TakeLab

Laboratorij za analizu teksta i inženjerstvo znanja
Text Analysis and Knowledge Engineering Lab
Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva
Unska 3, 10000 Zagreb, Hrvatska

Zaštićeno licencijom
Creative Commons Imenovanje-Nekomercijalno-Bez prerada 3.0 Hrvatska
https://creativecommons.org/licenses/by-nc-nd/3.0/hr/

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 3796

Postupak za polunadziranu akviziciju leksikona sentimenta

Matej Paradžik

Zagreb, lipanj 2014.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA
ODBOR ZA ZAVRŠNI RAD MODULA

Zagreb, 13. ožujka 2014.

ZAVRŠNI ZADATAK br. 3796

Pristupnik: **Matej Paradžik**
Studij: Računarstvo
Modul: Računarska znanost

Zadatak: **Postupak za polunadziranu akviziciju leksikona sentimenta**

Opis zadatka:

Porastom raspoloživih količina korisnički generiranog sadržaja povećalo se zanimanje za strojnom analizom mišljenja izraženog u tekstu. Jedan od pristupa analizi mišljenja jest analiza sentimenta, kojom se utvrđuje je li tekst usmjeren pozitivno, negativno ili neutralno. Uobičajeni postupci analize sentimenta temelje se na leksikonu apriornog sentimenta. Ručna izgradnja leksikona sentimenta odgovarajućeg opsega izuzetno je naporna i skupa. Stoga je u literaturi predložen niz postupaka za automatsku akviziciju sentimenta iz korpusa temeljenih na polunadziranome strojnom učenju.
U okviru završnoga rada potrebno je proučiti postupke za automatsku akviziciju sentimenta s naglaskom na polunadziranim metodama. Razraditi postupak za akviziciju sentimenta riječi hrvatskoga jezika koji će se oslanjati na informacije o odnosima između riječi dobivenima statističkom obradom korpusa, uključivo informacijama o njihovim sintaktičkim i semantičkim odnosima. Implementirati postupak u programskome jeziku po izboru, oslanjajući se na dostupne jezičnotehnološke alate za hrvatski jezik. Provesti iscrpno eksperimentalno vrednovanje na odgovarajućim ručno označenim skupovima podataka te detaljnu analizu pogrešaka. Radu priložiti izvorni i izvršni kod razvijenog sustava, skupove podataka i programsku dokumentaciju te citirati korištenu literaturu.

Zadatak uručen pristupniku: 14. ožujka 2014.
Rok za predaju rada: 13. lipnja 2014.

Mentor:

Doc. dr.sc. Jan Šnajder

Djelovođa:

Doc. dr.sc. Tomislav Hrkać

Predsjednik odbora za završni rad modula:

Prof. dr.sc. Siniša Srbljić

SADRŽAJ

1. Uvod

Od 2000. godine do danas, broj korisnika weba se povećao više od pet puta (ITU, 2014), a sve više korisnika sudjeluje u stvaranju sadržaja na webu kroz društvene mreže, forume, komentare aktualnih zbivanja i recenzije proizvoda. Jedna od stvari koja obilježava te sadržaje je da najčešće nose sentiment, odnosno mišljenje. Sentiment se odnosi na pozitivne ili negativne emocije, procjene i stavove prema nekome ili nečemu.

Ljudi često svoje odluke temelje na mišljenu i iskustvima drugih ljudi: koji automobil kupiti, gdje putovati, koji film gledati i slično. S druge strane, brojne organizacije paze na reputaciju, pa ulažu velike novce na praćenje javnog mijenja o sebi kako bi imale što bolju kontrolu nad slikom koju grade u javnosti. Prije weba do tih informacija se dolazilo razgovorom s prijateljima ili poznanicima u slučaju pojedinaca, odnosno provođenjem telefonskih i ostalih anketa (u slučaju organizacija). Međutim, sad su sve te i još mnoge informacije dostupne na webu. Zbog toga se javlja potreba za automatskim donešenjem zaključka o sentimentu koji neki sadržaj (tekst) nosi.

Područje obrade prirodnog jezika koje se bavi tom problematikom je analiza sentimenta (engl. *sentiment analysis*) koja je postala aktualna oko 2000. godine (kad je krenuo rast weba). Glavni problemi analize sentimenta su utvrđivanje je li sentiment izražen, a ukoliko jest utvrđivanje svojstava sentimenta:

- tko iznosi sentiment,
- na koga ili što se sentiment odnosi,
- orijentacija sentimenta (pozitivan, negativan, neutralan) te
- intezitet sentimenta.

Analiza sentimenta se može provoditi na više razina: na razini dokumenta, rečenice i na razini pojedinih riječi i fraza.

Primjene analize sentimenta su brojne. Tako Hu i Liu (2004) analiziraju recenzije proizvoda tako da za svaku značajku pojedinog proizvoda (npr. kvaliteta slike kod digitalne kamere) odrede sentiment kojim je ta značajka opisana te sumiraju recenzije

pojedinog proizvoda tako da za svaku njegovu značajku odrede broj pozitivnih i negativnih spominjanja te značajke u svim recenzijama tog proizvoda. Somasundaran et al. (2007) primjenjuju analizu sentimenta kod poslovnih sastanaka na problem odgovaranja na pitanja. O'Connor et al. (2010) analiziraju javno mijenje o političarima na temelju Twitter poruka.

Prethodni primjeri primjene analize sentimenta oslanjaju se na apriorno sastavljen leksikon sentimenta. Leksikon sentimenta je resurs koji nudi informacije o sentimentu koji pojedina riječ nosi, a eventualno i informaciju o intezitetu istog. Primjerice, riječ *dobar* nosi pozitivan, riječ *loš* negativan, a riječ *stol* neutralan sentiment.

Za mnoge jezike postoje izgrađeni i dostupni kvalitetni leksikoni sentimenta, npr. SentiWordNet (Baccianella et al., 2010) za engleski jezik. No, osim što za neke jezike leksikoni sentimenta nisu dostupni ili nisu dovoljno dobri, važno je primjetiti da iste riječi u različitim domenama primjene (kontekstima) mogu imati različit apriorni sentiment. Zbog toga je kod primjene analize sentimenta na određenoj domeni potreban leksikon sentimenta prilagođen toj istoj domeni.

Budući da je ručna izgradnja leksikona sentimenta dugotrajan i skup proces, na važnosti dobivaju automatski postupci izgradnje leksikona te će se u nastavku ovog rada dati pregled postupaka iz literature za akviziciju leksikona. Također, bit će opisana polunadzirana metoda izgradnje leksikona sentimenta za hrvatski jezik koja se temelji na odnosima između riječi dobivenih statističkom obradom tekstne zbirke.

2. Pregled postupaka iz literature

Tri glavna pristupa izgradnji leksikona sentimenta su:

- ručna izgradnja,
- izgradnja na temelju rječnika (engl. *dictionary-based*) te
- izgradnja na temelju tekstne zbirke (engl. *corpus-based*).

Ručna izgradnja, kao što je rečeno u uvodu, je dugotrajan i zahtjevan proces, stoga se obično ne koristi sama, već u kombinaciji s automatiziranim postupcima (druga dva pristupa), kao krajnja provjera dobivenog leksikona, budući da automatizirani postupci griješe.

Pristupi izgradnje temeljeni na rječniku i na tekstnim zbirkama najčešće rade tako da krenu od dva mala početna skupa riječi (skupa pozitivnih riječi i skupa negativnih riječi). Riječi u početnim skupu se biraju tako da njihova pozitivnost, odnosno negativnost, ne ovisi ili minimalno ovisi o kontekstu u kojem se te riječi nalaze. Potom na razne načine nastoje proširiti početne skupove i tako dobiti leksikon sentimenta.

U nastavku poglavlja dan je pregled polunadziranih postupaka izgradnje leksikona sentimenta temeljenih na rječniku i temeljenih na tekstnoj zbirci.

2.1. Pristupi temeljeni na rječniku

U ovom poglavlju dana je definicija rječnika u kontekstu izgradnje leksikona sentimenta te pregled pristupa izgradnje leksikona sentimenta temeljenih na rječnik.

2.1.1. Rječnik

U kontekstu izgradnje leksikona sentimenta, rječnik prestavlja resurs koji sadrži leksičke i semantičke podatke o riječima. To znači da osim klasičnih informacija o pojedinoj riječi kao što su vrsta riječi i značenje, nudi i informacije o semantičkim vezama među riječima. Neke se navedene u nastavku.

Sinonimija Dvije riječi pripadaju istoj vrsti riječi, a značenje im se poklapa (npr. tuga i žalost).

Antonimija Dvije riječi suprotne po značenju (npr. otvoriti i zatvoriti).

Hipernimija Jedna riječ po značenju sadrži drugu, odnosno nadređena je drugoj (npr. životinja je hiperonim od pas).

Hiponimija Jedna riječ je po značenju sadržana u drugoj, odnosno podređena je drugoj (npr. pas je hiponomin od životinja).

Polisemija Jedna riječ, ovisno o kontekstu, ima više značenja (npr. tuča).

WordNet (Miller et al., 1990) je jedan takav ručno izgrađeni resurs za engleski jezik koji se najčešće koristi u literaturi. WordNet je mreža riječi grupiranih u skupove sinonima (engl. *synsets*), odnosno u skupove istoznačnih riječi. Ti skupovi su međusobno povezani semantičkim relacijama. Jedna takva je veza *JE* (engl. *IS A*) koja spaja hiponim (specifičniji skup sinonima) s hipernimom (općenitijim skupom sinonima). Osnovna organizacija je stablo: u korijenu stabla nalazi se najopćenitiji skup sinonima, a na dnu najspecifičniji. Skupovi sinonima koji su suprotni po značenju povezani su relacijom antonimije. Osim toga WordNet vodi računa o višeznačnosti riječi, tako da se ista riječ može nalaziti u više skupova sinonima, ovisno u značenju.

Dakle, WordNet omogućava jednostavno dobivanje veza između riječi što je, kao što ćemo vidjeti, jako korisno u izgradnji sentimenta leksikona.

2.1.2. Osnovna ideja pristupa temeljenih na rječniku

Uzmimo dvije riječi i njihove rječničke definicije[1]:

1. odličan - koji se ističe najboljim osobinama, izvrstan, poseban, vrhunski
2. super - žarg. izvrstan, odličan, najbolji

Iz rječničkih definicija vidimo da su riječi sinonimi. Ako se uz to zapitamo kakav apriorni sentiment te riječi imaju, vidjet ćemo da je i on jednak: pozitivan.

Ovaj kratki primjer ocrtava osnovnu pretpostavku na kojoj se temelji izgradnja leksikona sentimenta na temelju rječnika: riječi koje imaju slično značenje imaju i jednako orijentiran sentimenta. Jednako tako, riječi koje imaju suprotno značenje imaju i suprotnu orijentaciju sentimenta.

[1] Definicije su uzete s http://hjp.novi-liber.hr/.

Ako se sjetimo rječnika u konteksu izgradnje leksikona sentimenta, vidimo da on nudi upravo takve informacije o riječima. Za svaku riječ možemo doznati njene sinonime (semantički jednake riječi) te njene antonime (semantički suprotne riječi). To ga čini idealnim resursom za izgradnju leksikona sentimenta.

2.1.3. Postupak temeljen na sinonimima i antonimima riječi

Najjednostavniji pristup izgradnji leksikona sentimenta na temelju rječnika uzima u obzir samo informacije o sinonimima i antonimima pojedinih riječi (Hu i Liu, 2004).

Prvi korak je ručno određivanje dva mala skupa riječi: skupa riječi s pozitivnim sentimentom i skupa riječi s negativnim sentimentom. Postupak potom proširuje te skupove tako da traži (na temelju WordNet-a ili nekog drugog rječnika) sinonime i antonime riječi iz početnih skupova. Sinonime riječi iz pozitivnog skupa dodaje u taj isti skup, a antonime u skup riječi s negativnim sentimentom. Jednako tako sinonime riječi iz negativnog skupa dodaje u taj isti skup, a antomnime u skup riječi s pozitivnim sentimentom. Postupak se iterativno ponavlja sve dok se u skupove dodaju riječi. Na kraju se ručno provede čišćenje eventualnih pogrešaka.

Sve riječi koje su završile u pozitivnom skupu su pozitivno usmjerene, a sve riječi koje su završile u negativnom skupu su negativno usmejerene. Primjetimo da nemamo informaciju o neutralno usmjerenim riječima.

Sličan pristup koristili su i Kim i Hovy (2004). Kreću od dva početna skupa, pozitivnog i negativnog, ali ih ne šire iterativno. Umjesto toga računaju vjerojatnost da riječ w pripada pozitivnom, odnosno negativnom skupu i to tako da izbroje sinonime riječi w koji se nalaze u određenom početnom skupu te ga podijele s veličinom istog skupa. Ovu vjerojatnost možemo poistovijetiti i s intezitetom pozitivnog, odnosno negativnog sentimenta koji određena riječ nosi te na temelju tih vjerojatnosti možemo donositi zaključke o orijentaciji sentimenta. Intuicija ovog pristupa je da će riječ w vjerojatnije pripadati skupu koji sadrži više njenih sinonima.

2.1.4. Postupak temeljen na udaljenosti riječi u grafu

Kamps et al. (2004) na temelju WordNeta grade neusmjereni i bezteżinski graf $\mathcal{G}(\mathcal{W}, \mathcal{S})$, gdje je $\mathcal{W}$ skup čvorova koji predstavljaju pridjeve iz WordNet-a, a $\mathcal{S}$ skup bridova koji spajaju svaki par sinonima među riječima. Definiraju udaljenost $d(w_1, w_2)$ kao najkraći put između riječi w_1 i w_2 u grafu $\mathcal{G}$, a ako takav put ne postoji uzima se da je udaljenost beskonačna. Budući da sinonimija između dvije riječi podrazumijeva slično, odnosno

isto značenje riječi, tako bi i udaljenost $d(w_1, w_2)$ mogla odgovarati mjeri sličnosti dviju riječi; što je udaljenost manja riječi w_1 i w_2 su sličnije.

Tako bi mogli pretpostaviti da za određivanje sentimenta neke riječi dovoljno izračunati njenu udaljenost od neke referentne riječi, odnosno riječi koja je poptuno pozitivna ili potpuno negativna. Međutim, pokazalo se da to nije baš tako jer je za, na primjer, riječi *good* (hrv. *dobar*) i *bad* (hrv. *loš*) najkraći put u grafu $\mathcal{G}$ samo 4, što implicira da riječi *good* i *bad* nose sličan sentiment, dok je zapravo suprotan.

Zbog toga odabiru dvije referentne riječi, jednu pozitivnu (*good*) i jednu negativnu (*bad*). Sada sentiment riječi w možemo određuju kao relativnu udaljenost između te riječi i referentnih riječi:

$$EVA(w) = \frac{d(w, bad) - d(w, good)}{d(good, bad)} \tag{2.1}$$

Zbog toga što maksimalna razlika udaljenosti d riječi w do referentnih riječi ovisi o udaljenosti d između dvije referentne riječi. Dijeljenjem te razlike s udaljenosti između referentnih riječi dobiva se broj u intervalu $[-1, 1]$ koji izražava intezitet sentimenta riječi w: -1 predstavlja potpuno negativnu riječ, dok 1 predstavlja potpuno pozitivnu.

Leksikon sentimenta možemo izgraditi tako da svaku riječ w označimo pozitivnom ako je $EVA(w) > 0$, negativnom ako je $EVA(w) < 0$ te neutralnom $EVA(w) = 0$. Dodatno, bolji rezultati postižu se uz proglašavanje riječi w neutralnom ako je $EVA(w) \in [-0.25, 0.25]$.

2.1.5. Polunadzirani postupak temeljen na grafu

Budući da u izgradnju leksikona sentimenta krećemo s malenim pozitivnim i negativnim početnim skupom i želimo odrediti orijentaciju sentimenta velikog broja riječi, kao jedna od primjenjivih metoda nameće se polunadzirano učenje. Jedan algoritam polunadziranog učenja koji omogućava upravo to i radi nad grafom, je propagacija labela (Zhu i Ghahramani, 2002). Za izgradnju leksikona na temelju rječnika koristili su ga Rao i Ravichandran (2009) te uz male preinake Blair-Goldensohn et al. (2008) (dodatno se koristi i početni skup neutralnih riječi). U nastavku je opisan postupak iz Rao i Ravichandran (2009).

Graf $G(V, E, W)$ se gradi na temelju WordNet-a. Vrhovi V grafa predstavljaju riječi. Bridovi E predstavljaju vezu između dvije riječi, a težina brida mjeru sličnosti riječi. Bridove stvaramo na temelju veza koje nudi WordNet: sinonimije i hipernimije. Sličnost također možemo odrediti na temelju veza iz WordNet-a: ako su riječi koje spaja brid sinonimi, sličnost je λ, ako su antonimi $-\lambda$, a inače 0. Parametar λ može

biti proizvoljan ili ekperimentalno određen. Blair-Goldensohn et al. (2008) uzimaju da je $\lambda = 0.2$. $W = [w_{ij}]$ je težinska matrica susjedstva s n redaka i stupaca, a $n = |V|$.

Uz svaki vrh vežemo labele, odnosno vjerojatnosti da riječ koju vrh predstavlja ima pozitivan ili negativan sentiment. Labele predstavljamo matricom Y koja sadrži 2 stupca i n redaka. Za riječ i, $Y_{i,1}$ predstavlja vjerojatnost da riječ nosi pozitivan sentiment, a $Y_{i,2}$ vjerojatnost da nosi negativan sentiment. Za riječi iz početnih skupova te su vjerojatnosti poznate. Matrica se inicijalizira tako da je za riječi iz pozitivnog skupa $Y_i =$ [1 0], za riječi iz negativnog skupa $Y_i =$ [0 1], a za ostale $Y_i =$ [0 0].

Algoritam propagacije labela zapravo minimizira kvadratnu funkciju energije:

$$\xi = \frac{1}{2} \sum_{(i,j) \in E} w_{ij}(y_i - y_j)^2 \tag{2.2}$$

gdje su y_i i y_j labele pridružene vrhovima i i j. Dakle, da bi dobili labele y_i, postavimo $\frac{\partial}{\partial y_i}\xi = 0$ i dobivamo izraz za labele y_i:

$$y_i = \frac{\sum\limits_{(i,j) \in E} w_{ij} y_j}{\sum\limits_{(i,j) \in E} w_{ij}} \tag{2.3}$$

U praksi se koristi iterativni algoritam dan u nastavku (algoritam 1) (Zhu i Ghahramani, 2002).

Algorithm 1 Algoritam propagacije labela

procedure PROPAGACIJA LABELA(W, Y)
    $T \leftarrow normaliziraj\_stupce(W)$
    $Y' \leftarrow Y$
    repeat
        $Y' \leftarrow TY'$
        $Y' \leftarrow normaliziraj\_retke(Y')$
        vrati početne vrijednosti iz Y u Y'
    until Y' ne konvergira
    return Y'
end procedure

Završetkom algoritma dobivamo za svaku riječ vjerojatnosti da nosi pozitivni, odnosno negativni sentiment. Na temelju tih vjerojatnosti možemo donositi zaključke o orijentaciji sentimenta pojedine riječi. Tako Rao i Ravichandran (2009) kao ispravnu, uzimaju onu orijentaciju koja ima veću vjerojatnost te uopće ne razmatraju neutralne riječi.

2.2. Pristupi temeljeni na tekstnim zbirkama

Glavni nedostatak leksikona sentimenta izgrađenih na temelju rječnika je što su neovisni o domeni, odnosno kontekstu. To otežava primjenu tako izgrađenog rječnika na specifičnoj domeni. Međutim, ako leksikon gradimo na temelju tekstne zbirke iz neke domene, dobit ćemo leksikon prilagođen toj domeni. Osim toga, leksikon temeljen na rječniku ne možemo izgraditi za jezik za koji takav rječnik ne postoji. Zbog toga je važno razmotriti i pristupe izgradnji leksikona temeljene na tekstnim zbirkama koji rješavaju navedene nedostatke.

Kod izgradnje leksikona na temelju tekstne zbirke nemamo unaprijed dostupne informacije o vezama između riječi koje nude rječnici (antonimija i sinonimija). Zato se kod takvog pristupa velika važnost pridaje pitanju kako doći do podataka o semantičkoj sličnosti riječi na temelju podataka koje možemo dobiti iz tekstne zbirke (broj pojavljivanja riječi, broj supojavljivanja riječi, okolina riječi).

U nastavku su razmotreni neki postupci temeljeni na tekstnim zbirkama.

2.2.1. Postupak temeljen na orijentaciji sentimenta pridjeva povezanih veznicima

Pogledajmo sljedeće dvije rečenice:

- Auto lijepo izgleda i prostran je.
- Dan je bio sunčan, ali more je bilo hladno.

Možemo vidjeti da pridjevi "lijep" i "prostran" u kontekstu prve rečenice nose jednak sentiment (pozitivan), a da pridjevi "sunčan" i "hladno" u kontekstu druge rečenice nose suprotan ("sunčan" pozitivan, a "hladno" negativan). Također, možemo primjetiti da su pridjevi u prvoj rečenici (nose jednak sentiment) povezani sastavnim veznikom "i", a u drugoj (nose suprotan sentiment) suprotnim veznikom "ali".

Iz ovog razmatranja možemo izvući dva zaključka:

1. Ako su dva pridjeva u rečenici povezana sastavnim veznikom (i, pa, te, ni, niti), tada ti pridjevi najčešće imaju sentiment iste orijentacije.
2. Ako su dva pridjeva u rečenici povezana suprotnim veznikom (a, ali, dok, nego), tada ti pridjevi najčešće imaju sentiment suprotne orijentacije.

Gornji zaključci su osnovna ideja koju u svom pristupu koriste Hatzivassiloglou i McKeown (1997). Prvo na temelju tekstne zbirke grade graf u kojem vrhovi predstav-

ljaju pridjeve, a bridovi sadrže informaciju o tome imaju li pridjevi koje ti bridovi povezuju jednako ili suprotno orijentiran sentiment (u tekstnoj zbirci pronalaze pridjeve koji su povezani sastavnim, odnosno suprotnim veznicima). Potom tako izgrađen graf dijele na dva podgrafa. Unutar svakog podgrafa vrhovi trebaju biti povezani sa što više bridova koji predstavljaju jednaku orijentaciju sentimenta. Sami podgrafovi trebaju biti povezani sa što više bridova koji predstavljaju suprotnu orijentaciju sentimenta. Tada pridjevi u prvom podgrafu imaju suprotnu orijentaciju sentimenta od pridjeva koji se nalaze u drugom podgrafu.

Hatzivassiloglou i McKeown (1997) pretpostavljaju da su pridjevi koji imaju pozitivni apriorni sentiment učestaliji u korištenoj tekstnoj zbirci.Zbog toga bi pridjevi iz podgrafa koji sadrži učestalije pridjeve trebali imati pozitivno orijentiran sentiment. Drugi podgraf bi tada trebao sadržavati pridjeve s negativno orijentiranim sentimentom.

2.2.2. Postupak temeljen na semantičkoj asocijaciji

Semantička asocijacija je pojam iz područja psiholingvistike koji se odnosi na činjenicu da ljudski um spomen jedne riječi često asocira na slične riječi. Jedan takav primjer je riječ bolnica koja asocira na npr. bolest. U nastavku ćemo za riječi w_1 i w_2 reći da su asocijativne ako spomen jedne budi asocijacije na drugu. Možemo i reći da ako su riječi semantički asocijativne, da su i semantički slične.

Turney i Littman (2003) nastoje iz semantičke asocijacije doći do orijentacije sentimenta. Orijentacija sentimenta dane riječi w se računa kao razlika sume snaga njezinih asocijacija s skupom pozitivnih riječi (P) i sume snaga asocijacije s skupom negativnih riječi (N):

$$OS - A(w) = \sum_{pozRijec \in P} A(w, pozRijec) - \sum_{negRijec \in N} A(w, negRijec) \qquad (2.4)$$

gdje je $A(w_1, w_2)$ neka mjera asocijacije između riječi $w1$ i $w2$ (realni broj). Kad je $A(w_1, w_2)$ pozitivan riječi su asocijativne (veća vrijednost znači jaču asocijaciju). Kad je $A(w_1, w_2)$ negativan, prisutnost jedne riječi obično znači odsustvo druge.

Riječ w ima pozitivnu orijentaciju sentimenta kad je $OS(w)$ pozitivan i negativnu kad je $OS - A(w)$ negativan. Iznos $OS - A(w)$ može se smatrati intezitetom orijentacije sentimenta.

Jedna od mjera asocijacije koju koriste je točkasta aproksimacija uzajamne zajedničke informacije (engl. *pointwise mutual information*), u nastavku PMI (Church i

Hanks, 1990). PMI između dvije riječi, w_1 i w_2 je definirana kao:

$$PMI(w_1, w_2) = \log_2 \frac{p(w_1, w_2)}{p(w_1)p(w_2)} \tag{2.5}$$

gdje je $p(w_1, w_2)$ vjerojatnost da se riječi w_1 i w_2 pojavljuju skupa, a $p(w_1)$ i $p(w_2)$ vjerojatnost da pojavljivanje riječi w_1 i w_2. Omjer vjerojatnosti $p(w_1, w_2)$ i vjerojatnosti $p(w_1)$ i $p(w_2)$ je mjera stupnja statiskičke ovisnosti između riječi w_1 i w_2, a logaritam tog omjera odgovara obliku korelacije: pozitivan je kad se riječi obično nalaze skupa, a negativan kad prisutnost jedne riječi znači vjerojatno odsustvo druge.

Turney i Littman (2003) PMI aproksimiraju korištenjem web-pretraživača i bilježenjem broja stranica na kojoj se nalazi tražena riječ. Dodatno, korišteni pretraživač (AltaVista) nudi operator NEAR koji pretragu ograničava na stranice koje sadrže tražene riječi u prozoru od 10 riječi. Tada izraz za PMI iz jednadžbe 2.5 prelazi u:

$$PMI(w_1, w_2) = \log_2 \frac{\frac{1}{N} hits(w_1 NEAR w_2)}{\frac{1}{N} hits(w_1) \frac{1}{N} hits(w_2)} \tag{2.6}$$

gdje je N broj stranica koje indeksira pretraživač, a $hits(upit)$ broj stranica vraćenih upitom.

Zamjenom općenite mjere asocijacije A s PMI, jedandžba 2.4 prelazi u:

$$OS - PMI(w) = \sum_{pozRijec \in P} PMI(w, pozRijec) - \sum_{negRijec \in N} PMI(w, negRijec) \tag{2.7}$$

Tada je za riječ w je orijentacija sentimenta pozitivna ako je $OS - PMI(w) > 0$, a negativna ako je $OS - PMI(w) < 0$.

2.2.3. Postupak temeljen na supojavljivanju riječi u tekstnoj zbirci

Velikovich et al. (2010) kao tekstnu zbirku koriste kolekciju web dokumenata. Slično kao Blair-Goldensohn et al. (2008) i Rao i Ravichandran (2009) grade graf $G = (V, E)$, gdje je $w_{ij} \in [0, 1]$ težina brida $(v_i, v_j) \in E$. Skup vrhova V predstavlja pojedine riječi iz tekstne zbirke. Težina brida bi trebala predstavljati semantičku sličnost između riječi koje taj brid spaja. Koriste modificiran algoritam propagacije labela iz Zhu i Ghahramani (2002) (opisan u potpoglavlju 2.1.5). To znači da imamo početne skupove riječi: početni skup pozitivnih riječi i početni skup negativnih riječi. Ovaj modificirani algoritam se pokreće dvaput: jednom s pozitivnim početnim skupom, a drugi put s negativnim. Izlaz algoritma je vektor $\mathbf{pol}$ takav da je $\mathbf{pol}_i$ intezitet pozitivnog, odnosno negativnog sentimenta riječi i (i-tog vrha u grafu). Ako je $\mathbf{pol}_i > 0$ i-ta riječ nosi

pozitivan sentiment, ako je $\mathbf{pol}_i < 0$ negativan, a ako je $\mathbf{pol}_i = 0$ i-ta riječ ne nosi sentiment (neutralna je).

Konačni sentiment se računa formulom $\mathbf{pol}_i = \mathbf{pol}_i^+ - \beta \mathbf{pol}_i^-$, gdje je $\beta = \frac{\sum \mathbf{pol}_i^+}{\sum \mathbf{pol}_i^-}$. β se uvodi kako bi se uravnotežili utjecaji sveukupnog pozitivnog i negativnog intezi-teta (ako tekstna zbirka velikim dijelom sadrži negativno orijentirane dokumente).

Kao što vidimo cijeli postupak je gotovo identičan kao i kod Blair-Goldensohn et al. (2008) i Rao i Ravichandran (2009). Dakle radi se o polunadziranom učenju nad grafom. Glavna razlika je način određivanja težina bridova u grafu, odnosno semantičke sličnosti riječi. Blair-Goldensohn et al. (2008) i Rao i Ravichandran (2009) koriste rječnik za njeno određivanje. Međutim, sada nemamo rječnika nego je potrebno na temelju informacija koje možemo dobiti iz tekstne zbirke odrediti semantičku sličnost riječi.

Velikovich et al. (2010) u tu svrhu koristi informaciju o supojavljivanju riječi u tekstnoj zbirci. Dvije riječi se supojavljuju ako se nalaze unutar prozora fiksne duljine. Velikovich et al. (2010) za svaku riječ w stvara njezin konteksni vektor koji sadrži sve riječi sa kojima se riječ w supojavljuje unutar prozora duljine 6 te za svaki brid (v_i, v_j) računa kosinusnu sličnost (kosinus kuta) između konteksnih vektora riječi w_i i w_j. Intuicija je da se slične riječi često supojavljuju.

3. Postupak izgradnje leksikona sentimenta

Kao što smo vidjeli u prethodnom poglavlju, semantički slične riječi najčešće imaju sentiment iste orijentacije. Te ideje ćemo se držati i u prikazanom postupku izgradnje leksikona sentimenta za hrvatski jezi te ćemo sentiment riječi određivati na temelju semantičke sličnosti između njih. Do semantičke sličnosti između riječi dolazit ćemo na temelju tekstne zbirke. Dakle, radi se o izgradnji leksikona sentimenta na temelju tekstne zbirke.

Također, graf se pokazao prikladnom strukturom za prikaz sličnosti između riječi: vrhovi grafa odgovaraju pojedinim riječima, a težina brida odgovara mjeri semantičke sličnosti između dvije riječi koje taj brid povezuje. Uz to, graf se pokazao dobrim za problem polunadziranog učenja na temelju malog broja označenih primjera. Tako će sljedeći postupak temeljiti na propagaciji labela, slično kako je opisano u prethodnom poglavlju.

Koristeći različite pristupe za određivanje sličnosti, primjenom dviju metoda propagacije labela poredati ćemo riječi po intezitetu kojim nose pozitivni, odnosno negativni sentiment. Na temelju razlike u pozitivnom i negativnom intezitetu pojedine riječi, istu ćemo pokušati odrediti kao pozitivnu, negativnu ili neutralnu.

3.1. Izgradnja grafa

Prvo, na temelju tekstne zbirke, gradimo graf $G(V, E)$. Skup vrhova V predstavlja sve pojedine riječi u tekstnoj zbirci. Težinu brida između dva vrha definiramo kao:

$$E(v_i, v_j) = max(0, slicnost(r_i, r_j)) \tag{3.1}$$

gdje je $slicnost(r_i, r_j)$ općenita mjera sličnosti između riječi r_i i r_j koje su predstavljene vrhovima v_i i v_j. Ako je mjera sličnosti manja od 0 (riječi imaju suprotno znače-

nje), težinu brida postavljamo na 0 jer će korišteni algoritmi propagacije labela očekivati bridove pozitivnih težina.

U nastavku ovog potpoglavlja dan je pregled korištenih mjera sličnosti.

3.1.1. Supojavljivanje

Najjednostavni indikator sličnosti dviju riječi jest informacija koliko često se jedna riječ nalazi u susjedstvu druge. Osim toga, pozitivne riječi se uglavnom nalaze u okolini pozitivnih, a negativne u okolini negativnih. Tako mjeru sličnosti između dvije riječi možemo definirati kao broj njihovih supojavljivanja unutar tekstne zbirke. Riječ r_i se supojavljuje s riječi r_j ako se riječ r_j najviše N riječi prije ili poslije riječi r_j, odnosno ako se obje riječi pojavljuju unutar prozora duljine $2N$.

3.1.2. Uzajamna zajednička informacija (PMI)

Nedostatak supojavljivanja kao mjere sličnosti je što ne vodi računa o učestalosti pojedinih riječi, nego samo njihovog supojavljivanja. Važno je primjetiti da se najfrekventnije riječi u tekstnoj zbirci (npr. glagol "biti") jednako često pojavljuju i uz pozitivne riječ i uz negativne riječi. Taj nedostatak rješava uzajamna zajednička informacija (PMI) Church i Hanks (1990) jer koristi i informaciju o frekvenciji pojedinih riječi u tekstnoj zbirci:

$$PMI(r_1, r_2) = log_2 \frac{p(r_1, r_2)}{p(r_1)p(r_2)}. \tag{3.2}$$

gdje $p(r_1, r_2)$ predstavlja vjerojatnost supojavljivanja riječi r_1 i r_2, a $p(r)$ predstavlja vjerojatnost pojavljivanja riječi r.

3.1.3. Latentna semantička analiza

Za određivanje semantičke sličnosti riječi na temelju tekstne zbirke često se koriste pristupi temeljeni na distribucijskim semantičkim modelima (engl. *distributional semantic models*). Glavna intuicija koja stoji iza distribucijskih semantičkih modela je da se do značenja riječi može doći preko konteksta u kojima se riječ nalazi. Kontekst se može različito definirati, pa se tako može raditi o nekoliko susjednih riječi, rečenici ili dokumentu. Na temelju svih konteksta u kojima se nalazi, za svaku riječ iz tekstne zbirke se gradi distribucijski vektor riječi (predstavlja distribuciju neke riječi po svim kontekstima) te se sličnost riječi definira kao sličnost njihovih distribucijskih vektora. Jedan od modela temeljenih na distribucijskoj semantici je latentna semantička analiza (engl. *latent semantic analysis* - LSA) (Dumais, 2004).

Prvi korak u primjeni LSA je izgradnja distribucijskih vektora riječi. Neka je D skup svih dokumenata (konteksta), a R skup svih riječi iz tekstne zbirke. Gradi se matrica M dimenzija $|R| \times |D|$ u kojoj retci stoje za riječi, a stupci za dokumente iz tekstne zbirke. Pojedini redak i predstavlja kontekstni vektor riječi R_i. Element M_{ij} predstavlja težinu koju nosi riječ R_i u dokumentu D_j. Ta težina se može definirati na razne načine (primjerice brojem pojavljivanja riječi u dokumentu).

U praksi se za određivanje elemenata matrice ne koristi pojavljivanje riječi u pojedinom dokumentu. Budući da su neke riječi izrazito česte i mogu se nalaziti u nepovezanim kontekstima (npr. glagol "biti"), intuicija je da bi takve riječi trebale imati manji utjecaj na kontekstne vektore. S druge strane, riječi koje se rijetko pojavljuju trebale bi imati veći utjecaj na kontekstni vektor jer upravo njihova najviše utječe značenje konteksta. Jedan način modeliranja upravo takvog ponašanja je definiranje težine kao vrijednosti *tf-idf* (engl. *term frequency - inverse document frequency*):

$$tfidf(R_i, D_j, D) = tf(R_i, D_i) \times idf(R_i, D) \tag{3.3}$$

gdje je $tf(R_i, D_i)$ broj pojavljivanja riječi R_i u dokumentu D_i, a idf je ocjena koliko je riječ R_i česta u tekstnoj zbirci te se definira kao:

$$idf(R_i, D) = log\frac{|D|}{|d \in D : r \in d)} \tag{3.4}$$

Težine definirane s *tf-idf* bit će veće što je neka riječ češća u danom dokumentu, a rjeđa u cjelokupnoj tekstnoj zbirci (što je i željeno ponašanje).

Sljedeći korak je primjena singularne dekompozicije (engl. *singular value decomposition* - SVD) na matricu M. SVD je matematički postupak kojim se matrica M rastavlja na dvije ortogonalne matrice: matricu U dimenzija $|R| \times |R|$ i matricu V dimenzija $|D| \times |D|$ te na dijagonalnu matricu D dimenzija $|R| \times |D|$. Matrica U u svojim stupcima sadrži lijeve, matrica V desne singularne vektore matrice M, dok matrica D na svojoj dijagonali sadrži odgovarajuće singularne vrijednosti. Matrica M se može rekonstruirati na sljedeći način:

$$M = UDV^T \tag{3.5}$$

Retci matrice U i dalje prikazuju distribuciju riječi, ali po novim kontekstima određenih matricom DV^T. Zato je moguće napraviti smanjenje dimenzionalnosti i to tako da iz matrice D uklonimo sve osim najvećih k singularnih vrijednosti te uklonimo odgovarajuće singularne vektore iz matrica U i V. Time ćemo dobiti matricu U' dimenzija $|R| \times k$ čiji retci predstavljaju nove kontekstne vektore pojedinih riječi na temelju kojih

ćemo određivati sličnost riječi. U praksi se pokazalo da je za dobre rezultate dovoljno uzeti 100-500 najvažnijih elemenata (iako se dimenzije početne matrice mjere u stotinama tisuća redaka i stupaca). Smanjenje dimenzionalnosti, osim zbog smanjivanja veličine matrice i bržeg izračuna, provodi se zbog smanjenja šuma u početnoj matrici: odbacuju se manje važni konteksti pojedine riječi kako bi se naglasilo njeno temeljno značenje.

Konačno, za mjeru sličnosti između dvije riječi R_i i R_j koristi se kosinusna sličnost između njihovih kontekstnih vektora $\mathbf{u}$ i $\mathbf{v}$ (odgovarajućih redaka u matrici U'):

$$slicnost(R_i, R_j) = \frac{\mathbf{u} \cdot \mathbf{v}}{||\mathbf{u}|| \times ||\mathbf{v}||} \tag{3.6}$$

Za razliku od supojavljivanja i PMI, kod LSA se sličnost riječi određuje usporedbom vektora.

3.2. Propagacija labela

Definiramo dva početna skupa P i N, gdje je P skup pozitivno usmjerenih riječi, a N skup negativno usmjerenih riječi. Budući da nam cilj nije ručno graditi leksikon sentimenta, skupovi P i N ne trebaju biti veliki (desetak riječi je dovoljno) te se mogu ručno odrediti. Pri odabiru riječi za početne skupove treba voditi računa o tome da iste riječi u različitim kontekstima mogu nositi suprotno orijentiran sentiment. Zbog toga bi sentiment koji nose riječi iz skupova P i N trebao biti što manje ovisan o kontekstu u kojima se te riječi nalaze.

Sljedeći korak je primjena propagacije labela na prethodno izgrađenom grafu. Propagacija labela obavlja se dva puta: jednom za pozitivno usmjerene riječi (uz početni skup P), a drugi put za negativno usmjerene riječi (uz početni skup N). Propagacija labela svakom vrhu $v_i \in V$ pridjeli labelu koja ima vrijednosti iz intervala $[0, 1]$. Vrijednost labele predstavlja vjerojatnost da riječ koju vrh predstavlja pripada početnom skupu za koji se radi propagacija, tj. da ima jednako orijentiran sentiment kao i riječi iz početnog skupa. Inicijalno se labele postavljaju tako da vrh v_i ima vrijednost labele 1, ako se riječ r_i nalazi u danom početnom skupu, a 0 ako se riječ r_i ne nalazi u danom početnom skupu. Početne labele se s vrhova koji predstavljaju riječi iz početnih skupova (i čije labele su inicijalno 1), propagiraju na ostale vrhove. Vrijednosti labela se propagiraju na temelju težina bridova između vrhova, odnosno na temelju semantičke sličnosti između riječi koje predstavljaju ti vrhovi. Što je težina brida između dva vrha veća, to će njihove labele biti sličnije. Dakle, što je veća semantička sličnost između dvije riječi, to će vjerojatnost da one nose sentiment kao i riječi iz početnog skupa biti

veća. Završetkom propagacije labela za pozitivni i negativni početni skup, za svaku riječ ćemo dobiti vjerojatnost da nosi pozitivni, odnosno negativni sentiment.

Za propagaciju labela koriste se dva algoritma:

- klasični algoritam propagacije labela te
- PageRank.

Oba su opisana u nastavku poglavlja.

3.2.1. Algoritam propagacije labela

Prvi korišteni algoritam za propagaciju labela je klasični algoritam propagacije labela (Zhu i Ghahramani, 2002) koji je već spomenut u potpoglavlju 2.1.5.

Cilj algoritma propagacije labela je na temelju težina bridova propagirati labele s vrhova kojima su labele unaprijed poznate na vrhove kojima labele nisu poznate. Pri tome se početne labele drže konstantnima kroz cijeli postupak.

Kako bi primjenili algoritam propagacije labela za početni skup PS na prethodno izgrađenom grafu G, koji predstavlja riječi i sličnosti između njih, prvo je potrebno odrediti matricu vjerojatnosti prijelaza W. Neka je M težinska matrica susjedstva grafa G s N vrhova. Matricu vjerojatnosti prijelaza W dobit ćemo tako da normaliziramo stupce matrice M. Dakle, elemente matrice W određujemo kao:

$$W_{ij} = \frac{M_{ij}}{\sum_{k=1}^{N} M_{kj}} \tag{3.7}$$

Element matrice W_{ij} odgovara vjerojatnosti prijelaza s vrh j na vrh i. Definiramo matricu Y s N redaka i jednim stupcem čiji su elementi definirani kao:

$$Y_i = \begin{cases} 1 & \text{ako je } r_i \in PS \\ 0 & \text{inače} \end{cases}$$

Tada je algoritam propagacije labela definiran kao:

1. Propagiraj labele sa svih čvorova na susjede: $Y \leftarrow WY$.
2. Svaki Y_i takav da je r_i iz početnog skupa postavi na 1.
3. Ponavljaj dok Y ne konvergira.

Kada izvođenje algoritma završi, i-ti element možemo interpretirati kao vjerojatnost da riječ r_i pripada početnom skupu.

Drugi korak algoritma je ključan jer ne dopušta rasipanje početno poznatih labela. Time osiguravamo da vrijednosti labela budu najveće upravo oko vrhova koji predstavljaju riječi oko početnih skupova. Tada će vjerojatnost da riječ pripada početnom skupu biti najveća upravo oko vrhova koji predstavljaju riječi iz početnih skupova.

3.2.2. PageRank

Drugi korišteni algoritam za propagaciju labela je PageRank (Page et al., 1999) koji je razvijen za rangiranje web stranica po njihovoj važnosti. Intuicija je da će web stranica biti važnija što više drugih (važnih) stranica ima poveznicu na nju. Dakle, možemo reći da se važnost pojedine web stranice određuje na temelju glasanja svih drugih web stranica. Kad neka web stranica A ima poveznicu na web stranicu B, tada web stranica A pridonosi važnosti stranice B.

Web stranice i njihove poveznice predstavljene su usmjerenim grafom u kojem brid $E(v_i, v_j)$ ima težinu koja predstavlja broj poveznica sa stranice A (predstavljene vrhom v_i) na stranicu B (predstavljene vrhom v_j).

U Page et al. (1999) PageRank vrijednost vrha v dana je kao:

$$PR(v) = (1 - d) + d \sum \frac{PR(v_i)}{C(v_i)} \tag{3.8}$$

gdje je:

- $PR(v)$ PageRank vrijednost vrha v,
- $PR(v_i)$ je PageRank vrijednost vrha v_i koji ima brid prema vrhu v,
- $C(v_i)$ je suma težina svih izlaznih bridova vrha v_i te
- d je faktor prigušenja (iz intervala $[0, 1]$).

Vidimo da je PageRank vrijednost vrha v rekurzivno definirana PageRank vrijednostima vrhova koji imaju bridove prema vrhu v. Također, doprinos vrha v_i PageRank vrijednosti vrha v ovisi o ukupnoj sumi svih izlaznih bridova vrha v_i. Što ima više izlaznih bridova i što su njihove težine veće, to će doprinos PageRank vrijednosti vrha v biti manji.

PageRank vrijednost vrha v zapravo predstavlja vjerojatnost da će slučajni šetač po grafu, krenuvši od bilo kojeg vrha, nakon dovoljno velikog broja slučajnih skokova doći do vrha v. Faktor prigušenja d (engl. *PageRank damping factor*) pri tome označava vjerojatnost sljedećeg skoka, dok $(1-d)$ označava vjerojatnost da će šetač prestati pratiti izlazne bridove,te skočiti na slučajno odabran vrh grafa.

Iterativni algoritam za računanje PageRank-vektora svih vrhova u grafu dan je izrazom:

$$\mathbf{PR}^{(i)} = dW\mathbf{PR}^{(i-1)} + (1 - d) * \mathbf{e} \tag{3.9}$$

gdje je:

- $\mathbf{PR}$ PageRank-vektor u kojem i-ti element predstavlja PageRank vrijednost vrha v_i,

– W je matrica vjerojatnosti prijelaza, gdje element W_{ij} predstavlja vjerojatnost skoka s vrha v_i na vrh v_j te

– vektor $\mathbf{e}$ definira distribuciju vjerojatnosti po kojoj se slučajni šetač, kad odluči prestati pratiti bridove, bira novi vrh na koji će skočiti.

PageRank algoritam u pravilu završava nakon konvergencije vektora PR ili (rjeđe) nakon nekog unaprijed zadanog broja iteracija. Svi elementi vektora $\mathbf{e}$, u uobičajenoj primjeni algoritma PageRank, iznose $1/N$, gdje je N broj vrhova u grafu. Dakle, vjerojatnost skoka na svaki vrh je jednaka.

Kako bi primjenili PageRank na problem propagacije labela nad prethodno izgrađenim grafom G, koji predstavlja riječi i sličnosti između njih, prvo je potrebno odrediti matricu vjerojatnosti prijelaza W. Neka je M težinska matrica susjedstva grafa G s N vrhova. Matricu vjerojatnosti prijelaza W dobit ćemo tako da normaliziramo stupce matrice M. Dakle, elemente matrice W određujemo kao:

$$W_{ij} = \frac{M_{ij}}{\sum_{k=1}^{N} M_{kj}} \tag{3.10}$$

S obzirom da znamo labele (PageRank vrijednosti) za riječi iz početnog skupa, vektor $\mathbf{e}$ određujemo kao:

$$e_i = \begin{cases} \frac{1}{|PS|} & \text{ako je } r_i \in PS \\ 0 & \text{inače} \end{cases}$$

gdje PS predstavlja početni skup za koji se radi propagacija. S ovako definiranim vektorom $\mathbf{e}$, kad šetač radi slučajni skok na bilo koji vrh u grafu, osiguravamo da taj skok bude moguć samo na vrhove koji predstavljaju riječi iz početnog skupa. Time osiguravamo da njihova PageRank vrijednost bude očuvana tijekom izvođenja algoritma.

Završetkom algoritma PageRank dobit ćemo vektor $\mathbf{PR}$. Elemente tog vektora skaliramo tako da njegov najveći element ima vrijednost 1. U tako skaliranom vektoru, i-ti element možemo interpretirati kao vjerojatnost da riječ r_i pripada početnom skupu.

3.3. Određivanje orijentacije sentimenta

Preostaje na temelju dobivenih labela svrstati svaku riječ u pozitivnu, negativnu ili neutralnu. Dobivene vjerojatnosti možemo interpretirati kao intezitet pozitivnog, odnosno negativnog sentimenta riječi. Neka su vektori dobiveni propagacijom labela $\mathbf{s}^+$ i $\mathbf{s}^-$ te neka i-ti element vektora $\mathbf{s}^+$ i $\mathbf{s}^-$, odgovara pozitivnom, odnosno negativnom intezitetu sentimenta koji nosi riječ r_i, a $D(r_i) = \mathbf{s}_i^+ - \mathbf{s}_i^-$. Sada riječ r_i možemo odrediti kao pozitivnu, negativnu ili neutralnu. Riječ r_i je:

- pozitivna ako je $D(r_i) > 0$ i ako je $D(r_i) > \gamma$,
- negativna ako je $D(r_i) < 0$ i ako je $D(r_i) < -\gamma$,
- inače je neutralna.

Parametar γ je prag koji definira kada je riječ neutralna te se određuje eksperimentalno.

4. Vrednovanje postupka izgradnje leksikona sentimenta

4.1. Obrada tekstne zbirke i izgradnja leksikona

Tekstna zbirka se sastoji od 43 456 novinskih članaka raznih tema skupljenih s web portala. Svim riječima iz tekstne zbirke odredimo leme, odnosno rječnički oblik riječi te za svaku riječ odredimo vrstu riječi. Tako dobijemo 249 417 različitih riječi. Kako bi ubrzali izračune, razmatrat ćemo samo one riječi koje se u tekstnoj zbirci pojavljuju najmanje 50 puta. Osim toga, od preostalih riječi uzet ćemo samo one koje su imenice, glagoli, prilozi i pridjevi. Nakon toga nam ostaje 14 149 riječi koje će činiti leksikon sentimenta. Za LSA koristimo na drugoj tekstnoj zbirci izgrađene kontekstne vektore riječi (dimenzije 300).

Na temlju riječi koje su preostale nakon obrade tekstne zbirke, gradi se graf sličnosti riječi i to tako da se brid stvara samo između riječi koje se supojavljuju unutar prozora duljine 5 riječi. Od tako stvorenog grafa izgradit ćemo tri nova grafa, po jedan za svaku mjeru sličnosti (supojavljivanje, PMI i LSA). Pri tome, nećemo stvarati nove bridove, nego ćemo postojećima samo dodati težine dobivene mjerom sličnosti. Tako ćemo dobiti rijetki graf (odnosno rijetku matricu susjedstva grafa) čime smanjujemo računsku složenost.

Na svakom od tri grafa se dva puta provodi propagacija labela (klasična propagacija labela i PageRank), jednom za pozitivni i jednom za negativni početni skup. Na temelju dobivenih labela, klasificiramo svaku riječ u pozitivnu, negativnu i neutralnu. Promotrit ćemo utjecaj učestalosti riječi iz početnih skupova na rezultate vrednovanja. Početni skupovi bit će prvo odabrani bez znanja o učestalosti riječi u tekstnoj zbirci, a potom uz uvid u učestalost riječi i to tako da se izaberu najučestalije riječi za svaki početni skup.

4.2. Vrednovanje

Za vrednovanje se koriste dva skupa unaprijed označenih riječi [1]. Prvi je skup s niskim slaganjem između označivača te je pojedina riječ svrstana u onu klasu (pozitivna, negativna i neutralna) u koju ju je svrstalo najviše ljudi. Drugi skup je skup s visokim slaganjem između označivača te su u njega uključene samo riječi koje je najmanje 10 od 12 ljudi svrstalo u istu klasu. Prvi skup se sastoji od 2500 riječi, a drugi od 1706 riječi.

Zbog određivanja parametra γ, svaki skup ćemo podijeliti na skup za vrednovanje i skup za provjeru i to tako da u svakom bude jednak broj riječi. Parametar γ odredit ćemo na temelju skupa za provjeru i to tako da uzmemo onaj γ koji daje najbolje rezultate na validacijskom skupu. Potom koristeći tako određeni parametar γ vrednujemo dobiveni leksikon na skupu za vrednovanje.

Pozitivni (P) i negativni (N) početni skupovi su odabrani kako slijedi:

$P =$ {dobar, uspjeh, sjajan, prekrasno, super, ljubav, sreća, fantastičan, zadovoljan, sposobnost, zdravlje}

$N =$ {sumnjati, zatvor, rat, kritika, kazna, umrijeti, nesreća, smrt, eksplozija, nezadovoljstvo, propust}

Riječi u početnim skupovima su određene bez znanja o njihovoj učestalosti u tekstnoj zbirci.

U tablicama 4.1 i 4.2 dani su rezultati vrednovanja nad skupovima s niskom i visokom razinom slaganja. Rezultati su dani po razredima: pozitivnom, negativnom i neutralnom. P predstavlja preciznost (engl. *precision*), a R odziv (engl. *recall*). Ukupno vrednovanje izgradnje leksikona dano je korištenjem makro F1 mjere.

Očekivano, vrednovanje nad skupom s visokom razinom slaganja daje bolje rezultate. Gledajući ukupne rezulate možemo primjetiti da PMI kao mjera sličnosti daje nešto bolje rezultate od čistog supojavljivanja. Generalno najbolje rezultate dobivamo kada kao mjeru sličnosti koristimo LSA, neovisno o načinu propagacije. To je i očekivano jer LSA, za razliku od supojavljivanja i PMI koji koriste broj supojavljivanja i frekvencije riječi, koristi puno više informacija za određivanje sličnosti riječi. PageRank i klasični algoritam propagacije labela, uz istu mjeru sličnosti, daju približno jednake rezultate na skupu za vrednovanje s niskom razinom slaganje, dok je algoritam propagacije labela nešto bolji na skupu s visokom razinom slaganja.

Vidimo da su za neutralni razred rezultati prilično dobri, no to je zbog toga što i u tekstnoj zbirci i u skupu za vrednovanje prevladavaju riječi s neutralnim aprior-

[1] http://takelab.fer.hr/data/sentilex/

		PageRank			Propagacija labela		
	Razred	P	R	F1	P	R	F1
Supojavljivanje	Pozitivni	0.30	0.16	0.21	0.28	0.22	0.25
	Negativni	0.21	0.26	0.23	0.22	0.23	0.23
	Neutralni	0.76	0.80	0.78	0.76	0.79	0.79
	Ukupno	0.42	0.41	0.41	0.42	0.41	0.42
PMI	Pozitivni	0.31	0.12	0.18	0.29	0.13	0.18
	Negativni	0.27	0.33	0.30	0.27	0.21	0.24
	Neutralni	0.77	0.84	0.81	0.76	0.87	0.82
	Ukupno	0.45	0.43	0.43	0.44	0.40	0.41
LSA	Pozitivni	0.34	0.27	0.30	0.31	0.3	0.31
	Negativni	0.34	0.41	0.37	0.28	0.40	0.33
	Neutralni	0.79	0.81	0.80	0.79	0.74	0.77
	Ukupno	0.49	0.49	0.49	0.46	0.48	0.48

Tablica 4.1: Rezultati vrednovanja na skupu za vrednovanje s niskim slaganjem

		PageRank			Propagacija labela		
	Razred	P	R	F1	P	R	F1
Supojavljivanje	Pozitivni	0.18	0.26	0.21	0.16	0.23	0.20
	Negativni	0.12	0.34	0.18	0.17	0.24	0.20
	Neutralni	0.86	0.68	0.76	0.86	0.79	0.82
	Ukupno	0.39	0.43	0.38	0.39	0.42	0.40
PMI	Pozitivni	0.25	0.14	0.18	0.21	0.14	0.17
	Negativni	0.21	0.28	0.24	0.31	0.24	0.27
	Neutralni	0.87	0.87	0.87	0.87	0.91	0.89
	Ukupno	0.44	0.43	0.43	0.46	0.43	0.44
LSA	Pozitivni	0.35	0.32	0.34	0.38	0.32	0.35
	Negativni	0.32	0.36	0.34	0.38	0.31	0.34
	Neutralni	0.88	0.88	0.88	0.88	0.91	0.90
	Ukupno	0.52	0.52	0.52	0.55	0.52	0.53

Tablica 4.2: Rezultati vrednovanja na skupu za vrednovanje s visokim slaganjem

nim sentimentom. Za izgradnju leksikona sentimenta važniji su rezultati za pozitivni i negativni razred, a oni su dosta loši. Možemo primjetiti da variraju ovisno o korištenoj mjeri sličnosti: kad se koristi LSA F1 rezultat je više od 10% bolji nego kad se

kao mjera sličnosti koriste supojavljivanje ili PMI (neovisno o korištenom algoritmu propagacije).

4.2.1. Utjecaj početnih skupova na rezultate

Prethodno korišteni početni skupovi su odabrani bez razmatranja njihove frekvencije pojavljivanja u tekstnoj zbirci. Zanimljivo je vidjeti rezultate kada kao početne skupove odaberemo najfrekventnije pozitivne, odnosno negativne riječi u tekstnoj zbirci. Uvidom u frekvencije riječi u tekstnoj zbirci odabrani su sljedeći početni skupovi:

$P =$ {dobar, dobro, uspjeti, snaga, poseban, nadati, slobodan, nagrada, siguran, pobjeda}

$N =$ {problem, teško, rat, težak, izgubiti, nesreća, kazna, zatvor, borba, kriza}

Budući da se LSA pokazao kao najbolja mjera sličnosti, izgradit ćemo leksikon koristeći LSA kao mjeru sličnosti. Evaluiramo tako izgrađen leksikon na skupu za vrednovanje s visokom razinom slaganja. Rezultati vrednovanja su dani na tablici 4.3.

		PageRank			Propagacija labela		
	Razred	P	R	F1	P	R	F1
LSA	Pozitivni	0.39	0.24	0.3	0.40	0.26	0.31
	Negativni	0.47	0.27	0.34	0.52	0.34	0.41
	Neutralni	0.87	0.94	0.91	0.88	0.94	0.91
	Ukupno	0.58	0.48	0.52	0.60	0.51	0.55

Tablica 4.3: Rezultati vrednovanja uz LSA i drugačiji izbor početnih skupova

Vidimo da uz visoko frekventne riječi u početnim skupovima preciznost za pozitivni i negativni razred raste. To je intuitivno jasno jer će visoko frekventne riječi imati više susjeda u grafu, pa će lakše prenositi svoj početni sentiment na ostale riječi što dovodi do boljeg diferenciranja između razreda. Odziv je ostao sličan, a za pretpostaviti je da bi i on porastao dodavanjem više riječi u početne skupove.

Možemo zaključiti da je dobar odabir početnih skupova važan, a jedno od mjerila dobrote je i učestalost riječi iz početnih skupova u tekstnoj zbirci: učestalije riječi dovode do veće preciznosti.

5. Zaključak

Porastom korisnički generiranog sadržaja, analiza sentimenta postala je poprilično popularno područje obrade prirodnog jezika. Uobičajeni postupci analize sentimenta koriste apriorno sastavljen leksikon sentimenta. Budući da je ručna izgradnja takvog leksikona skup i dugotrajan posao, postupci automatske akvizicije leksikona sentimenta također dobivaju na važnosti.

Obrađeno je nekoliko postupaka automatske izgradnje leksikona sentimenta dostupnih u literaturi. Dio tih postupaka temelji se na informacijama o vezama između riječi dobivenih iz rječnika. Zbog toga se bave samo iskorištavanjem tih informacija u svrhu izgradnje leksikona sentimenta. Tako najčešće nastoje proširiti početne skupove pozitivnih i negativnih riječi kako bi izgradili leksikon sentimenta. S druge strane, rječnici ne postoje za sve jezike, a izgradnja rječnika se radi ručno te je poput izgradnje leksikona sentimenta dugotrajan postupak. Zato je korisno vidjeti i postupke temeljene na tekstnoj zbirci. Drugi dio opisanih postupaka je upravo takav. Ti postupci uglavnom koriste slične metode za proširivanje početnih skupova u cilju izgradnje leksikona. Razlikuju se u tome što do informacija o vezama između riječi dolaze analizom tekstne zbirke.

Na temelju opisanih postupaka implementiran i vrednovan je postupak za izgradnju leksikona sentimenta za hrvatski jezik temeljen na tekstnoj zbirci. Postupak se temelji na polunadziranom učenju nad grafom. Svaki vrh u grafu predstavlja riječ, a brid između dva vrha predstavlja semantičku sličnost između riječi. Na temelju grafa, postupak nastoji na temelju malog broja poznatih pozitivnih, odnosno negativnih, riječi doći do sentimenta ostalih riječi iz tekstne zbirke. Korišteni algoritmi za polunadzirano učenje su algoritam propagacije labela i PageRank. Oba nastoje, na temelju semantičke sličnosti između riječi, propagirati sentiment s riječi čiji je sentiment poznat na ostale riječi čiji sentiment nije poznat. Intuicija je da će semantički slične riječi imati i jednako orijentiran sentiment. Sličnost riječi određena je na više načina:

- na temelju supojavljivanja riječi,
- na temelju točkaste procjene uzajamne zajedničke informacije (PMI) te

– na temelju latentne semantičke analize (LSA).

Vrednovanjem je ustvrđeno da se kao najbolja mjera sličnosti pokazala latentna semantička analiza. Rezultati postignuti korištenjem PageRank-a i algoritma propagacije labela su usporedivi, iako se malo bolja pokazala propagacija labela. Također, vidjeli smo da na rezultate ponešto utječu riječi koje se nalaze u početnim skupovima. Za bolje rezultate bi ih trebalo biti još više, trebale bi što frekventnije te sentiment koji nose ne bi smio ovisiti o kontekstu u kojem se riječi nalaze.

Konačno, možemo zaključiti da implementirani postupak ipak ne daje pretjerano dobre rezultate te ne bi bio pogodan za izgradnju kvalitetnog leksikona sentimenta.

LITERATURA

Stefano Baccianella, Andrea Esuli, i Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. U *LREC*, svezak 10, stranice 2200–2204, 2010.

Sasha Blair-Goldensohn, Kerry Hannan, Ryan McDonald, Tyler Neylon, George A Reis, i Jeff Reynar. Building a sentiment summarizer for local service reviews. U *WWW Workshop on NLP in the Information Explosion Era*, stranica 14, 2008.

Kenneth Ward Church i Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

Susan T Dumais. Latent semantic analysis. *Annual review of information science and technology*, 38(1):188–230, 2004.

Vasileios Hatzivassiloglou i Kathleen R McKeown. Predicting the semantic orientation of adjectives. U *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, stranice 174–181. Association for Computational Linguistics, 1997.

Minqing Hu i Bing Liu. [acm press the 2004 acm sigkdd international conference - seattle, wa, usa (2004.08.22-2004.08.25)] proceedings of the 2004 acm sigkdd international conference on knowledge discovery and data mining - kdd '04 - mining and summarizing customer reviews. 2004. ISBN 1581138889. doi: 10.1145/1014052.1014073.

Međunarodna telekomunikacijska unija ITU. Statistike, 2014. URL `http://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx`.

Jaap Kamps, MJ Marx, Robert J Mokken, i Maarten De Rijke. Using wordnet to measure semantic orientations of adjectives. 2004.

Soo-Min Kim i Eduard Hovy. Determining the sentiment of opinions. U *Proceedings of the 20th international conference on Computational Linguistics*, stranica 1367. Association for Computational Linguistics, 2004.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, i Katherine J Miller. Introduction to wordnet: An on-line lexical database*. *International journal of lexicography*, 3(4):235–244, 1990.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, i Noah A Smith. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11:122–129, 2010.

Lawrence Page, Sergey Brin, Rajeev Motwani, i Terry Winograd. The pagerank citation ranking: Bringing order to the web. 1999.

Delip Rao i Deepak Ravichandran. Semi-supervised polarity lexicon induction. U *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, stranice 675–682. Association for Computational Linguistics, 2009.

Swapna Somasundaran, Josef Ruppenhofer, i Janyce Wiebe. Detecting arguing and sentiment in meetings. 2007.

Peter D Turney i Michael L Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346, 2003.

Leonid Velikovich, Sasha Blair-Goldensohn, Kerry Hannan, i Ryan McDonald. The viability of web-derived polarity lexicons. U *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, stranice 777–785. Association for Computational Linguistics, 2010.

Xiaojin Zhu i Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report, Technical Report CMU-CALD-02-107, Carnegie Mellon University, 2002.

Postupak za polunadziranu akviziciju leksikona sentimenta

Sažetak

U ovom radu opisani su postupci automatske izgradnje sentimenta iz literature te je na temelju njih opisan i implementiran postupak akvizicije leksikona sentimenta na temelju tekstne zbirke. Postupak se temelji na polunadziranom učenju nad grafom i radi nad tekstnom zbirkom te je zbog toga primjenjiv na jezike za koje ne postoje izgrađeni jezični resursi poput rječnika. Za određivanje sličnosti između riječi koriste se tri mjere: supojavljivanje, uzajamna zajednička informacija i latentna semantička analiza. Za učenje se koriste dva algoritma: propagacija labela i PageRank. Vrednovanjem je utvrđeno da postupak nije dovoljno dobar da bi se njime izgradio dobar leksikon sentimenta.

Ključne riječi: analiza sentimenta, polunadzirano učenje, leksikon sentimenta

Semisupervised Acquisition of Sentiment Polarity Lexicon

Abstract

Firstly, we describe related methods of automatic sentiment lexicon acquisition. Based on these methods, we implement and evaluate corpus-based sentiment lexicon acquisition approach. The approach is based on semisupervised graph-based algorithms, which makes this approach suitable for languages lacking prebuilt lexical resources. For similarity measures we use raw co-occurrence, pointwise mutual information and latent semantic analysis. PageRank and label propagation are the two used algorithms for semisupervised graph-based learning. The approach is shown to have not so good results, so it would be inadvisable to use it for acquisition of good sentiment lexicon.

Keywords: sentiment analysis, semisupervised learning, sentiment lexicon