Laboratorij za analizu teksta i inženjerstvo znanja – TakeLab

Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva
Zavod za elektroniku, mikrolektroniku, računalne i inteligentne sustave
Unska 3, 10000 Zagreb, Hrvatska

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 456

STROJNA ANALIZA SENTIMENTA TEMELJENA NA APRIORNOJ POLARNOSTI RIJEČI

Ante Kegalj

Zagreb, rujan 2012.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA
ODBOR ZA DIPLOMSKI RAD PROFILA

Zagreb, 5. ožujka 2012.

DIPLOMSKI ZADATAK br. 456

Pristupnik: **Ante Kegalj**
Studij: Računarstvo
Profil: Računarska znanost

Zadatak: **Strojna analiza sentimenta temeljena na apriornoj polarnosti riječi**

Opis zadatka:

Porastom komunikacije putem Interneta povećao se interes za strojnom analizom mišljenja izraženog u korisnički generiranom tekstu. Jedan od pristupa analizi mišljenja jest analiza sentimenta, kojom se utvrđuje je li tekst pozitivno, negativno ili neutralno orijentiran. Analiza ukupnog sentimenta dokumenta može se temeljiti na analizi apriorne polarnosti pojedinačnih riječi.
U okviru diplomskog rada potrebno je proučiti postojeće postupke za određivanje sentimenta dokumenta te postupke za određivanje kontekstne polarnosti dijelova teksta temeljem apriorne polarnosti riječi. Razraditi postupak za određivanje polarnosti dokumenata i dijelova dokumenata na engleskome jeziku temeljem apriorne polarnosti riječi. Postupak se treba temeljiti na metodama nadziranoga strojnog učenja i na javno dostupnim leksikonima apriorne polarnosti rijeci, SentiWordNet i MPQA. Problem višeznačnosti treba pokušati riješiti razrješavanjem pomoću leksičke baze WordNet. Načiniti programsku izvedbu postupka i na odgovarajućem tekstnom uzorku provesti eksperimentalno vrednovanje na zadatcima određivanje osnovne polarnosti (klasifikacija) i određivanja stupnja polarnosti (regresija). Potrebno je ispitati nekoliko različitih metoda strojnog učenja te provesti detaljnu analizu parametara, značajki i pogrešaka. Razmotriti prilagodbu pristupa za primijenu na tekstovima na hrvatskom jeziku, uzevši u obzir ograničenost jezičnotehnoloških alata za hrvatski jezik. Radu priložiti izvorni programski kod, programsku dokumentaciju i označene skupove podataka.

Zadatak uručen pristupniku: 9. ožujka 2012.
Rok za predaju rada: 21. lipnja 2012.

Mentor:

Doc.dr.sc. Jan Šnajder

Djelovođa:

Prof.dr.sc. Domagoj Jakobović

Predsjednik odbora za diplomski rad profila:

Prof.dr.sc. Siniša Srbljić

Zahvaljujem doc. dr. sc. Janu Šnajderu na pruženoj prilici i povjerenju za izradu ovog rada.

Sadržaj

Poglavlje 1

Uvod

"Što drugi ljudi misle?" pitanje je koje si mnogi postavljaju prilikom donošenja nekih odluka. Puno prije razvoja weba, mnogima je bila potrebna informacija o preporukama za dobrog frizera, preporuci za kvalitetnog automehaničara ili informacija za koga će netko glasovati na sljedećim stranačkim izborima. Danas nam internet daje mogućnost saznati mišljenja i iskustva ljudi koji niti su nam bliži prijatelji niti moraju biti stručni u području koji komentiraju, nego su ljudi za koje nikada nismo niti čuli. Drugim riječima, sve više ljudi iznose svoja mišljenja na internetu dostupna svima. Sudeći po dvije ankete u kojima je sudjelovala više od 2000 ispitanika (comScore/the Kelsey group, 2007; Horrigan, 2008):

- 81% korisnika interneta su barem jednom pretraživali informacije o nekom proizvodu na webu;
- 20% korisnika čini to svaki dan;
- 73% do 87% ljudi koji čitaju recenzije ističu kako su recenzije imale značajan utjecaj na njihovu odluku;
- potrošači ističu kako bi platili 20% do 99% više za proizvod ocijenjen s pet zvjezdica nego za proizvod ocijenjen sa četiri zvjezdice;
- 30% korisnika ostavilo je komentar na proizvod.

Iz gore navedenih činjenica očita je motivacija i potreba za automatskom analizom komentara i mišljenjima koji se njima iznose.

U posljednjih desetak godina web je dramatično promijenio način na koji ljudi iznose svoja mišljenja. U mogućnosti su napisati osvrt na neki proizvod na stranicama

proizvođača i iznositi svoje stavove o raznim temama na internetskim forumima, blogovima, socijalnim mrežama, itd. Ovakav kontekst generiran od korisnika predstavlja vrijedan izvor informacija za mnoge korisne aplikacije. Razvijaju se brojne tehnike za analizu takvih informacija da bi se učinkovito mogle iskoristiti u poslovne ili privatne svrhe.

Web sadrži ogromne količine podataka u obliku nestrukturiranog teksta. Obrada tog teksta od velike je važnosti zbog broja korisnih informacija koje su u njemu sadržane. Također, udio nestrukturiranih informacija puno je veći od udjela strukturiranih informacija pohranjenih u bazama. Analiza nestrukturiranih informacija tehnički je zahtjevna zbog potrebe za obradom prirodnog jezika (engl. *naturall language processing*), ali i od velike koristi u praksi. Tako, na primjer, moguće je doznati kakvo je mišljenje korisnika o nekom novom proizvodu ili usluzi na tržištu. Ako rezultati pokažu da korisnici nisu zadovoljni, moguće je reagirati promjenom usluge ili poboljšanjem proizvoda na nekim segmentima.

Analiza sentimenta (engl. *sentiment analysis*) vrsta je analize subjektivnosti koja se fokusira na prepoznavanje pozitivnih i negativnih mišljenja. Primarni zadatak ovoga rada jest metodama nadziranog strojnog učenja razraditi postupak za određivanje polarnosti dokumenta i dijelova dokumenta na engleskome jeziku. Postupak se temelji na javno dostupnim leksikonima apriorne polarnosti riječi, SentiWordNet i MPQA, te mnogim drugim jezičnim značajkama. Mnogi pristupi automatskoj analizi sentimenta počinju s velikim rječnikom riječi koje su označene s obzirom na njihovu polarnost. Fokus ovoga rada upravo je istražiti utjecaj takvih rječnika na sveukupnu točnost prepoznavanja sentimenta. Napravljena je programska izvedba postupka te je provedeno vrednovanje na zadacima određivanja osnovne polarnosti (klasifikacija) i određivanja stupnja polarnosti (regresija). Ispitano je nekoliko različitih metoda strojnog učenja te je provedena detaljna analiza značajki i njihovog utjecaja za rješavanje polarnosti.

U sljedećem poglavlju će najprije biti opisana definicija problema. Zatim sljedi poglavlje u kojemu je opisana trenutna i moguća primjena ovakvih sustava. Četvrto poglavlje je kratak opis pristupa automatskoj analizi mišljenja. Peto poglavlje daje pregled korištenih resursa u ovome radu. Postupci određivanja polarnosti opisani su u šestom poglavlju . Poglavlje sedam donosi opis algoritama nadziranog strojnog učenja koji su korišteni u ovome radu. Značajke su opisane u osmom poglavlju. Zatim, u poglavlju devet, sljedi evaluacija sustava i eksperimenti. Deseto poglavlje, s naslovom "Osvrt na hrvatski jezik", donosi nam procjenu kakvi bi rezultati ovakvog sustava bili za hrvatski jezik. Dalje sljedi poglavlje o srodnim radobima iz ovog područja i konačno

zadnje poglavlje donosi zaključak rada i smjernice za budući rad.

Poglavlje **2**

Definicija problema

Analiza mišljenja svodi se na tri specifična manja pod problema (Liu, 2007):

Određivanje semantičke orijentacije: Bavi se klasifikacijom teksta. Evaluirani tekst klasificira se kao pozitivan ili negativan. Primjerice, za korisničku recenziju sustav određuje izražava li recenzija pozitivan ili negativan stav. Uglavnom se radi o klasifikaciji na razini dokumenta, no moguće je da se klasifikacija radi na razini rečenice ili na razini fraze[1]. Zadatak ovakvog sustava nije odrediti o čemu točno korisnik izražava mišljenje, već samo karakter tog mišljenja.

Izlučivanje značajki objekta iz teksta: Zadatak ovakvog sustava jest spustiti se na razinu rečenice da bi se dohvatili detalji o objektu o kojemu korisnik izražava mišljenje. Objekt može biti neki proizvod, usluga, tema rasprave, osoba, organizacija i sl. Na primjer, za recenziju korisnika o nekom proizvodu, ovim sustavom dobivamo značajke proizvoda koje je korisnik komentirao u svojoj recenziji te jesu li komentari pozitivni ili negativni. U rečenici, "Sjedalo njegovog bicikla je preusko", komentira se "Sjedalo bicikla" i mišljenje je negativno.

Izlučivanje značajki iz relacija među objektima: Usporedba objekata još je jedan od načina koji uvelike pridonosi razrješavanju analize subjektivnoga mišljenja. Zadatak ovog sustava jest odrediti koje rečenice sadrže relacije među objektima. Na primjer, "Sjedalo njegovog bicikla više je od sjedala na mome biciklu", iskazuje relaciju između dva bicikla.

U ovom radu razvijen je sustav koji se bavi prvom stavkom, određivanjem semantičke orijentacije. Ako nam je dan skup tekstova D, tada sustav koji određuje seman-

[1] Ili na razini prozora od nekoliko riječi

tičku orijentaciju klasificira svaki dokument $d \in D$ u jednu od dviju klasa, pozitivnu ili negativnu. Pozitivna klasifikacija znači da d izražava pozitivno mišljenje. Negativna klasifikacija znači da d izražava negativno mišljenje. Na primjer, ako ocjenjujemo kvalitetu nekog bicikla, sustav klasificira tekstove o biciklu u pozitivna i negativna mišljenja. Evaluiran je i sustav koji klasificira svaki dokument $d \in D$ u jednu od tri klase: pozitivnu, neutralnu ili negativnu.

Zadatak semantičke orijentacije sličan je zadatku klasifikacije teksta na teme o kojima se piše u tekstu. Sustavi koji klasificiraju kojoj temi pripada određeni tekst za zadatak imaju klasificirati tekst u jednu od klasa, npr. politika, znanost, itd. Ovakvi sustavi, kako bi pravilno odredili kojoj klasi pripada neki tekst, razmatraju riječi koje se često pojavljuju u tekstovima te klase. Sustav semantičke orijentacije također razmatra riječi, ali riječi kojima se izražava pozitivan ili negativan stav, kao što su dobar, odličan loš, zao, itd.

Razvoj sustava koji bi automatski analizirao nestrukturirani tekst od velike je važnosti. Takav sustav doveo bi do unaprjeđenja već postojećih tražilica kao što su google.com ili bing.com. No, primjena ovakvog sustava bila bi i puno šira od samih tražilica. Korist u ovakvom sustavu imale bi medicinske institucije, institucije koje se bave prikupljanjem i analizom statističkih podataka, poslovne organizacije, privatne svrhe i tako dalje.

2.1 Pretraživanje recenzija

Forumi, blogovi, socijalne mreže i drugi web-prostori koji omogućavaju dijalog ili komentiranje nepresušan su izvor informacija. Korisnici koriste takve usluge satima čitajući brojne komentare u potrazi za informacijom koja im je potrebna. Na taj način gube svoje dragocjeno vrijeme i novac. Uzmimo za primjer kupca koji traži novi bicikl. Kupnja bicikla po mjeri u posljednje vrijeme pretvorila se u pravu znanost. Proizvođači opreme za bicikle nude na izbor razne prednje i zadnje mjenjače, razne tipove kočnica, rame različitih veličina, materijala i geometrija, zatim tip guma, kvaliteta pogona i tako dalje. Običan korisnik, da bi odabrao kvalitetan bicikl po svojim potrebama i mogućnostima, ima na izbor nekoliko opcija. Najvjerojatnije će najprije posjetiti web-prostore trgovina koje prodaju bicikle i ostat će zateknut brojnom ponudom. Da bi suzio izbor posjetit će web-forume i blogove u potrazi za mišljenjima korisnika. Na forumima će naići na brojne korisne savjete, no kojem komentaru posvetiti najviše pažnje?

Tražilica koja bi automatski prikupljala komentare korisnika, obrađivala ih te ih

sažeto prikazala korisniku riješila bi ovaj problem. Kupci bi dobili listu bicikala poredanu po zadovoljstvu korisnika. Zahtjevniji kupci mogli bi čak i pročitati koji su komentari bili negativni, a koji pozitivni, te naučiti više o traženom upitu na ciljan način. Primjenjivost ovakve aplikacije puno je šira: traženje idealne lokacije za odmor, traženje rute putovanja, itd. – traženje bilo kakvih savjeta i mišljenja ljudi ovako postaje puno jednostavnije. Razvijen je sustav za određivanje subjektivnoga teksta o digitalnim kamerama, mobitelima, *DVD-player*-ima i *mp3-player*-ima. Preciznost tog sustava je 72% (Hu i Liu, 2004).

2.2 Ocjenjivanje recenzija

Socijalne mreže zavladale su svijetom interneta. Društvene mreže danas uglavnom prikupljaju mišljenja korisnika kroz sustave ocjenjivanja[2] ili sustav preporuka[3]; no još uvijek velik dio korisnika takvih mreža komentira, ali ne iznosi mišljenje u obliku ocjene ili preporuke. Sustav za pretraživanje recenzija mogao bi poslužiti kao osnova za kreiranje sustava za automatsko ocjenjivanje recenzija i sakupljanje mišljenja. Takav sustav mogao bi uzeti u obzir sva ona mišljenja koja su unesena samo u obliku teksta na način da procijeni kakvog je polariteta uneseni komentar. Također, mogao bi se razviti sustav za ispravljanje pogrešno rangiranih komentara: postoje slučajevi gdje su korisnici očito pogrešno rangirali svoj komentar (Cabral i Hortaçsu, 2006). Ako je korisnik pozitivno komentirao neki proizvod, no zabunom mu dao lošu ocjenu, bilo bi moguće pronaći i ispraviti ocjene tih komentara.

2.3 Primjena u poslovanju

Poslovne organizacije svakodnevno imaju potrebu biti informirane o tome kako njihov proizvod ili ponuda stoji na tržištu. Pad interesa javnosti za proizvod bitno utječe na budućnost i poslovanje takve organizacije. Kompanije bi automatski mogle pratiti popularnost svog proizvoda i reagirati na svaki pad ili rast popularnosti. Reakcije nakon toga bile bi brojne. Proizvod bi se mogao povući s tržišta. Moglo bi se proučiti s kojim značajkama proizvoda korisnici nisu zadovoljni te zatim reagirati mijenjajući komponente tog proizvoda. Također, mogli bi saznati koja skupina ljudi nije zadovoljna proizvodom, pa ciljano pristupiti korisnicima mijenjajući politiku poslovanja.

[2]Sustav ocjenjivanja ocjenama 1–5, ocjenom 1–0 (poput facebooka, "Sviđa mi se" ili "Ne sviđa mi se").

[3]Sustav "Preporuči prijatelju" kojim svoje mišljenje objavljujemo odabranom krugu prijatelja.

Brojne kompanije imale bi korist i prije proizvodnje samog proizvoda. Istraživanjem trendova tržišta (Mishne i Glance, 2006) dobio bi se niz vrijednih informacija što bi korisnicima bilo zanimljivo imati, a što još nije ni proizvedeno. Poslovne organizacije mogle bi saznati niz informacija o drugoj kompaniji s kojom surađuju ili namjeravaju surađivati.

Marketing bi na ovaj način bitno promijenio sliku svog poslovanja. Oglašivači bi imali pristup informaciji korisnika o svojoj reklami i njenoj popularnosti kroz vrijeme. Također, mogu reagirati na svaki lošiji komentar ili trend komentara korisnika.

U politici se analiza stavova može upotrijebiti kao jedan od načina ispitivanja javnog mnijenja. Prednost je ta što nije potrebno zamarati građane anketama nego se podaci prikupljaju iz već postojećih izvora. Analiza stavova izraženih u raspravama na internetu može dati i dobru procjenu o uspješnosti izborne kampanje neke stranke pa i predvidjeti pobjednika na predstojećim izborima. Sustav za procjenu popularnosti kandidata na predsjedničkim izborima, pod nazivom *CandyPop* (Akšamović et al., 2010), samo je jedan primjer kolika je vrijednost istraživanja ovog područja (Šolta, 2009; Efron, 2004; Hopkins i King, 2007; Laver et al., 2003; Mullen i Malouf, 2006).

2.4 Nove tehnologije

Sustavi za analizu mišljenja također imaju važnu ulogu u omogućavanju razvoja novih tehnologija. Jedna mogućnost je proširiti sustav za preporuku (Tatemura, 2000; Terveen et al., 1997) na način da ne daje preporuke za vrlo loše komentare. Detekcija neprimjerenog jezika u elektroničkoj pošti ili drugom tipu komunikacije (Spertus, 1997) još je jedan od primjera korištenja detekcije emocija.

Web-stranice koje prikazuju reklame na svome prostoru mogle bi koristiti sustav koji bi odredio je li reklama primjerena za sadržaj koji se prikazuje tom web-stranicom (Jin et al., 2007) ili pak složeniji sustav koji bi prikazivao određene reklame na pojavu pozitivnih emocija, a zabranjivao druge reklame na pojavu negativnih emocija.

Sustav odgovaranja na pitanja (engl. *question answering*) još je jedno područje u kojemu analiza sentimenta može biti korisna (Lita et al., 2005; Somasundaran et al., 2007; Stoyanov et al., 2005).

Poglavlje 3

Pristupi automatskoj analizi mišljenja

S obzirom na kontekst koji se analizira, automatsku analizu mišljenja možemo podijeliti na: analiza komentara, analiza rečenice i analiza fraza. U nastavku je dan primjer svakog slučaja.

3.1 Određivanje semantičke orijentacije dokumenta

Najjednostavniji pristup određivanja semantičke orijentacije jest da se promatra problem na razini dokumenta (recenzije, članka, komentara itd.). Tada se može iskoristiti bilo koja već poznata metoda klasifikacije teksta, na primjer Bayesov klasifikator, stroj s potpornim vektorima (engl. *support vector machine; SVM*), neuronske mreže i tako dalje.

Gamon (2004) postiže najbolje rezultate određivanja osnovne polarnosti dokumenta koristeći široki spektar značajki, uključujući bogate jezične značajke, kao što su značajke koje kombiniraju informacije o vrsti riječi i informacije o sintaktičkoj relaciji i značajke koje bilježe napete izraze (engl. *tense information*). Koppel i Schler (2006) pokazuju važnost neutralnih riječi u određivanju osnovne polarnosti dokumenta. Kennedy i Inkpen (2006) objavljuju da korištenje negacija i intenzifikatora poboljšava rezultate u velikoj mjeri. Das i Chen (2001), Pang et al. (2002) i Dave et al. (2003) također koriste negaciju. U njihovim eksperimentima, riječi koje slijede iza negacije su označene posebnom oznakom i tretiraju se kao jedna zasebna riječ. Pang et al. ističu kako korištenje ovakve značajke pospješuje rezultate u maloj mjeri, dok Dave et al. ističu pogoršanje rezultata pri korištenju iste značajke.

Pang et al. proveli su niz eksperimenata ovim metodama za klasifikaciju recenzija filmova. Recenzije filmova trebalo je klasificirati u dvije klase: pozitivnu ili negativnu. Pokazali su da i Bayesov klasifikator i stroj s potpornim vektorima daju dobre rezultate ako se radi o tekstu podijeljenom na unigrame. Testni uzorak sastojao se od 700 pozitivnih i 700 negativnih recenzija filmova te je Bayesov klasifikator postigao točnost 81%, dok je stroj s potpornim vektorima postigao točnost od 82.9%. Neutralni komentari nisu korišteni pri ovom eksperimentu, što je sam eksperiment učinilo mnogo jednostavnijim (Liu, 2007).

3.1.1 Određivanje semantičke orijentacije teksta pomoću funkcije cilja

Dave et al. predložu korištenje heurističke *funkcije cilja* za određivanje semantičke orijentacije teksta . Algoritam se sastoji od dva koraka:

Korak 1: Za svaki izraz računa se *funkcija cilja* sljedećom jednadžbom:

$$score(rijec) = \frac{p(rijec|K) - p(rijec|K')}{p(rijec|K) + p(rijec|K')}, \tag{3.1}$$

gdje je $rijec$ riječ za koju se računa *funkcija cilja*, K je klasa i K' je njen komplement. Izraz $p(rijec|K)$ označava uvjetnu vjerojatnost da riječ $rijec$ pripada klasi K. Računa se pomoću procjenitelja navjeće izglednosti, odnosno tako da se broj pojavljivanja riječi $rijec$ u klasi K podijeli s ukupnim brojem riječi u klasi K. Izrazom danim u 3.1 dobiva se rezultat između -1 i 1.

Korak 2: Da bi se klasificirao novi dokument $d_i = \{rijec_1, \ldots, rijec_n\}$, algoritam zbraja rezultate ciljanih funkcija svih riječi te određuje klasu prema formuli:

$$class(d_i) = \begin{cases} C & eval(d_i) > 0 \\ C' & \text{inače,} \end{cases} \tag{3.2}$$

gdje

$$class(d_i) = \sum_j score(j). \tag{3.3}$$

Istraživanja su se provodila na više od 13000 recenzija. Rezultati su pokazali kako se korištenjem bigrama i trigrama može ostvariti točnost sustava od $84.6\% - -88.3\%$.

3.2 Određivanje semantičke orijentacije rečenica

Neka istraživanja analize sentimenta klasificiraju sentiment rečenica (Morinaga et al., 2002; Yu i Hatzivassiloglou, 2003; Kim i Hovy, 2004; Hu i Liu, 2004; Grefenstette et al., 2004). Sva ova istraživanja najprije grade rječnik apriorno polarnih riječi, a tek nakon toga određuju polaritet rečenica. U ovakvim sustavima uglavnom se koristi ista inačica rješenja kao i kod određivanja semantičke orijentacije teksta. Jedina razlika jest što ulazni skup nije sastavljen od više rečenica, nego samo od jedne. Yu i Hatzivassiloglou određuju sentiment rečenice usrednjavajući apriorno polarne fraze rječnika koje su sadržane u rečenici. Morinaga et al. razmatraju samo pozitivne ili negativne riječi svake rečenice koje su najbliže ciljanoj rečenici. Kim i Hovy, Hu i Liu i Grefenstette et al. broje ili množe apriorno polarne vrijednosti riječi u rečenici. Ova istraživanja također uzimaju u obzir lokalnu negaciju za obrtanje polariteta, a Morinaga et al. uz to uzimaju u obzir i negativan utjecaj riječi poput nedovoljan (engl. *insufficient*). Kim i Hovy opisuju sustav koji određuje semantičku orijentaciju rečenice u više koraka. Najprije se određuje semantička orijentacija pridjeva, glagola i imenica, a zatim se pomoću prikupljenih polarnih riječi određuje semantička orijentacija cijele rečenice.

3.3 Određivanje semantičke orijentacije fraza

Istraživači koji su radili na određivanju polarnosti fraza su Yi et al. (2003), Popescu i Etzioni (2005) i Wilson et al. (2009). Yi et al. koriste rječnik i ručno napisane izraze za klasifikaciju polariteta konteksta. Njihovi ručno napisani regularni izrazi postižu veliku preciznost nad skupom tekstova koje evaluiraju. Popescu i Etzioni koriste tehniku nenadziranog učenja zvanu *meko označavanje* (engl. *relaxation labeling*) (Hummel i Zucker, 1983) u svrhu prepoznavanja polariteta riječi koji su na vrhu fraze kojom se izražava neko mišljenje. Koriste iterativan pristup u tri faze. Najprije koriste *meko označavanje* da bi označili polaritet svih riječi. U drugoj fazi označavaju polaritet riječi s obzirom na riječi na koje ciljaju u danom kontekstu. Treća faza *mekog označivanja* koristi se da bi se riječima dodijelio konačan polaritet, uzimajući u obzir polaritet drugih riječi iz konteksta (označen u drugoj fazi) i negaciju. Popescu i Etzioni,Wilson et al. koriste značajke koje predstavljaju relacije među polarnim riječima. Uz to, Wilson et al. koriste značajke za otkrivanje lokalne negacije i negacije veće udaljenosti. U otkrivanju negacije broje samo one negacije koje se koriste za negiranje fraze, a ne i negacije koje se koriste za pojačavanje fraza; primjerice izraz "ne samo" (engl. *"not only"*). Koriste i značajke za otkrivanje svojstava iz okolnih rečenica kao i značajke koje predstavljaju

pouzdanost informacija iz korištenih rječnika.

Turney (2002) predlaže algoritam koji klasificira recenzije korisnika na pozitivne ili negativne. Predloženo rješenje koristi metodu označivač vrste riječi (engl. *part-of-speach tagging; POS*) koja ima zadatak klasificirati riječi nekog teksta u kategorije: imenica, glagol, prilog, pridjev, prijedlog, zamjenica, broj i tako dalje.

Algoritam je podijeljen u tri faze:

Korak 1: Istraživanja (Liu, 2007) su pokazala da su pridjevi i prilozi dobri pokazatelji subjektivnosti u tekstu. Zbog toga se prvi korak bavi izvlačenjem takvih fraza iz teksta. No, iako pridjev kazuje da se radi o subjektivnom tekstu, takva informacija ne mora nužno određivati semantičku orijentaciju fraze. Na primjer, riječ "nepredvidiv" može biti negativnog konteksta ako se radi o upravljanju motornog vozila, no bila bi pozitivnog konteksta ako želimo kazati kako nam je radnja nekog filma nepredvidiva. Za razrješavanje ovog problema predloženi algoritam uz pridjev ili prilog uzima i sljedeću riječ koju nazivamo *kontekstnom* riječi.

Dvije riječi engleskog jezika izvlače se iz teksta ako označivač vrsta riječi potvrdi da se radi o nekom slučaju iz Tablice 3.3[1]. Na primjer, drugi redak tablice govori da su izvučene dvije riječi u nizu ako je prva riječ prilog, a druga riječ pridjev, dok treća riječ ne smije biti imenica. U rečenici "ova kamera generira lijepe slike", algoritam će izdvojiti izraz "lijepe slike" pošto zadovoljava prvi redak Tablice 3.3.

Prva riječ	Druga riječ	Treća riječ
Pridjev	Imenica (jednina ili množina)	bilo koja riječ
Prilog (pozitiv, komparativ ili superlativ)	Pridjev	nije imenica
Pridjev	Pridjev	nije imenica
Imenica (jednina ili množina)	Pridjev	nije imenica
Prilog (pozitiv, komparativ ili superlativ)	Glagol	bilo koja riječ

Tablica 3.1: Oznake vrsta riječi engleskog jezika za određivanje semantičke orijentacije riječi.

Korak 2: Određuje se semantička orijentacija izvučenih fraza korištenjem mjere zajedničke informacije PMI (engl. *pointwise mutual information*), dane jednadžbom 3.4:

$$PMI(rijec_1, rijec_2) = \log_2 \frac{P(rijec_1 \wedge rijec_2)}{p(rijec_1)p(rijec_2)}. \tag{3.4}$$

[1]Tablica je dana samo okvirno radi pregleda algoritma, točna tablica engleskog jezika je detaljnija te se može pronaći u (Liu, 2007).

Izraz $P(rijec_1 \wedge rijec_2)$ označava vjerojatnost zajedničke pojave riječi $rijec_1$ i riječi $rijec_2$. Izrazi $p(rijec_1)$ i $p(rijec_2)$ označavaju vjerojatnosti pojavu riječi $rijec_1$ odnosno pojavu riječi $rijec_2$ u jeziku. Prema tome izraz $\frac{P(rijec_1 \wedge rijec_2)}{p(rijec_1)p(rijec_2)}$ označava omjer zajedničke pojave dviju riječi i pojavu dviju riječi uzevši u obzir da su pojave riječi međusobno nezavisne.

Semantičku orijentaciju SO (engl. *semantic orientation*) fraze temeljenu na PMI određujemo prema sljedećem izrazu:

$$SO(rijec) = \sum_{p \in Prijeci} PMI(rijec, p) - \sum_{n \in Nrijeci} PMI(rijec, n). \tag{3.5}$$

$Prijec$ i $Nrijec$ označavaju skupove koje sadrže riječi (pridjeve ili priloge) pozitivne, odnosno negativne semantike. Skupovi $Prijeci$ i $Nrijeci$ nisu skupovi za treniranje, već skupovi koji definiraju pozitivnu, tj. negativnu orijentaciju te je postupak učenja koji predlažu u biti nenadzirani postupak. Kao referentni skup predlažu:

$$Prijeci = \{dobar, lijep, odličan, pozitivan, sretan, ispravan, superioran\}$$
$$Nrijeci = \{loš, ružan, jadan, negativan, nesretan, pogrešan, inferioran\}.$$

U slučaju da je rezultat dobiven izrazom 3.5 na nekoj riječi $rijec$ negativan, zaključujemo da je riječ $rijec$ negativne semantike. Ako je taj rezultat pozitivan, zaključujemo da je riječ $rijec$ pozitivne semantike. Apsolutna vrijednost SO jednadžbe označava u kojoj mjeri je riječ pozitivna, tj. negativna.

Za izračun vjerojatnosti iz izraza 3.4, Turney et al. (2003) predlažu korištenje dostupnih tražilica i prikupljanje broja pogodaka takvih riječi. Za svaki upit, tražilica obično vraća broj relevantnih dokumenta, što uzimamo kao broj pogodaka. Kako bismo dobili vjerojatnosti za izraz $P(rijec_1 \wedge rijec_2)$ kao upit tražilici šaljemo "$rijec_1$ $rijec_2$" dok za izraz $p(rijec_i)$ šaljemo upit "$rijec_i$". Također, predlažu korištenje tražilice AltaVista koja ima opciju pretraživanja operatorom BLIZU (engl. *NEAR*). Upit tražilici "$rijec_1$ $NEAR$ $rijec_2$" dao bi broj dokumenata u kojima se riječ $rijec1$ pojavljuje blizu riječi $rijec_2$ (unutar prozora od 10 riječi). Korištenjem operatora BLIZU i rezultata tražilice AltaVista, izraz 3.4 svodi se na :

$$PMI(rijec_1, rijec_2) = \log_2 \frac{bp(rijec_1 BLIZU rijec_2)}{bp(rijec_1)bp(rijec_2)} \tag{3.6}$$

gdje bp[2] (broj pogodaka) označava rezultat upita na tražilicu AltaVista.

Korak 3: Semantičku orijentaciju dokumenta ili rečenice možemo odrediti tako da izračunamo prosječnu vrijednost SO svih izvučenih fraza (Korak 1). Dobivenu prosječnu SO klasificiramo kao pozitivnu ako je veća od nule, odnosno negativnu ako je manja od nule.

Razni eksperimenti pokazali su da rezultati variraju ovisno o domeni na kojoj se pokušava odrediti semantička orijentacija. Za domenu automobilske industrije dobili su točnost klasifikacije 84%, dok je za recenzije filmova točnost bila 66%.

3.4 Ostala istraživanja

Određivanje polariteta konteksta samo je jedno područje istraživanja automatske analize sentimenta. Istražuje se sve od učenja polariteta apriorno-polarnih riječi i fraza (Kim i Hovy, 2004; Andreevskaia i Bergler, 2006; Hatzivassiloglou i McKeown, 1997; Turney et al., 2003; Hu i Liu, 2004; Esuli i Sebastiani, 2005; Popescu i Etzioni, 2005), prepoznavanja neprijateljskih poruka (Spertus, 1997), klasificiranja sentimenta *online* poruka (Das i Chen, 2001) ili recenzija proizvoda i filmova (Turney, 2002; Pang et al., 2002; Dave et al., 2003; Beineke et al., 2004; Bai et al., 2005; Koppel i Schler, 2006; Kennedy i Inkpen, 2006).

Određivanje apriornog polariteta drugačiji je zadatak od određivanja polariteta konteksta. Cilj određivanja apriornog polariteta je prikupiti polaritet riječi ili fraza za konstrukciju rječnika. Ovaj rad za određivanje polarnosti koristi unaprijed izgrađene rječnike apriorno polarnih riječi, *MPQA* i *SentiWordNet*, te pomoću njih određuje kakav je polaritet konteksta u kojemu se pojavljuje riječ iz rječnika apriorno polarnih riječi.

[2]Da bi se izbjeglo dijeljenje s nulom, dobivenom rezultatu može se dodati neki malen broj, npr. 0.01.

Poglavlje 4

Korišteni resursi

U okviru rada i izrade sustava za određivanje polarnosti, korišteni su dodatni resursi: korpusi komentara na engleskom jeziku te rječnici apriorno polarnih riječi.

4.1 Korpus

Za kreiranje sustava najprije je bilo potrebno sakupiti dovoljno velik broj komentara koji su označeni s obzirom na polarnost. Kao dobar izvor pokazao se upravo web-stranica operacijskog sustava za mobitele *Android*. Web-stranica *Androida* sadrži tržnicu (engl. *marketplace*)[1]. na kojoj je moguće skinuti aplikacije na operacijski sustav *Android* i komentirati tu aplikaciju. Također, svaka osoba koja komentira nužno mora dati i ocjenu **1, 2, 3, 4 ili 5**. Kako je korištenje mobitela u nezasitnom rastu, tako je i korišteni web dobar izbor za analizu komentara.

U svrhu izrade sustava skupljen je izvor od 2 067 724 komentara. Komentari su skupljeni od ukupno 16 691 komentiranih aplikacija, gdje svaka aplikacija pripada barem jednoj od 34 kategorije prikazane u Tablici 4.1.

Ovako dobiveni skup komentara nazvan je *korpus komentara*. Da bi se analiziralo koliko veličina komentara ovisi o radu sustava izrađen je korpus koji se sastoji samo od komentara koji sadrže jednu rečenicu. Ovaj korpus jednak je *korpusu komentara* uz tu razliku što su u njemu izbačeni komentari koji imaju dvije ili više rečenica. Ovakav sustav nazvan je *korpus rečenica*. Uz ova dva korpusa napravljen je i *korpus fraza*. Radi se o bigramskim frazama koje su dobivene postupkom objašnjenim u nastavku. Najprije, svakoj riječi u korpusu pridružena je njihova vrsta riječi koristeći *označivač vrste riječi* (engl. *part-of-speach tagger*) (Toutanova i Manning, 2000; Toutanova et al.,

[1] Tržnica dostupna na adresi `https://play.google.com` .

Aplikacije		Igre
pozadine	medicina	arkade
dodaci	muzika i audio	mozgalice
knjige i reference	vijesti i magazini	karte
posao	personalizacija	uobičajeno
stripovi	fotografija	pozadine
komunikacija	produktivnost	dodaci
edukacija	kupovina	utrke
zabava	društvenost	sportske igre
financije	sport	
fitnes i zdravlje	alati	
demo i knjižnice	transport	
životni stil	putovanja	
video i mediji	vrijeme	

Tablica 4.1: Kategorije aplikacija s Android-tržnice.

2003). Nadalje, određeno je da se dvije uzastopne riječi smatraju frazom ako se njihove vrste podudaraju s Tablicom 3.3(Turney, 2002).

U Tablici 4.2 prikazana je distribucija ocjena skinutih komentara s Android-tržnice.

ocjena	broj komentara	udio(%)	broj rečenica	udio(%)	broj fraza	udio(%)
1.0	283443	14	76645	15	15245	12
2.0	127202	6	33430	6	10110	8
3.0	214561	10	62443	12	19529	15
4.0	395956	19	103641	20	28927	23
5.0	1046562	51	250778	47	53775	42
Ukupno	2067724		526937		127586	

Tablica 4.2: Distribucija ocjena komentara (korpus komentara), rečenica (korpus rečenica) i fraza (korpus fraza) s Android-tržnice.

Od toga, *14%* komentara ima ocjenu 1, *6%* komentara ima ocjenu 2, *10%* komentara ima ocjenu 3, *19%* komentara ima ocjenu 4 te *51%* komentara ima ocjenu 5. Razlog ovakvoj distribuciji vjerojatno leži u činjenici da većina osoba skida aplikaciju koja im se svidjela. Ako se nekoj osobi aplikacija nije svidjela, mala vjerojatnost je da će

ju skinuti, a još manja da će ju komentirati. Iz Tablice 4.2 je vidljivo da ocjena 1 odstupa od ovog pravila. Više stvari ide tome u prilog. Prvo, ljudi koji iznose stavove imaju tendenciju iznijeti jako pozitivan ili jako negativan stav. Čim je raspon ocjena veći, manje je neslaganje među ocjenjivačima (Pang i Lee, 2005). Drugo, osoba koja je skinula neku aplikaciju vjerojatno je očekivala da će aplikacija biti dobra (inače ju ne bi skinula). Kako neke aplikacije ipak ne ispune očekivanja (greške, rušenje aplikacije) osobe općenito daju negativnije komentare od osoba koji nisu imali nikakva očekivanja jer kod tih osoba nastupa razočaranje.

Korpus rečenica skoro je četiri puta manji od *korpusa komentara* što upućuje na to da korisnici Android-tržnice ipak pišu podulja obrazloženja zašto im se neka aplikacija sviđa ili ne. *Korpus fraza* sadrži tek 127 586 fraza jer se radi samo o bigramima koji zadovoljavaju pravila iz Tablice 3.3.

Obrada svih komentara s trenutno dostupnim algoritmima strojnog učenja (sekvencijalni) bio bi vremenski predug proces. U tu svrhu izabran je podskup od 5000 komentara. Tih 5000 komentara izabrano je dalje na dva načina. Prvi način je izbor komentara iz skupa svih komentara tako da bude sačuvana njihova distribucija ocjena iz Tablice 4.2. Drugi način je izbor komentara uniformnom razdiobom. S obzirom na to, ukupno je konstruirano 6 korpusa: market, uniform, sent_market, sent_uniform, pra_market, pra_uniform. *Market* je *korpus komentara* sačinjen od 5000 komentara stvarne razdiobe. *uniform* je *korpus komentara* sačinjen od 5000 komentara uniformne razdiobe. Prefiks *sent*(engl. *sentence*) govori da se radi o *korpusu rečenica* dok prefix *pra*(engl. *phrase*) govori da se radi o *korpusu fraza.* Tako primjerice *sent_market* znači da se radi o korpusu rečenica stvarne razdiobe. Slično vrijedi i za drugih pet korpusa.

U Tablici 4.3 prikazan je broj tokena, broj jedinstvenih tokena i broj jedinstvenih lema svih šest korpusa. Zanimljivo je kako *rečenični korpusi* sadrže gotovo dvostruko manje riječi od *korpusa komentara.* Kako su komentari u *korpusu rečenica* sastavljeni od jedne rečenice, može se zaključiti da je *korpus komentara* u prosjeku sastavljen od skoro dvije rečenice.

4.2 Rječnik apriorno polarnih riječi

U razvoju sustava korištene su neke unaprijed sastavljene liste polarnih riječi. Tim riječima pridružene su informacije o njihovu polaritetu ili stupnju polariteta, kao i informacije o vrsti riječi i kojem kontekstu pripada ta riječ. Korištena su dva izvora apriorno polarnih riječi: *SentiWordNet* i *MPQA rječnik subjektivnosti* (engl. *subjectivity*

Korpus	Broj tokena	Broj jed. tokena	Jed. lema
market	106754	9363	6894
uniform	113561	9539	6924
sent_market	58938	6380	4822
sent_uniform	61568	6410	4756
pra_market	10002	4123	3449
pra_uniform	10008	4080	3388

Tablica 4.3: Distribucija tokena, jedinstvenih tokena i jedinstvenih lema po korpusima.

lexicon).

4.2.1 SentiWordNet

SentiWordNet (Baccianella et al., 2010; Esuli i Sebastiani, 2006) je rječnik koji se koristi za ekstrakciju mišljenja iz teksta. Naziv dolazi od *Sentiment WordNet* gdje je *WordNet* leksička baza engleskog jezika (Fellbaum, 1998; Miller, 1995) iz koje je konstruiran *SentiWordNet*. Svaka riječ sadrži ocjenu pozitivnosti, negativnosti i objektivnosti te zadovoljava izraz $pozitivno + negativno + objektivno = 1$. *Pozitivno*, *negativno*, *objektivno* oznake su redom za ocjene pozitivnosti, negativnosti i objektivnosti. Ove ocjene kreću se u rasponu od 0 do 1. Sve riječi koje nisu pronađene u *SentiWordNet* rječniku uzete su kao neutralne. Riječnik svakoj riječi pridružuje i oznaku koja nam govori u kojemu kontekstu je određena polarnost, oznaku vrste riječi i još nekoliko informacije koje nisu korištene u ovome radu. Ista riječ u različitim kontekstima može imati različita značenja. Tom prigodom svakoj takvoj riječi pridružen je prirodni broj određen u *WordNet-u*. U *SentiWordNet-u* ovaj se broj koristi kao oznaka konteksta.

Prvi stupac Tablice 4.4 prikazuje sve riječi iz korpusa koje su sadržane i u *SentiWordNet-u*. Drugi stupac prikazuje ukupan broj svih pozitivnih i negativnih riječi iz korpusa koje su i u *SentiWordNet* rječniku. Pokrivenost označava omjer $pokrivenost = \frac{\text{sve poz + neg}}{\text{sve riječi rječnika}}$. Na primjer, ukupan broj riječi *market* korpusa iznosi 106 754, a ukupan broj svih pozitivnih i negativnih riječi za *market* korpus je 36 643. Iz toga proizlazi da je pokrivenost 34%. Ostali stupci prikazuju broj jedinstvenih parova *(riječ, vrsta riječi)* redom po svim riječima, samo pozitivne, samo negativne, samo neutralne riječi. Pokrivenost pozitivnih i negativnih riječi opada što je veći broj riječi u korpusu. Tako najveći korpus, *korpus komentara*, ima najmanju pokrivenost; dok najmanji, *korpus fraza*, ima najveću pokrivenost. Informacija o pokrivenosti mogla bi nam okvirno pokazivati koliki će utje-

caj rječnik apriorno polarnih riječi imati na točnost sustava. Što je pokrivenost veća, to imamo više informacije o pozitivnim i negativnim riječima te bi time sustav raspolagao s većim apriornim znanjem i postizao veću točnost.

Korpus	sve riječi	sve poz + neg	pokrivenost(%)	riječi	poz	neg	neut
market	99825	36643	34	6802	1918	1486	3398
uniform	106497	38886	34	6983	1929	1557	3477
sent_market	54746	21586	37	4860	1446	1040	2347
sent_uniform	57521	22675	37	4953	1446	1124	2383
pra_market	8322	5685	57	2904	904	747	1253
pra_uniform	8385	5646	56	2883	911	763	1209

Tablica 4.4: Distribucija riječi korpusa sadržanih u *SentiWordNet* rječniku.

4.2.2 MPQA

Rječnik subjektivnosti *MPQA* (engl. *Multi-Perspective Question Answering*) sastavljen je od nekoliko izora. Neke riječi sakupljene su od ručno sastavljenih izvora. Druge su sakupljene koristeći označene i neoznačene podatke. Većina riječi skupljena je kao dio rada objavljenog u Riloff i Wiebe (2003). U radu su korištene sljedeće informacije iz rječnika: tip, vrsta riječi, zastavica o tome je li riječ korjenovana, polaritet. Tip govori je li riječ jakog polariteta (engl. *strong subject*) ili slabog polariteta (engl. *weak subject*). Ako je riječ jakog polariteta, to znači da ta riječ izražava emociju intenzivnije nego riječ slabog polariteta. Vrsta riječi u ovom rječniku može biti imenica, glagol, pridjev, prilog ili *bilo koja* (engl. *any pos*). Ako je riječ korjenovana tada je priložen korijen te riječi, inače je riječ u punom obliku. Polaritet riječi u korpusu *MPQA* može biti pozitivan, negativan, oboje ili neutralan. Pozitivan polaritet govori nam da riječ izražava nešto pozitivno dok negativan polaritet govori da izražava nešto negativno. Kada je riječ označena s *oboje*, to znači da može izražavati i pozitivan i negativan stav. Radi li se o pozitivnom ili negativnom ovisi o kontekstu u kojemu se ta riječ pojavljuje. Neutralna riječ koristi se za izražavanje niti pozitivnog niti negativnog stava.

Prvi stupac Tablice 4.5 prikazuje sve riječi iz korpusa koje su sadržane i u *MPQA*. Ostali stupci prikazuju broj jedinstvenih parova *(riječ, vrsta riječi)* redom po svim riječima, riječima pozitivnog tipa, riječima negativnog tipa, riječima oboje tipa i riječima neutralnog tipa. Drugi stupac prikazuje ukupan broj svih pozitivnih i negativnih riječi iz korpusa koje su i u rječniku *MPQA*. Pokrivenost označava omjer $pokrivenost = \frac{\text{sve poz + neg}}{\text{sve riječi rječnika}}$. Pokrivenost rječnika *MPQA* manja je od pokrive-

Korpus	Sve riječi	Sve poz + neg	pokrivenost (%)	riječi	poz	neg	oboje	neut
market	15883	11319	11	1435	620	576	0	239
uniform	15670	10988	10	1509	607	658	1	243
sent_market	9450	6700	11	1053	455	399	0	199
sent_uniform	9271	6432	10	1093	460	440	1	192
pra_market	3263	2596	26	767	326	310	0	131
pra_uniform	3135	2418	24	777	323	323	1	130

Tablica 4.5: Distribucija riječi korpusa sadržanih u rječniku *MPQA*.

nosti rječnika *SentiWordNet*. Iz toga je moguće zaključiti da će utjecaj rječnika *MPQA* na rad sustava vjerojatno biti od manjeg značaja od utjecaja rječnika *SentiWordNet*.

Poglavlje 5

Postupci određivanja polarnosti

Postupci određivanja polarnosti algoritmima nadziranog strojnog učenja mogu biti postupci koji određuju osnovnu polarnost ili postupci koji određuju stupanj polarnosti.

5.1 Određivanje osnovne polarnosti

Klasifikacija u obradi prirodnog jezika odnosi se na svrstavanje dokumenata u jednu ili više unaprijed definiranih kategorija. Slično tome, određivanje osnovne polarnosti jest svrstavanje dokumenata u jednu ili više unaprijed definiranih kategorija gdje su kategorije polarne klase. Polarne klase mogu biti *{pozitivan, negativan}*, *{pozitivan, neutralan, negativan}*, *{jako pozitivan, slabo pozitivan, neutralan, slabo negativan, jako negativan}* i slične. U ovome radu bit će razmotrena dva slučaja: određivanje osnovne polarnosti u dvije klase i određivanje osnovne polarnosti u tri klase. Klasifikacija u dvije klase je svrstavanje primjera u pozitivno ili negativno. Klasifikacija u tri klase je svrstavanje primjera u pozitivno, neutralno i negativno. Klasa *neutralno* nije neutralna u pravom smislu riječi. Više se radi o miješanim pozitivnim i negativnim komentarima, ali pozitivnih ili negativnih u manjoj mjeri nego kod klasa *pozitivno* ili *negativno*.

5.2 Određivanje stupnja polarnosti

Određivanje stupnja polarnosti ne koristi klase. Radi se zapravo o postupku regresije kojim se pokušava aproksimirati funkcija kroz prostor rješenja koja najbolje opisuje ulazne podatke, odnosno komentare. U ovome radu korištena je aproksimacija linearnom funkcijom. Na taj način pokušavamo kroz prostor rješenja postaviti hiperravninu koja najbolje aproksimira komentare.

Poglavlje 6

Algoritmi nadziranog strojnog učenja

Primarni zadatak ovoga rada jest metodama nadziranog strojnog učenja razraditi postupak za određivanje polarnosti dokumenta i dijelova dokumenta na engleskome jeziku. Eksperimenti su izvedeni koristeći sljedeće metode strojnog učenja: *naivan Bayesov klasifikator*, *SVM* i *SVR* (engl. *support vector regression*).

Za ekstrakciju značajki korišten je sljedeći algoritam. Neka $\{f_1, \ldots, f_m\}$ predstavlja skup od m značajki koje se pojavljuju u dokumentu. Neka $n_i(d)$ predstavlja koliko puta se značajka f_i pojavljuje u dokumentu d. Tada je svaki dokument d predstavljen kao vektor $\vec{d} := (n_1(d), n_2(d), \ldots, n_m(d))$. Binarni vektori su oni vektori koji bilježe prisustvo ili odsustvo značajke f_i. Takvi vektori mogu poprimiti vrijednosti 0 ili 1. Pobrojani vektori poprimaju vrijednosti iz unaprijed definiranog skupa vrijednosti. Numerički vektori mogu poprimiti bilo koju numeričku vrijednost. Većina vektora u ovome radu je binarnog tipa. Nekoliko njih je pobrojanog, a ostatak su numerički vektori.

6.1 Naivan Bayesov klasifikator

Jedan pristup klasifikaciji teksta je taj da se svakom dokumentu d pridruži klasa $c^* = \arg\max_c P(c \mid d)$. Izvest ćemo naivan Bayesov (NB) klasifikator iz Bayesovog pravila:

$$P(c \mid d) = \frac{P(c)P(d \mid c)}{P(d)}, \tag{6.1}$$

gdje $P(d)$ nema nikakvog utjecaja na izbor klase c^*. Da bi mogli procijeniti izraz $P(d \mid c)$, naivni Bayes pretpostavlja da su značajke f_i uvjetno nezavisne za zadanu

klasu c:

$$P_{NB}(c \mid d) := \frac{P(c)\left(\prod_{i=1}^{m} P(f_i \mid c)^{n_i(d)}\right)}{P(d)}. \tag{6.2}$$

Treniranje se svodi na procjenu vjerojatnosti $P(c)$ i $P(f_i \mid c)$.

Unatoč svojoj jednostavnosti i unatoč tome što pretpostavka uvjetne nezavisnosti očito ne stoji u stvarnome svijetu, naivni Bayes u klasifikaciji teksta daje iznenađujuće dobre rezultate. Naivan Bayesov klasifikator optimalan je izbor za određene probleme koji sadrže značajke velike međusobne zavisnosti (Domingos i Pazzani, 1997).

6.2 SVM

Pokazalo se da SVM daje odlične rezultate na zadacima klasifikacije teksta, uglavnom bolje nego naivan Bayesov klasifikator (Joachims, 1998; Pang et al., 2002; Pang i Lee, 2004). Algoritam funkcionira na način da se postavlja velika margina između podataka, za razliku od naivnog Bayesa koji problem rješava vjerojatnosnim modelom. Glavna ideja iza treniranja podataka je da se nađe hiperravnina $\vec{w}$, koja razdvaja ulazne primjere na dvije klase. Margina se postavlja na način da bude maksimalno udaljena od primjera jedne i druge klase. Pretraga gdje će se margina postaviti predstavlja problem optimizacije s ograničenjima. Neka $c_n \in \{1, -1\}$ predstavlja ispravnu klasu za uzorak d_j; tada rješenje možemo zapisati kao:

$$\vec{w} := \sum_j \alpha_j c_j \vec{d_j},\ \alpha_j \geq 0. \tag{6.3}$$

6.3 SVR

Smola i Vapnik (1997) predstavljaju verziju SVM-a namijenjenu za regresiju. SVM algoritam opisan gore ovisi samo o podskupu ulaznih podataka koji leže najbliže margini razdvajanja. Slično tome, SVR koji su predstavili Smola i Vapnik također ovisi o podskupu ulaznih podataka, ali ignorirajući podatke kojima je udaljenost od margine $\vec{w}$ manja od unaprijed određene vrijednosti ϵ.

6.4 Značajke

Jedan od glavnih zadataka pri rješavanju problema metodama strojnog učenja je izbor značajki. Značajke ili atributi zapravo su vektori prostora rješenja problema koji se rješava. Kvaliteta rješenja u najvećoj mjeri ovisi o kvalitetnom izboru značajki. Za

potrebe izrade ovoga rada razvijeno je 63 značajke prikazane u Tablici 6.2. Neke od njih su u većoj ili manjoj korelaciji s drugim značajkama, odnosno ne radi se o potpuno nezavisnim vektorima. Kako se uglavnom radi o vektorskim značajkama, Tablica 6.1 prikazuje ukupan broj značajki pronađenih u svakom korpusu.

Korpus	broj značajki
market	832401
uniform	883523
sent_market	513637
sent_uniform	532987
pra_market	103486
pra_uniform	103309

Tablica 6.1: Ukupan broj značajki svakog korpusa.

Da bi se lakše objasnile, značajke su podijeljene u grupe po tehnologiji ili resursu koji je potreban za njihovu ekstrakciju: osnovna, lema, vrsta riječi, *SentiWordNet*, sintaktičko stablo, *MPQA* i *SentiWordNet* + razrješavanje višeznačnosti.

Osnovna Skupina značajki koje koriste uglavnom prebrojavanje jedinki. Značajka *unigram riječi* je binarni vektor koji označava je li se određena riječ pojavila u primjeru koji ispitujemo ili nije. Sukladno tome *bigram riječi* i *trigram riječi* redom označavaju binarne vektore je li se određena uzastopna dvojka odnosno trojka riječi pojavila u primjeru. Značajke *frekvencija unigrama/bigrama/trigrama riječi* razlikuju se od značajki *frekvencija unigrama/bigrama/trigrama riječi* samo u tome što se radi o numeričkom vektoru, a ne o binarnom vektoru. Tako će vektor *frekvencija unigrama riječi* biti broj pojava svake pojedine riječi, a vektor *unigram riječi* je li se riječ pojavila ili nije. *Pozicija riječi* je vektor značajka koja svakoj riječi pridružuje informaciju u kojem dijelu teksta se pojavila. Tekst je podijeljen na četiri jednaka dijela po riječima. Ako tekst ima 10 riječi tada prva, druga i treća riječ ulaze u prvu četvrtinu teksta; četvrta, peta i šesta ulaze u drugu četvrtinu teksta; šesta, sedma i osma ulaze u treću četvrtinu teksta; deveta i deseta ulaze u zadnju četvrtinu teksta. Značajka *broj rečenica* numerička je vrijednost koja označava koliko rečenica ima u tekstu primjera.

Lema Značajke ove skupine jednake su redom prvim šest značajkama skupine *osnovna* uz tu razliku što umjesto izvornih oblika riječi koriste lematizirani oblik riječi.

Vrsta riječi Prvih šest značajki ove skupine iz Tablice 6.2 jednako je redom prvim šest značajkama skupine *osnovna* uz tu razliku što umjesto riječi prebrojavaju vrste riječi. *Pozicija vrste riječi* je vektor značajka koja svakoj vrsti riječi pridružuje informaciju u kojem dijelu teksta se pojavila.[1] *Riječi i vrsta riječi* slična je značajki *unigram riječi* samo što je svaka riječ stopljena s njenom vrstom. Sljedeće značajke određuju koliko u primjeru ima pridjeva, priloga, brojeva, modalnih glagola, negacija i zamjenica. To su redom značajke *broj pridjeva, broj priloga, broj brojeva, broj modalnih glagola, broj negacija, broj zamjenica.* Zadnja značajka u skupini *vrsta riječi* je *pridjevi.* Radi se o binarnom vektoru koji određuje je li se pridjev pojavio u primjeru teksta ili nije.

SentiWordNet Značajke ove skupine koriste informacije dostupne u *SentiWordNet* rječniku. Također, kako su izrazi u *SentiWordNet-u* svedeni na osnovni oblik, ove značajke koriste i lematizaciju. *SentiWordNet riječi* je binarni vektor koji označava je li određena riječ *SentiWordNet* rječnika sadržana u primjeru. *Frekvencija SentiWordNet riječi* je numerički vektor koji označava koliko se puta određena riječ pojavljuje u primjeru, a da je ujedno sadržana i u *SentiWordNet* rječniku. *Polarna vrijednost SentiWordNet riječi* značajka je koja govori u kojoj mjeri je neka riječ *SentiWordNet* rječnika pozitivna ili negativna. Vrijednost ove značajke računa se kao $vr = pozitivno - negativno$ gdje *pozitivno* i *negativno* označavaju vrijednosti sadržane u *SentiWordNet* rječniku koje redom određuju u kojoj mjeri je riječ pozitivna odnosno negativna. Ovo je numerički vektor pa se vrijednost vr računa za svaku riječ *SentiWordNet* korpusa sadržanu u primjeru teksta koji ispitujemo. *Objektivna vrijednost SentiWordNet riječi* značajka je koja računa u kojoj mjeri riječ iz *SentiWordNet* rječnika izražava objektivnost. Objektivna vrijednost pomoću *SentiWordNet* rječnika računa se kao $ob = 1 - (pozitivno + negativno)$ gdje izraz u zagradi $su = pozitivno + negativno$ određuje subjektivnu vrijednost. Značajka *polaritet SentiWordNet riječi* binarni je vektor koji svakoj izračunatoj vrijednosti vr pridružuje njen predznak. Ako je $vr < 0$ vrijednost riječi ovog vektora je 0, inače je 1. *Pozitivne/negativne/neutralne SentiWordNet riječi* binarni su vektori koji redom označavaju pojavu pozitivnih, negativnih ili neutralnih riječi.[2] *Broj pozitivnih/negativnih SentiWordNet riječi* numeričke su značajke koje određuju koliko neki primjer teksta sadrži pozitivnih odnosno negativnih riječi.

[1] Vidi opis značajke *pozicija riječi* iz skupine *osnovna.*

[2] Riječ smo u ovome slučaju proglasili neutralnom ako je $vr = pozitivno - negativno = 0$. Ovakve riječi većinom nisu neutralne jer uglavnom sadrže neku pozitivnu i negativnu vrijednost.

Zadnja značajka ove skupine je *diskretizirana polarna vrijednost SentiWordNet riječi*. Ova značajka polarnu vrijednost diskretizira tako što joj umjesto realnog broja pridruži pobrojani tip po vrijednostima: -1 do -0.75 (jako negativno), -0.75 do -0.25 (srednje negativno), -0.25 do 0 (malo negativno), 0 (neutralno), 0 do 0.25 (malo pozitivno), 0.25 do 0.75 (srednje pozitivno), 0.75 do 1 (jako pozitivno). Kako *SentiWordNet* rječnik sadrži riječi s obzirom na njihov kontekst, sve značajke ove skupine računaju polaritet kao srednju vrijednost svih polariteta u kojima se ta riječ može pojaviti.

Sintaktičko stablo Korištenjem *Stanfordovog parsera* (Klein i Manning, 2003) razvijeno je šest značajki ove skupine. *Stanfordov parser* gradi sintaktičko stablo kojim je određena gramatička struktura rečenice (služba riječi). Prva značajka, *relacije sintaktičkog stabla*, vektor je koji svakoj relaciji sintaktičkog stabla određuje je li se pojavio u primjeru teksta. Relacije sintaktičkog stabla mogu biti neke od 48 predstavljenih u radovima (De Marneffe et al., 2006; De Marneffe i Manning, 2008). *Frekvencija relacija sintaktičkog stabla* vektor je koji pobrojava koliko se puta pojavila određena relacija u primjeru. Relacije sintaktičkog stabla označavaju u kojem odnosu su dvije riječi unutar rečenice. Nazovimo te riječi roditelj, odnosno dijete. *Vektor roditelj-dijete sintaktičkog stabla* popisuje koje dvije riječi su u relaciji. Sljedeća značajka, *frekvencija roditelj-dijete sintaktičkog stabla*, slično prošloj značajki pobrojava koliko puta su određene dvije riječi bile u relaciji. Zadnje dvije značajke slične su prethodnim dvjema. Značajka *relacija-roditelj-dijete sintaktičkog stabla* označava je li se određena relacija sa svoje dvije riječi pojavila u primjeru, a zadnja značajka *frekvencija relacija-roditelj-dijete sintaktičkog stabla* pobrojava koliko se puta takva trojka pojavila u primjeru.

MPQA Značajke ove skupine koriste rječnik *MPQA*. *Tip MPQA-riječi* binarni je vektor riječi iz rječnika *MPQA* koji određuje je li neka riječ primjera jakog ili slabog polariteta. *Polaritet MPQA-riječi* pobrojani je vektor riječi iz rječnika *MPQA* koji određuje kakvog je polariteta neka riječ primjera: pozitivna, negativna, oboje ili neutralna[3]. *Pozitivne/negativne/neutralne MPQA-riječi* binarni su vektori riječi iz rječnika *MPQA* koji redom određuju je li riječ pozitivna, negativna ili neutralna. Neke riječi mogu određivati i pozitivnost i negativnost. Kakvog će polariteta biti ovisi o kontekstu u kojemu se pojavljuju. To su riječi dvojnog polariteta koje prikuplja binarni vektor *MPQA-riječi dvojnog polariteta*.

[3]Vidi poglavlje 4.2.2 za opis.

SentiWordNet + razrješavanje višeznačnosti Ova skupina koristi razrješavanje višeznačnosti (engl. ***W**ord **S**ense **D**isambiguation*) i rječnik *SentiWordNet*. Razrješavanje višeznačnosti (Pedersen i Kolhatkar, 2009; Pedersen et al., 2005) postupak je kojim riječima dodjeljujemo njihov točan smisao zavisno od konteksta u kojemu se nalaze. Sve značajke ove skupine najprije koriste razrješavanje višeznačnosti, a zatim nad tim riječima, kojima je određen smisao, primjenjuju postupke iz skupine *SentiWordNet*. Zbog toga će riječi ove skupine dobiti polaritete koji su im i namijenjeni rječnikom *SentiWordNet*, a ne uprosječene vrijednosti svih konteksta traženih riječi, kao što je to kod skupine *SentiWordNet*.

	SentiWordNet
	SentiWordNet riječi
	frekvencija *SentiWordNet* riječi
	polarna vrijednost *SentiWordNet* riječi
Osnovna	objektivna vrijednost *SentiWordNet* riječi
unigram riječi	binarni polaritet *SentiWordNet* riječi
bigram riječi	pozitivne *SentiWordNet* riječi
trigram riječi	negativne *SentiWordNet* riječi
frekvencija unigrama riječi	neutralne *SentiWordNet* riječi
frekvencija bigrama riječi	broj pozitivnih *SentiWordNet* riječi
frekvencija trigrama riječi	broj negativnih *SentiWordNet* riječi
pozicija riječi	diskretizirana polarna vrijednost *SentiWordNet* riječi
broj rečenica	
	Sintaktičko stablo
Lema	relacije sintaktičkog stabla
lematiziran unigram riječi	frekvencija relacija sintaktičkog stabla
lematiziran bigram riječi	vektor roditelj-dijete sintaktičkog stabla
lematiziran trigrami riječi	frekvencija roditelj-dijete sintaktičkog stabla
frekvencija lematiziranog unigrama riječi	relacija-roditelj-dijete sintaktičkog stabla
frekvencija lematiziranog bigrama riječi	frekvencija relacija-roditelj-dijete sintaktičkog stabla
frekvancija lematiziranog trigrama riječi	
	MPQA
Vrsta riječi	tip *MPQA* riječi
unigram vrste riječi	polaritet *MPQA* riječi
bigram vrste riječi	pozitivne *MPQA* riječi
trigram vrste riječi	negativne *MPQA* riječi
frekvencija unigrama vrste riječi	*MPQA* riječi dvojnog polariteta
frekvencija bigrama vrste riječi	neutralne *MPQA* riječi
frekvencija trigrama vrste riječi	
pozicija vrste riječi	*SentiWordNet* + razrješavanje višeznačnosti(*WSD*)
riječi i vrsta riječi	*WSD* + *SentiWordNet* riječi
broj pridjeva	*WSD* + frekvencija *SentiWordNet* riječi
broj priloga	*WSD* + polarna vrijednost *SentiWordNet* riječi
broj brojeva	*WSD* + objektivna vrijednost *SentiWordNet* riječi
broj modalnih glagola	*WSD* + binarni polaritet *SentiWordNet* riječi
broj negacija	*WSD* + pozitivne *SentiWordNet* riječi
broj zamjenica	*WSD* + negativne *SentiWordNet* riječi
pridjevi	*WSD* + neutralne *SentiWordNet* riječi
	WSD + broj pozitivnih *SentiWordNet* riječi
	WSD + broj negativnih *SentiWordNet* riječi
	WSD + diskretizirana polarna vrijednost *SentiWordNet* riječi

Tablica 6.2: Značajke za klasifikaciju i regresiju.

Poglavlje 7

Eksperimenti

Eksperimenti su provedeni za svih šest korpusa: *market, uniform, sent_market, sent_-uniform, pra_market, pra_uniform.* Svaki od sustava izgrađen je pomoću sljedećih algoritama nenadziranog učenja: *Naivan Bayes*, *SVM* i *SVR*. Pomoću *naivnog Bayesa* i *SVM-a* izgrađeni su sustavi koji određuju osnovnu polarnost teksta, a pomoću *SVR* izgrađen je sustav koji određuje stupanj polarnosti teksta. Najprije je potrebno ispitati kako svaka značajka zasebno utječe na točnost sustava. Zatim je potrebno grupirati značajke na način prikazan u Tablici 6.2 te vidjeti kakav utjecaj grupe zasebno imaju na kvalitetu sustava.

7.1 Postavke eksperimenata

Za potrebe određivanja osnovne polarnosti razmotrena su dva slučaja: određivanje osnovne polarnosti u dvije klase i određivanje osnovne polarnosti u tri klase. Kako su ulazni komentari ocijenjeni ocjenama 1 do 5, tako smo dobili sljedeće sustave određivanja osnovne polarnosti:

filter1-5 Filter je dobiven na način da se komentari koji sadrže ocjene 1 ili 5 zadrže u korpusu, dok se komentari koji sadrže ocjene 2, 3 ili 4 uklanjaju iz skupa. Na taj način dobiven je podskup s klasama "1" i "5".

filter12-45 Komentari koji sadrže ocjenu 3 uklonjeni su iz skupa dok su svi ostali komentari zadržani. Ocjene 1 i 2 spojene su u novu klasu "12", a ocjene 4 i 5 u klasu "45".

filter1-3-5 Komentari koji sadrže ocjenu 2 ili 4 uklonjeni su iz skupa dok su svi ostali komentari zadržani. Na ovaj način dobiven je podskup s klasama "1", "3" i "5".

filter12-3-45 Ocjene 1 i 2 spojene su u klasu "12" te su ocjene 4 i 5 spojene u klasu "45".

filter1-234-5 Ocjene 2, 3 i 4 spojene su u klasu "234". Ovaj podskup sadrži klase "1", "234" i "5".

Gledajući kontekst ocjena, filtri s dvije klase *filter1-5* i *filter12-45* pokušavaju zapravo donijeti odluku je li primjer koji ispitujemo pozitivan ili negativan. Pozitivnu ocjenu predstavljaju klase "5" i "45", dok negativnu ocjenu predstavljaju klase "1" i "12". Slično tome, filtri s tri klase *filter1-3-5*, *filter12-3-45* i *filter1-234-5* pokušavaju donijeti odluku je li primjer koji ispitujemo pozitivan, neutralan ili negativan. Pozitivnu ocjenu predstavljaju klase "5" i "45", neutralnu klase "3" i "234", a negativnu klase "1" i "12".

Na ovaj način dobili smo pet sustava za određivanje osnovne polarnosti i jedan sustav za određivanje stupnja polarnosti. Za svaki korpus od mogućih šest potrebno je provesti evaluaciju sa svakim navedenim sustavom. To čini ukupno 36 sustava.

7.2 Evaluacija

Evaluacija sustava podijeljena je na tri dijela. Bit će prikazani rezultati određivanja osnovne polarnosti u dvije klase, određivanja osnovne polarnosti u tri klase i određivanja stupnja polarnosti. Potrebno je evaluirati značajke opisane u Odlomku 6.4 i procijeniti koliko su zasebno korisne za rješavanje ovog problema. Zatim značajke treba skupiti u manje grupe te potom grupe evaluirati. Jedan od grupiranja značajki je grupiranje po Tablici 6.2. Na taj način trebali bismo dobiti uvid koliko korištenje tehnologija obrade prirodnoga jezika utječe na rješavanje ovoga problema.

Izgrađeni sustavi evaluirani su koristeći nekoliko algoritama učenja. Variranje raznim algoritmima učenja omogućava nam provjeru koliko su značajke sustava robusne i provjeru da rezultati nisu vezani isključivo za određeni algoritam. Eksperimentirano je korištenjem algoritama za klasifikaciju i regresiju. Za klasifikaciju su korišteni algoritmi naivni Bayes (John i Langley, 1995) i LibSVM (Joachims, 1998; EL-Manzalawy i Honavar, 2005). Za regresiju je korišten algoritam LibSVR (Smola i Vapnik, 1997; EL-Manzalawy i Honavar, 2005). Izabran je linearan kernel za LibSVM i LibSVR.

Svi eksperimenti provedeni su korištenjem dodatnog skupa za testiranje. Skupovi za testiranje su bili iste one razdiobe po ocjenama kao i skupovi nad kojima je sustav bio treniran. Točnost sustava koji određuju osnovnu polarnost mjerena je preciznošću,

odzivom i F-mjerom. Odziv, preciznost i F-mjera za klasu C definirani su na sljedeći način. Odziv(engl. ***R**ecall*) udio je svih primjera klase C koji su točno klasificirani.

$$R(C) = \frac{|\text{primjeri klase } C \text{ točno klasificirani}|}{|\text{svi primjeri klase } C|} \tag{7.1}$$

Preciznost (engl. ***P**recision*) udio je primjera klasificiranih u klasu C koji zaista jesu klasa C.

$$P(C) = \frac{|\text{primjeri klase } C \text{ točno klasificirani}|}{|\text{svi primjeri klasificirani u } C|} \tag{7.2}$$

F-mjera je harmonijska sredina odziva i preciznosti.

$$F(C) = \frac{2 \times R(C) \times P(C)}{R(C) + P(C)} \tag{7.3}$$

Točnost sustava koji određuju stupanj polarnosti mjerena je srednjom apsolutnom greškom i korijenom srednje kvadratne pogreške. Srednja apsolutna pogreška (engl. ***M**ean **A**bsolute **E**rror*) srednja je vrijednost apsolutnih odstupanja od predviđene vrijednosti.

$$MEA = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| \tag{7.4}$$

Korijen srednje kvadratne pogreške (engl. ***R**oot **M**ean **S**quare **E**rror*) definiran je na sljedeći način:

$$RMSE = \sqrt{\frac{\sum_{i=n}^{n}(f_i - y_i)^2}{n}}, \tag{7.5}$$

gdje f_i predstavlja predviđenu vrijednost, dok y_i predstavlja pravu vrijednost. Kako je RMSE stroža mjera od MEA mjere, pri određivanju koji sustav regresije je bolji, najprije će biti uzeta u obzir mjera RMSE te mjera MEA ako mjere RMSE između dva sustava budu jednake.

7.2.1 Referentni algoritam

Da bismo bili u mogućnosti utvrditi koliko određene značajke imaju utjecaj na određivanje polarnosti odabran je referentni (engl. *baseline*) sustav koji kao značajku koristi samo unigrame riječi. Tablica 7.1 prikazuje evaluaciju referentnog sustava. Svaki redak tablice predstavlja rezultate referentnog sustava po svim ulaznim skupovima. Prva dva retka prikazuju evaluaciju određivanja osnovne polarnosti u dvije klase, a zadnja tri

retka evaluaciju određivanja osnovne polarnosti u tri klase. Podebljanim slovima prikazani su najbolji rezultati u određivanju polarnosti u dvije klase i određivanju polarnosti u tri klase za svaki korpus. Vidimo da najbolje rezultate u određivanju polarnosti u dvije klase uglavnom daje sustav *filter1-5*, a najbolje rezultate određivanja polarnosti u tri klase sustav *filter1-3-5*.

		market			uniform			sent_market			sent_uniform			pra_market			pra_uniform		
Sustav	C	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
filter1-5	NB	82.2	82.5	78.5	81.1	80.2	80.1	81.1	81.2	77.0	**84.1**	83.8	83.7	60.7	77.9	68.2	**67.4**	**66.7**	66.3
	SVM	**88.9**	**89.2**	**89.0**	**84.7**	**84.7**	**84.6**	**86.4**	**86.6**	**86.4**	84.0	**84.0**	**83.9**	74.9	**78.2**	**75.3**	67.3	**66.7**	**66.4**
filter12-45	NB	80.1	81.7	78.6	77.2	76.4	76.3	80.0	81.0	77.6	78.8	78.7	78.7	75.6	76.3	66.8	65.3	65.0	64.7
	SVM	85.0	85.3	85.1	79.0	79.0	79.0	83.7	84.0	83.8	78.5	78.5	78.5	72.3	75.9	72.7	64.4	64.1	63.9
filter1-3-5	NB	68.0	71.5	64.2	65.1	**64.6**	64.1	66.8	69.1	61.5	**67.6**	**67.7**	**67.2**	37.4	61.1	46.4	47.1	46.9	46.7
	SVM	**77.1**	**77.3**	**77.1**	65.0	**64.6**	**64.8**	**72.0**	**73.4**	**72.6**	64.8	64.8	64.8	56.2	60.8	57.1	47.0	46.8	46.7
filter12-3-45	NB	67.0	73.1	67.2	57.1	61.3	56.1	65.4	71.3	64.4	58.2	62.7	58.0	**56.3**	**64.8**	51.6	46.0	51.9	46.2
	SVM	74.2	74.2	74.2	60.3	59.9	60.1	70.3	72.5	71.2	60.3	60.8	60.5	55.9	62.9	**57.5**	47.3	49.8	47.7
filter1-234-5	NB	63.9	63.9	59.5	65.7	63.9	55.7	62.6	63.1	59.3	66.4	64.2	56.1	47.0	52.7	48.7	**64.9**	**60.3**	45.8
	SVM	65.8	66.0	65.8	63.6	64.3	63.8	63.6	64.3	63.8	63.0	63.7	63.2	51.3	52.8	51.4	54.1	58.7	**54.0**

Tablica 7.1: Evaluacija filtera nad unigramom riječi svih korpusa.

Uspoređujući odziv, preciznost i F-mjeru po svim filtrima vidljivo je da SVM daje bolje rezultate od naivnoga Bayesa u gotovo svim testovima. Gledajući F-mjeru, naivni Bayes daje bolje rezultate u 13% testova, odnosno u samo četiri testa. Značajno bolji, čak za 2.4% od SVM-a, samo je u *korpusu komentara* uniformne razdiobe (sent_-uniform) za sustav određivanja osnovne polarnosti u tri klase *filter1-3-5*.

Prosječna razlika F-mjere između naivnoga Bayesa i SVM-a za korpus *market* po svim filtrima iznosi 8.64%. Drugim riječima, SVM je od naivnoga Bayesa bolji 8.64% gledajući korpus *market*. U *uniform*nom korpusu SVM također prednjači za 4%. Za korpuse *sent_market*, *sent_uniform*, *pra_market* i *pra_uniform* SVM je bolji od naivnoga Bayesa redom za 7.6%, 1.44%, 6.46% i 1.8%. Gledajući sveukupno, SVM u prosjeku prednjači nad Naivnim Bayesom za sve referentne sustave za 4.99%. Nešto manje, no ipak zamjetno, SVM prednjači nad naivnim Bayesom i za ostale mjere: odzivom i preciznošću.

Korpusi stvarne distribucije ocjena imaju veće odstupanje u rezultatima naivnoga Bayesa i SVM-a od korpusa uniformne distribucije ocjena. SVM pred naivnim Bayesom u stvarnoj distribuciji ima veće poboljšanje od istih klasifikatora u uniformnoj distribuciji. To poboljšanje za *korpus komentara*, *korpus rečenica* i *korpus fraza* re-

dom iznosi 4.64%, 5.12% i 4.66%. Ako pak usporedimo rezultate naivnoga Bayesa na korpusu *market* s rezultatima naivnoga Bayesa na korpusu *uniform* vidjet ćemo da je korpus stvarne distribucije bolji za 3.14%. Naivni Bayes uniformne distribucije *korpusa rečenica* bolji je od stvarne distribucije istog korpusa za 0.78% dok je naivni Bayes uniformne distribucije *korpusa fraza* lošiji od stvarne distribucije *korpusa fraza* za 2.4%. Iz priloženih rezultata je vidljivo da naivni Bayes stvarne distribucije postiže bolje rezultate od naivnoga Bayesa uniformne distribucije za prosječno 1.59%. Preciznosti stvarne i uniformne distribucije skoro pa su jednaki, ali odziv stvarne distribucije veći je za 5.13%. Preciznost, odziv i F-mjera SVM-a stvarne distribucije bolja je od SVM-a uniformne distribucije redom za 6.23%, 7.53% i 6.74%. SVM i naivni Bayes stvarne distribucije bolji su od SVM-a i naivnoga Bayesa uniformne distribucije. Puno veće poboljšanje vidljivo je na odzivu, preciznosti i F-mjeru SVM-a. Tako je poboljšanje SVM-a među distribucijama veće od poboljšanja naivnoga Bayesa među distribucijama za 5.15%. Poboljšanja preciznosti i odziva SVM-a također su veća redom za 6.15% i 2.41% od poboljšanja naivnoga Bayesa.

Dva su očita razloga zašto su klasifikatori stvarne razdiobe bolji od klasifikatora uniformne razdiobe ocjena. Prirodna distribucija u filtrima s dvije klase ima veći broj komentara za učenje. Tako primjerice korpus *market* za filter1-5 sadrži 3231 komentar, dok korpus *uniform* istog filtra sadrži 2000 komentara. Što imamo više komentara za učenje, to nam uglavnom raste kvaliteta sustava. Također, kako se radi o stvarnoj distribuciji ocjena iz Tablice 4.2 i činjenici da su ulazni i testni skupovi jednakih distribucija, jasno je da će sustav koji je treniran i validiran s većim brojem klasa "5" imati veću točnost.

Analizirajući referentne sustave potvrđujemo rezultate ostalih radova (Pang et al., 2002; Pang i Lee, 2004), u kojima je otkriveno da SVM generalno daje bolje rezultate od ostalih klasifikatora. U daljnjem radu, prilikom evaluiranja određivanja osnovne polarnosti, sustavi će biti evaluirani algoritmom SVM.

7.2.2 Evaluacija sustava sa svim značajkama

Tablica 7.2 prikazuje rezultate evaluacije sustava učenih na svim značajkama. U prva dva retka prikazani su sustavi koji određuju osnovnu polarnost u dvije klase, dok su u zadnja tri retka prikazani sustavi koji određuju osnovnu polarnost u tri klase. *Filter1-5* postiže bolje rezultate u odzivu, preciznosti i F-mjeri od sustava *filter12-45*. Preciznost, odziv i F-mjera sustava *filter1-5* u prosjeku po svim korpusima bolje su od sustava *filter12-45* za redom 3.75%, 3.63% i 3.65%. Kod sustava s tri klase najveću točnost

postiže *filter1-3-5*. Odstupanje je vidljivo tek na korpusu *pra_uniform*, gdje najveću točnost, iako samo 52.6%, ima sustav *filter1-234-5*.

	market			uniform			sent_market			sent_uniform			pra_market			pra_uniform		
Sustav	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
filter1-5	91.0	91.1	91.0	86.2	86.2	86.1	87.9	88.2	88.0	85.0	85.0	84.9	75.8	78.1	76.4	66.0	66.0	65.9
filter12-45	87.2	87.5	87.3	82.3	82.3	82.2	85.2	85.6	85.4	79.3	79.3	79.3	72.1	74.8	72.9	63.3	63.3	63.3
filter1-3-5	78.6	79.6	79.0	69.0	68.8	68.9	74.6	76.1	75.1	66.3	66.5	66.4	56.5	60.2	57.7	45.7	45.9	45.8
filter12-3-45	75.5	77.3	76.3	63.3	64.4	63.8	72.1	74.7	73.1	61.1	61.9	61.4	56.1	61.0	57.8	46.8	48.4	47.4
filter1-234-5	66.6	66.7	66.6	65.2	66.1	65.3	64.9	65.2	65.0	63.2	64.3	63.5	50.2	51.2	50.5	51.7	55.4	52.6

Tablica 7.2: Evaluacija sustava sa svim značajkama.

Kao i kod referentnih sustava, sustavi trenirani na stvarnoj distribuciji ocjena postižu bolje rezultate nego sustavi trenirani na uniformnoj distribuciji. Poboljšanje stvarne nad uniformnom distribucijom u preciznosti, odzivu i F-mjeri redom iznosi 6.67%, 7.57% i 7.02%. Poboljšanje između distribucija sustava sa svim značajkama veće je od poboljšanja između distribucija referentnog sustava u preciznosti za 0.44%, u odzivu za 0.03% i F-mjeri za 0.28%.

Sustavi *market*, *uniform*, *sent_market*, *sent_uniform* i *pra_market* u F-mjeri redom su se poboljšali za 1.8%, 2.8%, 1.76%, 0.92% i 0.26%, dok se *pra_uniform* pogoršao za 0.74%. Vidimo da se najviše poboljšao korpus komentara, zatim korpus rečenica, a korpus fraza bilježi pogoršanje od 0.24%. Zanimljivo je kako sustav *pra_uniform* bilježi pogoršanje po svim filtrima.

Uspoređujući poboljšanja nad referentnim sustavom po filtrima, poboljšanja sustava *filter1-5*, *filter12-45*, *filter1-3-5*, *filter12-3-45* i *filter1-234-5* redom iznose 1.12%, 1.23%, 1.63%, 1.43% i 0.25%. Poboljšanja sustava koji određuju osnovnu polarnost u dvije klase veća su u odnosu na rezultat referentnog sustava od poboljšanja sustava koji određuju osnovnu polarnost u tri klase. Pomalo neočekivano, sustav *filter12-45* postiže veće poboljšanje od sustava *filter1-5*. Najveće poboljšanje u odnosu na referentni sustav napravio je *filter1-3-5*. Skup svih značajki najmanje je doprinio sustavu *filter1-234-5* koji bilježi i degradaciju odziva u odnosu na referentni sustav za 0.15%.

7.2.3 Frekvencijske vs. binarne značajke

Svaki komentar d predstavili smo kao vektor pobrojanih značajki $(n_1(d), \ldots, n_m(d))$. Svaka značajka $n_j(d)$ primjerice broji koliko puta se određena riječ j pojavila u komentaru d. U tome slučaju govorimo o frekvencijskom vektoru značajki. Ako samo

gledamo je li se određena riječ j pojavila u dokumentu ili nije, govorimo o binarnom vektoru značajki. Referentni sustav prikazan Tablicom 7.2 predstavlja unigrame riječi, odnosno osnovne binarne značajke. Zanima nas hoće li frekvencijski vektori postići veću točnost od binarnih vektora. Tablicom 7.3 prikazani su sustavi evaluirani nad značajkom frekvencijski unigram riječi.

	market			uniform			sent_market			sent_uniform			pra_market			pra_uniform		
Sustav	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
filter1-5	88.8	89.1	88.9	85.2	85.2	85.1	86.5	86.6	86.5	83.1	83.1	83.0	75.0	78.2	75.5	67.2	66.6	66.3
filter12-45	85.3	85.5	85.4	78.9	78.9	78.9	83.4	83.7	83.5	78.3	78.3	78.3	72.6	76.1	72.9	64.4	64.1	63.9
filter1-3-5	77.0	77.3	77.1	65.2	65.0	65.1	71.7	72.9	72.2	64.8	64.7	64.7	56.3	60.9	57.1	47.0	46.8	46.7
filter12-3-45	74.3	74.2	74.2	60.1	59.9	60.0	70.0	72.2	70.9	59.7	60.2	60.0	56.0	62.9	57.6	47.3	49.7	47.8
filter1-234-5	66.1	66.2	66.1	64.0	64.6	64.2	63.6	64.2	63.8	63.1	63.8	63.4	51.3	52.8	51.3	54.2	58.7	54.1

Tablica 7.3: Evaluacija filtera nad frekvencijskim unigramom riječi svih korpusa.

Frekvencijski unigrami pokazuju bolje rezultate nad unigramima (referentni sustav) u *korpusu komentara* za 0.15% (5 od 10 sustava), *korpusu fraza* za 0.05% (5 od 10 sustava) dok u *korpusu rečenica* pokazuju lošije rezultate za 0.24% (2 od 10 sustava). Po filtrima, frekvencijski unigrami pokazuju lošije rezultate nad referentnim sustavom po svim filtrima osim za *filter1-234-5* koji pokazuje poboljšanje od 0.15%. Od svih 30 sustava prikazanih u tablici, frekvencijski unigrami daju bolje rezultate u 12 sustava što čini 40%. Sustavi stvarne i uniformne razdiobe imaju svaki po 6 sustava u kojima frekvencijski unigrami daju bolje rezultate. *filter1-234-5* filter pokazuje poboljšanje u 4 od 6 sustava, a *filter1-5* poboljšanje u 3 od 6 sustava. Ostali filtri *filter12-45*, *filter1-3-5* i *filter12-3-45* redom pokazuju poboljšanje u dva sustava, jednom sustavu i dva sustava.

Odnos frekvencijskih bigrama i bigrama riječi vidljiv je u Tablici 7.4. *Korpus fraza* ne pokazuje nikakva poboljšanja rezultata između frekvencije bigrama i bigrama riječi. *Korpus rečenica* postiže bolje rezultate koristeći frekvenciju bigrama za 0.7% (5 od 10 sustava), dok *korpus komentara* također postiže bolje rezultate za 0.6% (4 od 10 sustava). Gledajući prosjek, frekvencijski bigrami bilježe bolje rezultate od bigrama za 0.03%. *Filter1-3-5* postiže jednake rezultate s bigramom i frekvencijom bigrama. *Filter1-5* postiže bolje rezultate za bigrame i to za 0.03%, dok *filter12-45*, *filter12-3-45* i *1-234-5* postižu bolje rezultate koristeći frekvencijski bigram i to redom za 0.07%, 0.08% i 0.05%.

Uspoređujući rezultate lematiziranih unigrama s rezultatima frekvencija lematiziranih unigrama iz Tablice 7.5 vidjet ćemo da frekvencijski vektor postiže bolje rezultate

	market			uniform			sent_market			sent_uniform			pra_market			pra_uniform		
Sustav	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
f1-5 bigram	85.9	86.5	85.0	81.8	81.7	81.7	83.3	84.1	82.2	77.6	77.4	77.3	60.7	77.9	68.2	25.0	50.0	33.3
f1-5 f.bigram	85.7	86.3	84.9	81.8	81.7	81.7	83.2	84.0	82.2	77.6	77.3	77.2	60.7	77.9	68.2	25.0	50.0	33.3
f12-45 bigram	83.1	84.2	82.7	78.3	78.3	78.2	80.5	81.8	79.9	75.1	75.1	75.1	57.8	76.0	65.6	25.0	50.0	33.3
f12-45 f.bigram	82.9	84.1	82.7	78.5	78.5	78.4	80.5	81.8	79.9	75.4	75.3	75.3	57.8	76.0	65.6	25.0	50.0	33.3
f1-3-5 bigram	72.7	75.6	72.6	64.2	64.5	64.2	69.0	72.0	68.3	58.7	59.0	58.5	37.4	61.1	46.4	11.1	33.3	16.7
f1-3-5 f.bigram	72.5	75.5	72.5	64.1	64.5	64.2	68.7	71.8	68.1	59.1	59.4	58.8	37.4	61.1	46.4	11.1	33.3	16.7
f12-3-45 bigram	71.2	74.9	71.9	59.3	61.8	60.1	67.0	71.7	67.6	56.3	58.4	57.0	41.7	64.6	50.7	16.0	40.0	22.9
f12-3-45 f.bigram	71.3	75.0	72.0	59.6	62.1	60.3	67.0	71.7	67.6	56.6	58.6	57.2	41.7	64.6	50.7	16.0	40.0	22.9
f1-234-5 bigram	61.8	62.6	61.4	62.4	64.1	62.0	60.1	61.0	59.5	60.0	62.1	60.0	20.3	45.0	28.0	36.0	60.0	45.0
f1-234-5 f.bigram	61.9	62.7	61.5	62.2	63.9	61.9	60.4	61.2	59.7	60.1	62.2	60.1	20.3	45.0	28.0	36.0	60.0	45.0

Tablica 7.4: Evaluacija bigrama i frekvencija bigrama riječi svih korpusa.

u čak 15 od 30 sustava. Najveće poboljšanje postiže u *korpusu komentara* i to 0.29% (7 od 10 sustava). *Korpus rečenica* i *korpus fraza* također postižu poboljšanja od 0.03% i 0.02%. Svi filtri osim *filter1-5* postižu bolje rezultate kada koriste frekvencijski vektor.

	market			uniform			sent_market			sent_uniform			pra_market			pra_uniform		
Sustav	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
f1-5 lema	89.3	89.4	89.3	85.1	85.1	85.1	87.3	87.3	87.3	84.3	84.3	84.2	75.4	78.1	76.0	66.6	66.2	66.0
f1-5 f.lema	89.2	89.3	89.2	85.5	85.5	85.4	87.1	87.0	87.0	83.8	83.8	83.7	75.5	78.2	76.1	66.6	66.2	66.0
f12-45 lema	85.8	85.9	85.8	79.5	79.5	79.5	84.4	84.4	84.4	78.9	78.9	78.9	73.3	76.4	73.8	64.5	64.5	64.4
f12-45 f.lema	86.0	86.2	86.1	80.1	80.1	80.1	84.3	84.4	84.3	78.9	78.9	78.9	73.4	76.4	73.8	64.5	64.5	64.4
f1-3-5 lema	77.8	77.6	77.6	66.3	65.9	66.0	72.4	73.9	73.0	64.6	64.7	64.6	56.3	60.7	57.3	47.1	47.1	47.0
f1-3-5 f.lema	77.6	77.6	77.6	67.3	66.9	67.0	72.7	73.9	73.2	64.7	64.8	64.7	56.5	60.8	57.5	47.1	47.1	46.9
f12-3-45 lema	74.6	74.6	74.6	60.8	60.9	60.9	70.5	72.8	71.5	60.2	61.1	60.6	56.8	63.4	58.3	47.5	50.2	48.2
f12-3-45 f.lema	75.1	74.9	75.0	61.3	61.4	61.3	70.5	72.9	71.5	60.3	61.3	60.8	56.8	63.4	58.4	47.4	50.2	48.1
f1-234-5 lema	66.6	66.9	66.7	64.3	64.9	64.5	64.8	65.4	65.0	63.6	64.6	63.9	51.1	52.4	51.2	54.3	58.8	54.1
f1-234-5 f.lema	66.5	66.7	66.5	64.5	65.2	64.7	65.3	65.8	65.5	63.8	64.8	64.1	51.1	52.4	51.2	54.3	58.8	54.1

Tablica 7.5: Evaluacija lematiziranih unigrama i frekvencija lematiziranih unigrama riječi svih korpusa.

Sljedećim dvjema tablicama prikazana je usporedba frekvencijskih i nefrekvencijskih značajki. Frekvencijske značajke su sve one značajke iz Tablice 6.2 koje broje koliko puta se neka značajka pojavila u primjeru. To su frekvencija unigrama, bigrama ili trigrama riječi; frekvencija lematiziranog unigrama, bigrama ili trigrama riječi; frekvencija

unigrama, bigrama ili trigrama vrste riječi; frekvencija SentiWordNet-riječi; frekvencija SentiWordNet-riječi s razrješenom višeznačnosti; frekvencija relacija, roditelj-dijete ili relacija-roditelj-dijete sintaktičkog stabla. Nefrekvencijske značajke su binarne značajke iz Tablice 6.2 koje označavaju prisustvo ili odsustvo svojstva koje bilježe. Sve mjere u Tablici 7.6 F-mjere. Stupac I1 (engl. *include*) označava prisustvo frekvencijskih značajki. To je sustav treniran samo sa frekvencijskim značajkama. Stupac I2 označava prisustvo samo svih nefrekvencijskih značajki. Stupac I12 označava prisustvo i frekvencijski i nefrekvencijskih značajki.

	market			uniform			sent_market			sent_uniform			pra_market			pra_uniform		
Sustav	I1	I2	I12	I1	I2	I12	I1	I2	I12	I1	I2	I12	I1	I2	I12	I1	I2	I12
filter1-5	90.6	90.4	90.5	86.2	85.8	86.2	87.5	87.9	87.5	84.7	85.1	85.4	75.9	75.9	75.9	66.8	66.7	66.8
filter12-45	87.3	87.2	87.2	82.2	82.5	82.2	85.2	84.9	85.1	79.7	79.4	79.4	73.3	73.3	73.3	64.0	64.0	64.0
filter1-3-5	78.8	78.5	78.8	68.6	68.6	68.9	74.5	74.5	74.4	66.7	66.5	66.9	57.4	57.5	57.4	47.9	47.8	47.9
filter12-3-45	76.3	76.2	76.3	63.7	63.7	63.8	72.5	72.2	72.3	61.4	61.5	61.7	57.9	57.9	57.9	47.4	47.4	47.4
filter1-234-5	66.6	66.7	67.3	66.1	65.6	66.2	66.1	65.4	66.2	64.7	64.2	64.4	51.6	51.6	51.6	53.8	53.8	53.8

Tablica 7.6: Usporedba dodavanja frekvencijskih i nefrekvencijskih značajki.

Za korpus komentara, korpus rečenica i korpus fraza frekvencijske značajke postižu bolje rezultate redom za 0.12%, 0.14% i 0.01%. Neobično, no frekvencijske značajke najviše su pridonijele korpusu rečenica. Po filtrima frekvencijske značajke doprinose više od nefrekvencijskih najviše za filter1-234-5 i to za 0.27%. Ostali filtri također su pobošljani, svi osim filter1-5 koji bilježi 0.02% lošije rezultate za frekvencijske značajke. U prosjeku, frekvencijske značajke postižu 0.09% bolje rezultate od nefrekvencijskih značajki.

Zanimljivo za primjetiti, dodavanje nefrekvencijskih značajki na frekvencijske u nekim sustavima pogoršava rezultate. Tako je primjerice za filter1-5 korpusa market, gdje je dodavanjem nefrekvencijskih značajki F-mjera narušena za 0.1%, a za filter12-45 korpusa sent_uniform za 0.3%. Za filter12-45 svih korpusa dodavanje nefrekvencijskih značajki na frekvencijske pogoršava rezultate za 0.08%. Kod ostalih filtera rezultati su poboljšani dodavanjem nefrekvencijskih značajki. Dodavanje nefrekvencijske značajke na frekvencijske prosječno poboljšava F-mjeru za 0.04%. Frekvencijske značajke postižu bolje rezultate od referentnog algoritma za prosječno 1.35%. Frekvencijske značajke postižu bolje rezultate čak i od sustava u kojemu su uključene sve značajke, iz Tablice 7.2 i to za 0.22%. Tome najviše pridonose varirajući rezultati korpusa fraza kod kojih dodavanje svih značajki značajno narušava performanse.

Stupci E1, E2 i E12 (engl. *exclude*) u Tablici 7.7 redom označavaju odsustvo frekvencijskih značajki, odsustvo nefrekvencijskih značajki i odsustvo i frekvencijskih i nefrekvencijskih značajki od svih značajki iz Tablice 7.2. Na ovaj način pokušat ćemo procijeniti i koliki utjecaj imaju frekvencijske i nefrekvencijske značajke na ostale značajke sustava.

	market			uniform			sent_market			sent_uniform			pra_market			pra_uniform		
Sustav	E1	E2	E12	E1	E2	E12	E1	E2	E12	E1	E2	E12	E1	E2	E12	E1	E2	E12
filter1-5	90.7	90.7	89.1	85.1	85.7	81.6	87.8	87.9	86.0	84.2	84.1	82.4	76.2	76.1	74.9	65.8	65.8	65.0
filter12-45	86.9	86.9	85.8	81.7	81.6	76.6	85.0	85.3	82.4	79.0	79.0	74.2	72.3	72.3	71.2	63.0	63.1	62.5
filter1-3-5	78.7	78.8	76.7	67.6	68.0	63.9	74.4	74.6	71.2	65.2	65.5	62.7	57.1	57.1	55.5	45.4	45.4	44.4
filter12-3-45	75.9	76.0	74.1	63.3	63.4	58.8	72.5	72.7	69.4	61.4	61.0	56.2	57.3	57.3	56.0	47.1	47.2	46.3
filter1-234-5	66.0	66.0	61.5	64.8	64.9	60.7	63.9	63.6	59.2	62.5	62.8	58.3	50.1	50.1	49.1	52.3	52.3	52.1

Tablica 7.7: Usporedba izbacivanja frekvencijskih i nefrekvencijskih značajki.

Izbacivanje frekvencijskih značajki šteti performansama sustava više nego izbacivanje nefrekvencijskih i to za 0.07%. Izbacivanje frekvencijskih značajki u usporedbi s izbacivanjem nefrekvencijskih značajki najviše šteti korpusu komentara uniformne razdiobe (uniform), čak 0.22% u F-mjeri. Svi filtri postižu lošije rezultate izbacivanjem frekvencijskih značajki nego izbacivanjem nefrekvencijskih. Iz Tablice 7.6 možemo vidjeti da filter1-5 postiže bolje rezultate korištenjem nefrekvencijskih značajki, što kod izbacivanja nije slučaj. U prosjeku, izbacivanjem frekvencijskih značajki postižu se za 0.07% lošiji rezultati nego kada se izbace nefrekvencijske značajke. Ako usporedimo ovaj rezultat s ubacivanjem frekvencijskih i nefrekvencijskih značajki, vidjet ćemo da je njihova razlika manja korištenjem ostalih značajki (dakle kod izbacivanja). Možemo reći da među ostalim značajkama postoje one koje imaju svojstva frekvencijskih, ali tek u maloj mjeri, poboljšavajući rezultate za 0.02%. Izbacimo li frekvencijske značajke, točnost sustava svih značajki pada za 0.52%. Izbacimo li nefrekvencijske značajke, točnost sustava pada za 0.45%. Izbacimo li i frekvencijske i nefrekvencijske značajke, točnost sustava pada za 3.04%. Najveći pad vidljiv je na korpusu komentara i korpusu rečenica, dok je nešto manji pad F-mjere za korpusu fraza.

Pokazano je da frekvencijski unigrami pokazuju poboljšanja po nekim sustavima, no generalno ne doprinose kvaliteti rješenja, što je u skladu s rezultatima objavljenim u Pang et al. (2002). Frekvencijski bigrami ne pokazuju značajno poboljšanje u usporedbi s bigramima. Slično, frekvencijski lematizirani unigrami pokazuju malo poboljšanje u odnosu s lematiziranim unigramima riječi. Zadnje dvije tablice pokazuju nešto opširnija

mjerenja. Pokazano je da korištenje većeg broja frekvencijskih značajki ipak poboljšava rezultate naspram korištenja nefrekvencijskih značajki, što se kosi s Pang et al. koji su koristili samo frekvencijske unigrame. Ovo otkriće potvrđuje nam da ne bismo trebali ignorirati frekvencijske značajke u zadacima određivanja osnovne polarnosti teksta.

7.2.4 Osnovne vs. lematizirane značajke

Želimo vidjeti kakav je odnos značajki iz grupe *osnovna* i značajki iz grupe *lema* Tablice 6.2. Vidimo da je *lema* grupa sadrži lematizirane značajke grupe *osnovna*[1]. Ovime ćemo pokušati dati odgovor na pitanje koliko postupak lematizacije utječe na kvalitetu određivanja osnovne polarnosti.

Stupci I1, I2 i I3 Tablice 7.8 označavaju redom prisustvo osnovne grupe značajki, prisustvo lematizirane grupe značajki i prisustvo i osnovne i lematizirane grupe. Sve mjere u tablici su F-mjere.

	market			uniform			sent_market			sent_uniform			pra_market			pra_uniform		
Sustav	I1	I2	I12	I1	I2	I12	I1	I2	I12	I1	I2	I12	I1	I2	I12	I1	I2	I12
filter1-5	89.6	90.5	90.5	86.0	87.3	87.0	87.5	88.9	88.4	84.8	86.2	85.8	75.5	76.2	76.3	66.6	65.4	66.3
filter12-45	86.5	87.4	87.5	82.0	82.7	82.4	85.1	85.8	85.5	80.2	80.5	80.4	73.0	73.6	73.4	63.6	64.5	64.5
filter1-3-5	78.2	79.1	78.9	68.9	69.9	70.1	74.5	75.0	75.2	67.2	67.9	67.9	57.1	57.0	57.5	46.6	47.0	46.9
filter12-3-45	75.3	76.4	76.5	63.3	63.8	64.0	72.2	72.8	72.6	62.0	62.4	62.7	58.2	58.5	58.5	47.6	48.2	48.1
filter1-234-5	67.3	67.9	67.8	66.2	66.2	67.0	65.7	65.4	65.7	64.6	64.8	65.0	50.7	51.6	51.1	53.9	54.3	54.3

Tablica 7.8: Usporedba dodavanja osnovne grupe značajki i lematiziranih značajki.

Korištenje lematiziranih značajki bolje je od korištenja značajki osnovne grupe. Korpus fraza bilježi 0.79% bolje rezultate, korpus rečenica 0.59%, a korpus fraza 0.35% bolje rezultate. Svi filtri bilježe poboljšanje od minimalno 0.3% za filter1-234-5 do maksimalno 0.75% za filter1-5. Filtri koji klasificiraju u dvije klase imaju veću korist korištenjem lematiziranih značajki od filtera koji klasificiraju u tri klase. Bolji su u F-mjeri za 0.23%. Dodavanje lematiziranih značajki na značajke osnovne grupe poboljšava F-mjeru za 0.6%, dok dodavanje značajki osnovne grupe na lematizirane značajke poboljšava rezultate za tek 0.02%. Dodatno, dodavanje značajki osnovne grupe na lematizirane značajke pogoršava rezultate filtera koji klasificiraju u dvije grupe, dok su filtri koji klasificiraju u tri grupe neznačajno poboljšani. Vidimo da lematizirane značajke pokazuju bolje rezultate od značajki osnovne grupe pa čak i dodavanje značajki osnovne grupe na lematizirane značajke minimalno poboljšava rezultate. Osnovna

[1]Izuzevši značajke o poziciji riječi i broju rečenica.

grupa značajki pokazuje bolje rezultate od referentnog sustava za 1.17%, dok lematizirana grupa pokazuje bolje rezultate za 1.74%. Najmanje poboljšanje osnovne grupe značajki u odnosu na referentni sustav postiže filter1-5 i to 0.73%, a najviše filter1-3-5 za 1.57%. Slična situacija je za lematizirane značajke gdje najmanje poboljšanje postiže filter1-234-5 i to 1.37%, a najviše filter1-3-5, čak 2.13%. U odnosu na sustav sa svim značajkama, korpusi stvarne razdiobe osnovne grupe značajki postižu lošije rezultate za 0.38%, a korpusi uniformne razdiobe iste grupe postižu bolje rezultate za 0.45%. Najveće poboljšanje u odnosu na sustav sa svim značajkama bilježi filter1-234-5, 0.82% u F-mjeri. Lematizirane značajke postižu 0.61% bolje rezultate od sustava sa svim značajkama. Najveće poboljšanje bilježe korpusi uniformne razdiobe, koji su bolji za 0.95% od sustava sa svim značajkama. Korpusi stvarne razdiobe bolji su za 0.27% od sustava sa svim značajkama. Najveće poboljšanje sustava s lematiziranim značajkama u odnosu na sustav sa svim značajkama bilježi filter1-234-5. To poboljšanje iznosi 1.12%.

Stupci E1, E2 i E12 u Tablici 7.9 prikazuju redom odsustvo osnovne grupe značajki, odsustvo lematiziranih značajki i odsustvo obje navedene grupe. Zanima nas hoće li lematizirane značajke jednako dominirati koristeći i ostale značajke sustava.

	market			uniform			sent_market			sent_uniform			pra_market			pra_uniform		
Sustav	E1	E2	E12	E1	E2	E12	E1	E2	E12	E1	E2	E12	E1	E2	E12	E1	E2	E12
filter1-5	90.7	90.8	89.8	85.4	85.3	84.6	87.5	87.8	86.3	83.8	84.6	82.4	76.0	76.1	74.9	66.0	65.7	65.0
filter12-45	87.0	86.9	86.1	81.9	81.6	80.7	85.3	84.9	84.5	78.6	78.5	76.7	72.4	72.4	71.6	63.2	63.0	62.5
filter1-3-5	78.7	78.7	78.0	68.6	67.8	66.8	74.7	74.5	73.4	65.2	65.8	63.8	57.2	57.3	56.1	45.8	45.4	44.6
filter12-3-45	76.0	76.0	75.0	63.4	63.4	62.1	72.6	72.4	72.0	60.8	61.0	58.9	57.3	57.4	56.6	47.2	47.3	46.2
filter1-234-5	65.7	66.3	64.2	65.1	65.0	63.7	63.9	64.2	62.2	61.8	62.9	60.4	50.7	50.2	49.4	52.3	52.1	51.5

Tablica 7.9: Usporedba izbacivanja osnovne grupe značajki i izbacivanja lematiziranih značajki.

Izbacivanje osnovne grupe više je naštetilo sustavu nego izbacivanje lematizirane grupe, ali tek za 0.02% više. Time vidimo da su ostale značajke koje se koriste u sustavu svih značajki smanjile onu dominaciju kakvu su imale lematizirane značajke. Izbacivanje i osnovne i lematizirane grupe najviše šteti korpusu sent_uniform. Kvaliteta tih sustava u slučaju izbacivanja osnovne i lematizirane skupine pada za 2.66%. Kvaliteta sustava sent_market pada za 1.64%. Kvaliteta korpusa komentara i korpusa fraza u prosjeku pada redom za 1.55% i 1.19%. Korpus filter1-234-5 bilježi najveći gubitak kvalitete ako se izbace obje grupe iz sustava svih značajki. Filtri koji klasificiraju

u tri klase u tome slučaju bilježe 0.27% veći gubitak od filtera koji klasificiraju u dvije klase.

Pokazali smo da korištenje lematiziranih značajki značajno poboljšava rezultate u odnosu na osnovnu grupu značajki. Također, ti rezultati bolji su i od referentnog sustava i od sustava koji koristi sve značajke sustava. Lematizirane značajke najviše su doprinjele kvaliteti sustava s korpusima uniformne razdiobe: uniform, sent_uniform i pra_uniform. Filtri koji određuju osnovnu polarnost u dvije klase pokazuju veću zavisnost o lematiziranim značajkama od filtera koji određuju osnovnu polarnost u tri klase.

7.2.5 Značajke vrsta riječi

U ovome odjeljku dati ćemo odgovor na pitanje koliki utjecaj na određivanje osnovne polarnosti imaju značajke vrsta riječi iz Tablice 7.2. Da bi dobili bolji uvid u njihov utjecaj, usporedit ćemo ih s značajkama prve dvije skupine iz navedene tablice. *Osnovna2* grupa neka bude unija značajki osnovne grupe i lematizirane grupe značajki. U Tablici 7.10 prikazani su rezultati korištenja osnovne2 grupe značajki i značajki vrsta riječi. Stupci I1, I2 i I12 označavaju redom korištenje osnovne2 grupe značajki, korištenje značajki vrsta riječi i korištenje obje navedene grupe. Sve mjere u tablici su F-mjere.

	market			uniform			sent_market			sent_uniform			pra_market			pra_uniform		
Sustav	I1	I2	I12	I1	I2	I12	I1	I2	I12	I1	I2	I12	I1	I2	I12	I1	I2	I12
filter1-5	90.5	80.5	90.5	87.0	70.0	86.2	88.4	76.8	87.4	85.8	70.5	84.8	76.3	69.5	76.1	66.3	51.1	65.6
filter12-45	87.5	76.8	87.1	82.4	67.4	81.8	85.5	74.0	84.7	80.4	64.7	79.4	73.4	65.6	73.4	64.5	52.8	64.3
filter1-3-5	78.9	67.2	78.3	70.1	51.8	68.9	75.2	61.9	74.0	67.9	51.3	66.9	57.5	49.2	56.9	46.9	31.6	46.8
filter12-3-45	76.5	65.7	76.1	64.0	50.6	63.4	72.6	61.6	72.3	62.7	48.2	61.2	58.5	50.6	58.6	48.1	37.6	48.1
filter1-234-5	67.8	54.7	66.5	67.0	53.7	65.8	65.7	52.6	65.7	65.0	54.1	63.7	51.1	43.4	51.1	54.3	45.4	53.8

Tablica 7.10: Usporedba korištenja osnovne2 grupe značajki i značajki vrsta riječi.

Na prvi je pogled očito da grupa značajki osnovna2 poprima puno bolje rezultate naspram značajki vrsta riječi. Razlika njihove kvalitete u F-mjeri iznosi 12.23%. Iz ovih rezultata je jasno da značajke vrsta riječi nije potrebno dodatno uspoređivati s grupom značajki osnovna2 . Minimalnu točnost grupe značajki osnovna2, s F-mjerom 61.82%, poprima filter1-234-5. Maksimalnu točnost iste skupine značajki, s F-mjerom 82.38%, očekivano poprima filter1-5. Kod značajki vrsta riječi minimalna vrijednost za filter1-234-5 iznosi 50.65%, a maksimalna za filter1-5 iznosi 69.73%. Razmotrimo

kako će se ponašati sustavi kada su obje grupe značajki uključene. Dodavanje značajki vrsta riječi na skupinu značajki osnovna2 u prosjeku pogoršava F-mjeru za 0.61%. Na pogoršanje najvećim dijelom utjecali su korpusi uniformne razdiobe s F-mjerom od 0.78%, a manjim dijelom korpusi stvarne razdiobe od 0.45%. Značajke vrsta riječi lošije su od referentnog sustava od minimalno 9.48% za filter12-3-45 do maksimalno 11.68% za filter1-3-5.

U Tablici 7.11 prikazani su rezultati isključivanja skupine značajki osnovna2 i značajki vrsta riječi od skupa svih značajki. Stupci E1, E2 i E12 označavaju redom izbacivanje skupine značajki osnovna2, izbacivanje značajki vrsta riječi i izbacivanje obje skupine značajki.

	market			uniform			sent_market			sent_uniform			pra_market			pra_uniform		
Sustav	E1	E2	E12	E1	E2	E12	E1	E2	E12	E1	E2	E12	E1	E2	E12	E1	E2	E12
filter1-5	89.8	91.3	90.1	84.6	85.8	83.5	86.3	88.8	87.1	82.4	84.6	82.2	74.9	76.6	75.5	65.0	66.8	65.3
filter12-45	86.1	87.4	86.3	80.7	82.0	79.2	84.5	85.7	84.3	76.7	79.5	77.1	71.6	72.8	71.5	62.5	63.5	62.9
filter1-3-5	78.0	79.7	77.9	66.8	68.4	66.2	73.4	75.8	73.2	63.8	66.3	64.3	56.1	57.8	55.9	44.6	45.5	44.7
filter12-3-45	75.0	76.7	74.9	62.1	63.9	61.4	72.0	72.8	71.2	58.9	61.5	59.3	56.6	57.4	56.7	46.2	47.3	46.4
filter1-234-5	64.2	67.4	64.7	63.7	65.3	62.3	62.2	64.6	62.5	60.4	63.4	60.7	49.4	50.4	49.2	51.5	52.5	51.5

Tablica 7.11: Usporedba izbacivanja grupe značajki osnovna2 i značajki vrsta riječi.

Izbacivanje grupe značajki osnovna2 naštetilo je kvaliteti sustava za 1.72% manje nego izbacivanje značajki vrsta riječi. Razlika u informacijskoj vrijednosti ove dvije grupe smanjila se kod ubacivanja sa 12.23% na 1.72% kod izbacivanja. Izbacivanje obje značajke, u usporedbi sa sustavom svih značajki, pogoršava F-mjeru sustava za 1.70%, a najveće pogoršanje od 2.1% vidljivo je za filter1-234-5. Prilikom izbacivanja obje skupine značajki najveću degradaciju u usporedbi s referentnim sustavom bilježi filter1-234-5. To pogoršanje iznosi 1.85%.

Pokazali smo da su nam značajke vrsta riječi doprinjele degradaciji performansi sustava.

7.2.6 Značajke vrsta riječi vs. značajke sintaktičkog stabla

U ovome odjeljku osvrnut ćemo se na usporedbu između značajka vrsta riječi i značajka sintaktičkog stabla. U Tablici 7.12, s oznakom I1 prikazano je dodavanje samo značajki vrsta riječi, s oznakom I2 prikazano je dodavanje samo značajki sintaktičkog stabla i s oznakom I12 prikazano je dodavanje i značajka vrsta riječi i značajki sintaktičkog stabla.

	market			uniform			sent_market			sent_uniform			pra_market			pra_uniform		
Sustav	I1	I2	I12	I1	I2	I12	I1	I2	I12	I1	I2	I12	I1	I2	I12	I1	I2	I12
filter1-5	80.5	83.5	84.6	70.0	75.9	77.1	76.8	81.6	82.3	70.5	76.7	75.9	69.5	68.3	68.9	51.1	51.0	53.3
filter12-45	76.8	81.3	81.8	67.4	73.1	74.6	74.0	79.4	80.3	64.7	71.6	71.6	65.6	66.0	67.0	52.8	53.7	52.7
filter1-3-5	67.2	71.1	72.3	51.8	57.3	59.3	61.9	66.7	68.5	51.3	58.3	58.1	49.2	48.3	48.8	31.6	35.2	33.6
filter12-3-45	65.7	70.2	71.0	50.6	55.8	57.7	61.6	66.3	67.3	48.2	54.2	54.6	50.6	51.1	51.8	37.6	38.4	37.8
filter1-234-5	54.7	60.7	60.7	53.7	60.5	59.9	52.6	57.8	60.1	54.1	58.9	59.1	43.4	42.3	43.8	45.4	45.2	45.6

Tablica 7.12: Usporedba korištenja značajki vrsta riječi i značajki sintaktičkog stabla.

Značajke sintaktičkog stabla pokazuju 3.65% bolje rezultate od značajki vrsta riječi. Poboljšanje za korpus komentara iznosi 5.1%, za korpus rečenica 5.58%, a za korpus fraza samo 0.27%. Poboljšanje je podjednako raspoređeno po filtrima; od minimalnog 3.1% za filter1-5 do maksimalnog 3.98% za filter1-3-5. Referentni sustavi postižu bolje rezultate od sustava sa značajkama sintaktičkog stabla za prosječno 6.82%. Najveća poboljšanja od sustava koji klasificiraju u dvije klase postiže filter1-5, dok najveće poboljšanje od sustava koji klasificiraju u tri klase postiže filter1-3-5. Njihova poboljšanja redom iznose 8.1% i 7.7%.

Dodavanjem značajki sintaktičkog stabla na značajke vrsta riječi poboljšava se F-mjera za 4.31%. Korpus komentara bilježi poboljšanje od 6.06%, korpus rečenica poboljšanje od 6.21%, a korpus fraza od 0.65%. Prilikom dodavanja značajki vrsta riječi na značajke sintaktičkog stabla poboljšanje za F-mjeru iznosi tek 0.66%.

Tablica 7.13 prikazuje rezultate izbacivanja značajki sintaktičkog stabla i značajki vrsta riječi. Stupci E1, E2 i E12 označavaju redom izbacivanje značajki vrsta riječi, izbacivanje značajki sintaktičkog stabla i izbacivanje obiju ovih značajki iz sustava svih značajki.

	market			uniform			sent_market			sent_uniform			pra_market			pra_uniform		
Sustav	E1	E2	E12	E1	E2	E12	E1	E2	E12	E1	E2	E12	E1	E2	E12	E1	E2	E12
filter1-5	91.3	90.8	91.3	85.8	86.0	85.1	88.8	87.6	88.2	84.6	84.7	84.2	76.6	76.6	76.1	66.8	65.9	66.3
filter12-45	87.4	87.2	87.2	82.0	81.8	81.7	85.7	85.4	85.3	79.5	79.3	79.4	72.8	72.5	72.4	63.5	62.8	63.1
filter1-3-5	79.7	78.8	79.3	68.4	68.6	67.8	75.8	74.6	74.9	66.3	65.9	65.7	57.8	57.5	57.0	45.5	45.5	44.7
filter12-3-45	76.7	76.1	76.2	63.9	63.2	63.3	72.8	72.8	72.6	61.5	61.3	61.6	57.4	57.5	57.4	47.3	47.1	47.2
filter1-234-5	67.4	66.0	66.5	65.3	65.0	65.0	64.6	64.0	64.3	63.4	62.7	63.0	50.4	49.8	50.0	52.5	52.2	52.2

Tablica 7.13: Usporedba izbacivanja značajki vrsta riječi i značajki sintaktičkog stabla.

Izbacivanje značajki vrsta riječi pokazuje rezultate bolje od referentnog algoritama

za prosječno 1.22%. Izbacivanje značajki sintaktičkog stabla također pokazuje rezultate bolje od referentnog algoritama za prosječno 0.81%. Izbacivanje značajki vrsta riječi za 0.41% više povećava F-mjeru od izbacivanja značajki sintaktičkog stabla. Povećanje je najviše vidljivo na korpusima market i sent_market, gdje ono prosječno iznosi 0.69%. Na ostalim korpusima poboljšanje je prosječno 0.28%. Izbacivanje obiju značajki postiže bolje rezultate od sustava sa svim značajkama u korpusu market za 0.06%. U ostalim korpusima bilježe se lošiji rezultati. Korpus uniform lošiji je za 0.68%, korpus sent_-market i sent_uniform za redom 0.26% i 0.32%, a korpus pra_market i pra_uniform za redom 0.48% i 0.3%. Izbacivanje samo značajki vrsta riječi poboljšava rezultate u odnosu na sustav sa svim značajkam za 0.09%. Najveća poboljšavanja vidljiva su opet u korpusu market za 0.46% i korpusu sent_market za 0.22%. Izbacivanjem samo značajki sintaktičkog stabla postižu se lošiji rezultati od sustava sa svim značajkama.

Možemo zaključiti da značajke vrsta riječi narušavaju performanse sustava kada su sve značajke iz Tablice 7.2 uključene. Sustav bez obiju ovih značajki postiže tek nešto niže rezultate, dok s uključenim značajkama sintaktičkog stabla bilježi malo poboljšanje.

7.2.7 SentiWordNet vs. SentiWordNet + razrješavanje višeznačnosti

Ovim testom cilj nam je pokazati koliki utjecaj ima razrješavanje višeznačnosti na zadatak određivanja osnovne polarnosti. I1, I2 i I12 stupci Tablice 7.14 prikazuju značajke bez razrješavanja višeznačnosti, značajke s razrješavanjem višeznačnosti i korištenje obiju značajki.

	market			uniform			sent_market			sent_uniform			pra_market			pra_uniform		
Sustav	I1	I2	I12	I1	I2	I12	I1	I2	I12	I1	I2	I12	I1	I2	I12	I1	I2	I12
filter1-5	87.4	83.1	88.5	81.4	74.5	80.9	84.5	80.1	85.0	77.0	75.2	80.3	74.2	72.7	74.8	63.0	57.8	64.5
filter12-45	83.6	80.4	84.7	76.1	71.8	75.7	80.9	78.7	80.9	74.1	69.8	73.4	72.7	70.5	71.2	62.3	57.7	61.6
filter1-3-5	74.4	70.2	75.4	62.4	56.5	63.1	69.0	66.5	70.2	59.1	56.9	60.9	55.5	53.3	55.1	43.5	39.0	43.3
filter12-3-45	72.1	68.2	72.8	57.4	55.0	57.4	67.9	67.0	68.3	55.7	52.6	55.5	57.8	55.3	55.9	45.7	42.6	44.8
filter1-234-5	61.3	57.3	60.8	59.7	56.4	59.9	57.2	56.4	58.0	57.6	58.5	57.8	49.0	45.4	48.7	52.7	50.8	51.9

Tablica 7.14: Evaluacija SentiWordNet značajki bez i uz korištenje razrješavanja višeznačnosti.

Iz gornje tablice vidljivo je da sustav koji ne koristi razrješavanje višeznačnosti postiže bolje rezultate od sustava koji ga koristi. U prosjeku, F-mjera sustava koji ne koristi razrješavanje višeznačnosti veća je za 3.17%. Jedini sustav u kojemu je razrješavanje višeznačnosti postiglo bolje rezultate je za filter1-234-5 korpusa sent_uniform.

Kod tog sustava F-mjera pri korištenju razrješavanja višeznačnosti veća je za 0.9% od istog sustava koji ne koristi razrješavanje višeznačnosti. Korištenje unije značajki SentiWordNet-a koji koriste razrješavanje višeznačnosti i značajki SentiWordNet-a koji ne koriste razrješavanje višeznačnosti povećava F-mjeru za 0.2%. Ovo povećanje osobito je vidljivo kod sustava filter1-5 gdje iznosi 1.08% i sustava filter1-3-5 gdje iznosi 0.68%.

Neočekivano, pokazalo se da razrješavanje višeznačnosti narušava performanse sustava.

7.2.8 SentiWordNet + razrješavanje višeznačnosti vs. MPQA

Tablicom 7.15 prikazana je usporedba značajki SentiWordNet s razrješavanjem višeznačnosti i MPQA-značajki.

	market			uniform			sent_market			sent_uniform			pra_market			pra_uniform		
Sustav	I1	I2	I12	I1	I2	I12	I1	I2	I12	I1	I2	I12	I1	I2	I12	I1	I2	I12
filter1-5	83.1	82.8	85.2	74.5	78.1	77.3	80.1	78.8	81.4	75.2	76.6	76.7	72.7	74.0	73.9	57.8	59.6	61.6
filter12-45	80.4	79.7	81.8	71.8	74.4	74.4	78.7	76.2	79.2	69.8	72.9	71.2	70.5	71.5	71.8	57.7	56.4	60.8
filter1-3-5	70.2	69.3	71.1	56.5	58.4	58.4	66.5	64.8	67.4	56.9	56.7	57.8	53.3	53.0	54.6	39.0	39.2	41.1
filter12-3-45	68.2	68.4	69.6	55.0	55.9	56.7	67.0	62.8	66.9	52.6	54.3	54.2	55.3	55.8	56.7	42.6	41.1	45.4
filter1-234-5	57.3	61.0	58.6	56.4	59.6	58.2	56.4	56.6	57.2	58.5	55.2	58.0	45.4	46.5	47.2	50.8	51.8	51.3

Tablica 7.15: Usporedba korištenja značajki SentiWordNet + razrješavanje višeznačnosti i MPQA-značajki.

MPQA-značajke u prosjeku daju bolje rezultate od SentiWordNet-značajki s razrješenom višeznačnosti za 0.37%. Poboljšanje je najveće za korpus komentara i iznosi 1.42%, zatim za korpus fraza i iznosi 0.38%, dok je za korpus rečenica F-mjera manja za 0.68%. Koristeći i SentiWordNet i MPQA-rječnik F-mjera raste za 1.52% u usporedbi sa sustavom kada se koristi samo SentiWordNet i razrješavanje višeznačnosti.

Rječnik MPQA postiže manju točnost od rječnika SentiWordNet, što je bilo i za očekivati ako uzmemo u obzir njihove veličine i koliki udio riječi iz korpusa sadrže.

7.2.9 Osnovne + lematizirane značajke vs. polarne značajke

Osnovna2 grupa neka je unija značajki iz osnovne skupine i značajki iz lematizirane skupine iz Tablice 6.2. *Polarna* grupa neka je unija značajki SWNWSD[2] (engl. *SentiWordNet*

[2]SentiWordNet + razrješavanje višeznačnosti.

Word Sense Disambiguous) i MPQA-značajki. Cilj nam je pokazati kakav utjecaj imaju polarne značajke na kvalitetu određivanja osnovne polarnosti.

Tablica 7.16 prikazuje rezultate dodavanja grupe značajki osnovna2 i dodavanja polarne grupe značajki. Stupci I1, I2 i I12 redom označavaju dodavanje grupe značajki osnovna2, dodavanje polarne grupe značajki i dodavanje obiju navedenih grupa značajki.

	market			uniform			sent_market			sent_uniform			pra_market			pra_uniform		
Sustav	I1	I2	I12	I1	I2	I12	I1	I2	I12	I1	I2	I12	I1	I2	I12	I1	I2	I12
filter1-5	90.5	85.2	91.3	87.0	77.3	85.9	88.4	81.4	88.2	85.8	76.7	84.9	76.3	73.9	76.2	66.3	61.6	66.7
filter12-45	87.5	81.8	87.5	82.4	74.4	81.4	85.5	79.2	85.5	80.4	71.2	80.0	73.4	71.8	73.3	64.5	60.8	63.4
filter1-3-5	78.9	71.1	79.3	70.1	58.4	68.4	75.2	67.4	75.3	67.9	57.8	67.0	57.5	54.6	57.4	46.9	41.1	46.0
filter12-3-45	76.5	69.6	76.4	64.0	56.7	63.3	72.6	66.9	72.9	62.7	54.2	62.6	58.5	56.7	57.8	48.1	45.4	46.9
filter1-234-5	67.8	58.6	66.2	67.0	58.2	65.2	65.7	57.2	64.9	65.0	58.0	63.4	51.1	47.2	50.6	54.3	51.3	52.3

Tablica 7.16: Usporedba korištenja grupe značajki osnovna2 i polarne grupe značajki.

Polarne značajke u korpusima stvarne distribucije pokazuju u F-mjeri za 7.97% bolje rezultate od korpusa uniformne distribucije. Prosječna F-mjera polarnih značajki za sustave koji određuju osnovnu polarnost u dvije klase iznosi 74.61%, a za sustave koji određuju osnovnu polarnost u tri klase iznosi tek 57.24%. Sustavi polarnih značajki postižu nižu F-mjeru od referentnog sustava za prosječno 4.64%. Dodavanje polarnih značajki na skupinu značajki osnovna2 u prosjeku narušava F-mjeru za 0.59%. U tome najviše šteti korpus uniform s narušavanjem F-mjere za 1.8%, a najmanje šteti korpus market koji narušava F-mjeru za 0.1%. Korpus market za filter1-5 i filter1-3-5 bilježi poboljšanje F-mjere dodatkom polarnih značajki redom za 0.8% i 0.4%. Ukupan dojam spuštaju ostali filtri, pogotovo filter1-234-5 korpusa market, koji narušava F-mjeru za 1.6%. Gledano kroz za sve filtre, dodavanje polarnih značajki najmanje šteti sustavima filter1-5, prosječno 0.18%, a najviše sustavima filter1-234-5, prosječno 1.38%.

Tablicom 7.17 prikazana je usporedba izbacivanja dviju grupa značajki. Tako stupci E1, E2 i E12 redom označavaju izbacivanje grupe značajki osnovna2, izbacivanje polarne grupe značajki i izbacivanje obiju grupa.

Gornja tablica pokazuje da izbacivanje polarne grupe značajki narušava performanse sustava svih značajki za korpuse stvarne razdiobe. Sustavi korpusa market, sent_market i pra_market time su redom pogoršali F-mjeru za 0.22%, 0.43% i 0.24%. Korpusi uniformne razdiobe izbacivanjem polarnih značajki povećali su svoju F-mjeru. Korpus uniform, sent_uniform i pra_uniform tako su redom povećali točnost za 0.38%,

	market			uniform			sent_market			sent_uniform			pra_market			pra_uniform		
Sustav	E1	E2	E12	E1	E2	E12	E1	E2	E12	E1	E2	E12	E1	E2	E12	E1	E2	E12
filter1-5	89.8	90.5	89.6	84.6	86.6	84.5	86.3	87.9	86.5	82.4	85.2	82.5	74.9	75.4	74.7	65.0	65.7	65.5
filter12-45	86.1	86.6	85.8	80.7	82.8	80.7	84.5	84.9	83.9	76.7	79.4	77.4	71.6	72.8	72.2	62.5	63.9	62.8
filter1-3-5	78.0	79.1	77.3	66.8	68.7	66.4	73.4	74.5	73.0	63.8	66.2	64.1	56.1	57.4	56.5	44.6	46.1	45.2
filter12-3-45	75.0	75.7	74.8	62.1	64.3	61.8	72.0	72.2	71.1	58.9	61.7	59.4	56.6	57.9	57.1	46.2	47.8	46.5
filter1-234-5	64.2	67.2	65.2	63.7	65.8	63.4	62.2	65.0	62.9	60.4	64.1	61.2	49.4	50.6	49.4	51.5	53.3	52.6

Tablica 7.17: Usporedba izbacivanja grupe značajki osnovna2 i polarne grupe značajki.

0.22% i 0.36%. Također, izbacivanje polarne grupe prosječno je naštetilo svim filtrima osim sustavu filter1-234-5 koji je time povećao F-mjeru za 0.42%.

Puno veću korist od polarnih značajki imaju sustavi koji klasificiraju u dvije klase za razliku od sustava koji klasificiraju u tri klase. Dodavanje polarnih značajki grupi značajki osnovna2 uglavnom narušava točnost, makar neki sustavi pokazuju poboljšanje. Navedeni rezultati pokazuju kako polarne značajke doprinose sustavima stvarne razdiobe ocjena po komentarima, dok oni uniformne razdiobe bilježe pad F-mjere.

7.2.10 Polarne značajke vs. nepolarne značajke

U ovome odjeljku pokazat ćemo usporedbu polarnih i nepolarnih značajki. U skupinu polarnih značajki spadaju sve one značajke koje se nalaze u grupi *SentiWordNet + razrješavanje višeznačnosti* i grupi *MPQA*. Stupci I1, I2 i I12 Tablice 7.18 prikazuju redom rezultate evaluacije polarnih značajki, nepolarnih značajki i sustava svih značajki.

	market			uniform			sent_market			sent_uniform			pra_market			pra_uniform		
Sustav	I1	I2	I12	I1	I2	I12	I1	I2	I12	I1	I2	I12	I1	I2	I12	I1	I2	I12
filter1-5	85.2	90.5	91.0	77.3	86.6	86.1	81.4	87.9	88.0	76.7	85.2	84.9	73.9	75.4	76.4	61.6	65.7	65.9
filter12-45	81.8	86.6	87.3	74.4	82.8	82.2	79.2	84.9	85.4	71.2	79.4	79.3	71.8	72.8	72.9	60.8	63.9	63.3
filter1-3-5	71.1	79.1	79.0	58.4	68.7	68.9	67.4	74.5	75.1	57.8	66.2	66.4	54.6	57.4	57.7	41.1	46.1	45.8
filter12-3-45	69.6	75.7	76.3	56.7	64.3	63.8	66.9	72.2	73.1	54.2	61.7	61.4	56.7	57.9	57.8	45.4	47.8	47.4
filter1-234-5	58.6	67.2	66.6	58.2	65.8	65.3	57.2	65.0	65.0	58.0	64.1	63.5	47.2	50.6	50.5	51.3	53.3	52.6

Tablica 7.18: Usporedba evaluacije polarnih i nepolarnih značajki.

Korištenje samo polarnih značajki postiže lošije rezultate od korištenja samo nepolarnih značajki[3] za sve korpuse. Korpusi komentara i rečenica postižu lošije rezultate koristeći polarne značajke redom za 7.6% i 7.11%, dok korpus fraza postiže pogoršanje

[3]Svih ostalih značajki.

od 2.65%. Ovo je bilo i za očekivati jer se korpus fraza uglavnom sastoji od vrsta riječi koje su sadržane u rječnicima SentiWordNet i MPQA. U prošlom odjeljku pokazali smo kako dodavanje polarnih značajki na nepolarne povećava F-mjeru za sve korpuse stvarne distribucije, dok za one uniformne distribucije F-mjera pada.

7.2.11 Pozitivne značajke vs. negativne značajke

U Tablici 7.19 prikazana je usporedba pozitivnih i negativnih značajki iz Tablice 6.2. Stupci I1, I2 i I12 redom označavaju pozitivne značajke, negativne značajke i pozitivne i negativne značajke.

	market			uniform			sent_market			sent_uniform			pra_market			pra_uniform		
Sustav	I1	I2	I12	I1	I2	I12	I1	I2	I12	I1	I2	I12	I1	I2	I12	I1	I2	I12
filter1-5	83.1	83.1	87.5	80.1	75.9	82.0	81.2	82.7	86.4	78.3	74.5	82.6	70.7	74.0	75.4	61.5	58.1	64.4
filter12-45	80.8	80.8	84.4	74.9	72.2	78.0	78.0	80.6	83.3	73.5	69.6	76.9	68.0	71.9	72.2	57.9	57.2	62.2
filter1-3-5	70.3	69.9	75.1	59.8	53.7	61.5	65.1	68.1	71.1	57.3	52.4	61.2	49.8	54.1	55.9	39.4	38.0	44.2
filter12-3-45	69.5	69.4	72.7	57.2	53.8	59.2	65.0	67.5	70.0	55.3	52.5	58.3	53.2	56.1	56.7	42.4	41.8	45.8
filter1-234-5	59.8	58.0	62.5	60.4	53.9	62.3	56.9	55.8	61.5	57.7	49.9	61.0	47.2	45.3	48.8	50.9	48.6	53.1

Tablica 7.19: Usporedba evaluacije pozitivnih i negativnih značajki.

Referentni sustavi postižu za 5.32% bolje rezultate od pozitivnih značajki i 6.52% bolje rezultate od negativnih značajki. Korištenjem i pozitivnih i negativnih značajki F-mjera se jako poboljšala, no i dalje referentni sustavi postižu za 1.62% bolje rezultate. Korištenje pozitivnih značajki poboljšava rezultate za 1.19% više nego korištenje negativnih značajki. Tome najviše doprinose uniformni korpusi. Korpusi uniform, sent_uniform i pra_uniform redom postižu bolje rezultate koristeći pozitivne značajke za 4.58% 4.64% i 1.68% što je prosječno 3.63%. Market korpus također postiže bolje rezultate koristeći pozitivne značajke uz poboljšanje F-mjere nad negativnim značajkama za 0.46%. Korpusi sent_market i pra_market postižu bolje rezultate koristeći negativne značajke i to redom 1.7% i 2.5%. Kako korpusi stvarne razdiobe imaju manji broj negativnih komentara, logično je da su negativne značajke u tim sustavima zapravo najviše doprinjele pravilnoj klasifikaciji jer je upravo njih i trebalo nekako prepoznati. Po tome negativne značajke u korpusima stvarne razdiobe imaju veću važnost od pozitivnih značajki. Dodavanje negativnih značajki na pozitivne najviše pogoduje sustavu filter1-3-5, koji time poboljšava F-mjeru za 4.55%, a najmanje sustavu filter1-234-5, koji F-mjeru poboljšava za 2.72%. Dodavanje pozitivnih značajki na negativne najviše pogoduje sustavu filter1-234-5, koji time poboljšava F-mjeru za 6.28%, a najmanje

sustavu filter12-3-45 koji poboljšava F-mjeru za 3.6%.

Pokazali smo kako korištenje pozitivnih značajki doprinosi više nego korištenje negativnih značajki. Također, pokazali smo i da korpusima stvarne razdiobe više pridonose negativne značajke, dok korpusima uniformne razdiobe više pridonose pozitivne značajke.

7.2.12 Pozitivne + negativne značajke vs. neutralne značajke

Pokazat ćemo utjecaj neutralnih značajki na zadacima određivanja osnovne polarnosti. I1, I2 i I12 stupci Tablice 7.20 prikazuju redom uniju pozitivnih i negativnih značajki, neutralne značajke, i uniju pozitivnih, negativnih i neutralnih značajki.

	market			uniform			sent_market			sent_uniform			pra_market			pra_uniform		
Sustav	I1	I2	I12	I1	I2	I12	I1	I2	I12	I1	I2	I12	I1	I2	I12	I1	I2	I12
filter1-5	87.5	87.7	89.4	82.0	79.7	82.3	86.4	83.9	85.5	82.6	79.8	80.7	75.4	73.6	74.5	64.4	62.8	65.6
filter12-45	84.4	84.3	85.3	78.0	77.0	77.7	83.3	82.2	82.9	76.9	75.3	75.8	72.2	70.8	72.0	62.2	61.3	62.8
filter1-3-5	75.1	75.1	76.6	61.5	63.8	64.3	71.1	70.8	72.1	61.2	62.3	62.1	55.9	54.6	56.2	44.2	43.2	44.8
filter12-3-45	72.7	72.5	74.0	59.2	59.8	60.3	70.0	69.8	70.1	58.3	57.5	57.6	56.7	56.1	57.2	45.8	45.3	46.4
filter1-234-5	62.5	62.4	62.8	62.3	60.9	61.0	61.5	60.3	60.3	61.0	61.2	60.0	48.8	49.7	50.6	53.1	50.7	51.9

Tablica 7.20: Usporedba evaluacije pozitivnih + negativnih značajki i neutralnih značajki.

Pozitivne i negativne značajke analizirane su u prošlom odjeljku. Korištenjem samo neutralnih značajki postiže se F-mjera za 0.73% manja nego kada se koriste samo pozitivne i negativne značajke. Korištenje samo neutralnih značajki poboljšava rezultate naspram korištenja pozitivnih i negativnih za sustav filter1-3-5 i to za 0.13%, dok ostali filtri pokazuju puno bolje rezultate koristeći pozitivne i negativne značajke. To se pogotovo odnosi na sustave koji klasificiraju u dvije klase, koji koristeći pozitivne i negativne značajke postižu 1.41% veću F-mjeru nego kada koriste neutralne značajke. Dodavanje neutralnih značajki na pozitivne i negativne povećava F-mjeru za prosječno 0.22%. Korpus komentara time povećava F-mjeru za 0.85%, korpus fraza za 0.33%, dok korpusu rečenica time pada F-mjera za 0.52%. Očekivano, veću korist pri dodavanju neutralnih značajki postižu sustavi koji određuju osnovnu polarnost u tri klase. Filter1-3-5 dodavanjem neutralnih značajki povećao je svoju F-mjeru za 1.18%, filter12-3-45 za 0.48%, a sustav filter1-234-5 bilježi pogoršanje F-mjere za 0.43%. Filter1-5 i filter12-45 redom ovime bilježe pad F-mjere redom za 0.05% i 0.08%.

Dodavanjem i neutralnih značajki na pozitivne i negativne nije dovoljno da se pre-

maši referetni sustav. Unatoč tome, pokazali smo da dodavanje neutralnih značajki generalno povećava kvalitetu sustava, pogotovo za sustave koji imaju zadatak odrediti osnovnu polarnost u tri klase.

7.2.13 Pozitivne + negativne značajke vs. broj pozitivnih + negativnih značajki

Pozitivne i negativne značajke u ovom slučaju označavaju samo četiri značajke iz Tablice 6.2. To su značajke: pozitivne i negativne SentiWordNet riječi, i pozitivne i negativne SentiWordNet-riječi kojima je razrješena višeznačnost. Broj pozitivnih i negativnih značajki su značajke istih grupa koje broje koliko puta se pojavila pozitivna i koliko puta se pojavila negativna značajka. To su značajke: broj pozitivnih i negativnih SentiWordNet-riječi, i broj pozitivnih i negativnih SentiWordNet-riječi kojima je razrješena višeznačnost. Rezultati ove evaluacije prikazani su u Tablici 7.21.

	market			uniform			sent_market			sent_uniform			pra_market			pra_uniform		
Sustav	I1	I2	I12	I1	I2	I12	I1	I2	I12	I1	I2	I12	I1	I2	I12	I1	I2	I12
filter1-5	86.7	69.5	87.2	82.0	67.7	82.6	86.3	67.2	86.3	82.5	70.1	82.6	74.8	68.2	74.9	63.9	61.9	64.5
filter12-45	84.0	68.4	84.1	77.9	64.7	77.9	83.1	66.1	82.9	76.4	64.2	76.9	72.1	65.6	72.2	61.2	56.7	61.5
filter1-3-5	74.2	55.2	74.1	61.2	43.7	62.1	71.5	51.2	71.2	61.5	45.0	61.1	55.5	46.4	55.5	44.0	38.1	43.3
filter12-3-45	72.6	57.8	72.5	59.0	46.0	59.1	69.8	54.2	69.7	58.3	45.7	58.7	56.8	50.7	56.7	45.0	40.3	45.2
filter1-234-5	62.1	40.3	62.6	62.6	45.0	62.9	61.2	43.1	61.3	61.1	45.0	61.1	48.8	40.8	48.6	52.8	45.0	52.9

Tablica 7.21: Evaluacija pozitivnih i negativnih SentiWordNet-značajki, i evaluacija broja pozitivnih ili negativnih SentiWordNet-značajki.

Iako rezultati korištenja samo broja pozitivnih i negativnih riječi nema veliku F-mjeru, prilikom dodavanja tih značajki na pozitivne i negativne SentiWordNet-značajke postiže se povećanje F-mjere za prosječno 0.11%. Nije puno, no uzmimo u obzir da se radi o tek četiri numeričke značajke koje drže brojeve pojavljivanja pozitivnih, odnosno negativnih riječi u primjeru teksta. Dodavanjem tih značajki filter1-5 i filter12-45 povećavaju svoju F-mjeru za prosječno 0.23%, a filtri koji klasificiraju u tri klase povećavaju F-mjeru za 0.03%. Filtri s tri klase postižu manje poboljšanje zato što ove značajke mjere samo broj pojava pozitivnih i negativnih riječi, dok neutralne ne uzimaju u obzir.

Vidimo da jednostavne značajke poput broja pozitivnih i negativnih riječi u rečenici ne mogu kvalitetno zasebno riješiti problem određivanja osnovne polarnosti, no njihovo dodavanje ostalim značajkama doprinosi ukupnoj kvaliteti sustava.

7.2.14 Određivanje stupnja polarnosti

Tablicom 7.22 prikazan je referentni sustav koji određuje stupanj polarnosti teksta. Kao i kod određivanja osnovne polarnosti, referentan sustav kao značajku sadrži samo unigrame riječi. Sustav je evaluiran kroz svih šest korpusa.

	market		uniform		sent_market		sent_uniform		pra_market		pra_uniform	
Sustav	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
regresija	0.97	1.28	1.06	1.34	0.96	1.26	1.00	1.26	1.08	1.35	1.12	1.36

Tablica 7.22: Evaluacija sustava stupnja polarnosti nad unigramom riječi svih korpusa.

Prosječna srednja apsolutna pogreška (MAE) sustava stvarne razdiobe iznosi 1.00. Korijen srednje kvadratne pogreške (RMSE) sustava stvarne razdiobe iznosi 1.30. MAE i RMSE mjere sustava uniformne razdiobe redom iznose 1.06 i 1.32. Zaključujemo da stvarna razdioba ima manju pogrešku od uniformne razdiobe. Najveću pogrešku pokazuju sustavi *korpusa fraza*, dok najmanju pogrešku pokazuju sustavi *korpusa rečenica*. Ovi rezultati drugačiji su od rezultata određivanja osnovne polarnosti. Najbolji korpus u sustavima klasifikacije bio je *korpus komentara*.

Sustav određivanja stupnja polarnosti uključujući sve značajke prikazan je Tablicom 7.23. *Korpus komentara* kao i *korpus rečenica* prikazuju poboljšanja u odnosu na referentne sustave, a *korpus fraza* pokazuje pogoršanje. Kako je RMSE stroža mjera od mjere MAE, možemo zaključiti kako *korpus komentara* ima nešto manju grešku od *korpusa rečenica* što je u suprotnosti s referentnim sustavima istih korpusa.

	market		uniform		sent_market		sent_uniform		pra_market		pra_uniform	
Sustav	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
regresija	0.90	1.16	0.96	1.20	0.93	1.20	0.96	1.20	1.14	1.43	1.17	1.43

Tablica 7.23: Evaluacija sustava stupnja polarnosti sa svim značajkama.

Uspoređujući refentne sustave, *korpus rečenica* nije se puno poboljšao. Greška se smanjila samo 0.06 za mjeru RMSE i 0.04 za mjeru MAE. *Korpus komentara* bilježi poboljšanje mjere RMSE za 0.13 i poboljšanje mjere MAE za 0.09. Poboljšanje *korpusa komentara* s obzirom na referentni sustav dvostruko je veće od poboljšanja *korpusa rečenica* s obzirom na referentni sustav. *Korpus fraza* povećao je RMSE za 0.08, a MAE za 0.06. Ovi rezultati u skladu su s usporedbom referentnog sustava i sustava određivanja osnovne polarnosti sa svim značajkama.

Poglavlje 8

Osvrt na hrvatski jezik

Kako obrada prirodnog jezika za hrvatski još uvijek nije na stupnju razvoja kao što je to za engleski jezik, podjela značajki na grupe definirane u Tablici 6.2 omogućuje nam procjenu kako bi se ovakav sustav ponašao za hrvatski jezik. Značajke su grupirane po tehnologiji ili resursu koji koriste. Tako osnovna grupa značajki ne koristi nikakav dodatan resurs osim ulaznog korpusa. Lematizirana grupa koristi alat i algoritme pomoću kojih je moguće odrediti osnovni oblik riječi. Grupa vrsta riječi koristi alat i algoritme koji određuju vrstu riječi u rečenici. Skupina značajki SentiWordNet koristi rječnik apriorno polarnih riječi. Slično, SentiWordNet razrješavanje višeznačnosti grupa koristi dodatno i postupke koji riječima dodijeljuje njen pravi smisao ovisno u kojem kontekstu se nalaze. Grupa MPQA koristi rječnik sličan rječniku SentiWordNet. Grupa sintaktičkog stabla koristi algoritme koji generiraju sintaktičko stablo koje sadrži relacije među riječima rečenice.

Odjeljak 7.2.4 pokazuje da se korištenjem osnovne grupe značajki postižu zadovoljavajući rezultati za korpus komentara i rečenica sustava koji određuju osnovnu polarnost u dvije klase, odnosno tri klase. Korištenje algoritama za lematizaciju značajno poboljšava rezultate u odnosu na osnovnu grupu značajki. Osnovna grupa značajki pokazuje bolje rezultate od referentnog sustava za 1.17%, dok lematizirana grupa pokazuje bolje rezultate za 1.74%.

Odjeljkom 7.2.6 istražili smo kako se ponašaju značajke koje koriste vrstu riječi. Rezultati pokazuju kako koristeći značajke točnost sustava pada, stoga pretpostavljamo da bi se slično ponašao i za hrvatski jezik.

Korištenje riječnika apriorno polarnih riječi u odjeljku 7.2.9 u prosjeku narušava točnost. Korpusi stvarne razdiobe zabilježili su korist od apriorno polarnih riječnika, dok su uniformni korpusi zabilježili lošije rezultate. Pokazalo se također da korištenje

algoritama za razrješavanje višeznačnosti narušava rezultate.

Rezultati pokazuju kako bi za hrvatski jezik bilo dovoljno korištenje lematizirane grupe značajki. Takav sustav postiže rezultate bolje od referentnog sustava pa čak i od sustava u kojemu su uključene sve značajke iz Tablice 6.2.

Poglavlje 9

Zaključak i budući rad

Mogućnost automatskog određivanja polarnosti konteksta važan je problem strojne analize mišljenja korisnika. Pokazalo se kako je određivanje polarnosti konteksta puno kompleksniji problem nego što se to čini na prvi pogled.

Fokus rada je na razumjevanju koje značajke su bitne za određivanje polarnosti konteksta. Razvijen je niz jezičnih značajki koje su potom evaluirane koristeći nekoliko algoritama nadziranog strojnog učenja. Eksperimenti su provedeni za cijele komentare, rečenice i fraze da bi se utvrdilo kakav utjecaj ima veličina komentara. Otkriveno je kako korpus komentara daje rezultate bolje od korpusa rečenica i korpusa fraza. Osim tipa korpusa, ispitano je kakav utjecaj ima distribucija komentara po ocjenama u korpusu na rezultate sustava. Dobiveno je da stvarna distribucija komentara postiže puno bolje rezultate od uniformne distribucije komentara po ocjenama. Također, pokazano je da sustavi koji određuju osnovnu polarnost u dvije klase postižu veću točnost od sustava koji određuju osnovnu polarnost u tri klase. Kod obje varijante najveću točnost postižu oni sustavi koji imaju veću udaljenost među susjednim ocjenama. Pokazalo se i to da vektori jednostavnih značajki postižu zadovoljavajuće dobra rješenja te bi se ovakve metode mogle primijeniti i na hrvatski jezik. Korištenje složenijih značajki kao i korištenje apriorno polarnih riječnika samostalno pokazuju dobre rezultate, no zajedno sa svim značajkama pokazuju lošije rezultate.

Budući nastavak ovoga rada zahtjevat će povećanje skupa za učenje. Algoritmima izbora značajki je potrebno detaljnije istražiti koji podskup značajki daje bolje rezultate. Bilo bi korisno provesti izbor značajki za svaki vektor značajki zasebno i izbor značajki nakon što vektore grupiramo u smislene grupe. Time bismo dobili odgovor koje značajke su međusobno zavisne, pa njihovom eliminacijom moguće povećali kvalitetu sustava. Zanimljivo bi bilo vidjeti utjecaj povećanja korpusa na kvalitetu rada

određenih grupa značajki.

Literatura

Marin Akšamović, Fran Dragomanović, Marin Japec, Matea Mužek, Vjekoslav Osmann, i Ivan Šolta. Presidential candidate popularity analysis - candypop, 2010.

A. Andreevskaia i S. Bergler. Mining wordnet for fuzzy sentiment: Sentiment tag extraction from wordnet glosses. U *Proceedings of EACL*, svezak 6, stranice 209–216, 2006.

S. Baccianella, A. Esuli, i F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. U *Seventh conference on International Language Resources and Evaluation, Malta. Retrieved May*, svezak 25, stranica 2010, 2010.

X. Bai, R. Padman, i E. Airoldi. On learning parsimonious models for extracting consumer opinions. U *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*, stranice 75b–75b. IEEE, 2005.

P. Beineke, T. Hastie, i S. Vaithyanathan. The sentimental factor: Improving review classification via human-provided information. U *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, stranica 263. Association for Computational Linguistics, 2004.

L. Cabral i A. Hortaçsu. The dynamics of seller reputation: Evidence from ebay. *Working Papers*, 2006.

comScore/the Kelsey group. Online consumer-generater reviews have significant impact on offline purchese behavior. `http://www.comscore.com/press/relese.asp?press=1928`, November 2007.

S. Das i M. Chen. Yahoo! for amazon: Sentiment parsing from small talk on the web. U *EFA 2001 Barcelona Meetings*, 2001.

K. Dave, S. Lawrence, i D.M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. U *Proceedings of the 12th international conference on World Wide Web*, stranice 519–528. ACM, 2003.

M.C. De Marneffe i C.D. Manning. The stanford typed dependencies representation. U *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, stranice 1–8. Association for Computational Linguistics, 2008.

M.C. De Marneffe, B. MacCartney, i C.D. Manning. Generating typed dependency parses from phrase structure parses. U *Proceedings of LREC*, svezak 6, stranice 449–454, 2006.

P. Domingos i M. Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine learning*, 29(2):103–130, 1997.

M. Efron. Cultural orientation: Classifying subjective documents by cociation analysis. U *AAAI Fall Symposium on Style and Meaning in Language, Art, and Music*, 2004.

Yasser EL-Manzalawy i Vasant Honavar. *WLSVM: Integrating LibSVM into Weka Environment*, 2005. Software available at `http://www.cs.iastate.edu/~yasser/wlsvm`.

A. Esuli i F. Sebastiani. Determining the semantic orientation of terms through gloss classification. U *Proceedings of the 14th ACM international conference on Information and knowledge management*, stranice 617–624. ACM, 2005.

A. Esuli i F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. U *Proceedings of LREC*, svezak 6, stranice 417–422. Citeseer, 2006.

C. Fellbaum. ed. wordnet: an electronic lexical database, 1998.

M. Gamon. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. U *Proceedings of the 20th international conference on Computational Linguistics*, stranica 841. Association for Computational Linguistics, 2004.

G. Grefenstette, Y. Qu, J.G. Shanahan, i D.A. Evans. Coupling niche browsers and affect analysis for an opinion mining application. U *Proceedings of RIAO*, svezak 4, stranice 186–194. Citeseer, 2004.

V. Hatzivassiloglou i K.R. McKeown. Predicting the semantic orientation of adjectives. U *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, stranice 174–181. Association for Computational Linguistics, 1997.

D. Hopkins i G. King. Extracting systematic social science meaning from text. *Manuscript available at http://gking. harvard. edu/files/words. pdf*, 2007.

J.A. Horrigan. Online shopping. *Pew Internet & American Life Project Report*, 36, 2008.

M. Hu i B. Liu. Mining opinion features in customer reviews. U *Proceedings of the National Conference on Artificial Intelligence*, stranice 755–760. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004.

R.A. Hummel i S.W. Zucker. On the foundations of relaxation labeling processes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, stranice 267–287, 1983.

X. Jin, Y. Li, T. Mah, i J. Tong. Sensitive webpage classification for content advertising. U *Proceedings of the 1st international workshop on Data mining and audience intelligence for advertising*, stranice 28–33. ACM, 2007.

T. Joachims. Text categorization with support vector machines: Learning with many relevant features. *Machine Learning: ECML-98*, stranice 137–142, 1998.

G.H. John i P. Langley. Estimating continuous distributions in bayesian classifiers. U *Proceedings of the eleventh conference on uncertainty in artificial intelligence*, stranice 338–345. Morgan Kaufmann Publishers Inc., 1995.

A. Kennedy i D. Inkpen. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22(2):110–125, 2006.

S.M. Kim i E. Hovy. Determining the sentiment of opinions. U *Proceedings of the 20th international conference on Computational Linguistics*, stranica 1367. Association for Computational Linguistics, 2004.

D. Klein i C.D. Manning. Accurate unlexicalized parsing. U *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, stranice 423–430. Association for Computational Linguistics, 2003.

M. Koppel i J. Schler. The importance of neutral examples for learning sentiment. *Computational Intelligence*, 22(2):100–109, 2006.

M. Laver, K. Benoit, i J. Garry. Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(2):311–331, 2003.

L.V. Lita, A.H. Schlaikjer, W.C. Hong, i E. Nyberg. Qualitative dimensions in question answering: Extending the definitional qa task. U *PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE*, svezak 20, stranica 1616. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005.

Bing Liu. *Web Data Mining: exploring hyperlinks, contents, and usage data.* Springer Verlag, 2007.

G.A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38 (11):39–41, 1995.

G. Mishne i N. Glance. Predicting movie sales from blogger sentiment. U *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, 2006.

S. Morinaga, K. Yamanishi, K. Tateishi, i T. Fukushima. Mining product reputations on the web. U *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, stranice 341–349. ACM, 2002.

T. Mullen i R. Malouf. A preliminary investigation into sentiment analysis of informal political discourse. U *AAAI symposium on computational approaches to analysing weblogs (AAAI-CAAW)*, stranice 159–162, 2006.

B. Pang i L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. U *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, stranica 271. Association for Computational Linguistics, 2004.

B. Pang i L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. U *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, svezak 43, stranica 115, 2005.

B. Pang, L. Lee, i S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. U *Proceedings of the ACL-02 conference on Empirical*

methods in natural language processing-Volume 10, stranice 79–86. Association for Computational Linguistics, 2002.

T. Pedersen i V. Kolhatkar. Wordnet:: Senserelate:: Allwords: a broad coverage word sense tagger that maximizes semantic relatedness. U *Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics, companion volume: Demonstration session*, stranice 17–20. Association for Computational Linguistics, 2009.

T. Pedersen, S. Banerjee, i S. Patwardhan. Maximizing semantic relatedness to perform word sense disambiguation. *Supercomputing institute research report umsi*, 25, 2005.

A.M. Popescu i O. Etzioni. Extracting product features and opinions from reviews. U *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, stranice 339–346. Association for Computational Linguistics, 2005.

E. Riloff i J. Wiebe. Learning extraction patterns for subjective expressions. U *Proceedings of the 2003 conference on Empirical methods in natural language processing*, stranice 105–112. Association for Computational Linguistics, 2003.

A. Smola i V. Vapnik. Support vector regression machines. *Advances in neural information processing systems*, 9:155–161, 1997.

S. Somasundaran, T. Wilson, J. Wiebe, i V. Stoyanov. Qa with attitude: Exploiting opinion type analysis for improving question answering in on-line discussions and the news. U *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2007.

E. Spertus. Smokey: Automatic recognition of hostile messages. U *Proceedings of the National Conference on Artificial Intelligence*, stranice 1058–1065. JOHN WILEY & SONS LTD, 1997.

V. Stoyanov, C. Cardie, i J. Wiebe. Multi-perspective question answering using the opqa corpus. U *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, stranice 923–930. Association for Computational Linguistics, 2005.

J. Tatemura. Virtual reviewers for collaborative exploration of movie reviews. U *Proceedings of the 5th international conference on Intelligent user interfaces*, stranice 272–275. ACM, 2000.

L. Terveen, W. Hill, B. Amento, D. McDonald, i J. Creter. Phoaks: A system for sharing recommendations. *Communications of the ACM*, 40(3):59–62, 1997.

K. Toutanova i C.D. Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. U *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*, stranice 63–70. Association for Computational Linguistics, 2000.

K. Toutanova, D. Klein, C.D. Manning, i Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. U *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, stranice 173–180. Association for Computational Linguistics, 2003.

P. Turney, M.L. Littman, et al. Measuring praise and criticism: Inference of semantic orientation from association. 2003.

P.D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. U *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, stranice 417–424. Association for Computational Linguistics, 2002.

T. Wilson, J. Wiebe, i P. Hoffmann. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational linguistics*, 35(3):399–433, 2009.

J. Yi, T. Nasukawa, R. Bunescu, i W. Niblack. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. U *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, stranice 427–434. IEEE, 2003.

H. Yu i V. Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. U *Proceedings of*

the 2003 conference on Empirical methods in natural language processing, stranice 129–136. Association for Computational Linguistics, 2003.

Ivan Šolta. *Ispitivanje stavova dubinskom analizom teksta.* Fakultet elektrotehnike i računarstva, 2009.

Sažetak

Naslov: Strojna analiza sentimenta temeljena na apriornoj polarnosti riječi

Autor: Ante Kegalj

Porastom komunikacije putem Interneta povećao se interes za strojnom analizom mišljenja izraženog u korisnički generiranom tekstu. Jedan od pristupa analizi mišljenja jest analiza sentimenta, kojom se utvrđuje je li tekst pozitivno, negativno ili neutralno orijentiran. Analiza ukupnog sentimenta dokumenta može se temeljiti na analizi apriorne polarnosti pojedinačnih riječi.

U okviru rada proučeni su postojeći postupci za određivanje sentimenta dokumenta te postupci za određivanje kontekstne polarnosti dijelova teksta temeljem apriorne polarnosti riječi. Razrađen je postupak za određivanje polarnosti dokumenta i dijelova dokumenta na engleskome jeziku temeljem apriorne polarnosti riječi. Postupak se temelji na metodama nadziranog strojnog učenja i na javno dostupnim leksikonima apriorne polarnosti riječi, SentiWordNet i MPQA. Napravljena je programska izvedba postupka te je provedeno vrednovanje na zadacima određivanja osnovne polarnosti (klasifikacija) i određivanja stupnja polarnosti (regresija). Ispitano je nekoliko različitih metoda strojnog učenja te je provedena detaljna analiza parametara, značajki i pogrešaka.

Ključne riječi: Obrada prirodnog jezika, Ekstrakcija informacija, Analiza sentimenta, Analiza mišljenja, Semantička orijentacija, Nenadzirano strojno učenje

Abstract

Title: Sentiment Analysis Based on Prior Word Polarity

Author: Ante Kegalj

From the intensified activity in online communication arose the interest for an automated opinion analysis from a user generated text. One of the approaches in opinion analysis is sentiment analysis, which aims to determine wether a presented text is sentimentally positive, negative or neutral. Analysis of a document's sentiment can be based on a priori polarity of individual words.

In the thesis we discuss the existing approaches for determining a document's sentiment and the approaches for determining context polarity for text segments based on a priori word polarity. We also devised the approach for determining the polarity of a document and document's segments written in English. This approach is based on a priori word polarity extracted from the public dictionaries containing word polarities, SentiWordNet and MPQA. It uses supervised machine learning techniques. The software implementation of this approach is also presented along with extensive evaluation which includes classification of text segments based on their sentiment polarity and determining the level of polarity using regression analysis. Several machine learning methods have been tested and a detailed analysis of methods' parameters, features and errors has been conducted.

Keywords: Natural language processing, Information extraction, Sentiment analysis, Opinion mining, Semantic orientation, Unsupervised machine learning

Životopis

Rođen sam 22. svibnja 1987. godine u Zagrebu. Godine 2006. završio sam XV. gimnaziju u Zagrebu. Iste godine upisao sam se na studij na Fakultetu elektrotehnike i računarstva u Zagrebu. Na drugoj godini studija upisao sam smjer Računarstvo, a na trećoj godini studija odabrao sam modul Računarska znanost.