Laboratorij za tehnologije znanja (KTLab)
Zavod za elektroniku, mikrolektroniku, računalne i inteligentne sustave
Fakultet elektrotehnike i računarstva
Sveučilište u Zagrebu

Interni dokument

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 288

Sažimanje korisničkih recenzija metodama strojnog učenja

Slavko Kručaj

Zagreb, lipanj 2011.

SADRŽAJ

1. Uvod **1**

2. Pregled postupaka sažimanja i analize sentimenta **4**
2.1. Analiza sentimenta 4
2.2. Sažimanje 6
2.3. Povezani radovi 7
2.4. Primjeri sličnih sustava 9
2.5. Sličnosti s drugim radovima 12

3. Postupak sažimanja recenzija **13**
3.1. Prikupljanje podataka 13
3.2. Obrada podataka 14
3.3. Označavanje podataka 16
3.4. Učenje 18
3.5. Naknadna obrada 19

4. Eksperimenti **21**
4.1. Rezultati 26
4.2. Eksperimenti 28
4.2.1. Atributi 29
4.2.2. Rezultati nakon redukcije atributa 30

5. Prilagodba sustava za hrvatski jezik **33**
5.1. Prikupljanje ulaznih podataka 33
5.2. Obrada podataka 33
5.3. Rezultati 34
5.3.1. Klasifikator učen na engleskom skupu 35
5.3.2. Klasifikator učen na hrvatskom skupu 36

6. Zaključak **37**

Literatura **38**

1. Uvod

Proteklog desetljeća, pojavom Web-a 2.0 i generiranjem korisničkih sadržaja te popularizacijom interneta kao sredstva komunikacije, počeo se stvarati veliki broj korisničkih recenzija. S jedne strane to znaci velik izvor informacija pri kupovini nekih proizvoda, a s druge strane nemogućnost da se u razumnom vremenu prouče sve te recenzije na temelju kojih bi se donijela određena odluka. Za ilustraciju na slici 1.1 prikazano je nekoliko različitih stranica i servisa koji sadrže recenzije i kritike.

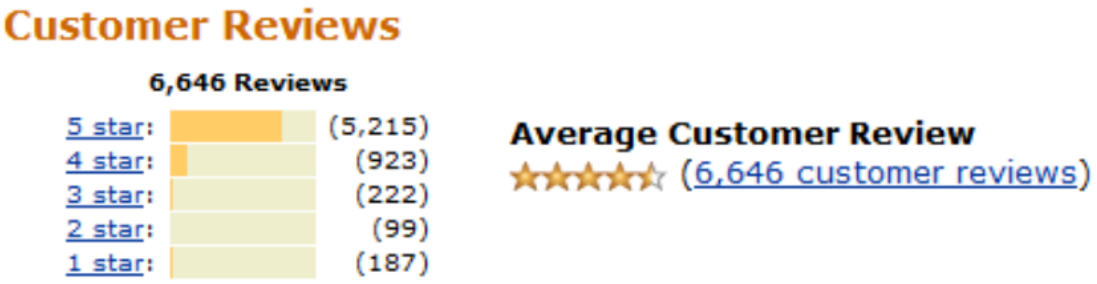
Uzmimo za primjer organizaciju puta u London. Ako želimo organizirati siguran i bezbrižan put, vjerojatno ćemo htjeti unaprijed rezervirati karte za avion, hostel/hotel, ulaznice za muzeje. Tijekom tog cijelog postupka, komentari i recenzije drugih ljudi koji su već putovali u London će nam služiti kao pomoć pri organizaciji našeg puta. Tako ćemo pri odabiru hostela birati one koji imaju bolje recenzije. Pri odabiru turističkih atrakcija birat ćemo one koji su ocijenjeni pozitivno. Isto vrijedi i za odabir prijevoznog sredstva, izložbi, muzeja, itd.

Kao druge primjere korištenja recenzija možemo uzeti odabira filma koji želimo gledati ili knjige koju želimo citati. Osim toga također danas pri popularnoj kupovini preko Interneta uvelike koristimo recenzije drugih ljudi, kao npr. na ebayu prije ćemo se odlučiti za kupovinu nekih artikala od ljudi koji imaju bolje recenzije od drugih, također pri kupovini proizvoda koje prije nikada nismo vidjeli pouzdat ćemo se u ocjene drugih ljudi koji iste već imaju.

Dakako prilikom svega toga ne trebaju nam nužno recenzije, ali one pružaju dosta dobar uvid u ono za što se zanimamo. Problem leži u tome što broj recenzija može biti golem. Npr. za pojedine proizvode na Amazonu broj recenzija može biti oko tristotinjak i ako želimo usporediti nekoliko sličnih proizvoda taj broj se istovremeno povećava. Nažalost u većini recenzija ima i dosta suhoparnog, nebitnog teksta koji nije važan za pozitivan ili negativan dojam o nekom proizvodu, a kupac nema vremena sve temeljito pročitati ili mu se jednostavno ne da.

Tema ovoga rada jest upravo izrada sustava koji će iz velikog broja recenzija (slika 1.2 dočarava veličinu broja recenzija) omogućiti izvlačenje liste pozitivnih, odnosno

Slika 1.1: Različite web stranice i servisi koji sadrže recenzije



Slika 1.2: Broj recenzija za *Amazon kindle*

1

negativnih dijelova recenzija. Time bi se omogućilo jednostavno i brzo pregledavanje recenzija. Prilikom izrade tog sustava koristit će se određene metode i postupci obrade prirodnog jezika (engl. Natural Language Processing, NLP), izvlačenja informacija (engl. Information Extraction, IE), koji će biti detaljnije objašnjeni u daljnjem tekstu.

Ovaj rad je sastavljen od nekoliko cjelina. U poglavlju 2 dan je detaljan pregled postojećih postupaka sažimanja i analize teksta. U poglavlju 3 opisan je postupak sažimanja recenzija koji se koristio u ovom radu. Zatim su u poglavlju 4 opisani eksperimenti koji su napravljeni te su prikazani rezultati istih. U poglavlju 5 opisan je postupak prilagodbe sustava na hrvatski jezik te su prikazani eksperimenti s rezultatima. Na kraju, u poglavlju 6 nalazi se zaključak ovog rada.

2. Pregled postupaka sažimanja i analize sentimenta

Sažimanje sentimenta zapravo je spoj dvaju područja: analize sentimenta i sažimanja teksta. Oba ta područja pripadaju obradi prirodnog jezika, no sažimanje sentimenta se kao jedinstvena cjelina počelo razvijati nedavno. U daljnjem tekstu ćemo zasebno opisati te čitatelju ukratko približiti područja analize sentimenta i sažimanja teksta. Nakon toga ćemo dati pregled istraživanja u području sažimanja teksta te pokazati sličnosti i razlike ovog rada i radova koji su poslužili kao inspiracija.

2.1. Analiza sentimenta

Analiza sentimenta je usko povezana s računalnom lingvistikom (engl. Computational linguistics), obradom prirodnog jezika (engl. Natural Language Processing) i dubinskom analizom teksta (engl. Text mining). Analiza sentimenta se u literaturi može naći pod raznim imenima: analiza subjektivnosti (engl. subjectivity analysis), dubinska analiza mišljenja (engl. Opinion mining), afektivno računarstvo (engl. Affective computing) i sl.

Područje analize sentimenta proučava subjektivne elemente, koje su Wiebe et al. definirali kao "lingvističke izraze osobnog stanja u kontekstu" (Wiebe, 1999). Uobičajeno su to rijeci, fraze ili rečenice. Ponekad se čak i cijeli dokumenti uzimaju kao jedinica sentimenta (Turney, 2003). No, ipak prevladava mišljenje da se sentiment pronalazi u manjim jezičnim cjelinama (Pang, 2008). Također valja napomenuti da se pojmovi "analiza sentimenta" i "analiza mišljenja" koriste kao sinonimi.

Sentiment koji se pojavljuje u tekstu, pretežno dolazi u dva oblika: eksplicitni gdje su subjektivne rečenice direktno iznesene kao mišljenje, npr. "Danas je lijep dan.", implicitni gdje sam sadržaj teksta implicira mišljenje, npr. "Slušalice su se odmah pokvarile." (Mejova, 2009). Dosadašnji radovi pretežno su orijentirani na proucavanje eksplicitnih subjektivnih rečenica, iz praktičnog razloga – jednostavnije ih je analizi-

rati.

Polarnost sentimenta je svojstvo teksta, uobičajeno dihotomizirano u pozitivno i negativno, no isto tako može biti i raspon od pozitivnog do negativnog. Rečenice mogu sadržavati više različitih sentimenata i tada ćemo za njih reci da imaju miješanu polarnost, što je različito od toga kad uopce nemaju polarnost (objektivne recenice). Osim toga postoji razlika i u jačini izraženog sentimenta.

Analiza sentimenta se sama po sebi može smatrati klasifikacijskim zadatkom, jer je za pojedinu rečenicu potrebno pronaći klasu kojoj pripada: pozitivna, negativna ili neutralna. Strojno učenje nudi mnoštvo algoritama kojima se taj problem može pokušati riješiti. S obzirom da ponašanje značajki frekvencije izraza i negacije prilikom računanja na prvi pogled nije očita, u narednom tekstu cemo ih malo detaljnije obrazložiti.

Frekvencija izraza

TF-IDF se računa po formuli 2.1. Ona se računa kao umnožak dvaju vrijednosti: frekvencije izraza (engl. term frequency, TF) i inverzne frekvencije dokumenta (engl. inverse document frequency, IDF). TF nam daje mjeru koliko je neki izraz t_i važan unutar dokumenta d_j. TF je definiran formulom 2.1 gdje je $n_{i,j}$ broj ponavljanja izraza t_i u dokumentu d_j, dok je nazivnik zbroj ponavljanja svih izraza u dokumentu d_j. IDF, također definiran formulom 2.1. $|D|$ je kardinalnost skupa dokumenata, a $|j : t_i \in d_j|$ broj dokumenata u kojima se pojavljuje izraz t_i.

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

$$idf = \log \frac{|D|}{|j : t_i \in d_j|} \tag{2.1}$$

$$Tf - Idf = Tf * Idf$$

Značenje TF-IDFa je sljedeće: ako se neki izraz u dokumentu pojavljuje cesto, a rijetko u cijeloj kolekciji, onda je taj izraz informativniji nego izraz koji se spominje samo jednom. U radu (Pang, 2002) pokazano je da ukoliko se prati jedinstveni izrazi a ne oni učestali može povećati uspješnost sustava, laički rečeno; “ljudi su kreativniji kada izražavaju svoje mišljenje” (Wiebe, 1999).

Negacije

Negacije su sastavni dio analize sentimenta. Pristupi koji promatraju podatke kao "vreće" rijeci dosta se loše nose s negacijama. Primjera radi, rečenice "Volim ovu knjigu." i "Ne volim ovu knjigu." teško je razlikovati ako promatramo svaku riječ zasebno, jer je razlika u samo jednoj rijeci. Kod analize sentimenta, negacije okreću polaritet cijelog izraza. No da stvari ne bi bile previše jednostavne, postoje isto tako primjeri gdje negacije ne okreću polaritet i te probleme se u nekim radovima rješavalo tako da su se tražili uzorci u oznakama vrste rijeci (engl. POS-tag) kako bi se otkrile negacije koje utječu na polarnost same rečenice (Potts, 2010).

U analizi sentimenta uvelike pomažu i gotovi resursi koji se mogu pronaći na Internetu. Jedan takav resurs je SentiWordNet (sen, 2011), koji se koristio prilikom izrade ovog rada. SentiWordNet je zapravo korpus rijeci na engleskom jeziku, koji za svaku riječ sadrži dvije brojke, koje govore kolika je vjerojatnost da je neka riječ pozitivna, odnosno negativna. Ovisno o pozitivnosti i negativnosti rijeci može se izračunati i njezina objektivnost.

2.2. Sažimanje

U ovom radu sažimanje teksta je "banalizirano". Drugim riječima, ne pokušava se sažimati tekst nego pozitivni i negativni isječci koje smo izvukli iz recenzija. Stoga ćemo opisat što se i kako danas radi na području sažimanja teksta iz više dokumenata.

Sažetak teksta definiran je kao tekst koji sadrži važne dijelove i informacije originalnog teksta, a da pritom nije duži od pola originalnog teksta (Hovy, 2005). Samim time, sažimanje teksta je figurativno govoreći proces filtriranja važnih informacija kako bi se proizveo sažetak.

Izlaz sustava za sažimanje, odnosno sažetak, može biti u nekoliko oblika. Jedna od podjela bila bi na one koji kao izlaz nude popis bitnih rečenica te na one koji na izlazu daju apstraktan tekst koji može poslužiti kao zamjena za originalni tekst. Također, sažetke se može podijeliti na one genericke i na one koji su fokusirani na upite korisnika. Prvi služi kao zamjena za originalni tekst i cilj je preuzeti sve važnije značajke iz originala, dok se drugi fokusiraju na potrebe korisnika i u skladu s tim kreiraju sažetak koji bi mogao odgovoriti na te upite.

Iako se u literaturi za sažimanje teksta spominje puno različitih pristupa, opisat ćemo jedan od načina, koji je predložen u (Mani, 1999). Tamo se sažimanju teksta pristupa u nekoliko razina: površinskoj, na razini entiteta i na razini diskursa.

U površinskoj razini, pokušava se uzimanjem plitkih značajki oblikovati funkcija koja bi se kasnije koristila u izvlačenju informacija. Neke od tih značajki su: tematska obilježja (značajnost rijeci, statistike pojavljivanje, tf-idf), lokacijske značajke (paragraf) ili neko drugo obilježje koje govori gdje se u tekstu pojavila bitna rečenica, sažetak ili pak kljucna rijec. Na drugoj razini, razini entiteta, modeliramo entitete i relacije medu njima. Time pokušavamo otkriti što je ključno u samom tekstu. Relacije koje se mogu pojaviti medu entitetima su sličnost (koliko je neka riječ slična nekoj drugoj, ukoliko dijele korijen), gdje se sličnost mjeri kao preklapanje znakova ili pak nekom drugom lingvističkom metodom. Zatim, udaljenost rijeci (leksikografska), sintaksne relacije, zatim kolokacije, sinonimi, antonimi i sl.

Na posljednjoj razini, razini diskursa,[1] se modelira globalna struktura teksta, u skladu s područjem kojim se tekst bavi te samim oblikom teksta.

Postoji mnogo sustava za izradu sažetaka, na slici 2.1 prikazani su samo neki od sustava i njihove karakteristike.

2.3. Povezani radovi

U zadnje vrijeme područje analize sentimenta postaje popularno, upravo zbog gomile "neobrađenih" sadržaja na Internetu. U nastavku je dan tablicni prikaz povezanih radova.

Rad koji je poslužio kao inspiracija prilikom izrade ovog rada je (Feczko, 2010). Cilj tog rada bio je sažeti recenzije s web stranica Amazona, odnosno automatsko sažimanje, filtriranje i sažimanje recenzija korisnika u kratku preglednu listu pozitivnih i negativnih mišljenja.

SentiSummary[2] pri tome koristi višekoračni pristup rješavanju tog problema, uz upotrebu nekoliko vanjskih alata i resursa za pripremu podataka za učenje, npr. LingPipe, Stanford NLP Toolkit i SentiWordNet. SentiSummary proces učenja započinje čitanjem XML-datoteka u kojima su zapisani podaci, tj. recenzije. Nakon toga započinje proces obrade podataka, koji se odvija u nekoliko koraka. Prvi korak je racunanje statistickih parametara, frekvencija rijeci u cijelom korpusu te broj dokumenata u kojima su one nađene. Te mjere ce biti potrebne prilikom računanja vrijednosti Tf-Idf. Nakon toga slijedi analiziranje polarnosti pri čemu se koristi unaprijed ucen sustav iz LingPipea, koji se temelji na radu (Pang, 2008). Pomoću njega analizira se po-

[1] diskurs - način i stil izlaganja s obzirom na temu ili podrucje djelatnosti u kojem se ostvaruje (hjp, 2011)

[2] Sustav razvijen u sklopu rada (Feczko, 2010)

Studies	Polarity mining techniques used	Text granularity	Features	Data sources/Domains	Performance (accuracy)
Wilson et al. (2005)	AdaBoost	phrase	subjectivity lexicon	multiperspective Question, Answering Opinion Corpus	contextual polarity: 65.7%
Kennedy and Inkpen (2006)	support vector machines, term-counting method, a combination of the two	document	term frequencies	General Inquirer dictionary, CTRWdictionary & Adj, IMDb (Movie review)	enhanced combined method: 86.2%
Chesley et al. (2006)	Support Vector Machines Wiktionary	document	textual features (e.g., exclamation points and question marks) and lexical semantics	Web sites of CNN, NPR, Atlanta Journal and Constitution, newspaper columns, reviews, political blogs, etc.	positive: 84.2% negative: 80.3% objective: 72.4%
Thomas and B. Pang (2006)	support vector machines	speech segment	reference classification	2005 U.S. floor debate in the House of Representatives	with same-speaker links and agreement links: 71.16%
Kaji and Kitsuregawa (2007)	phrase trees and word co-occurrence, Pointwise Mutual Information	phrase	lexical relationships, word co-occurrence	HTML documents	62.7–92.9%
Blitzer et al. (2007)	Structural Correspondence Learning	document	word frequencies and co-occurrences, part-of-speech	book, DVD, electronics and kitchen appliance product reviews	66.1–86.6%
Godbole et al. (2007)	lexical (WordNet)	word	graph distance measurements between words based on relationships of synonymity and anonymity, commonality of a words	newspapers, blog posts	82.7–95.7%
Annett and Kondrak (2008)	lexical (WordNet) & Support Vector Machines	document	number of positive/negative adjectives/adverbs, presence, absence or frequency of words, minimum distance from pivot words in WordNet	movie reviews, blog posts	65.4–77.5%
Zhou and Chaovalit (2008)	ontology-supported polarity mining	document	n-grams, words, word senses	movie reviews	72.2%
Hou and Li (2008)	Conditional Random Fields	sentence	POS tags, comparative sentence elements	product reviews, forum discussions; labeled manually and automatically	precision: man.: 89% aut.: 75% recall: man.: 81%, aut.: 71%
Ferguson et al. (2009)	Multinomial Naive Bayes (MNB)	phrase	binary word feature vectors	financial blog articles	75.25%
Tan et al. (2009)	Naive Bayes Classifier with feature adaptation using Frequently Co-occuring Entropy	document	words	Education reviews, stock reviews, and computer reviews	F1 score: 69–91%
Wilson et al. (2009)	boosting, memory-based learning, rule learning, and support vector learning	phrase	words, negation, polarity modification features	MPQA Corpus	83.6%
Melville et al. (2009)	Bayesian classification with lexicons and training documents	document	words	Blog posts reviewing software, political blogs, movie reviews	Blogs: 91.21% Political: 63.61% movies: 81.42%

Slika 2.1: Pregled radova na podrucju analize i sažimanja setnimenta

larnost, unigrama, bigrama i trigrama iz rečenica, iz cega se već dobiva neka ocjena pozitivnosti/negativnosti svake recenice. U trećem koraku koristi se Stanfordov POS tagger pomoću kojega dobivamo vrste rijeci, koje ce biti potrebne prilikom izvlacenje vrijednosti sentimenta iz korpusa SentiWordNet (sen, 2011). Nakon što se napravi i posljednji korak, imamo spremne značajke za učenje. Ukupno ih se u ovom radu koriste 29, a njihov popis dan je u tablici 2.1 (Feczko, 2010).

Tablica 2.1: Značajke korištene u radu (Feczko, 2010)

Polarnost unigrama, bigrama i trigrama
Amazon ocjena komentara
Duljina rečenice
Prosječan i log-normaliziran tf-idf
Prosječan i log-normaliziran maksimalan tf-idf
Prosjek i zbroj log-vjerojatnosti
5 najvećih tf-idfa i log-normaliziranih tf-idfa
Zbroj i prosjek vrijednosti za sve imenice, pridjeve, priloge i glagole

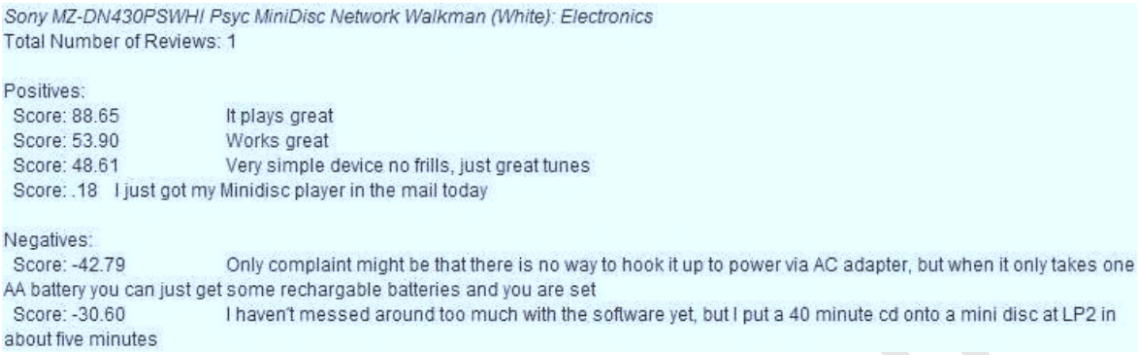
Strojno učenje se provodi s nekoliko klasifikatora na uzorku od 1000 pozitivno ili negativno označenih rečenica (neutralne recenice nisu se uzimale u obzir), čije se vrijednosti na kraju uspoređuju. Prikaz rezulata za pojedine klasifikatore prikazan je u tablici 2.2, koja je preuzeta iz (Feczko, 2010).

Tablica 2.2: SentiSummary točnost klasifikacije za nekoliko algoritama

Algoritam	Točnost
Osnovna metoda[3]	51.8%
Ridge regression	66.4%
Naivan Bayes	55.4%
Streamwise regression	65.8%
Stepwise regression	64.3%
Perceptron	56.9%

2.4. Primjeri sličnih sustava

U današnjem svijetu teško je pronaći nešto što nije izmišljeno ili već napravljeno. Isti slučaj je ovdje, tako da već postoji nekoliko sustava namijenjenih analizi sentimenta u



Slika 2.2: Prikaz rezultata sustava SentiSummary, preuzeto iz (Feczko, 2010)

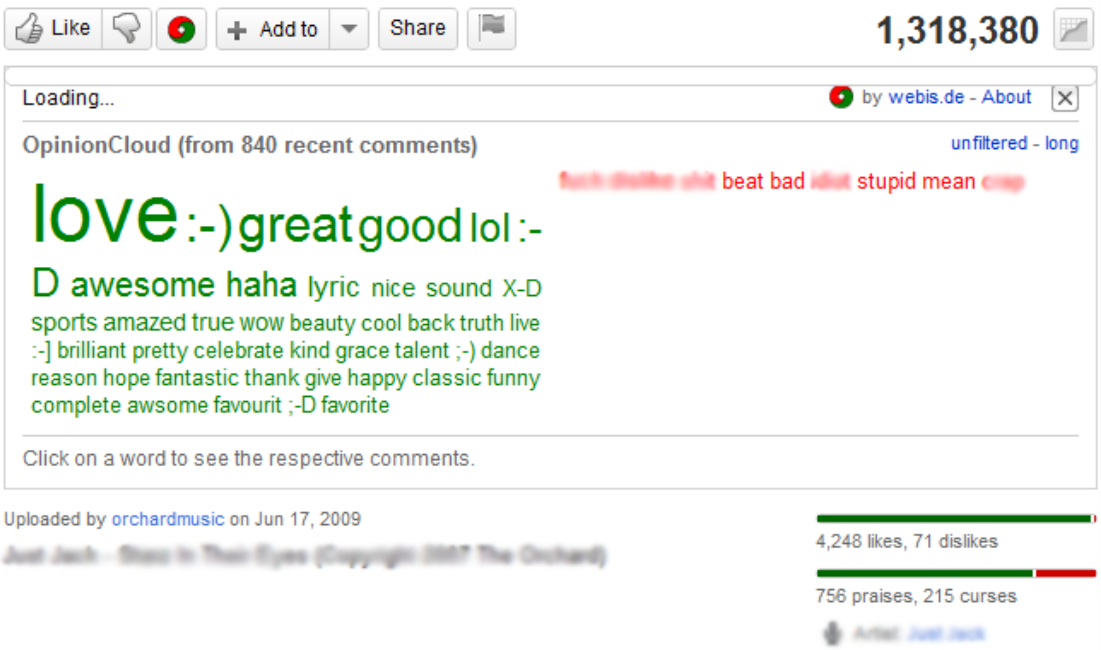
jednom ili drugom obliku.

Kao primjer takvih sustava možemo izdvojiti SentiSummary, koji je razvijen u sklopu članka na kojem je temeljen ovaj rad. Doduše taj sustav nije javno dostupan, tako da nismo mogli niti ispitati njegovu stvarnu korisnost. Prikaz sustava vidljiv je na 2.2.

Drugi sustav koji valja istaknuti jest OpinionCloud[4]. Taj sustav dolazi u obliku dodatka (engl. plugin) za različite preglednike. Kada se jednom instalira, on računa jednostavnu statistiku u realnom vremenu za komentare na Youtubeu ili Flickru. Iako rezultati rada tog sustava nisu dostupni može se primjetiti da sustav radi podjednako dobro i za različite jezike koji se mogu naći u komentarima. Jednostavne statistike koje se računaju iz komentara su: omjer pozitivnih i negativnih komentara, omjer pohvala i psovki te kratka lista pozitivnih i negativnih rijeci. Na slici 2.3 je dan prikaz rada sustava na jednom videu sa Youtubea.

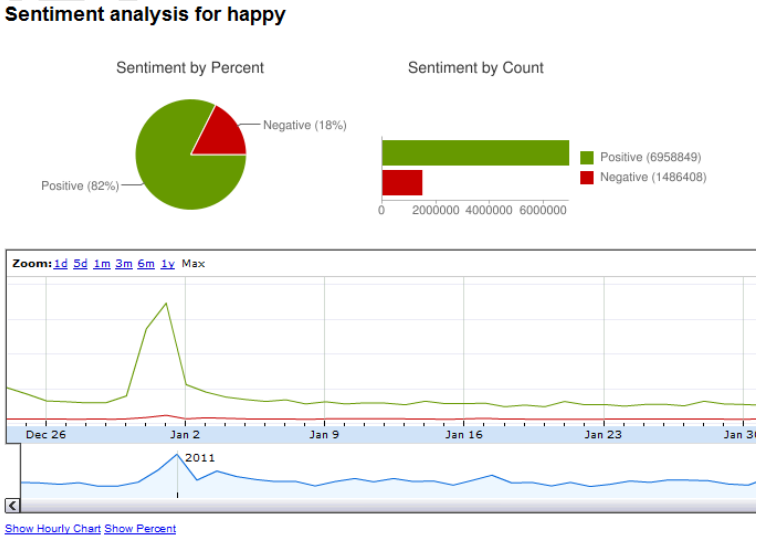
Još jedna zanimljiva aplikacija koju smo izdvojili je TwitterSentiment.[6] Stranica nudi malo drugačiji pristup od gornja dva sustava. To je ukratko pretraživac, koji za zadanu rijec ili frazu pretražuje statuse korisnika Twittera te daje statistiku, koliko puta se tražena riječ ili fraza spominje u pozitivnom ili negativnom kontekstu. Osim te informacije, također se daje vremenski prikaz, odnosno koliko se puta ta riječ spominjala u negativnom i pozitivnom kontekstu. Prikaz tog sustava dan je na slici 2.4. Ovaj sustav je naročito zanimljiv jer bi se mogao koristiti kao zamjena SentiWordNeta ili kao neki novi izvor informacija o sentimentu pojedine rijeci.

[4] `https://chrome.google.com/webstore/detail/jobpaepjhflihdcgajlbmkipfdmjmkda`
[6] `http://twittersentiment.appspot.com`



Slika 2.3: Prikaz rezultata sustava OpinionCloud

5



Slika 2.4: Prikaz rezultata sustava TwitterSentiment

7

2.5. Sličnosti s drugim radovima

Prilikom izrade ovog rada kao inspiracija je poslužio rad (Feczko, 2010), no ipak s manjim promjenama.

Neke od promjena su: ne korištenje lingpipe alata, izmijenjene značajke koje su se koristile pri učenju (smanjen je njihov broj). Osim toga, kao izvor recenzija nije korišten Amazon, nego su se koristile recenzije hostela preuzete sa servisa za rezerviranje hostela.[8] Motivacija za te promjene bila je bolje prilagođavanje potrebama ovog rada te iz jednostavnog razloga što u izvornom radu svi postupci nisu bili dovoljno dobro opisani. Opis rada sustava s tim izmjenama bit ce dan u kasnijem tekstu.

[8] `http://www.hostels.com`.

3. Postupak sažimanja recenzija

U ovom dijelu opisan je razvijeni sustav, počevši od samog prikupljanja podataka, obrade i pripreme tih istih podataka za strojno učenje pomoću kojih bismo trebali doći do jednostavnog sažetka u obliku pozitivne i negativne liste. Razvijeni sustav kao ulazne podatke za učenje koristi recenzije hostela.[1] Tocnije, koriste se recenzije svih hostela u Engleskoj i SAD-u. Sustav se sastoji od tri faze, prva je ekstrakcija podataka i oznacavanje istih, druga je priprema podataka u kojoj se koriste neke od već gotovih biblioteka za obradu prirodnog jezika, te zdanja faza je strojno učenje koje će se obavljati u alatu RapidMiner. U daljnjem tekstu detaljnije ćemo opisati svaku fazu u radu sustava.

3.1. Prikupljanje podataka

Kao prvi korak u radu sustava potrebno je ekstrahirati podatke (u daljnjem tekstu recenzije) i prevesti ih u oblik koji će biti pogodan za učenje.

Napravljen je modul za dohvat recenzija s Interneta koji radi u dva koraka. Prvi korak je automatsko dohvaćanje linkova na sve hostele koji se nalaze u Engleskoj i SAD-u. U drugoj fazi pristupa se pojedinom hostelu i s njegove stranice dohvacaju se svi komentari. Osim za hostele, napravljen je također ekstraktor filmskih recenzija na hrvatskom jeziku.

Prilikom izrade ekstraktora podataka koristili su se oblikovni obrasci(Erich Gamma i Vlissides, 1994), tako da je izvorni kod organiziran tako da omogućuje jednostavno nadogradnju s ekstraktorima za druge izvore na internetu.

Podaci se spremaju u formatu JSON.[2] Struktura koja se koristi za pohranu podataka prikazana je na slici 3.1

[1] Komentari se dohvaćaju sa stranice www.hostels.com, za sve hostele u Engleskoj i SAD-u, a radi na principu čišćenja web stranica.

[2] JSON je jednsotavan otvoreni standard za razmjenu podataka. Čitljiv je ljudima i za razliku od xml-a koristi manje zalihosti, `http://www.json.org`.

Slika 3.1: Struktura podataka koja predstavlja znacajke za učenje

3.2. Obrada podataka

Nakon što smo dobili potrebne podatke u obliku veće količine tekstova, potrebno je te tekstove "prevesti" u oblik pogodan za ulaz u modul strojnog učenja odnosno izvlačenje značajki koje ćemo dalje koristiti (drugim riječima treba popuniti strukturu prikazanu na slici 3.1). Kao prvo, potrebno je komentare podijeliti na rečenice koje će biti osnovni podatak iz kojeg će se izračunavati značajke.

Kako bi olakšali računanje značajki korišteni su neki već gotovi sustavi te neki postojeći resursi.

- Stanford NLP – sustav gdje smo koristili POS-tagger,[3] kako bismo za svaku riječ u rečenici mogli saznati njenu vrstu, što ce nam biti važno prilikom korištenja drugih resursa. Osim toga, korišten je i algoritam korjenovanja (engl. stemmer) kako bi se svaka riječ svela na osnovni oblik, što je pomoglo pri računanju tf-idfa.
- Google SOAP Search API – koji se koristi za provjeru pravopisa riječi, koji smo također koristili kako bismo prilikom računja tf-idf mogli bolje ocijeniti

[3]Oznacivac vrste rijeci – kao izlaz dobivamo vrstu rijeci (imenica, pridjev, glagol, itd.)

važnost neke riječi.[4]

- SentiWordNet[5], koristi se za dirketno računanje nekih značajki.
- LingPipe – također sustav za obradu prirodnog jezika. Ovaj sustav, odnosno podsustav PolarityAnalyzer, u završnoj verziji nije korišten jer je ocijenjeno da ne daje dovoljno dobre rezultate za podatke koji se koriste u ovom radu. U okviru ovog rada razvijena je vlastita programska podrška za zadaću koju je trebala ispuniti ova biblioteka.
- Google Translate API – koristi se za prevođenje hrvatski riječi na engleski prilikom prilagodbe sustava za hrvatski jezik.[6]

Znacajke za učenje

Korištenjem navedenih sustava i resursa izračunavamo značajke. Značajke koje se koriste i način na koji se izračunavaju opisan je u daljnjem tekstu. Nakon što smo podijelili komentare na rečenice, ostaje nam popuniti strukturu prikazanu na slici 3.1. Polarnost unigrama, bigrama i trigrama, trebale su se računati pomocu LingPipeovog podsustava PolarityAnalyzera, no kao što je već prije receno zbog nedovoljno dobrih rezultata koristili smo vlastitu implementaciju koja ce također biti opisana u daljnjem tekstu. Korištene su sljedeće značajke:

- Broj rijeci u rečenici;
- Polarnost unigrama – Realni broj koji daje opis koliko je rečenica pozitivna, odnosno negativna. Vrijednost smo računali tako da se za svaku rijec uzimala vrijednost koja se dobila korištenjem SentiWordNeta te smo na kraju samo izračunali prosjek za cijelu rečenicu;
- Polarnost bigrama – Slično kao i unigram polarnost, no ovdje su se kao ulaz uzimali svi bigrami riječi, (eg. good old house, daje bigrame good old i old house), za svaki taj bigram je zbrojena vrijednost iz SentiWordNet-a te je izračunat prosjek za svaki bigram. Važno je napomenuti da ukoliko je prva riječ u bigramu negativna, polaritet sljedeće riječi se obrće (Pang, 2002).
- Polarnost trigrama – Računa se isto kao i bigram polarnost, samo što su ulaz trigrami, a ne bigrami. Također negativna riječ obrće polaritet sljedeće ili sljedeće dvije riječi.

[4] `http://code.google.com/apis/soapsearch/reference.html`

[5] SentiWordNet (`http://sentiwordnet.isti.cnr.it`)) – resurs s vrijednostima sentimenta za veliki broj rijeci

[6] `http://code.google.com/apis/language/translate/overview.html`

– Zbroj, prosjek te razlika između najmanje i najveće polarnosti za svaku od sljedećih vrsta riječi: imenica, pridjev, prilog i glagol. Korištenjem resursa SentiWordNet i Stanfordovog označivača vrste rijeci za svaku grupu rijeci izračunamo gore navedene mjere.

Neki od parametara iz originalnog članka, tj. oni vezani uz račun tf-idf, nismo koristili jer je primjećeno da su zbog velikog broja pravopisnih grešaka u recenzijama, rezultati puni krivih vrijednosti, tako da su značajke, prosječan tf-idf, najboljih 5 tf-idf te njihove log-normalizirane vrijednosti izbačene iz značajki koje idu na ulaz strojnog ucenja.

Nakon što smo opisali sve značajke koje nam trebaju i gotove sustave koje koristimo, nadalje ćemo opisati proces pripreme skupa podataka za učenje. Priprema podataka se obavlja prilikom dohvata samih recenzija s Interneta. Čim se recenzija dohvati s Interneta dijeli se na rečenice. Te se rečenice pomoću Stanfordovog alata podijele na rijeci s pripadajućom vrstom. Zatim se za svaku riječ napravi provjera pravopisa te se ukoliko nije ispravna, postavi na najvjerojatniju ispravku. Nakon što je riječ prošla pravopisnu provjeru, za istu se riječ u SentiWordNetu traži njena polarnost te se s tom vrijednosti dalje računaju zbroj, prosjek i razlika najpozitivnije i najnegativnije rijeci. Kada je i taj korak prošao i kada imamo vrijednosti senitmenta za sve rijeci, pristupamo računanju unigramske, bigramske i trigramske polarnosti. Nakon što smo popunili strukturu podataka s izračunatim značajkama, još jedino moramo ručno ocijeniti polarnost recenice, kako bismo imali pripremljene podatke za strojno učenje. Tablica značajki zajedno s primjerom za rečcenicu “I wish all places were as comfortable, central, clean and friendly as this one.” dana je u tablici 3.1.

3.3. Označavanje podataka

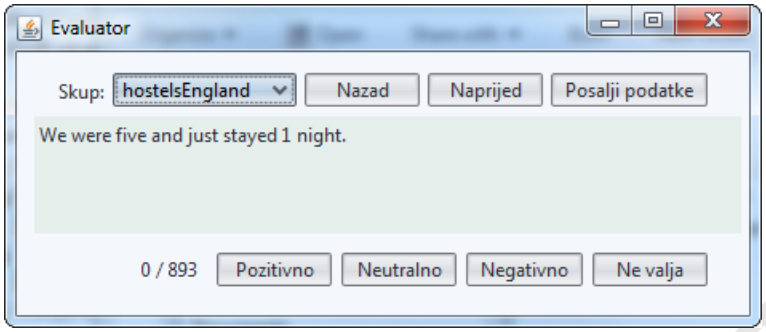
Kako bi se moglo napraviti strojno učenje nad recenzijama, potrebno je te iste i označiti. Drugim rječima, strukture podataka koje smo dobili nakon obrade podataka i koje smo popunili izračunatim značajkama koje će se dalje koristiti u treniranju različitih postupaka strojnog učenja, moramo i označiti. Svaku rečenicu iz svake recenzije ćemo označiti kao pozitivnu, negativnu ili neutralnu. Dodatno smo za slučaj da je rečenica loša ili da je na nekom drugom jeziku, uveli i oznaku “ne valja”. Nakon što se sve rečenice iz skupa za učenje oznace, izbacuju se neutralne rečenice te one koje su označene kao one koje ne valjaju. Prikaz sustava za ocjenjivanje dan je na slici 3.2, i tamo se vidi jednostavna funkcionalnost koja je podržana.

Tablica 3.1: Značajke za recenicu "I wish all places were as comfortable, central, clean and friendly as this one."

Značajka	Vrijednost
Broj riječi	17
Polarnost unigrama	0.08775
Polarnost bigrama	-0.97375
Polarnost trigrama	-1.97159
Zbroj sentimenta imenica	0
Prosjek sentimenta imenica	0
Zbroj sentimenta glagola	-0.0869
Prosjek sentimenta glagola	-0.0869
Zbroj sentimenta pridjeva	0.17468
Prosjek sentimenta pridjeva	0.05822
Zbroj sentimenta priloga	0
Prosjek sentimenta priloga	0
Razlika najveće i najmanje vrijednosti sentimenta imenice	0
Razlika najveće i najmanje vrijednosti sentimenta glagola	0.086
Razlika najveće i najmanje vrijednosti sentimenta pridjeva	0
Razlika najveće i najmanje vrijednosti sentimenta priloga	0.027
Ocjena	1

Kako je ocjenjivanje sentimenta subjektivan zadatak, čak i za osobu označavanje istog skupa rečenica obavile su dvije osobe, kako bi se provjerilo koliko je podudaranje između označivača. Rezultati koji su dobiveni tom analizom prikazani su na slici 3.2.

Iz tih rezultata može se vidjeti da se čak i pri ručnom ocjenjivanju pojavljuju razlike, iako možda ne toliko velike opet su dobar pokazatelj koliko je ovaj na očigled jednostavan zadatak kompliciran. Na slici 3.2 se vidi koliko rečenica u skupu za ocjenjivanje ocijenjen pozitivno odnosno negativno te koliko je podudaranje. Računanjem κ-statistike, koja nam govori koliko se ocjene podudaraju doći cemo do podatka koji govori podudaraju li se ocjenjivači ili ne. Ukoliko je vrijednost κ manja od 0.69 (heuristička vrijednost) može se reci da se ocjenjivači ne podudaraju te da nema smisla dalje pokušavati strojno učenje na tom skupu. Tablicu cemo označiti na sljedeći način: $tablica_{(1,1)} = a$, $tablica_{(1,2)} = b$,$tablica_{(2,1)} = c$ i $tablica_{(2,2)} = d$. Sada možemo κ vrijednost izračunati po formuli 3.1. U našem je slučaju vrijednost $\kappa = 0.92$ i iz toga možemo zaključiti da se ocjenjivači podudaraju te da možemo pristupiti strojnom



Slika 3.2: Evaluator rečenica, kojim se označava skup dobivenih podataka

učenju.

$$\kappa = \frac{N * (a+d) - [(a+c) * (a+b) + (b+d) * (c+d)]}{N^2 - [(a+c) * (a+b) + (b+d) * (c+d)]} \tag{3.1}$$

Tablica 3.2: Razlike medu ocjenjivačima

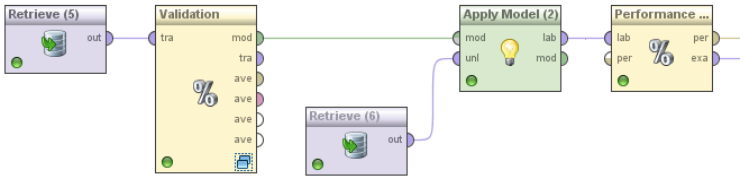
		Ocjenjivač 1	
		Pozitivno	Negativno
Ocjenjivač 2	Pozitivno	395	27
	Negativno	19	270

3.4. Učenje

Nakon što smo označili skupove podataka i izbacili neutralne rečenice te one koje ne valjaju, potrebno je pokušati naučiti razlikovati pozitivni odnosno negativni razred. Kao što je već prethodno najavljeno, ne treba očekivati izvanredne rezultate zbog subjektivnosti samog zadatka. Pri učenju je korišten sustav RapidMiner.[7] Taj je sustav omogućio da se na jednostavan način ispita više različitih algoritama strojnog učenja. U idućem poglavlju detaljnije ce biti opisani rezultati dobiveni strojnim učenjem korištenjem gore navedenih algoritama te će se detaljnije testirati kako koji parametar utječe na rezultate strojnog učenja.

Rapid miner sadrži gotove komponente s algoritmima strojnog učenja, komponente za učitavanje podataka te za njihovu provjeru i ispitivanje. RapidMiner čita većinu datoteka, no ne i JSON format koji je korišten za zapisivanje vektora, tako da su podaci

[7] `http://rapid-i.com/`



Slika 3.3: Tok podataka koji je kreiran u alatu RapidMiner

Would totally recommend it.
really nice staff!!!!!!!
Peaceful and beautiful site.
Very nice and especially good located hostel.
It is really nice hostel!!
Very nice hostel!
Very nice and clean place, kind staff
Room, bed, bath room and kitchen are also good!!!
:)
Clean and just 10 minutes from city center.
Thanks guys!

Slika 3.4: Pozitivno ocijenjene rečenice

(vektori za učenje), nakon što su označeni prebačeni u format CSV.[8] Kako bi rapid miner mogli koristiti potrebno je napraviti tijek podataka, odnosno posložiti i povezati komponente učitavanja, učenja i provjere. Tijek podataka koji se koristio prilikom učenja i ocjenjivanja uspješnosti dan je na slici 3.3.

3.5. Naknadna obrada

Nakon što je model naučen, potrebno ga je primjeniti na nepoznati skup podataka. Kao izlaz iz strojnog učenja, osim predviđene klase na izlazu dobivamo i mjeru pouzdanosti da ta rečenica pripada tom skupu. Te rečenice zatim podijelimo u dva razreda, pozitivan i negativan te svaki skup posebno poredamo po vjerojatnosti pripadanja toj klasi. Nakon toga imamo uređen skup rečenica te kao sažetak odabiremo N rečenica s najvecom pouzdanosti iz oba razreda. Motivacija za takav pristup je upravo u tome što ćemo na taj način dobiti one rečenice koje su ocijenjene kao najpozitivnije odnosno najnegativnije, što je upravo ono što smo htjeli.

Prikaz nekoliko sažetaka za određeni tekst dani su na slikama, slika 3.4 prikazuje rečenice koje su prepoznate kao pozitivne dok su na slici 3.5 prikazane rečenice koje su prepoznate kao negativne. Na toj slici se vidi da su rezultati relativno dobri i da na ovaj način dobivamo bolji, brži, čitljiviji pregled recenzija nekog hostela.

[8] csv (engl. comma separated value), gdje su pojedine vrijednosti polja odvojene zarezom ili nekim drugim znakom koji služi kao graničnik

It is annoying to say the least.
Loud music in the club next door, but my bed was changed when I worried about not being able to sleep.
It seems they only pick up in advertised time slots in spite of email saying they would do so out of advertised hours.
The people next door put their dog out in the tiny cement enclosure out back and the poor thing yaps for hours on end.
The city is dead and virtually not attractions, except the castle which I found out is actually even there anymore, it just a little museum they created.
the breakfast is worth the price but i prefer simple and free breakfast.
the staff is stiff.
The location, the hostel itself, GREAT, nothing to complain....BUT....the club nearby was so noisy, it was like somebody banging the window!!!
It is not a hostel.
It stoped at 2am, and the guy sleeping next was snoring loudly, I still couldn't sleep....All in all, the place is great if you don't mind the noise at all.
the bathroom is not cleanthe price is a little high for people who are not member of YHA.

Slika 3.5: Negativno ocijenjene rečenice

4. Eksperimenti

Klasifikatori koji su se koristili prilikom strojnog učenja bili su SVM (engl. Support Vector Machine), neuronska mreža, perceptron, LDA, naivan Bayesov klasifikator, naivan Bayesov klasifikator s jezgrama, k-NN, logistička regresija, linearna regresija. Sve implementacije gore navedenih klasifikatora su u Rapid mineru. U daljem tekstu ukratko ćemo opisati pojedine klasifikatore.

Stroj s potpornim vektorima

Ako imamo linearno razdvojive razrede, onda postoji i beskonačno mnogo mogućnosti na koje se razredi mogu odvojiti. Medu beskonačno mnogo klasifikatora postoji razlika u tome kako će se ponašati na dosad neviđenim podacima. Stroj s potpornim vektorima (engl. Support Vector Machine, SVM) traži takvu granicu koja će maksimizirati minimalnu udaljenost medu razredima (slika 4.1). Matematičkim rječnikom traži se rješenje sljedeće jednadžbe (Theodoridis i Koutroumbas, 2009)

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}||\mathbf{w}||_2^2 + \frac{C}{K}||\xi||_1 \tag{4.1}$$

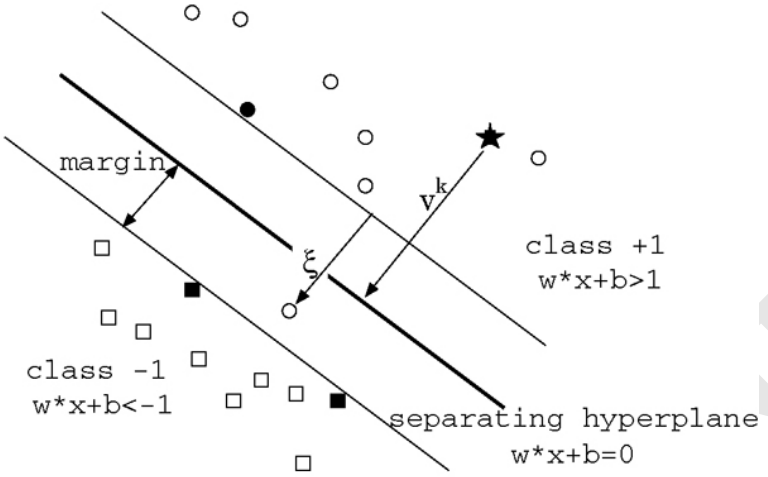
uz uvjete $y_k(\mathbf{w}^T\mathbf{x} + b) \geq 1 - \xi_k$ i $\xi_k \geq 0$ za $k = 1, \ldots, K$.

Perceptron

Perceptron je binaran klasifiaktor koji preslikava ulazni vektor na binarnu varijablu na izlazu.

$$f(\mathbf{x}) = \begin{cases} 1 & \text{ako } \mathbf{w} * \mathbf{x} + b > 0 \\ 0 & \text{inače} \end{cases} \tag{4.2}$$

Gornja formula opisuje rad perceptorna, $\mathbf{w}$ je vektor realnih vrijednosti, $\mathbf{x}$ je vektor težinskih vrijednosti a b je konstanta koja ne ovisi o ulazni vrijednostima. Funkcija poprima vrijednost 0 ili 1 i to je zapravo izlaz klasifikatora, odnosno perceptrona. Perceptron je najjednostavniji oblik unaprijednih neuronskih mreža (Theodoridis i Koutroumbas, 2009).

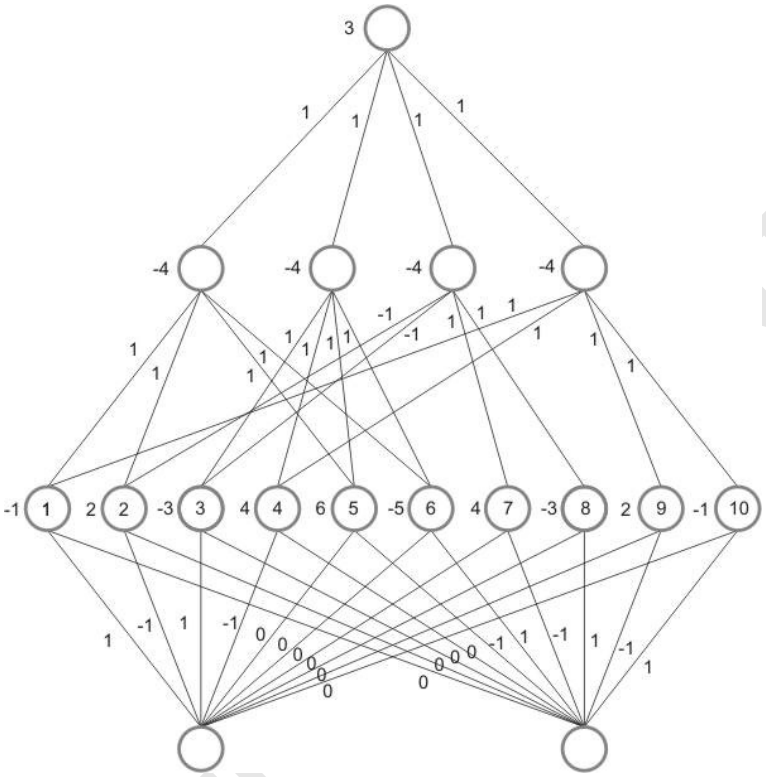


Slika 4.1: Ilustracija stroja sa potpornim vektorima (SVM) preuzeto iz (Theodoridis i Koutroumbas, 2009)

Umjetna neuronska mreža

Umjetne neuronske mreže inspirirane su načinom rada ljudskog mozga. Rad neuronske mreže nije pretjerano kompliciran. U najosnovnijem obliku mreža je sastavljene od mnogo procesnih jedinica, tzv. neurona. Ti neuroni nekim su načinom povezani sa drugim neuronima i za neke svoje ulaze oni daju neki izlaz ovisno o svojoj aktivacijskoj funkciji i taj izlaz prosljeđuju dalje kroz mrežu.

Tipično je mreža podijeljena u ulazni, skriveni i izlazni sloj. Učenjem mreže mi zapravo učimo optimalne težinske parametre na ulazu u svaki neuron (slika 4.2). Kako bismo mrežu mogli učiti moramo prilikom učenja podešavati težine na ulazu u pojedine neurone. Za rješavanje tog problema koristi se algoritam s prostiranjem pogreške unatrag (engl. backpropagation algorithm). Algoritam je prikazan u tablici 4.3, a radi na način da svakom neuronu malo po malo podešava težinu u ovisnosti o označenoj i klasificiranoj vrijednosti. Nakon učenja mreže, ona bi uz veliku količinu podataka za učenje morala naučiti ispravno preslikavanje između ulaznih podataka, značajki ulaznih vektora i izlaznog sloja neurona koji bi te vektore ispravno klasificirao. Nedostatak neurosnkih mreža je taj što pri visokoj dimenzionalnosti ulaznih vektora s jedne strane i malog skupa za učenje s druge strane može dovesti do slabije naučene mreže. Također tipični problemi koji se pojavljuju u umjetnim neuronskim mrežama su mogućnosti prenaučenosti, odlazaka u lokalni optimum i sl. (Theodoridis i Koutroumbas, 2009; Bojana Dalbelo-Bašić)



Slika 4.2: Ilustracija neuronske mreže (Bojana Dalbelo-Bašić)

Inicijaliziraj težinske faktore slučajne vrijednosti
Dok nije ispunjen uvjet zaustavljanja **čini**
  Za svaki $(\boldsymbol{x}, \boldsymbol{t})$ iz D **čini**
    Izračunaj izlaz o_u za svaku jedinicu u
    Za svaku **izlaznu** jedinicu k izračunaj pogrešku

$$\delta_k \leftarrow o_k(1-o_k)(t_k-o_k)$$

    Za svaku **skrivenu** jedinicu izračunaj pogrešku

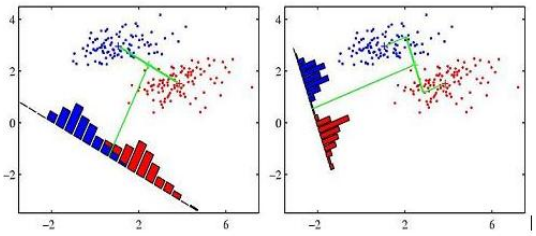
$$\delta_h \leftarrow o_h(1-o_h)\sum_{s\in Downstream(h)}\omega_{hs}\delta_s$$

    Ugodi svaki težinski faktor w_{ij}

$$\omega_{ij} \leftarrow \omega_{ij} + \Delta\omega_{ij}$$

    gdje je $\Delta\omega_{ij} = \eta\delta_j x_{ij}$
  Kraj
Kraj

Slika 4.3: Pseudokod algoritma sa prostiranjem pogreške unazad (Bojana Dalbelo-Bašić)



Slika 4.4: Ilustracija linearne diskriminantne analize (LDA)

[1]

Linearna diskriminantna analiza

Iako prvenstveno služi kao sredstvo za smanjivanje dimenzionalnosti (Theodoridis i Koutroumbas, 2009) vrlo lako se koristi i kao klasifikator. Osnovna ideja linearne diskriminantne analize je N-dimenzionalan vektor značajki reducirati na jednu dimenziju i tada ga upotrijebiti za klasifikaciju. Drugim riječima potrebno je pronaći orijentaciju pravca na koji se projiciraju N-dimenzionalni uzorci $\mathbf{x_i}$ pri čemu je $i = 1, 2, \ldots, N$, ali tako da su projicirani uzorci odvojivi (slika 4.4). To postižemo tako da minimiziramo raspršenje projeciranih uzoraka unutar razreda i maksimiziramo razliku srednjih vrijednosti projiciranih uzoraka.

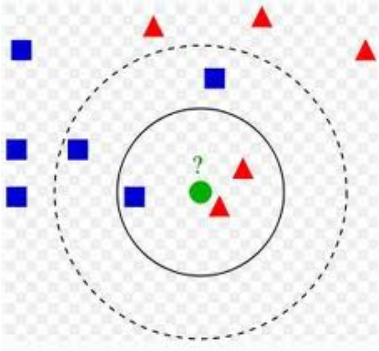
Naivan Bayesov klasifikator

Naivni Bayesov klasifikator jednostavni je vjerojatnosni klasifikator koji se temelji na primjeni Bayesovog teorema i uvjetnoj neovisnosti izmedu značajki ulaznih vektora. Ovisno o preciznosti vjerojatnosnog modela, Bayesov klasifikator može naučiti vrlo efikasno, a prednost mu je i to što mu je za učenje dovoljan manji skup podataka. Neka su ω_1 i ω_2 dvije klase u koje klasificirmao uzorke. Također pretpostavljamo da su poznate *a priori* vjerojatnosti: $P(\omega_1)$ i $P(\omega_2)$.[2] Osim toga još nam je poznata i funkcija gustoće razdiobe za svaki od razreda, $p(\mathbf{x}|\omega_i), i \in 1, 2$. Kada su nam poznate sve vrijednosti može se izračunati uvjetna vjerojatnost 4.3. $p(\mathbf{x})$ je funkcija gustoće razdiobe za $\mathbf{x}$, a računa se kao 4.4. Na kraju se uzorke klasificira u razred ω_1 ako je $P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x})$ ili u razred ω_2 ako je $P(\omega_1|\mathbf{x}) < P(\omega_2|\mathbf{x})$

$$P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i) * P(\omega_i)}{p(\mathbf{x})} \tag{4.3}$$

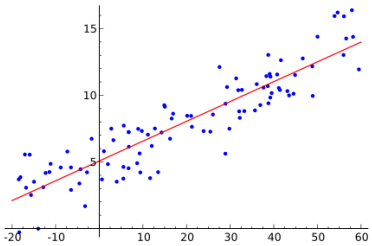
$$p(\mathbf{x}) = \sum_{i=1}^{2} p(\mathbf{x}|\omega_i) * P(\omega_i) \tag{4.4}$$

[2] a priori vjerojatnost računa se iz broja označenih uzoraka, $P(\omega_1) \approx \frac{N_1}{N}$.



Slika 4.5: Ilustracija algoritma k-NN

3



Slika 4.6: Ilustracija linearne regresije

4

k-NN

Algotritam k-NN ili algoritam k-najbližih susjeda je "lijena" metoda kojom se ulazi klasificiraju u ovisnosti o rezultatima k najbližih susjeda u vektorskom prostoru značajki (slika 4.5). Drugim riječima, objekt je klasificiran u ovisnosti o većinskom glasu njegovih k-najbližih susjeda. k-NN je "lijena" metoda jer se svo računanje odgađa do trenutka odluke o klasifikaciji ulaza. Također k-NN je jedan od najjednostavnijih algoritama strojnog učenja.

Linearna i logisticka regresija

Linearna regresija je postupak kojim se pokušava modelirati veza između skalarne varijable y i vektora $\mathbf{x}$ (slika 4.6). U lineranoj regresiji se za modeliranje koristi linearna funkcija, dok se u logističkoj regresiji za modeliranje koristiti sigmoidalna funkcija. Primjer linearne regresije dan je na slici 4.6.

4.1. Rezultati

Svaki od prethodno navedenih algoritama ima svoje prednosti i nedostatke. Osim toga svaki od njih se bolje ili lošije ponaša u nekom skupu podataka, što će se i vidjeti iz dobivenih rezultata.

Ulazni skup podataka sastojao se od 815 pozitivnih ili negativnih rečenica. Taj skup je podijeljen u dva dijela, skup za učenje i skup za ispitivanje u omjeru približno $75 : 25$.

Skup za učenje sastojao se od 599 pozitivnih ili negativnih rečenica. U skupu za učenje bilo je 333 pozitivnih (55.59%) i 266 negativnih (44.41%) rečenica. Ulazni vektor sastojao se od 16 značajki, navedenih u prethodnom tekstu, i oznake razreda kojem pripada. Skup za ispitivanje sastojao se od ukupno 216 rečenica, od toga 138 pozitivnih (63.88%) i 78 negativnih (36.12%) rečenica. Ti podaci su korišteni za ispitivanje svih prethodno navedenih algoritama.

Prilikom učenja korištena je unakrsna provjera u 10 koraka, kako bismo mogli odabrati model s optimalnim hiperparametrima algoritama[5] strojnog učenja. Na izlazu smo dobili optimalan model koji je dalje testiran sa skupom za ispitivanje, nakon čega su dobiveni rezultati prikazani u tablici 4.1.

Tablica 4.1: Preciznost, odziv, F_1-mjera i točnost za sve algoritme

Algoritam	Preciznost	Odziv	F_1-mjera	Točnost
SVM	82.17%	76.26%	79.10%	74.19%
Logistička regresija	74.15%	78.24%	76.14%	68.66%
Linearna regresija	82.76%	51.80%	63.71%	68.66%
Neuronska mreža	75.97%	84.78%	80.13%	73.14%
LDA	75.69%	78.98%	77.30%	70.37%
Naivan Bayes	79.66%	68.11%	73.43%	68.51%
Naivan Bayes (jezgre)	78.29%	73.18%	75.64%	69.91%
Perceptron	64.18%	100.00%	78.18%	64.35%
k-NN	75.18%	74.63%	74.91%	68.05%

Iz tablice 4.1 vidljivo je da su najbolje rezultate dali SVM (74.19%), zatim logis-

[5]Hiperparametri su parametri algoritma, koji se postavljaju "ručno" i ne mijenja ih algoritam učenja. Ti parametri su npr. broj slojeva i neurona u neuronskoj mreži, broj najbližih susjeda (k) u k-NN, itd. Za njihovo optimalno određivanje koristi se unakrsna validacija koja testira više različitih vrijednosti i kao optimalan hiperparametar daje onaj koji se pokazao kao najbolji na provjeri.

tička regresija (68.66%), pa neuronska mreža i ostale metode koje daju rezultate koji se kreću oko 60% tocnosti.

Rezultate ćemo uspoređivati korištenjem matrice zabune i vrijednostima koji se iz te tablice izvode. Tablica 4.2 prikazuje matricu zabune, a mjere koje ćemo koristiti bit će: preciznost ($P = \frac{a}{a+b}$), odziv ($R = \frac{a}{a+c}$), F_1-mjera ($F_1 = \frac{2*P*R}{P+R}$) i točnost ($T = \frac{a+d}{a+b+c+d}$).

Tablica 4.2: Matrica zabune

	Istina: pozitivno	Istina: negativno
Model: pozitivno	a	b
Model: negativno	c	d

Sve mjere za sve algoritme strojnog učenja prikazane su u tablici 4.1. Za SVM, logističku i linearnu regresiju ćemo detaljnije prikazati rezultate, F_1-mjeru, preciznost, odziv te matrice zabune. Za SVM matrica zabune prikazana je u tablici 4.3.

Tablica 4.3: SVM - matrica zabune

	Istina: pozitivno	Istina: negativno
Model: pozitivno	106	23
Model: negativno	33	55

Tablica 4.4: Logistička regresija - matrica zabune

	Istina: pozitivno	Istina: negativno
Model: pozitivno	109	38
Model: negativno	30	40

Za logističku regresiju matrica zabune vidljiva je u tablici 4.4, dok tablica 4.5 prikazuje tablicu zabune za linearnu regresiju.

Tablice koje 4.4 i 4.5 daju rezultate za cijeli skup podataka za testiranje. Kako bi provjerili rezultate tih algoritama nakon sažimanja, odnosno uzimanja 15% najpozitivnijih i najnegativnijih recenzija ispitali smo stroj s potpornim vektorima (SVM) i logističku regresiju, na tom smanjenom skupu podataka. Rezultati su ocekivano bolji iz razloga što neke rečenice mogu biti neutralne, a njih se mora svrstati u pozitivne i negativne te tu dolazi do grešaka. Te greške se smanjuju jer uzimamo najpouzdanije rezultate iz oba razreda. Rezultati tih testiranja prikazani su u tablici 4.6 za SVM

Tablica 4.5: Linearna regresija - matrica zabune

	Istina: pozitivno	Istina: negativno
Model: pozitivno	72	15
Model: negativno	67	63

te u tablici 4.7 za logističku regresiju. Za SVM izračunali smo mjere i one iznose $P = 82.19\%$, $R = 76.26\%$, $F_1 = 79.10\%$ i $T = 74.19\%$. Iste te mjere za logisticku regresiju su $P = 74.15\%$, $R = 78.24\%$, $F_1 = 76.14\%$ i $T = 68.66\%$.

Tablica 4.6: SVM - matrica zabune, za manji skup podataka

	Istina: pozitivno	Istina: negativno
Model: pozitivno	49	15
Model: negativno	1	30

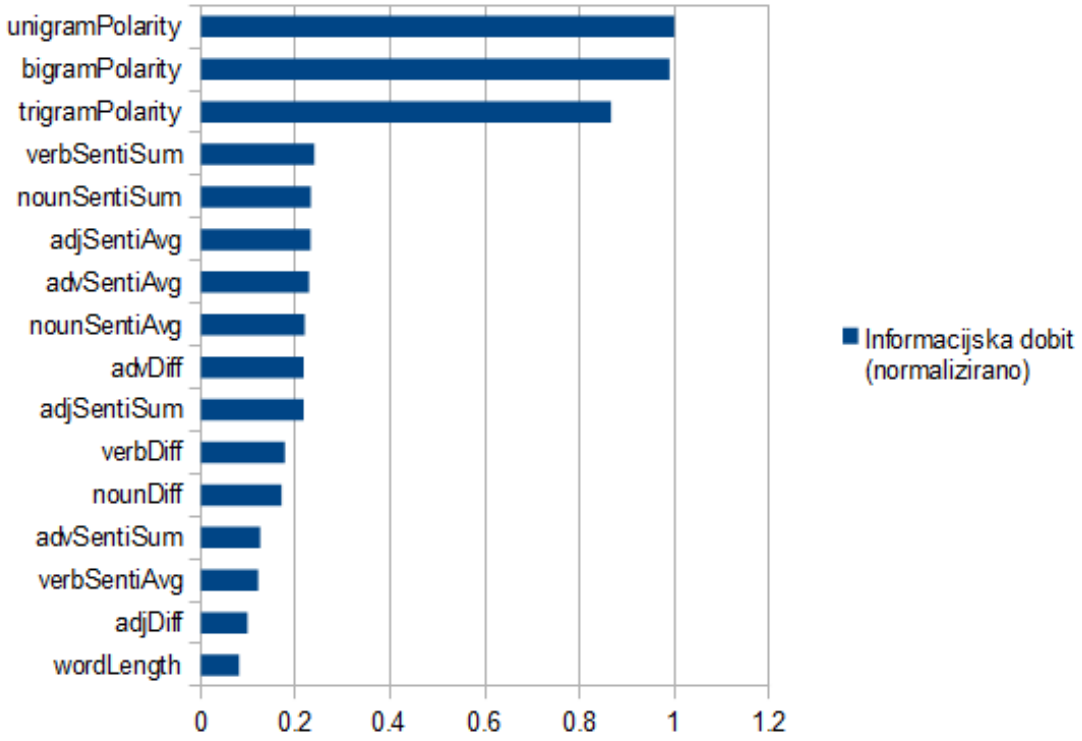
Tablica 4.7: Logistička regresija - matrica zabune, za manji skup podataka

	Istina: pozitivno	Istina: negativno
Model: pozitivno	50	16
Model: negativno	0	24

4.2. Eksperimenti

Iako smo dosad proveli samo jednostavna testiranja, možemo izvuci neke zaključke. Jedna od bitnih stvari jest ta da metode koje se temelje na pouzdanosti daju puno bolje rezultate.

Razlog tom ponašanju je to što smo učili sve rečenice klasificirati samo u pozitivne i negativne, dok realno postoje minimalno četiri klase (iste one koje smo koristili pri označavanju). Zbog toga dolazi i do lošijih rezultata svih klasifikatora. Ta razlika možda nije toliko očita nakon samog testiranja, no u naknadnoj obradi podataka kada se uzima 15% najpozitivnijih odnosno najnegativnijih, do izražaja dolazi pouzdanost jer ona daje neku mjeru po kojoj možemo naći ekstreme s obje strane. Kod binarnog klasifikatora (npr. perceptron) imamo za svaki vektor na izlazu samo predikciju u obliku pozitivno ili negativno, tako da ne možemo po ničemu naći ekstreme.



Slika 4.7: Informacijska dobit atributa

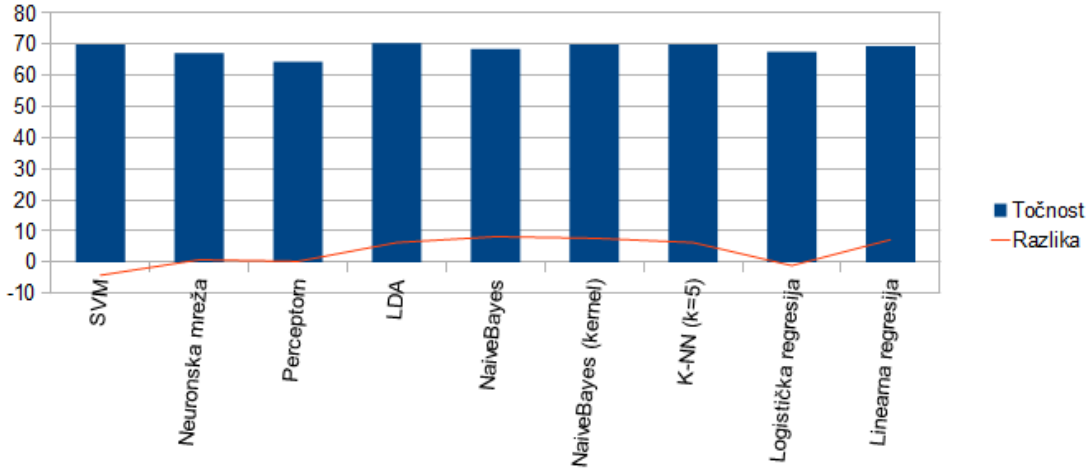
4.2.1. Atributi

Nakon tih jednostavnih testiranja odlučili smo se poigrati s atributima, odnosno testirati koji atributi donose najviše i najmanje informacija te kako se sustav ponaša ukoliko ih maknemo.

$$IG = -\sum_{i=1}^{M} P(\omega_i|t) * \log_2 P(\omega_i|t) \tag{4.5}$$

Pomoću formule 4.5, gdje je M broj razreda (u našem slučaju samo dva, pozitivni i negativni) i $P(\omega_i|t)$ vjerojatnost da vektor u skupu podataka koji sadrži vrijednost t pripada klasi i (Theodoridis i Koutroumbas, 2009). Jednostavnom analizom informacijske dobiti za svaki atribut dobili smo uređeni normalizirani prikaz atributa, prikazan na slici 4.7.

Iz ovoga vidimo nekoliko zanimljivih stvari. Možemo primjetiti da postoje tri atributa koja u najvećoj mjeri donose informaciju dok ostali donose puno manje. Takve stvari događaju se vjerojatno zbog pravopisnih grešaka, koje se ne mogu ispraviti. Takve se rijeci ne mogu ispravno koristiti u resursu SentiWordNet te otuda dolazi mala informacijska dobit za te atribute. No usprkos tome testirat ćemo kako se sustav po-



Slika 4.8: Točnost algoritama sa reduciranim skupom na tri najbolja atributa

naša kada ostavimo samo tri atributa koja unose najviše informacija te kako se ponaša kada maknemo tri koja donose najmanje informacija.

4.2.2. Rezultati nakon redukcije atributa

Za sve klasifikatore smo ponovili testove ali samo s različitim skupom za učenje te smo izračunali matrice zamjene za SVM i logističku regresiju kako bi mogli bolje usporediti rezultate.

Samo tri najbolja atributa

Za ovaj slučaj izbacili smo sve atribute osim unigramske, bigramske i trigramske polarnosti. Rezultati koje smo dobili prikazani su na slici 4.8

Crvenom linijom prikazana je razlika u odnosu na točnost koja je računata sa svim atributima. Zanimljivo je primjetiti da se u prosjeku točcnost i povećala, iako je svm-u točnost pala za 4.2% posto. Osim toga izracunate su i matrice zabune za SVM kada imamo cijeli skup (tablica 4.8) i na skupu 20% najpozitivinijih i najnegativnijih. Iz tih vrijednosti vidimo da je SVM-u pala točnost nakon što smo ostavili samo tri najinformativnija atributa.

Bez tri najlošija atributa

Osim analize skupa atributa sa samo tri najbolja atributa, napravili smo i analizu svih atributa osim tri najlošija, broja rijeci, razlike između pridjeva te razlike između gla-

Tablica 4.8: SVM - matrica zabune

	Istina: pozitivno	Istina: negativno
Model: pozitivno	107	34
Model: negativno	31	44

Tablica 4.9: SVM - matrica zabune, za sažeti skup podataka

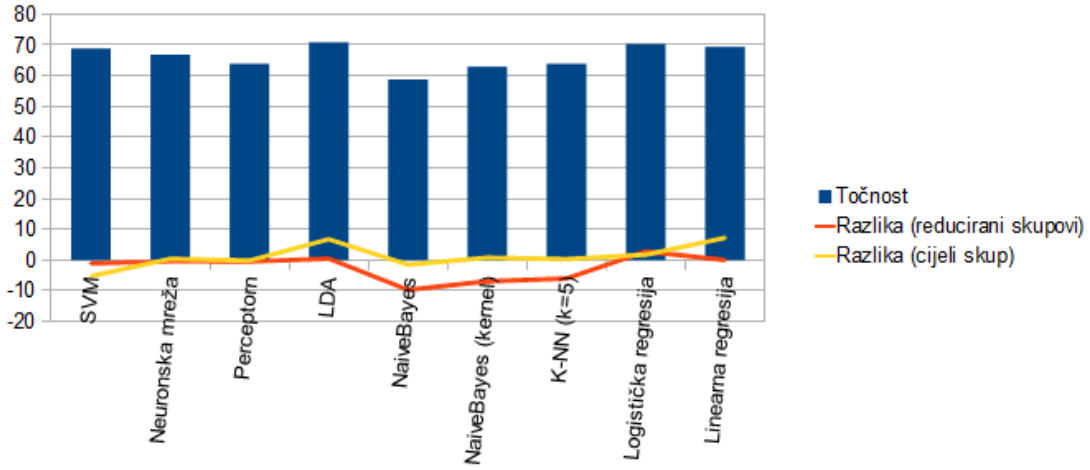
	Istina: pozitivno	Istina: negativno
Model: pozitivno	49	16
Model: negativno	1	24

gola. Dobivena točnost algoritama prikazana je na slici 4.9, zajedno s razlikama između točnosti algoritama s atributima iz prošlog skupa te s točnosti algoritama s cijelim skupom atributa.

Iz slika je vidljivo da uklanjanje atributa donekle ima utjecaja na rezultate performansi algoritama strojnog učenja. Također, stroj s potpornim vektorima, koji je na cijelom skupu davao najbolje rezultate nakon redukcije skupa atributa, pokazao je nešto lošije performanse. Pri tom su se pojavili neki drugi klasifikatori koji su relativno dobro obavljali posao. Tu ćemo izdvojiti linearnu regresiju, koja uz manji skup atributa, pokazuje jedne od boljih performansi (točnost se kreće oko 70%). Takoder je pokazano da u ovom specifičnom zadatku, analize i sažimanja sentimenta u recenzijama hostela, sentiment možemo dovoljno dobro aproksimirati s tri atributa, polarnost unigrama, bigrama i trigrama. Linearna regresija je zadatak s tim reduciranim skupom obavila najbolje, iako na cijelom skupu ima nešto lošije rezultate. U nastavku je dana matrica zabune za diskriminantnu analizu na reduciranom skupu atributa sa samo tri najbolja, i to za cijeli skup 4.10 i za samo 15% najboljih odnosno najlošijih, tablica 4.11.

Tablica 4.10: Linearna regresija - matrica zabune

	Istina: pozitivno	Istina: negativno	Preciznost
Model: pozitivno	104	32	76.47%
Model: negativno	34	46	57.50%
Odziv	75.36%	58.97%	F_1-mjera: 75.91%



Slika 4.9: Točnost algoritama s reduciranim skupom bez tri najbolja atributa

Tablica 4.11: Linearna regresija - matrica zabune, za sažeti skup podataka

	Istina: pozitivno	Istina: negativno	Preciznost
Model: pozitivno	46	18	68.18%
Model: negativno	0	20	100%
Odziv	100%	53.33%	F_1-mjera: 81.07%

5. Prilagodba sustava za hrvatski jezik

Jedna od dodatnih tema ovoga rada bila je ispitati ponašanje sustava na hrvatskom jeziku. Naravno alati koji su se koristili prilikom obrade ulaznih podataka, nisu izravno iskoristivi na hrvatskom jeziku, npr. označivanje vrste rijeci, provjera pravopisa. Upravo zbog toga sustav će trebati prilagoditi hrvatskom jeziku te nakon toga ponovno provesti ispitivanje. Sustav ce se prilagoditi hrvatskom jeziku *ad hoc* metodama koje će biti opisane u nastavku.

5.1. Prikupljanje ulaznih podataka

Internet je, iako globalni proizvod, ipak najviše zastupljen na engleskom, tako su i podaci koje smo prikupili za sažimanje recenzija na engleskom. Također većina sadržaja koja sadrži recenzije ili komentare je na engleskom te nisu primjenjivi za hrvatski jezik, također tu se ubraja većina resursa koji sadrže komentare i recenzije u većem broju. Ali ipak je pronađen resurs na hrvatskom koji sadrži recenzije u većim količinama. Web stranica[1] sadrži recenzije i komentare filmova, kako službene tako i anonimne te ce poslužiti kao odličan izvor podataka.

5.2. Obrada podataka

Kao što smo vec naglasili, alati koje smo koristili za engleski jezik neće biti primjenjivi na hrvatskom, tako da će se morati pribjeći nekim jednostavnijim metodama te nekim ugađanjima.

Recenzije su dohvaćene sa Interneta te se također nalaze u JSON datotekama u istom formatu kao i recenzije na engleskom. Kako bi smo se detaljnije upoznali sa problematikom prilagodbe sustava hrvatskom jeziku, detaljnije ćemo opisati koji se gotovi alati neće moći koristiti te što ćemo koristiti kao njihovu zamjenu.

[1] `www.filmski.net`

– Razdvajanje komentara na rečenice - za englesku verziju za to je korišten Stanford NLP, dok ćemo za hrvatsku verziju komentar dijeliti na recenice tako da svaku točku označimo kao kraj rečenice te dodatno izbacimo prljave podatke koji se mogu stvoriti na taj način (npr. ... će biti u tom postupku izbačene).

– Označavanje vrste rijeci - zbog nedostupnosti alata za označavanje vrste rijeci ovaj dio se neće koristiti na hrvatskom već će se vrsta rijeci doznati kroz resurs SentiWordNet.[2] On u svom korpusu za svaku riječ sadrži i njenu vrstu tako da ako uspijemo pronaci rijec u većini slučajeva ćemo doznati i vrstu rijeci.

– SentiWordNet - problem kod korištenja ovog možda i najbitnijeg resursa je u tome što je korpus na engleskom jeziku, a u verziji sustava za hrvatski jezik sve je na hrvatskom jeziku. Rješavnju problema ćemo priskočiti tako što ćemo koristiti Googleov API za prevodenje.[3] Dakako neće sve riječi biti ispravno prevedene, no to ce biti dovoljno dobra početna pozicija.

– Također nećemo biti u mogućnosti koristiti Googleov API za provjeru pravopisa jer trenutno nije podržan hrvatski jezik. No taj problem će se riješiti kroz prevođenje riječi, jer se prilikom prevođenja ispravljaju pravopisne greške.

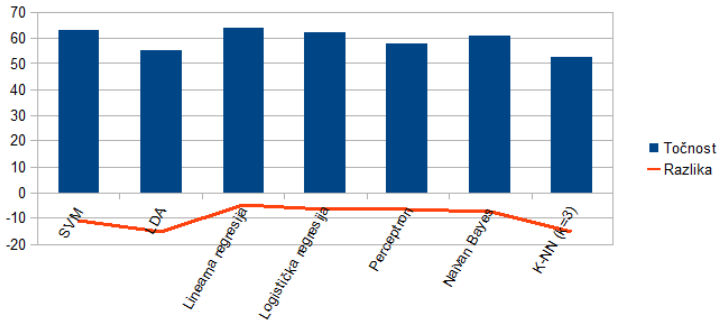
Pošto smo za većinu alata našli zamjenu, skup značajki ostati ce nepromijenjen, tako da ćemo prilikom testiranja klasifikatora moći koristiti i ona koji su učili značajke na engleskom jeziku te one koji su ućili specifično za hrvatski kako bi mogli usporediti rezultate.

5.3. Rezultati

Skup za učenje na hrvatskom jeziku bio je u odnosu na skup na engleskom duplo manji zbog toga što smo za hrvatski skup za učenje uzimali recenzije filmova koje same po sebi ne sadrže toliko sentimenta koliko prepričavanja sadržaja filmova. Iako je ocijenjen skoro dvostruko veći skup (1600 rečenica) nego na engleskom, pozitivno odnosno negativno ocijenjenih rečenica ima svega 411. Od cijelog skupa pozitivno ih je 226 (54.98%), a negativno 185 (45.02%). Tako malen broj recenzija dovest će do znatno slabijih rezultata klasifikatora koji će biti učeni na hrvatskom jeziku.

[2] `http://sentiwordnet.isti.cnr.it/`

[3] `http://code.google.com/apis/language/translate/overview.html`



Slika 5.1: Usporedba tocnosti klasifikatora ucenog na engleskom izmedu ispitnog skupa na engleskom i hrvatskom

5.3.1. Klasifikator učen na engleskom skupu

U prvom koraku testiranja sustava na hrvatskom jeziku uzeli smo klasifikator koji je naučen na engleskom jeziku te ga primjenili na hrvatski. Sustav smo ispitivali sa cijelim skupom na hrvatskom jeziku. Dobiveni rezultati prikazani su u tablici 5.1. Iz rezultata je vidljivo da, iako su originalne recenzije bile na hrvatskom prethodno opisanim *ad hoc* zahvatima uspjeli smo dovoljno dobro prilagoditi sustav za hrvatski jezik te da su se opet kao najbolji klasifikatori pokazali SVM te linearna i logistička regresija. Također tu valja istaknuti da je ovdje prikazano ne samo da se sustav može nositi sa više jezika, nego i sa različitim stilovima recenzija (za dobar rad sustava nije nužna samo jedna vrsta recenzija). Na slici 5.1 prikazana je usporedba točnosti klasifikatora učenih na engleskom jeziku, kada je ispitni skup bio na hrvatskom odnosno, engleskom.

Tablica 5.1: Preciznost, odziv, F_1-mjera i točnost za sve algoritme

Algoritam	Preciznost	Odziv	F_1-mjera	Točnost
SVM	68.37%	61.19%	64.58%	63.16%
LDA	54.89%	100%	70.88%	54.89%
Linearna regresija	61.9%	89.04%	73.03%	63.91%
Logistička regresija	65.32%	66.21%	65.76%	62.16%
Naivan Bayes	69.81%	50.68%	58.73%	60.9%
Perceptron	76.6%	32.88%	46.01%	57.64%
k-NN	54.14%	89.5%	67.47%	52.63%

5.3.2. Klasifikator učen na hrvatskom skupu

Kao što je već rečeno, zbog malog skupa na hrvatskom jeziku, rezultati su vidljivo lošiji. Skup smo kao i kod učenja klasifikatora na engleskom jeziku podijelili na skup za učenje i skup za ispitivanje u omjeru $75\% : 25\%$. Takva podjela je dovela do smanjenja ionako malog skupa na hrvatskom. No usprkos tome proveli smo učenje i ispitivanje, a rezultati su vidljivi u tablici 5.2.

Tablica 5.2: Preciznost, odziv, F_1-mjera i točnost za sve algoritme

Algoritam	Preciznost	Odziv	F_1-mjera	Tocnost
SVM	48.15%	73.58%	58.21%	50.44%
LDA	58.33%	13.21%	21.54%	54.87%
Linearna regresija	70%	13.21%	22.23%	56.64%
Logistička regresija	50%	67.92%	57.6%	53.1%
Perceptron	46.9%	75.71%	57.92%	40.77%
Naivan Bayes sa jezgrama	44.9%	83.02%	58.28%	44.25%

6. Zaključak

U ovom rad razvijen je sustav za sažimanje korisničkih recenzija za engleski i hrvatski jezik. Sustav obuhvaća dohvaćanje skupova za učenje, izvlačenje značajki, učenje, ispitivanje te na kraju sažimanje recenzija. Sustav je detaljno ispitan na hrvatskom i engleskom jeziku korištenjem različitih klasifikatora.

Iz rezultata koji su dobiveni može se zaključiti da sustav iako ne radi savršeno ipak radi dovoljno dobro za svrhu koju je i namijenjen, a to je prvenstveno kao pomoć korisnicima u donošenju odluka. Dobiveni rezultati također prikazuju da se ovaj sustav ponaša podjednako kao i većina sustava razvijenih u sklopu radova sa sličnom tematikom.

Ispitano je i ponašanje sustava na više jezika: engleskom i hrvatskom. Prikazano je da i kada se sustav nauči na engleskom jeziku on se može koristiti i za sažimanje recenzija na hrvatskom jeziku, doduše, sustav će imati lošije performanse, no ipak dovoljno dobre za svoju svrhu. Uz to prikazano je i ponašanje sustava na dvije vrste recenzija. Na engleskom smo imali skup recenzija hostela, dok smo na hrvatskom imali skup recenzija filmova.

Jedan od mogućih budućih koraka u unaprijeđenju sustava moglo bi biti izvlačenje ključnih riječi u cijelim recenzijama ili rečenicama. Na taj način moglo bi se omoguciti sažimanje recenzija oko ključnih rijeci npr. kod recenzija kamera, moglo bi se posebno sažimati recenzije koje se bave objektivima, rezolucijom itd. Još jedno od mogućih unaprijedenja moglo bi biti podržavanje višejezičnosti uz korištenje dostupnih alata. Na isti način na koji se radila prilagodba na hrvatski, mogla bi se raditi prilagodba na druge jezike. Time bi ovaj sustav mogao postati globalno tražen i zanimljiv proizvod.

Na kraju treba zaključiti da u današnjem svijetu, gdje svijet postaje "sve manji i manji", broj informacija sve veći, a tempo života sve brži postoji realna potreba za ovakvim sustavom, pogotovo iz razloga što smo u ovom radu pokazali da je izrada takvog sustava moguca.

LITERATURA

Hrvatski jezični portal, 2011. URL http://hjp.srce.hr.

Sentiwordnet, 2011. URL http://sentiwordnet.isti.cnr.it/.

Marko Čupić i Jan Šnajder Bojana Dalbelo-Bašić. *Umjetne neuronske mreže*. URL http://www.fer.hr/_download/repository/UmjetneNeuronskeMreze.pdf.

Ralph Johnson Erich Gamma, Richard Helm i John Vlissides. *Design patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley, 1994.

Schaye A. Marcus M. Nenkova A. Feczko, M. Sentisummary: Sentisummarization for user product review, 2010.

E. H. Hovy. *The Oxford Handbook of Computational Linguistics, chapter 32, pages 583-598*. Oxford University Press, 2005.

Maybury M. T. Mani, I. *Advances in Automatic Text Summarization*. The MIT Press, 1999.

Yelena Mejova. *Sentiment Analysis, An Overview*, 2009. URL http://uiowa.academia.edu/YelenaMejova/Papers/241860/Sentiment_Analysis_An_Overview.

Lee L. Pang, B. Thumbs up?: sentiment classification using machine learning techinques, 2002.

Lee L. Pang, B. *Opinion mining and sentiment analysis. Foundation and Trends in Information Retrieval*. 2008.

C. Potts. *Textual Sentiment Summarization*, 2010. URL http://www.stanford.edu/class/cs424p/materials/ling287-handout-10-05-textual-summarization.pdf.

Sergios Theodoridis i Konstantions Koutroumbas. *Pattern Recognition*. Academic Press, 2009.

Litman M. L. Turney, P. D. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Inforamtion Systems*, 2003.

Bruce R. O'Hara T. Wiebe, J. Development and use of a gold standard data set for subjectivity classifications. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguists*, 1999.

Sažimanje korisničkih recenzija metodama strojnog učenja

Sažetak

Automatska analiza mišljenja (engl. opinion mining) izraženog u tekstu važno je i sve popularnije područje istraživanja s brojnim mogućnostima primjene. Od posebnog su interesa postupci strojnog sažimanja recenzija koje je pisalo više korisnika (npr. recenzije filmova ili proizvoda). U okviru diplomskog rada proučeni su postupci za određivanje semantičke orijentacije rijeci , fraza i rečenica te postupci strojnog sažimanja. Razrađen je postupak za sažimanje recenzija na engleskom jeziku temeljen na metodama strojnog ucenja, a koji u obzir uzima semantičku orijentaciju rečenica. Razvijena je programska implementacija postupaka i provedeno je eksperimentalno vrednovanje postupka na odgovarajućem uzorku. Postupak je također primjenjen i ispitan na sažimanju recenzija na hrvatskom jeziku.

Ključne riječi: sentiment, sažimanje, recenzije, hrvatski, engleski

Applying machine learning methods to user review summarization

Abstract

Automated opinion-mining is important and increasingly popular field of study with a wide range of applications. Of particular interest are the processes of machine review summarization written by multiple users (npr. movie or product reviews). In this thesis, the procedures for determining the semantic orientations of words, phrases and sentences and the machine summarization have been studied. A review summarization process in English language based on machine learning methods taking into account the semantic orientation of the sentence has been developed. A software implementation of described process has been developed and an experimental process evaluation on an adequate sample has been conducted. The process has also been applied and tested on review summarization in the Croatian language.

Keywords: Keywords.