ali fraza neovisno o tome ima polaritet (negativan) koji je u ovom slučaju neovisan o kontekstu i objektu.

...[omogućiti povratak kosovskih izbjeglica]...
...[omogućiti povratak izbjeglica] sa Kosova.

Gornji primjeri ilustriraju određivanje granica uz očuvanje značenja. Ukoliko bi u prvom primjeru bio označen manji podskup dobivena fraza *omogućiti povratak* izvan konteksta bi imala potpuno drugačije značenje. Kako i ta fraza ima polaritet to ne bi bilo pogrešno. S ciljem prikupljanja što raznovrsnijih uzoraka labavo pravilo za označavanje granica bi glasilo: granica obuhvaća minimalan broj riječi koji nosi maksimalno specifično značenje. U drugom primjeru granica ne uključuje cijeli odsječak budući da leksikon ne sadrži vlastita imena. Proširivanjem granice do kraja rečenice ne bi se uključila nikakva dodatna informacija.

spreman prihvatiti načela
reagirali su skeptično
ne vide znakove
želi povući snage
zaustavljanje bombardiranja

Ukupno je prikupljeno 426 subjektivnih fraza. Polaritet je označen nakon ekstrakcije fraza iz članka. Ukupno 229 fraza je označeno kao negativno, a 197 fraza kao pozitivno. Prilikom označavanja polariteta izvan konteksta ne postoji mogućnost neslaganja. Označene su samo subjektivne fraze čiji je polaritet, za razliku od polariteta jedne riječi, jednoznačno određen. Iz tog razloga označavanje je provela jedna osoba.

6 Evaluacija

Zasebna evaluacija leksikona i značajki kojima su opisane subjektivne fraze provedena je nad odgovarajućim skupovima uzoraka opisanim u prethodnom poglavlju. Rezultati evaluacije leksikona prikazani su u odjeljku 6.1, a rezultati analize utjecaja značajki na postupak klasifikacije u odjeljku 6.2.

6.1 Performanse izgrađenog leksikona

Svi testovi provedeni su na uzorcima opisanim u prethodnom poglavlju. Iz skupa uzoraka izbačene su riječi kojima su sudci dodijelili oznaku *oboje* (+-). Te riječi izvan konteksta mogu poprimiti pozitivno ili negativno značenje i ne mogu biti korištene za evaluaciju postupka koji riječima u leksikonu dodjeljuje jednu ocjenu u intervalu [-1, 1]. Time je odbačeno 166 riječi čime je dobiven skup riječi za testiranje koji se sastoji od 1334 riječi svrstanih u tri kategorije (*pozitivno*, *negativno*, *neutralno*).

Učinkovitost postupka izražena je mjerama točnosti, preciznosti, odziva i F-mjerom. Točnost je omjer ispravno klasificiranih riječi i ukupnog broja riječi u skupu za testiranje. Ostale mjere izračunate su za svaku klasu *C* zasebno. Preciznost je postotak uzoraka klasificiran kao klasa *C* koji zaista i pripadaju klasi *C*.

$$prec(C) = \frac{|ispravno\ klasificirani\ uzorci\ klase\ C|}{|uzorci\ klasificirani\ kao\ C|}$$

Odziv je postotak ispravno prepoznatih uzoraka klase C.

$$odz(C) = \frac{|ispravno\ klasificirani\ uzorci\ klase\ C|}{|ukupan\ broj\ uzoraka\ klase\ C|}$$

F-mjera je harmonična srednja vrijednost odziva i preciznosti.

$$F(C) = \frac{2 \times odz(C) \times prec(C)}{odz(C) + prec(C)}$$

Kako je ocjena dodijeljena riječima u leksikonu kontinuirana vrijednost klasifikacija riječi na tri klase radi se usporedbom sa pragovima *lo* i *hi*. Ako je ocjena riječi u intervalu [-1, *lo*]

riječ je klasificirana kao negativna. Neutralne riječi su one kojima je dodijeljena ocjena iz intervala [*lo*, *hi*]. Pozitivne riječi su one sa ocjenom iz intervala [*hi*, 1].

U nastavku odjeljka prikazani su rezultati testova provedenih nad tri različita leksikona. Moguće je da se neka riječ iz skupa za testiranje ne javlja u nekom od leksikona. Test za svaki leksikon je proveden na temelju onih riječi iz skupa za testiranje koje se javljaju u leksikonu. Donja granica je dobivena nasumičnim dodjeljivanjem kategorija riječima iz tog skupa.

Tablica 6.1 prikazuje rezultate testa za leksikon izgrađen na temelju 80 000 dokumenata uz upotrebu referentnih riječi navedenih u odjeljku 3.1.3. Prilikom izgradnje leksikona u obzir su uzete riječi koje se javljaju u barem 150 dokumenata, ali ne u više od 60 000. Leksikon sadrži 965 riječi iz skupa za testiranje.

Tablica 6.1 Rezultati ispitivanja za leksikon

			pozitivno			**negativno**			**neutralno**		
lo	hi	toč.	prec.	odz.	f	prec.	odz.	f	prec.	odz.	f
donja granica		32.5	34.7	32.2	33.4	20.0	35.5	24.8	43.3	32.8	37.3
-0.1	0.1	44.5	45.4	44.9	45.1	35.8	64.5	46.0	55.3	34.4	42.5
-0.05	0.1	41.9	45.4	44.9	45.1	32.7	**69.5**	44.5	55.8	26.1	35.6
-0.1	0.05	43.0	43.6	**52.5**	**47.6**	35.8	64.5	46.0	**55.9**	24.9	34.5
-0.125	0.125	46.1	46.9	41.9	44.3	37.5	60.0	46.2	53.7	42.8	47.6
-0.15	0.15	**46.4**	**47.3**	39.3	42.9	**38.8**	58.1	**46.6**	51.9	**46.6**	**49.1**

U rezultatima se jasno vidi da pomicanje granice *hi* ne utječe na performanse u klasifikaciji negativnih riječi budući da je razlika u omjeru pozitivno i neutralno klasificiranih riječi. Vrijedi i obratno. Prilikom podešavanja parametara valja imati na umu da se radi o učinkovitosti nad skupom za testiranje. Pomicanje granica utječe na broj riječi koje su klasificirane kao neutralne. Prevelika apsolutna vrijednost granica *hi* i *lo* može rezultirati velikim brojem neutralnih riječi u leksikonu što možda i nije poželjno. Nad promatranim leksikonom uz odabir vrijednosti granica -0.15 i 0.15 od ukupno 8525 riječi njih 3327 je

klasificirano kao neutralno. Ovaj poprilično konzervativan leksikon postiže relativno visoke vrijednosti F-mjere za sve klase.

U tablici 6.2 prikazani su rezultati testa za leksikon izgrađen na temelju istog skupa referentnih riječi kao i u prethodnom slučaju, ali nad manjim skupom članaka. Upotrijebljeno je 50 000 dokumenata s granicama pojava riječi od 100 odnosno 38 000 dokumenata. Leksikon sadrži 925 riječi iz skupa za testiranje.

Tablica 6.2 Rezultati ispitivanja za leksikon izgrađen nad manjim brojem dokumenata

			pozitivno			**negativno**			**neutralno**		
lo	hi	toč.	prec.	odz.	f	prec.	odz.	f	prec.	odz.	f
donja granica		34.1	35.6	3.0	34.3	22.1	36.5	27.5	45.2	33.7	38.6
-0.1	0.1	45.5	47.5	50.5	48.9	35.9	61.5	45.3	55.8	33.8	42.1
-0.05	0.1	42.8	47.5	50.5	48.9	33.4	**66.7**	44.5	52.7	24.8	33.7
-0.1	0.05	45.6	46.5	**58.3**	**51.7**	35.9	61.5	45.3	**61.5**	27.5	38.0
-0.125	0.125	46.4	48.3	46.8	47.6	37.0	58.9	45.5	53.9	40.0	45.9
-0.15	0.15	**48.2**	**51.0**	45.6	48.2	**38.5**	56.8	**45.9**	53.8	**46.3**	**49.7**

Metoda LSA koja se koristi u izgradnji leksikona je statistička metoda koja daje bolje rezultate ako radi nad većim brojem dokumenata. U konkretnom slučaju leksikon izgrađen nad manjim brojem dokumenata u korištenje istog skupa referentnih riječi dao je nešto bolje rezultate uz iste iznose granica *hi* i *lo*. Kako se i dalje radi o velikom broju dokumenata moguće je da je razlog ovim neočekivanim rezultatima manja frekvencija pojave referentnih riječi u proširenom skupu dokumenata. Tema dokumenata koji su obuhvaćeni proširenim skupom također može utjecati na rezultate.

Osim utjecaja veličine skupa dokumenata nad kojima je izgrađen leksikon zanimljivo je ispitati i utjecaj odabira referentnih riječi. Početni skup je malo izmijenjen dodavanjem riječi za koje se pretpostavlja da u novinskim člancima jednoznačno nose polaritet. Tako su u skup *nWords* dodane riječi kao što su *kriza* i *tragičan*, dok su u skup *pWords* dodane riječi *napredak*, *bogat* itd. Izmijenjeni skupovi su:

nWords = {dobar, uspješan, bogat, izvrstan, pozitivan, sretan, točan, pametan, napredak }
pWords = {loš, žalostan, uništen, siromašan, negativan, nesretan, pogrešan, tragičan, kriza}

Za izgradnju leksikona, osim navedenih skupova referentnih riječi, korišteni su isti parametri kao i za izgradnju prvog leksikona. Rezultati testa prikazani su u tablici 6.3. Rezultati su dosta lošiji nego za originalan skup referentnih riječi. Iako se možda tako ne čini na prvi pogled, to je pozitivna stvar. Da se dodavanjem riječi koje su specifične za neki izvor teksta ili stil pisanja učinkovitost sustava povećala to bi značilo da je metoda pristrana. Rezultati pokazuju da metoda daje najbolje rezultate uz upotrebu riječi koje su univerzalno i neosporno pozitivne, tj. negativne.

Tablica 6.3 Rezultati ispitivanja za leksikon izgrađen nad velikim skupom dokumenata uz dodavanje specifičnih referentnih riječi

			pozitivno			**negativno**			**neutralno**		
lo	hi	toč.	prec.	odz.	f	prec.	odz.	f	prec.	odz.	f
donja granica		32.4	35.0	34.0	34.5	20.2	33.5	25.2	43.3	30.6	35.9
-0.1	0.1	**44.4**	**48.6**	54.5	51.4	35.3	68.5	46.6	55.3	24.7	34.2
-0.05	0.1	43.4	48.6	54.5	51.4	34.1	**72.9**	46.5	**57.4**	20.2	29.9
-0.1	0.05	43.5	47.0	**58.1**	**51.9**	35.3	68.5	**46.6**	55.3	19.7	29.1
-0.125	0.125	44.1	48.6	51.3	49.9	35.7	65.0	46.1	50.6	28.3	36.3
-0.15	0.15	44.2	48.0	46.9	47.5	**36.2**	62.6	45.8	49.8	**33.3**	**39.9**

6.2 Učinkovitost određivanja semantičke orijentacije fraze

U svim ispitivanjima korišten je skup od 426 ručno označenih subjektivnih fraza. Skup se sastoji od 229 fraze sa negativnom orijentacijom i 197 pozitivno orijentiranih fraza. Za određivanje *a priori* orijentacije riječi korišten je leksikon čije su performanse prikazane u tablici 6.1. Za izvlačenje značajki n-grama granice *hi* i *lo* su postavljene na vrijednosti 0.15 i -0.15.

U tablici 6.4 prikazani su rezultati klasifikacije dobiveni korištenjem pojedine metode zasebno. Vidimo da sve metode glasanja daju vrlo slične rezultate. Razlog tome je što se metode ne razlikuju značajno pa u većini slučajeva daju istu odluku.

Tablica 6.4 Analiza učinkovitosti pojedine metode glasanja

	točnost	precizno	odziv	f mjera
vote	**64.4**	**71.9**	39.1	50.5
Neg(1)	62.3	64.8	41.1	50.3
Neg(*n*)	62.3	64.8	41.1	50.3
NegEx(1)	62.5	64.4	43.1	51.7
NegEx(*n*)	62.3	64.9	43.1	51.8
DKA	63.4	62.8	**52.3**	**57.1**

Metode Neg(1) i Neg(n) daju identične rezultate jer niti jedna fraza u skupu za testiranje ne sadržava dvije negacije. Uključivanjem riječi koje imaju ulogu negacije rezultat se neznatno popravlja. To daje naslutiti da bi se rezultati metode glasanja mogli popraviti proširivanjem liste riječi koje mijenjaju orijentaciju fraze. Metoda koja koristi konačni automat ima znatno veći odziv od 52.3 u odnosu na najbolji odziv među metodama temeljenim na glasanju od 43.1. Nažalost, ova metoda ima i znatno manju preciznost. Niti jedna od metoda zasebno nije dovoljno dobra da bi donijela odluku o orijentaciji subjektivne fraze.

Tablica 6.5 Analiza utjecaja skupova značajki na postupak klasifikacije

	greška	preciznost	odziv	f mjera
glasanje	36.56	68.75	39.09	49.84
glasanje + DKA	36.56	68.75	39.09	49.84
n-grami	**33.49**	**75.23**	41.62	53.59
n-grami + DKA	33.96	65.50	56.85	**60.86**
n-grami + glasanje	34.67	43.15	**70.83**	53.62
sve	34.20	45.18	70.63	55.1

Tablica 6.5 prikazuje rezultate klasifikacije algoritmom SVM uz postupak *leave-one-out*. Zadnji redak u tablici dobiven je korištenjem cijelog skupa značajki. Zanimljivo je da su rezultati klasifikacije bolji ako se izostave značajke dobivene glasanjem. Rezultati pokazuju da su n-grami najbolja skupina značajki i da se dodavanjem odluke konačnog automata F-mjera značajno popravlja.

7 Zaključak

Određivanje semantičke orijentacije fraze je zahtjevan problem u području strojne analize subjektivnosti. Dodatno, potrebno je odrediti granice subjektivnih fraza, prepoznati značajku na koju se fraza odnosi kao i kome pripada izraženo mišljenje. Nije posve krivo reći da je unatoč tome određivanje semantičke orijentacije ključan problem. Kako se radi o subjektivnom tekstu rješenje nije uvijek jednoznačno što problem čini još težim. Sveukupno, strojna analiza subjektivnosti predstavlja značajan izazov u obradi prirodnog jezika.

Automatski postupak izgradnje leksikona subjektivnih riječi daje solidne rezultate. Kako se radi o temeljnom resursu o čijim performansama direktno ovise ostali koraci od velike je važnosti da su ocjene čim preciznije. Prednost opisanog postupka je svakako ta što je moguće izgraditi različite leksikone, ovisno o potrebama pojedinog zadatka, bez intervencije ljudskih sudaca. Ipak, s ciljem povećavanja preciznosti leksikona, a time i unošenja manje greške u cijeli sustav, bilo bi dobro imati na raspolaganju ručno izgrađen leksikon.

Opis subjektivne fraze vektorom značajki zahtjeva postojanje određenih jezičnih alata za ciljni jezik. N-grami vrsta riječi opisani u radu pokazuju obećavajuće rezultate. Korištenjem označivača vrste riječi koji u obzir uzima cijelu rečenicu izbjegla bi se pogreška uzrokovana istim oblicima različitih lema. Isto tako, postojanje alata za analizu sintakse rečenice omogućilo bi iskorištavanje konteksta u kojem se fraza nalazi. Intuitivno je jasno, a istraživanja su dokazala da kontekst ima veliku ulogu u određivanju semantičke orijentacije fraze. Značajke opisane u radu maksimalno iskorištavaju postojeće resurse, no određivanje orijentacije je složen problem čije uspješno rješavanje nije moguće bez kvalitetnih alata na nižim razinama. Izgradnja tih alata nije jednostavna, ali je neophodan korak u uspješnoj izgradnji sustava za analizu subjektivnosti.

8 Literatura

Agarwal, A., Biadsy, F. and Mckeown, K. R. (2009). Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams, Proceedings of the 12th Conference of the European Chapter of the ACL, pages 24-32, Athens, Greece.

Chio, Y. and Cardie, C. (2008). Learning with compositional semantics as structural inference for subsentenial sentiment analysis, Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 793-801, Honolulu, Hawaii.

Church, K.W., Hanks, P. (1989). Word association norms, mutual information and lexicography. Proceedings of the 27th Annual Conference of the Association of Computational Linguistics. Association for Computational Linguistics, pages 76-83, New Brunswick, NJ.

Cohen, J. (1960). A coefficient of agreement for nominal scales, Educational and Psychological Measurement, vol. 20, pages 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological Bulletin, vol. 70, pages 213-220.

Collins, M. (1997). Three generative, lexicalized models for statistical parsing, In proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97), pages 16-23, Madrid.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S. and Fellenz, W. (2001). Emotion recognition in human-computer interaction. In IEEE Signal Processing Magazine, vol. 1, pages 32-80.

Esuli, A. and Sebastiani, F. (2005). Determining the semantic orientation of terms through gloss analysis. In Proceedings of CIKM-05, the ACM SIGIR Conference on Information and Knowledge Management, Bremen, DE.

Esuli, A. and Sebastiani, F. (2006). SentiWordNet: A publicly available lexical resource for opinion mining, In Proceedings of the 5th Conference on Language Resources and Evaluation, pages 417-422

Fellbaum, C. (1998). WordNet: An electronic lexical database, MIT Press, Cambridge, MA.

Joachims, T. (1999). Making large-Scale SVM Learning Practical: Advances in Kernel Methods - Support Vector Learning, MIT Press, Cambridge, MA.

Landauer, T.K., and Dumais, S.T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review, vol. 104, pages 211-240.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data, Biometrics, vol. 33(1), pages 159-174.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis, Foundations and trends in information retrieval, vol. 2, pages 1-135.

Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of ACL-04, 42nd Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, pages 271–278, Barcelona, ES.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of EMNLP-02, the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pages 79–86, Philadelphia, US.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. (1985). A comprehensive grammar of the English language, Longman, New York.

Rillof, E. and Wiebe, J. (2003). Learning extraction patterns for subjective expressions, In proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2003), pages 105-112, Sapporo.

Stone, P. J., Dunphy, D. C., Smith, M. S. and Ogilvie, D. M. (1966). The General Inquirer: A computer approach to content analysis, MIT Press, Cambridge, MA.

Šnajder, J., Dalbelo-Bašić, B. and Tadić, M. (2008). Automatic Acquisition of Inflectional Lexica for Morphological Normalization, Information Processing and Management, vol. 44(5), pages 1720-1731.

Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, pages 417–424.

Turney, P. D. and Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems, vol. 21(4), pages 315–346.

Vapnik, V. N., Boser, B. E. and Guyon, I. M. (1992) A training algorithm for optimal margin classifiers, In proceedings of 5th annual workshop on Computational learning theory, pages 144-152, Pittsburg, Pennsylvania

Whisel, C. M. (1989). The dictionary of affect in language, In R. Plutchik and H. Kellerman, editors, Emotion: theory research and experience, vol. 4, Academic Press, London.

Wiebe, J., Wilson, T. and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. Language Resources and Evaluation (formerly Computers and the Humanities), vol. 39(2/3):164-210.

Wilson, T., Wiebe, J., and Hoffmann, P. (2009). Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis, Association for Computational Linguistics, MIT Press, Cambridge, MA.

DODATAK A

Matrice zabune prilikom označavanja subjektivnih riječi

A.1 Matrica zabune za sudce 1 i 2

	neutralno	pozitivno	negativno	oboje	
neutralno	78	0	0	5	83
pozitivno	47	28	0	11	84
negativno	20	1	31	13	65
oboje	11	3	1	3	18
	156	30	32	32	250

A.2 Matrica zabune za sudce 1 i 3

	neutralno	pozitivno	negativno	oboje	
neutralno	30	38	4	11	83
pozitivno	9	65	6	4	84
negativno	9	11	39	6	65
oboje	5	5	1	7	18
	53	119	50	28	250

A.3 Matrica zabune za sudce 1 i 4

	neutralno	pozitivno	negativno	oboje	
neutralno	59	13	1	10	83
pozitivno	19	56	1	8	84
negativno	19	4	28	14	65
oboje	7	5	0	6	18
	104	78	30	38	250

A.4 Matrica zabune za sudce 1 i 5

	neutralno	pozitivno	negativno	oboje	
neutralno	42	17	6	18	83
pozitivno	17	42	4	21	84
negativno	5	2	39	19	65
oboje	4	6	0	8	18
	68	67	49	66	250

A.5 Matrica zabune za sudce 2 i 3

	neutralno	pozitivno	negativno	oboje	
neutralno	44	87	13	18	156
pozitivno	2	36	1	1	30
negativno	1	0	28	3	32
oboje	6	12	8	6	32
	53	119	50	28	250

A.6 Matrica zabune za sudce 2 i 4

	neutralno	pozitivno	negativno	oboje	
neutralno	87	47	4	18	156
pozitivno	1	24	1	4	30
negativno	5	0	22	5	32
oboje	11	7	3	11	32
	104	78	30	38	250

A.7 Matrica zabune za sudce 2 i 5

	neutralno	pozitivno	negativno	oboje	
neutralno	60	41	15	40	156
pozitivno	2	19	2	7	30
negativno	1	1	24	6	32
oboje	5	6	8	13	32
	68	67	49	66	250

A.8 Matrica zabune za sudce 3 i 4

	neutralno	pozitivno	negativno	oboje	
neutralno	39	8	1	5	53
pozitivno	38	61	4	16	119
negativno	18	3	23	6	50
oboje	9	6	2	11	28
	104	78	30	38	250

A.9 Matrica zabune za sudce 3 i 5

	neutralno	pozitivno	negativno	oboje	
neutralno	28	8	4	13	53
pozitivno	24	53	10	32	119
negativno	9	2	29	10	50
oboje	7	4	6	11	28
	68	67	49	66	250

A.10 Matrica zabune za sudce 4 i 5

	neutralno	pozitivno	negativno	oboje	
neutralno	51	20	13	20	104
pozitivno	9	38	4	27	78
negativno	0	0	23	7	30
oboje	8	9	9	12	38
	68	67	49	66	250

DODATAK B

Primjeri ručno označenih riječi

zadovoljavati	+	samac	−	otimati	−
šarmantan	+	doktrina	−	gradski	0
prvoligaš	+	dvogodišnji	0	ovoj	0
načelnik	+	autobus	0	javljati	+−
nezasluženo	−	natjerati	−	nijansa	0
isporučiti	0	otkrivanje	+−	obazirati	0
hvaliti	+	hladan	−	koristan	+
pužnica	0	obujam	0	pripadnost	+
svod	0	pravnik	−	motiviran	+
odbiti	−	legalizirati	+	odazvati	+
kulminacija	+−	dosje	−	iskušati	+−
zvati	0	ukras	+	ponedjeljak	0
uplitanje	−	protestirati	+−	odvratiti	0
prikupiti	+−	relacija	0	glad	−
zapaziti	0	četvoran	0	doživljavati	+
kolektor	+−	nedopušten	−	dopisnica	0
potpunost	+	hvarski	0	manevar	0
medvjed	0	preusmjeriti	0	nadanje	+
poslušati	+	ovisnica	−	evidentiran	0
nagovještaj	+−	kršćanin	+	upasti	0
upropastiti	−	barokan	+	referendum	0
skretati	+−	destinacija	0	stenogram	0
dozvoljen	+	suza	−	željeznički	0
tripartitan	0	košarica	+−	umjeren	+
dirljiv	+	služba	0	preostao	0
gudački	+	prevladan	−	glas	0
rimski	+	produžetak	+	neprirodan	−
skupljanje	+	zadušnica	−	protektorat	0
ansambl	0	srednji	0	tužiteljski	−
paleta	0	suditi	0	uređenje	0
neiskustvo	−	dobitak	+	zabranjivati	+−
uplaćivati	−	razvidan	0	parnica	0
pretvoren	0	dobitan	+	držati	0
iscrpljivati	−	minijatura	0	igralište	+
lančan	−	osnova	+	nafta	0
četvorka	+	zavjera	−	poznati	0
uhvatiti	+	nazočiti	+	okupljalište	0
kršćanski	+	paničan	−	odoljeti	+−
trovanje	−	povezati	+	počekati	0
goveđi	0	vreća	0	domoći	+
sjednica	0	parafiran	0	hvala	+

DODATAK C

Primjeri označenih i klasificiranih fraza

nije eliminirao opasnost	−
procedura završena	+
nedostaje potpis	−
postigli cilj	+
čuo sam vrlo jasnu izjavu	+
potpunosti podržava	+
odradi svoj dio posla	+
bio vrlo važan	+
odlično snašla	+
sigurno da smo dobro iskoristili	+
pridonesemo rješavanju svjetskih kriza	+
mnogi su izrazili zadovoljstvo	+
pridonosi rješavanju gorućih tema u svijetu	+
nešto što će iznimno cijeniti	+
nije bilo prigovora	+
jesmo i da ćemo ostati prijateljske zemlje	+
nastojati zaštititi svoje nacionalne interese	+
prijateljski i u dobrosusjedskoj atmosferi	+
ostati pri blokadi	−
to neće biti dobro	−
kvara na elektroinstalacijama	−
kasnila je sjednica	−
diplomatsko osoblje je evakuirano	+
otklonjena opasnost od požara	+
prihvaćeni su i financijski planovi	+
sabor je prihvatio	+
tražimo poništavanje predugovora i ugovora	−
trajnoj zabrani probnog rada tvornice	−
smanjen je deficit	+
najviše su smanjeni rashodi	+
žestoka rasprava	−
reagirali su skeptično	−
ne vide znakove	−
želi povući snage	+
jedna je od najprljavijih industrija	−
neprimjerena lokaciji na kojoj se nalazi	−

SAŽETAK

Određivanje semantičke orijentacije subjektivnih riječi i fraza

Strojna analiza mišljenja izraženog u subjektivnom tekstu važno je i sve popularnije područje istraživanja s brojnim mogućnostima primjene. U većini slučajeva subjektivni tekst opisuje više značajki od kojih su neke pozitivne, a neke negativne. Analizu subjektivnosti zato je potrebno provesti na razini subjektivnih fraza. Kao glavni problem javlja se utvrđivanje semantičke orijentacije fraze, koja klasificira frazu kao negativnu ili pozitivnu. Automatski postupak izgradnje leksikona sa *a priori* orijentacijama riječi dao je dobre, ali ne posve zadovoljavajuće rezultate. Leksikon i označivač vrste riječi korišteni su za izvlačenje značajki koje opisuju subjektivnu frazu izvan konteksta rečenice. N-grami vrsta riječi uz oznake *a priori* orijentacije pokazali su se kao obećavajuća skupina značajki.

Ključne riječi:

analiza subjektivnosti, semantička orijentacija, subjektivne fraze, kompozicija sentimenta, obrada prirodnog jezika

Determining semantic orientation of subjective words and phrases

Sentiment analysis of subjective text is important and growing area of research with numerous applications. In most cases subjective text contains both positive and negative opinions on different features. In tasks such as this, therefore, it is important that subjectivity analysis be done at the phrase level. Main challenge is to determine whether semantic orientation of a subjective phrase is positive or negative. Automatically compiled *a priori* lexicon produced promising but not great results. Lexicon and POS-tagger were used to extract phrase features. Classification using part of speech n-grams with annotated *a priori* orientation performed significantly better over the baseline.

Keywords:

subjectivity analysis, semantic orientation, subjective phrases, sentiment composition, natural language processing