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

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 4273

Klasifikacija i analiza stavova u korisničkim komentarima na internetu

Ivan Paljak

Zagreb, srpanj 2015.

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

Zagreb, 13. ožujka 2015.

ZAVRŠNI ZADATAK br. 4273

Pristupnik: **Ivan Paljak (0036474073)**
Studij: Računarstvo
Modul: Računarska znanost

Zadatak: **Klasifikacija i analiza stavova u korisničkim komentarima na internetu**

Opis zadatka:

Korisnički komentari na internetu vrijedan su izvor informacija za analizu stavova i mišljenja ljudi o događajima i njihovim protagonistima, političkim odlukama i političkim subjektima, ideološkim pitanjima, kontroverznim temama itd. Računalna analiza stavova razmjerno je novo područje u okviru analize sentimenta koje se bavi automatskom klasifikacijom i analizom stavova izraženih u tekstu, primjerice korisničkih komentara na internetu. Riječ je o posebno izazovnom zadatku, dodatno otežanom zbog vrlo niske jezične kvalitete korisničkih komentara.

U okviru završnoga rada potrebno je proučiti pristupe za automatsku klasifikaciju stavova, s naglaskom na pristupe temeljene na strojnom učenju. Osmisliti model za klasifikaciju i analizu stavova korisničkih komentara na hrvatskome jeziku. Pored klasifikacije stavova, model treba omogućiti grubu analizu stavova u okviru odabrane teme u vidu jednostavne analiza glavnih argumenata kojima korisnici potkrjepljuju svoje stavove. Izgraditi odgovarajući skup tekstnih podataka na hrvatskome jeziku za razvoj i ispitivanje modela. Razviti programsku implementaciju modela i primijeniti ga na korisničke komentare na hrvatskome jeziku za neke odabrane teme. Provesti iscrpno vrednovanje modela, usporedbu s referentim modelom, statističku obradu rezultata te analizu pogrešaka. Radu priložiti izvorni i izvršni kod razvijenog sustava, označene skupove podataka i potrebnu dokumentaciju te citirati korištenu literaturu.

Zadatak uručen pristupniku: 13. ožujka 2015.
Rok za predaju rada: 12. lipnja 2015.

Mentor:

Doc. dr. sc. Jan Šnajder

Djelovođa:

Doc. dr. sc. Tomislav Hrkać

Predsjednik odbora za završni rad modula:

Prof. dr. sc. Siniša Srbljić

Veliko hvala najdražim označivačima – Antei, Filipu, Frani, Ivni, Kaji, Pauli i Seketu što su nesebično izdvojili dio svoga vremena kada je bilo najpotrebnije.

Veliko hvala i mentoru Janu koji me beskonačno puta uputio u pravom smjeru, Domagoju na brojnim savjetima i iscrpnoj recenziji rada, te Goranu i Mladenu kojima nikad nije bilo teško otvoriti vrata zavoda.

Najveće hvala svim članovima obitelji na ogromnoj potpori i razumijevanju tijekom ovog semestra i čitavog studija.

Sadržaj

1. Uvod

Korisnički komentari na internetu vrijedan su izvor informacija za analizu stavova i mišljenja ljudi o događajima i njihovim protagonistima, političkim odlukama i političkim subjektima, ideološkim pitanjima, kontroverznim temama itd. Računalna analiza stavova razmjerno je novo područje u okviru analize prirodnog jezika (engl. *natural language processing*) koje se bavi automatskom klasifikacijom i analizom stavova izraženih u tekstu, primjerice korisničkih komentara na internetu. Riječ je o posebno izazovnom zadatku, dodatno otežanom zbog vrlo niske jezične kvalitete korisničkih komentara. Klasifikacija i analiza korisničkih komentara najčešće se temelje se na metodama strojnog učenja (engl. *machine learning*), jednog od glavnih problema umjetne inteligencije (engl. *artificial intelligence*).

Cilj ovoga rada jest klasificirati korisničke komentare s obzirom na njihove stavove spram događaja koji su prodrmali hrvatsku javnost ostavivši raskol među mišljenjima građana te, analizom komentara koji zagovaraju pojedini stav, ekstrahirati najčešće argumente kojima ga javnost podupire. U okviru rada, problem klasifikacije riješen je nadziranim učenjem, odnosno metodom potpornih vektora (engl. *support vector machines*), dok se provedena analiza stavova temelji na metodi grupiranja (engl. *clustering*), jednoj od metoda nenadziranog strojnog učenja. Skup korisničkih komentara preuzet je s hrvatskog web–portala *index.hr*[1], a spomenuti su komentari vezani uz unaprijed odabrane događaje.

Važno je istaknuti da su preuzeti korisnički komentari uglavnom slabe jezične i pravopisne kvalitete. Ova je činjenica, uz nedodstatak relevantnih radova koji se bave problemima klasifikacije i analize stavova iznesenih u tekstovima na hrvatskom jeziku, u velikoj mjeri negativno utjecala na konačan ishod ovog rada.

U poglavljima koja slijede detaljno su opisana dva temeljna problema kojima se bavi ovaj rad – automatska klasifikacija i analiza stavova u korisničkim komentarima. Za svaki su problem najprije navedena saznanja prikupljena iz srodnih

[1] http://www.index.hr

znanstvenih radova te je napravljena usporedba između izučenih i korištenih metoda. Zatim je naveden i detaljno objašnjen model koji taj problem rješava te je provedeno njegovo iscrpno vrednovanje. Rad završava prezentacijom te detaljnim opisom dobivenih rezultata uz analizu pogrešaka i eventualnih poboljšanja.

2. Klasifikacija stavova u korisničkim komentarima na internetu

Problem računalne klasifikacije stavova u korisničkim komentarima poseban je slučaj općeg problema statističke klasifikacije. Intuitivno, za rješavanje problema potrebno je, pomoću računala, odrediti stav izražen u svakom korisničkom komentaru. Formalnije, neka je K skup svih korisničkih komentara, a S skup svih mogućih stavova izraženih u komentarima iz K, tada svaka funkcija $f : K \mapsto S$ predstavlja rješenje problema klasifikacije stavova u korisničkim komentarima.

2.1. Srodni radovi

Iako se u okviru ovoga rada isključivo koncentriramo na stavove koji proizlaze iz korisničkih komentara na internetu, znanstveni se radovi uglavnom bave klasifikacijom stavova na općenitim tekstovima. Usprkos tome, većina je korištenih metoda, uz iznimke onih koje ovise o jezičnoj kvaliteti teksta, primjenjiva i na korisničke komentare s interneta. Također, valja naglasiti da se spomenuti radovi bave klasifikacijom stavova iz tekstova pisanih engleskim jezikom. Shodno tome, zanemarujemo metode koje se oslanjaju na karakterističnosti engleskoga jezika.

Glavni motivator za odabir korištene metode klasifikacije rad je dvoje autora, Bo Pang i Lillian Lee, sa sveučilišta Cornell iz SAD–a (Pang et al., 2002). Ovaj se rad, između ostalog, bavio usporedbom triju klasifikacijskih metoda – naivne Bayesove klasifikacije, klasifikacije maksimalne entropije i metode potpornih vektora. Skup podataka za klasifikaciju sastojao se od filmskih recenzija preuzetih s internetske baze filmova IMDB-a (engl. *internet movie database*), a recenzije su klasificirane u dvije ciljne klase s obzirom na to jesu li film ocijenile pozitivno ili negativno. Unatoč tome što su filmske recenzije na mnogo većoj jezičnoj razini od

korisničkih komentara na hrvatskim web–portalima, mišljenja smo da su problemi dovoljno slični. Sami autori rada ističu kako „metode strojnog učenja prikazane u ovom radu nisu usko vezane uz područje filmskih recenzija te su primijenjive na drugim domenama“. U zaključku rada autori navode da naivna Bayesova klasifikacija u većini slučajeva daje najgore rezultate, dok metoda potpornih vektora daje najbolje. Usprkos tome, valja naglasiti da razlike u performansama i nisu prevelike.

2.2. Metoda potpornih vektora

Motivirani zaključkom rada iz prethodnog poglavlja, problem klasifikacije korisničkih komentara odlučili smo riješiti metodom potpornih vektora koja služi za klasifikaciju vektora koji se nalaze u nekom hiperprostoru.

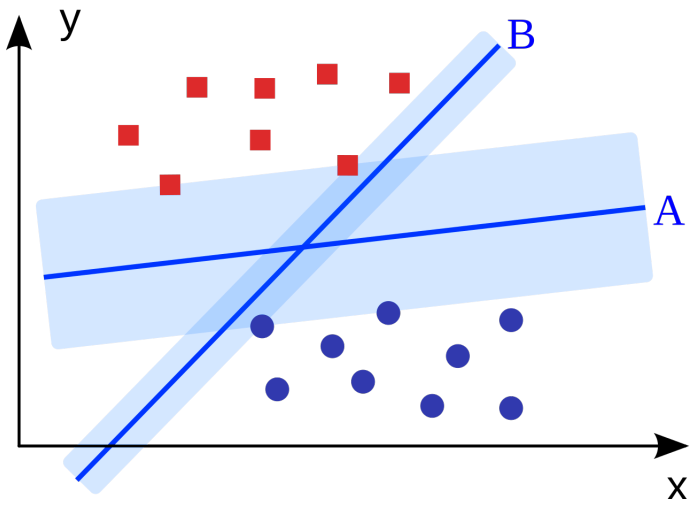
2.2.1. Osnovni oblik

U osnovnom obliku, metoda potpornih vektora rješava problem binarne klasifikacije, odnosno, zasad pretpostavljamo $|S| = 2$. Treniranje modela skupom unaprijed označenih[1] primjera učenja svodi se na ručno pridruživanje jedne od dviju klasa svakom od vektora koji se nalazi u skupu primjera. Tada se treniranje modela u suštini svodi na optimalan pronalazak hiperravnine koja razdvaja dvije klase označenih vektora (Meyer i Wien, 2014). Optimalnost podjele vektorskog prostora proizlazi iz maksimizacije minimalne udaljenosti između neka dva vektora koja su s različitih strana ravnine i ne pripadaju istoj klasi, a ta se udaljenost naziva margina (slika 2.1). Vektori koji se nalaze na rubu margine i ustvari definiraju njen položaj nazivaju se potpornim vektorima. Nakon što je model istreniran, klasificiranje vektora svodi se na provjeru položaja vektora u odnosu na ravninu podjele. Na temelju tog odnosa pridružujemo vektoru jednu od klasa.

2.2.2. Metoda meke margine

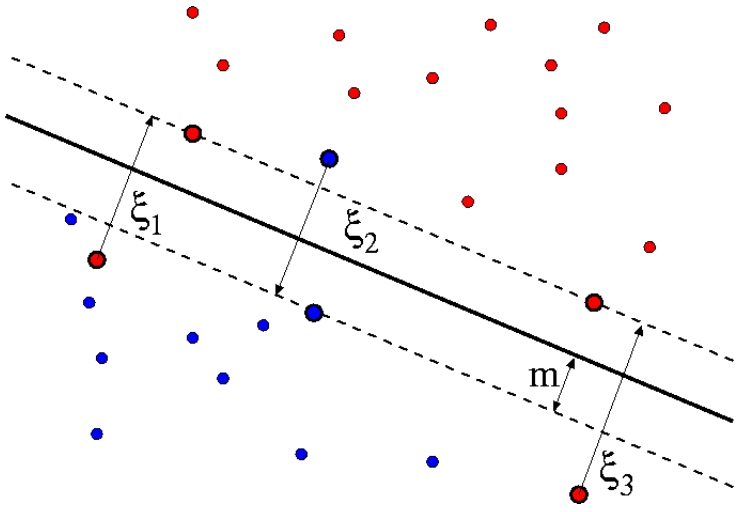
Prirodno se nameće pitanje – „Što ako ravninom nije moguće razdvojiti vektore različitih klasa?”. Problemu pristupamo na način da dopustimo određenu pogrešku pri presijecanju prostora ravninom. Konkretnije, za neku fiksnu ravninu

[1] skup primjera kojim je pridijeljena ispravna klasa.



Slika 2.1: Vizualni prikaz treniranja modela. Ravnina A predstavlja optimalnu podjelu te je očevidno bolji odabir od ravnine B zbog veće margine.

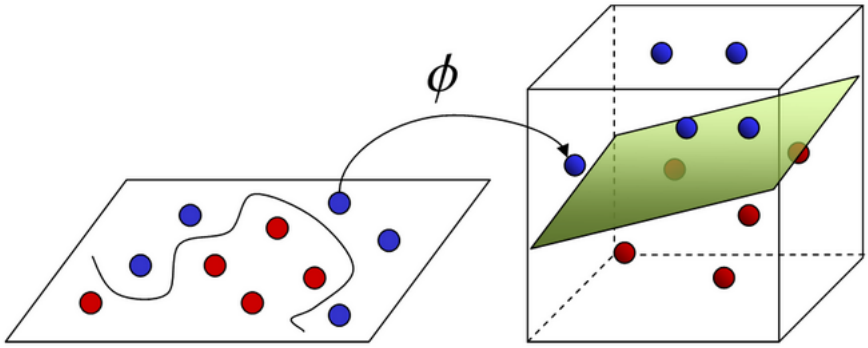
podjele i svaki, s obzirom na nju, pogrešno klasificirani vektor $\vec{x_i}$ možemo odrediti pogrešku pri klasificiranju ξ_i (engl. *slack variable*) koja raste s udaljenošću tog vektora i ravnine podjele (Cortes i Vapnik, 1995). Jasno, vrijednost ξ_i za točno klasificiran vektor iznosi 0. Standardno izračunatoj margini tada pridodajemo vrijednost $C \sum_{i=1}^{|K|} \xi_i$ gdje je parametar C realan broj pomoću kojeg reguliramo količinu dopuštene pogreške pri klasifikaciji. Očita mana ove metode jest da model neće uspjeti savršeno klasificirati čak ni skup podataka na kojima je istreniran. S druge pak strane, uspjeli smo pronaći dovoljno dobru linearnu podjelu prostora. Ovaj postupak nazivamo metodom meke margine (engl. *soft–margin SVM*).



Slika 2.2: Vizualni prikaz treniranja modela metodom meke margine.

2.2.3. Nelinearni klasifikator

Drugi se pristup temelji na transformaciji vektora u prostor većih dimenzija te linearnoj klasifikaciji vektora u tom prostoru koristeći neku od gore opisanih metoda. Transformacijska funkcija naziva se još i jezgrena funkcija (engl. *kernel function*), a njen izvod izlazi van okvira ovoga rada. Ovaj postupak opravdan je Coverovim teoremom koji nam govori kako je skup linearno nerazdvojivih vektora moguće s velikom vjerojatnošću transformirati u skup linearno razdvojivih vektora nekog više dimenzijskom prostoru. Budući da granica podjele u originalnom prostoru vrlo vjerojatno neće biti linearna, ovu metodu svrstavamo u nelinearne metode klasifikacije.



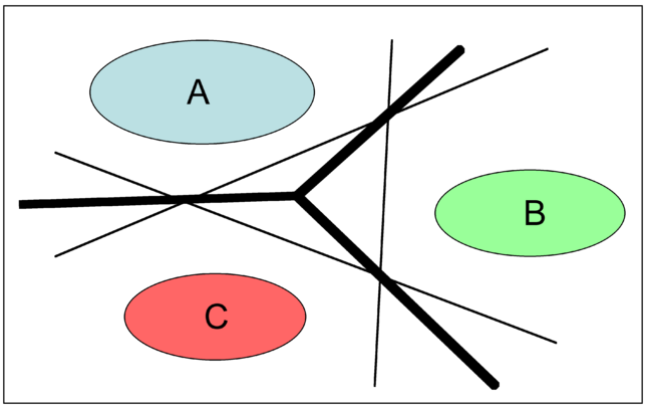
Slika 2.3: Prikaz transformacije ulaznog, linearno nerazdvojivog prostora u višedimenzijski linearno razdvojiv prostor pomoću transformacijske funkcije Φ.

2.2.4. Višeklasno klasificiranje

Na prvi se pogled čini kako metode predstavljene u prethodnim poglavljima nije moguće primijeniti ako vektore želimo klasificirati u više od dvije klase, odnosno ako vrijedi $|S| > 2$. No, ovaj ćemo problem svesti na već opisane metode, odnosno, višeklasni klasifikator izgradit ćemo pomoću više binarnih klasifikatora.

Koristeći metodu *jedan protiv svih* (engl. *one–against–all*), gradimo onoliko binarnih klasifikatora koliko ima klasa. Svaki je binarni klasifikator istreniran na način da odijeli vektore koji pripadaju jednoj od klasa od vektora koji pripadaju ostalim klasama. Konačna klasifikacija vektora ovisi o položaju vektora u odnosu na svih $|S| = n$, $(n > 2)$ ravnina podjele binarnih klasifikatora.

Često se u praksi koristi i metoda *jedan protiv jedan* (engl. *one–against–one*) u kojoj se, za svaki par klasa, gradi poseban binarni klasifikator. Dakle, gradi se



Slika 2.4: Vizualni prikaz treiranja troklasnog klasifikatora metodom *jedan protiv svih*. Tanjim linijama prikazane su ravnine podjele svih triju binarnih klasifikatora, dok se masno otisnutom linijom jasno vidi konačna granica podjele među klasama.

ukupno $\frac{n(n-1)}{2}$ binarnih klasifikatora. Za potrebe konačnog klasificiranja koristi se glasačka strategija, odnosno, ona klasa u koju je vektor najviše puta upao uzima se kao ispravna.

Obje prezentirane metode, iako se u drugačije ponašaju u određenim situacijama, u praksi daju gotovo jednake rezultate (Duan i Keerthi, 2005). Za potrebe višeklasnog klasificiranja u ovom smo radu koristili metodu *jedan protiv jedan*.

2.3. Model

Kako bismo iskoristili metodu potpornih vektora, potrebno je naš problem, klasifikaciju stavova u korisničkim komentarima, svesti na problem klasifikacije vektora opisanog u prethodnom poglavlju. Dakako, klase će u našem slučaju odgovarati izraženim stavovima, dok će vektori odgovarati korisničkim komentarima.

2.3.1. Korisnički stavovi

Komentari preuzeti s interneta za potrebe ovog rada vezani su uz sljedeće, kronološki navedene, događaje.

- Skandiranje kontroverznog pozdrava „Za dom, spremni!” od strane Josipa Šimunića, hrvatskog nogometnog reprezentativca.
- Pokrenuta građanska inicijativa „U ime obitelji” s ciljem ustavne definicije braka kao zajednice muškarca i žene.

– Pokrenuta inicijativa „Ne damo naše AUTOCESTE!" s ciljem sprječavanja monetizacije hrvatskih autocesta.

– Nekolicina hrvatskih branitelja započinje prosvjed u šatoru u Savskoj ulici.

Navedeni događaji namjerno su odabrani na način da se izneseni stavovi u korisničkim komentarima, sa strane sentimenta, mogu okarakterizirati kao:

– ZA!

– PROTIV!

– Neutralan komentar

Poznavajući uobičajen tijek komunikacije na hrvatskim web–portalima, potrebno je oformiti još jednu klasu koji obuhvaća sve one komentare koji skreću s teme. Ovu smo klasu prozvali *off–topic*. Važno je naglasiti kako se ova klasa uvelike razlikuje od prethodne tri jer je vezana uz tematiku komentara, a ne sentiment. Tematska klasifikacija je, u okviru umjetne inteligencije, dosta jednostavniji problem.

Detaljniji opis svake klase (oznake) nalazi se u dodatku A *(Upute za označivače)*.

2.3.2. Vektor značajki

Ostaje nam još problem preslikavanja korisničkog komentara u vektorski oblik. U te svrhe, svaki ćemo korisnički komentar predstaviti vektorom značajki (engl. *feature vector*). Svaka značajka brojčana je reprezentacija neke karakteristike komentara za koju vjerujemo da na neki način doprinosi iznesenom stavu u komentaru.

Prije nego navedemo korištene značajke, definirat ćemo *vektor riječi*. Vektor riječi jest vektorska reprezentacija sadržaja komentara pri čemu svaka komponenta vektora predstavlja frekvenciju pojavljivanja odgovarajuće riječi u komentaru. Dimenzija vektora riječi odgovara broju riječi u rječniku s kojim baratamo.

Naš se rječnik sastoji od svih riječi iz preuzetih komentara bez *zaustavnih riječi* (engl. *stopwords*). Formalno, neka W_i predstavlja skup riječi i–tog preuzetog komentara, a W_{stop} predstavlja skup zaustavnih riječi. Rječnik D tada definiramo kao

$$D = \bigcup_{i=1}^{i \leq |K|} W_i \setminus W_{stop}$$

Zaustavne su riječi sve riječi hrvatskoga jezika koje ne emitiraju semantičko ili tematsko značenje, primjerice čestice. Skup zaustavnih riječi hrvatskoga jezika

ustupio je *TakeLab*.[2] Također, kako bismo postigli jednačenje riječi neovisno o njihovim nastavcima uzrokovanim pravilima jezika (deklinacije, konjugacije, itd.), uspoređivat ćemo korjenovane riječi (engl. *stemmed words*). Korjenovanje riječi provodimo na način da odbacimo sva slova osim prvih pet.

Definiramo li funkciju $freq : K \times D \mapsto \mathbb{N}_0$ koja vraća broj pojavljivanja riječi $w \in D$ u i–tom komentaru $k_i \in K$, vektor riječi spomenutog komentara iznosi

$$\vec{v_{k_i}} = (freq(k_i, w_1), freq(k_i, w_2), freq(k_i, w_3), \ldots, freq(k_i, w_{|D|}))$$

Prvih $|D|$ komponenata vektora značajki komentara k_i odgovara upravo komponentama vektora $\vec{v_{k_i}}$. Komponente vektora riječi logičan su izbor značajki jer poneke riječi, odnosno skupovi riječi, snažno emitiraju korisnikov stav.

Preostale korištene značajke indeksirane su od $|D| + 1$ do $|D| + 5$ i to redom:

- Kosinus kuta između korisničkog komentara i centroida svih preuzetih članaka vezanih uz odgovarajući događaj.
- Redni broj komentara.
- Duljina komentara.
- Događaj na koji se komentar odnosi.
- Postojanje citata u komentaru.

Prva je značajka svojevrsna mjera sličnosti komentara s tekstovima članaka. Pretpostavlja se da će komentari koji sadrže imena osoba i lokacija vezanih uz događaj, koriste relevantne pojmove ili pak citiraju tekst članka, najvjerojatnije emitirati stav vezan uz temu. Centroid članaka vezanih uz neki događaj definiramo kao aritmetičku sredinu vektora riječi svih preuzetih članaka vezanih uz taj događaj. Kosinus kuta između vektora možemo izraziti pomoću popularne formule za skalarni produkt vektora:

$$cos(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\|\|\vec{b}\|} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{(\sum_{i=1}^{n} a_i^2)(\sum_{i=1}^{n} b_i^2)}} \tag{2.1}$$

Nadalje, redni broj komentara, naizgled pomalo beskorisna značajka, opravdana je činjenicom da korisnički komentari na hrvatskim web–portalima nakon nekog vremena imaju sve veću tendenciju skretanja s teme.

Čitajući preuzete komentare subjektivnog smo dojma da korisnici čiji se komentari drže teme i emitiraju stav imaju tendenciju ostavljati duže komentare od prosjeka. Shodno observaciji, uvrstili smo duljinu komentara u popis značajki.

[2]Text Analysis and Knowledge Engineering Lab (takelab.fer.hr)

Sam događaj na koji se komentar odnosi također može uzrokovati skretanje s teme među komentarima. Primjerice, politički nabijeni događaji često rezultiraju svađom među korisnicima na osobnoj razini. Također, događaj na koji se komentar odnosi u kombinaciji s vektorom riječi može jasnije emitirati stav.

Za kraj, postojanje citata u komentaru daje nam do znanja da se komentar odnosi na neki prethodno ostavljeni komentar. Ova značajka u kombinaciji s vektorom riječi može biti indikacija argumentirane rasprave ili pak privatne svađe među korisnicima.

2.3.3. Skup za treniranje

Da bismo uspješno istrenirali klasifikacijski model, moramo najprije generirati skup podataka za treniranje. Taj bi se skup trebao sastojati od značajnog broja primjera za koje znamo kojoj klasi pripadaju. Odnosno, potrebno je odrediti iznesene stavove u određenom broju korisničkih komentara. Budući da se komentari odnose na jedan od četiri odabrana događaja, važno je da je svaki od događaja jednako zastupljen u skupu za treniranje. Iz istih razloga, prilikom generiranja skupa za treniranje velik je naglasak stavljen na podjednaku zastupljenost:

- članaka s kojih su preuzeti komentari
- duljine komentara
- starosti komentara

Logično je da će stavove korisničkih komentara evaluirati čovjek. Točnije, evaluirat će ih grupa ljudi kako bismo preciznije odredili korisnikov stav i dobili procjenu koliko je problem ustvari težak. Za svaki događaj uzeli smo uzorak od 730 komentara pa se skup za treniranje sastoji od ukupno $4 \cdot 730 = 2920$ korisničkih komentara. Ovaj je skup ravnomjerno raspodijeljen na osmero označivača od kojih je svaki označio 800 komentara pri čemu je 40 komentara zajedničko svim označivačima. Ovaj nam presjek služi kao statistički pokazatelj te je njegova uloga objašnjenja u poglavlju *Eksperimentalno vrednovanje*. Ispravnim stavom na komentarima koje su označili svi označivači smatra se onaj stav kojeg je identificiralo najviše označivača.

2.4. Implementacija

Budući da smo na teoretskoj razini uspješno sveli problem klasifikacije korisničkih komentara na problem klasifikacije vektora koji znamo riješiti metodom potpornih

vektora, vrijeme je da taj postupak provedemo na računalu.

Metoda potpornih vektora poznat je i višestruko puta uspješno implementiran algoritam. Za potrebe ovog rada korištena je javno dostupna implementacija `libSVM` koja sadrži implementacije svih navedenih metoda klasifikacije osim višestrukog klasificiranja metodom *jedan protiv svih* (Chang i Lin, 2011). Budući da je metoda *jedan protiv jedan* u velikoj većini slučajeva jednako dobra kao i metoda *jedan protiv svih*, mišljenja smo da ovaj "nedostatak" ne predstavlja problem. Treniranju modela idejno smo pristupili na dva načina:

1. Četveroklasnim klasificiranjem.

2. Kaskadom binarnog klasifikatora između *off–topic* klase i ostalih klasa te troklasnog klasifikatorom između preostalih klasa.

Svaki od temeljnih klasifikatora pokušali smo izgraditi metodom meke margine te kombinacijom nelinearnog klasifikatora i metode blage margine. Za prvi je slučaj potrebno specificirati parametar `C` pa se treniranje klasifikatora svodi na pokretanje naredbe `svm-train -C <vrijednost> train.in` iz ljuske operacijskog sustava. U slučaju nelinearnog klasifikatora, potrebno je specificirati parametre `C` i γ, a tada se treniranje klasifikatora svodi na upisivanje naredbe `svm-train -C <vrijednost> -g <vrijednost> train.in`.

Postavlja se pitanje kako odabrati parametre `C` i γ? Unutar `LibSVM` paketa nalazi se skripta `grid.py` pomoću koje možemo odabrati optimalne vrijednosti parametara za naš skup podataka. Koristimo li linearni klasifikator metodom meke margine, optimalnu vrijednost parametra `C` dobit ćemo pokretanjem naredbe

```
python grid.py -log2c -5,15,2 -log2g null -v 5 -s 0 -t 0 train.in
```

Prvom zastavicom i njenom vrijednošću (`-log2c -5,15,2`) definirali smo interval na kojem skripta traži parametar `C`. Drugom zastavicom i njenom vrijednošću definirali bismo interval na kojem skripta traži parametar γ, no u slučaju linearnog klasifikatora taj parametar ne koristimo pa vrijednost postavljamo na `null`. Trećom smo zastavicom i njenom vrijednošću `-v 5` dali do znanja da želimo paremetre odrediti uz peterostruku kros–validaciju. Općenito, k–strukom unakrsnom validacijom (engl. *k–fold cross validation*) nasumično dijelimo ulazni skup podataka na k dijelova jednake veličine. Jednu od particija koristimo u validacijske svrhe, dok parametre odabiremo pomoću preostalih $k-1$ particija. Ovaj se postupak ponavlja k puta kako bi svaka particija imala priliku biti validacijska. Konačno, zastavice `-t` i `-s` definiraju tip stroja potpornih vektora i

korištene jezgre (u ovom slučaju linearna). Koristimo li nelinearni klasifikator i metodu meke margine, optimalne vrijednosti parametara dobivamo pokretanjem naredbe

```
python grid.py -log2c -5,15,2 -log2g 3,-15,-2 -v 5 -s 0 -t 2 train.in
```

Budući da nam je zbog nelinearnog klasifikatora potreban parametar γ navodimo odgovarajući interval. Također, umjesto linearne jezgre koristimo RBF (engl. *radial basis function*) jezgrenu funkciju koja je definirana kao

$$K(x_i, x_j) = e^{-\gamma(\|x_i - x_j\|^2)}, \gamma > 0 \tag{2.2}$$

Po odabiru optimalnih parametara uspoređujemo dva predložena modela klasifikatora te, na temelju evaluacijskih mjera opisanih u poglavlju *Eksperimentalno vrednovanje*, odabiremo onaj koji daje bolje rezultate.

Sve preostale zadatke vezane uz klasifikaciju komentara, primjerice, preuzimanje komentara i tekstova članaka, izrada web–aplikacija za označavanje komentara, generiranje vektora riječi i vektora značajki te ulaznih datoteka u `libSVM` implementirali smo ručno koristeći programski jezik *Python*.

3. Analiza stavova u korisničkim komentarima na internetu

Nakon klasifikacije, potrebno je analizirati komentare koje smo klasificirali kao "ZA!" ili "PROTIV" te za svaku od tih klasa izvući najčešće argumente koji brane taj stav. Ovakav postupak naziva se dubinska analiza argumenata (engl. *argumentation mining*) te predstavlja relativno nov i dosta zahtjevan problem u okviru obrade prirodnog jezika. Cilj argumentnog rudarenja jest pronaći argumente i njihovu strukturu unutar tekstnog dokumenta.

U okviru ovog rada, rudarenju argumenata pristupamo na dosta pojednostavljen način. Zanemarit ćemo pronalaženje strukture argumenata i ekstrakciju samog argumentiranog dijela iz teksta. Spomenuti zadaci vrlo su sofisticirani te zahtijevaju visoku jezičnu kvalitetu teksta za uspješno rješavanje. Budući da jezična kvaliteta često nije odlika korisničkih komentara, zadovoljit ćemo se u cijelosti izdvojenim komentarima koje smatramo zastupnicima pojedinog argumenta.

Pretpostavimo li uspješnu klasifikaciju stavova opisanu u prethodnim poglavljima, trebali bismo postupak analize provoditi nad komentarima za koje smo sigurni da emitiraju stav. Budući da korisnici komentiraju potresne događaje koji dijele građane s obzirom na njihova mišljenja i to u okruženju koje potiče diskurs, smatramo da je valjano pretpostaviti kako većina komentara sadrži nekakav oblik argumentacije. Također, u okviru ovog zadatka cilj nam je odrediti predstavnike *najčešće* korištenih argumenata koji brane neki stav. Zbog navedenih pretpostavki i dobivenih saznanja iz relevantnih radova, za rješavanje ovog problema odabrali smo nenadziranu metodu grupiranja (engl. *unsupervised clustering*). Odnosno, nadamo se da će podjela komentara u sadržajno slične grupe rezultirati filtracijom glavnih argumenata. Dakako, pretpostavlja se da svaka grupa odgovara jednom argumentu dok broj elemenata grupe odgovara broju korisnika koji se tim argumentom služe.

3.1. Srodni radovi

Nažalost, najuspješnije metode argumentnog rudarenja prezentirane u raznim znanstvenim radovima pretežito ovise vrlo visokoj jezičnoj kvaliteti komentara te specifičnostima jezika kojim su tekstovi pisani. Analiza argumenata nad tekstovima pisanim hrvatskim jezikom, koliko je autoru poznato, nije opisano ni u kakvoj znanstvenoj literaturi.

Primjerice, autorice članka *Analiza argumenata: detekcija, klasifikacija i struktura argumenata u tekstovima* (Palau i Moens, 2009) problemu detekcije argumenata pristupile su nadziranim strojnim učenjem u obliku izgradnje binarnih klasifikatora. Točnije, dijelove teksta svrstavale su u argumentne i neargumentne klase, a neke od komponenata vektora značajki su tip subjekta, tip predikata, glagolsko vrijeme predikata, vektor retoričkih fraza i vektor argumentativnih fraza. Da bismo ostvarili prve tri značajke potrebno je parsanjem rečenica identificirati glavne rečenične dijelove i njihova svojstva. Ovaj postupak je vrlo teško, a ponekad i nemoguće, provesti nad tekstovima korisničkih komentara koji su uglavnom pisani na raznim dijalektima, ne sadrže pravilnu interpunkciju te sadrže razne jezične pogreške. Također, zadnje dvije značajke zahtjevaju popis retoričkih i argumentativnih fraza hrvatskoga jezika kojima nemamo pristup. Zbog navedenih ograničenja, u okviru ovoga rada ne možemo se oslanjati na opisanu metodu.

Sukladno prethodnom odlomku, odlučili smo problem riješiti nekom od metoda grupiranja. Relevantan odgovor na pitanje koju metodu odabrati daje članak *Evaluacija algoritama grupiranja nad mrežama proteinskih interakcija* (Brohee i Van Helden, 2006), gdje autori uspoređuju uspješnost četiri popularna algoritma grupiranja.

- grupiranje RNSC (engl. *restricted neighbourhood search clustering*)
- Markovljevo grupiranje (MCL) (engl. *Markov clustering*)
- grupiranje MCODE (engl. *molecular complex detection*)
- grupiranje SPC (engl. *super–pragmatic clustering*)

Usporedba algoritama provedena je na 42 primjera od kojih je 41 dobiven nasumičnim dodavanjem i oduzimanjem veza početnog grafa koji odgovara proteinskim reakcijama kako bi se dodatno smanjila vjerojatnost pogreške pri interpretaciji rezultata. Svakim su algoritmom grupirani svi primjeri te je analizirana njihova osjetljivost na vrijednosti parametara te su utvrđene njihove optimalne vrijednosti. U zaključku stoji kako je algoritam Markovljevog grupiranja izuzetno

robustan spram alternacijama grafa te dobiveni rezultati dodatno potvrđuju njegovu superiornost u odnosu na konkurenciju.

3.2. Markovljevo grupiranje

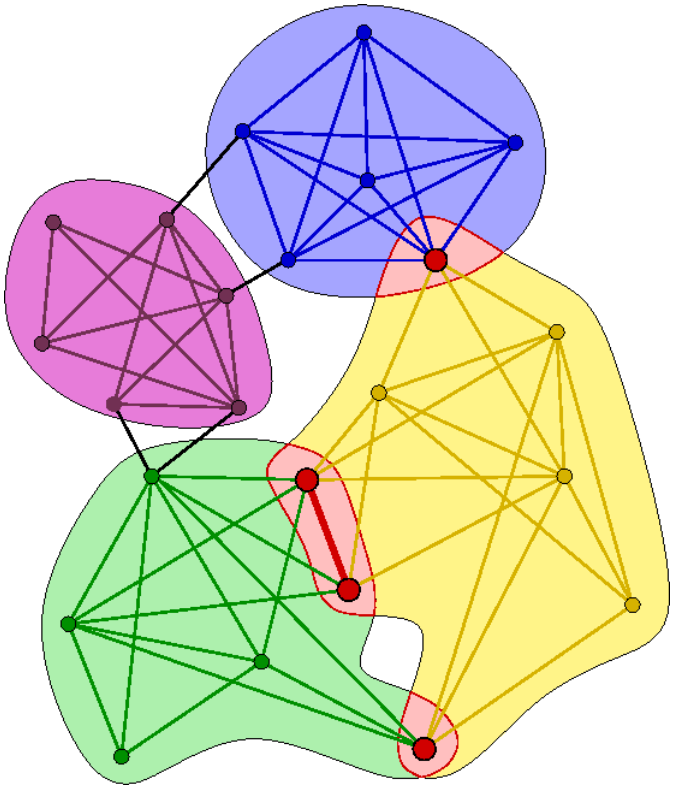
Osmišljen od strane nizozemca Stijn van Dongena, Markovljevo grupiranje (engl. *Markov cluster algorithm, MCL*) efikasan je algoritam grupiranja čvorova u grafu koji se temelji na simulaciji protoka kroz graf. Jedna od glavnih karakteristika algoritma jest da, za razliku od većine algoritama grupiranja, automatski određuje broj grupa u grafu. U kontekstu ovog algoritma, veze među čvorovima označavaju njihovu sličnost, a u slučaju težinskog grafa, težinska funkcija $w : (V \times V) \mapsto \mathbb{R}$ mjera je sličnosti između čvorova.

Zapitajmo se najprije kakve karakteristike posjeduje dobro odabrana grupa čvorova u grafu. Dongen u svom doktorskom radu (van Dongen, 2009), uz pretpostavku bestežinskog grafa, navodi sljedeća svojstva:

- Broj šetnji između dva čvora iz iste grupe uglavnom će biti veći od broja šetnji između dva čvora koji nisu u istoj grupi uz pretpostavku da su šetnje jednako duge.
- Nasumična šetnja na grafu koja posjeti neku grupu najvjerojatnije će posjetiti velik broj njenih čvorova prije no što napusti tu grupu.
- Uzmemo li sve najkraće puteve između svih parova čvorova u grafu, vrlo je vjerojatno da će veze koje spajaju dvije različite grupe biti često dio takvih putova.

Za izvedbu Markovljevog grupiranja, ključno je drugo navedeno svojstvo pa slijedi detaljnija definicija *nasumične šetnje* (engl. *random walk*). Općenito, niz posjećenih čvorova $(v_1, v_2, ..., v_n)$ pri čemu postoji veza između susjednih čvorova v_i i v_{i+1} nazivamo šetnjom. Nalazimo li se usred šetnje u čvoru v_i, u slučaju nasumične šetnje s jednakom ćemo vjerojatnošću šetnju nastaviti prema svakom od čvorova v_j (ako postoji veza između v_i i v_j). Prisjetimo li se definicije Markovljevih lanaca (engl. *Markov chain*), jasno je da zbog neovisnosti budućih stanja o prošlim stanjima, svaka nasumična šetnja odgovara nekom konačnom Markovljevom lancu. Ime algoritma slijedi iz spomenute ekvivalencije.

Prisjetimo se i matrice susjedstva grafa G u kojoj element r–tog retka i s–tog



Slika 3.1: Vizualni prikaz prirodne podjele čvorova u četiri grupe. Usprkos relativno malenom broju čvorova, graf uspijeva demonstrirati opravdanost Dongenovih svojstava.

stupca poprima vrijednost

$$(A_G)_{r,s} = \begin{cases} w(v_r, v_s), & v_r \text{ i } v_s \text{ su povezani} \\ 0, & \text{inače} \end{cases}$$

Odnosno, iz r–tog retka i s–tog stupca matrice susjedstva možemo očitati mjeru sličnosti između r–tog i s–tog čvora u grafu. Definirat ćemo i Markovljevu matricu M_G čiji s–ti stupac odgovara normiranom vektoru s–tog stupca matrice susjedstva. Primijetimo da se iz r–tog retka i s–tog stupca Markovljeve matrice može iščitati vjerojatnost prelaska s r–tog čvora u s–ti čvor usred nasumične šetnje. Ova je interpretacija valjana uz pretpostavku bestežinskog grafa, dok će u težinskom grafu naš zamišljeni šetač s većom vjerojatnošću krenuti putem koji odgovara većoj sličnosti između čvorova, a to je upravo ono što bismo željeli.

Možemo li odrediti vjerojatnost da će nasumična šetnja duljine dva koraka koja započinje u čvoru r završiti u čvoru s? Vjerojatnost da ćemo do čvora s doći preko čvora i iznosi $(M_G)_{r,i} \cdot (M_G)_{i,s}$. Traženu vjerojatnost dobivamo sumiranjem svih prethodno dobivenih vjerojatnosti za sve potencijalne međučvorove pa ona

iznosi

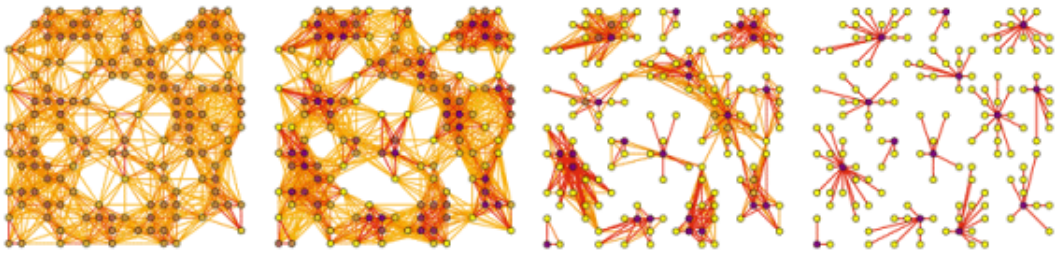
$$\sum_{i=1}^{i\leq|V|} (M_G)_{r,i} \cdot (M_G)_{i,s}$$

što odgovara elementu na r–tog stupca i s–tog retka matrice M_G^2. Odmah slutimo kako $(M_G)_{r,s}^n$ predstavlja vjerojatnost da će nasumična šetnja duljine $n-1$ koraka koja započinje u čvoru r završiti u čvoru s. Tvrdnja se vrlo jednostavno dokazuje metodom matematičke indukcije koristeći prethodno dobiveni rezultat kao bazu i istovjetan postupak kao korak indukcije. Prisjetimo li se drugog Dongenovog svojstva, uvidjet ćemo važnost dobivenog rezultata. Dignemo li Markovljevu matricu na dovoljno veliku, ali ne preveliku, potenciju, veze unutar grupa poprimat će veće vrijednosti od veza između grupa. Ovaj je postupak temelj Markovljevog grupiranja i nazivamo ga ekspanzijom (engl. *expansion*).

Nažalost, ovaj efekt se gubi nakon što šetnje dostignu dovoljno velik broj koraka. Kako bismo taj efekt održali što dulje, koristimo postupak inflacije (engl. *inflation*). Ideja je da nakon određenog broja koraka dodatno učvrstimo jake veze (unutar grupe) i dodatno oslabimo slabije veze (između grupa). Postupak inflacije provodimo tako da elemente Markovljeve matrice dignemo na neku potenciju te ponovo normaliziramo svaki stupac. Formalno, nad matricom $M \in \mathbb{R}^{n\times m}$ za realan, nenegativan broj k definiramo matrični operator inflacije $\Gamma_k : \mathbb{R}^{n\times m} \mapsto \mathbb{R}^{n\times m}$ pri čemu

$$(\Gamma_k M)_{r,s} = (M_{r,s})^k / \sum_{i=1}^{n} (M_{i,s})^k$$

Postupci ekspanzije i inflacije primjenjuju se naizmjence sve dok Markovljeva matrica ne konvergira, odnosno, dosegne stabilno stanje. Iz završnog stanja radi se konstrukcija grupa shodno drugom Dongenovom svojstvu.



Slika 3.2: Prikaz postepene kristalizacije grupa dobivene naizmjeničnom ekspanzijom i inflacijom Markovljeve matrice.

Algoritam je moguće implementirati u vremenskoj složenosti $\mathcal{O}(|V|r^2)$, gdje $|V|$ predstavlja broj čvorova, a r predstavlja prosječan broj resursa potrebnih

nekom čvoru.

3.3. Model

Da bismo uspješno iskoristili metodu Markovljevog grupiranja, jedino što je potrebno definirati jest mjera sličnosti između čvorova u grafu. Jasno je da u našem slučaju čvorovima u grafu poredstavljamo komentare korisnika koji imaju isti stav prema nekoj temi (”ZA!” ili ”PROTIV!”). U našem smo slučaju odlučili mjerom sličnosti okarakterizirati svaki par korisničkih komentara pa gradimo potpuni težinski graf. Logičan odabir mjere sličnosti jest kosinus kuta između vektora riječi dvaju komentara kojeg, prisjetimo se, računamo pomoću

$$cos(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\|\|\vec{b}\|} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{(\sum_{i=1}^{n} a_i^2)(\sum_{i=1}^{n} b_i^2)}}$$

.

Dakle, da bismo napravili odgovarajuće grupe vezane uz dva različita stava o svakom od četiri odabrana događaja, potrebno je izgraditi ukupno $2 \cdot 4 = 8$ grafova nad kojima radimo grupaciju. Budući da želimo ekstrahirati samo najčešće korištene argumente, u obzir ćemo uzeti samo one grupe koje sadrže određen broj komentara. Na taj ćemo način, u teoriji, izbaciti slabije argumente i neke pogrešno klasificirane komentare.

3.4. Implementacija

Za potrebe ovog rada koristili smo javnu implementaciju[1] Markovljevog grupiranja samog autora algoritma. Za uspješno korištene ove biblioteke, potrebno je ulazni graf predstaviti putem matrice susjedstva ili datotekom u `-abc` formatu. Odlučili smo se na `-abc` format u kojem broj redaka datoteke predstavlja broj veza u grafu, a svaki redak opisuje jednu vezu pomoću oznaka `a` i `b` te numeričke vrijednosti `c`. Oznake `a` i `b` jednoznačno označavaju čvorove između kojih je mjera sličnosti jednaka `c`. U našem slučaju, oznake su bile indeksi komentara, a broj `c` bio je kosinus kuta između vektora riječi tih komentara.

Sam se postupak grupiranja tada svodi na pokretanje naredbe `mcl graf.in -abc -o grupe` iz ljuske operacijskog sustava. Prvi je argument naziv datoteke

[1] preuzeto s *http://micans.org/mcl/*

koja sadrži definiciju grafa u `abc` formatu, dok je četvrti argument ime izlazne datoteke u koju će program ispisati sadržaj svake grupe. Izlazna se datoteka sastoji od onoliko redaka koliko je pronađeno grupa, a u svakom se retku, razdvojene tabulatorom, nalaze oznake čvorova koje pripadaju toj grupi. Generiranje odgovaraućih datoteka u `abc` formatu implementirano je korištenjem programskog jezika *Python*.

4. Eksperimentalno vrednovanje

U ovom se poglavlju nalaze detalji o preuzetom skupu podataka, izneseni su rezultati klasifikacije i argumentnog rudarenja nad korisničkim komentarima, navedene su i objašnjene metode vrednovanja rezultata te je iznesena njihova interpretacija uz moguće metode poboljšanja.

4.1. Skup podataka

Naš se početni skup podataka sastoji od korisničkih komentara preuzetih s hrvatskog web-portala index.hr. Prisjetimo se, svi su komentari vezani uz članke koji su pak vezani uz jedan od sljedeća četiri događaja:

- Skandiranje kontroverznog pozdrava „Za dom, spremni!" od strane Josipa Šimunića, hrvatskog nogometnog reprezentativca.
- Pokrenuta građanska inicijativa „U ime obitelji" s ciljem ustavne definicije braka kao zajednice muškarca i žene.
- Pokrenuta inicijativa „Ne damo naše AUTOCESTE!" s ciljem sprječavanja monetizacije hrvatskih autocesta.
- Nekolicina hrvatskih branitelja započinje prosvjed u šatoru u Savskoj ulici.

U tablici 4.1 prikazani su podaci o broju članaka s kojih su preuzeti komentari, prosječnom broju komentara po članku te ukupnom broju preuzetih komentara za svaki od događaja. Događaji su numerirani rednim brojevima od 1 do 4 redom kako su gore navedeni.

Rečeno je da su odabrani događaji izazvali burnu reakciju ostavivši raskol među stavovima hrvatskih građana, a činjenica da članci u prosjeku generiraju više stotina korisničkih komentara sasvim sigurno idu tome u prilog.

	1. događaj	2. događaj	3. događaj	4. događaj
Broj članaka	16	12	3	12
Ukupan broj komentara	5145	6458	736	10057
Prosječan broj komentara	321.56	538.17	245.3	838.10

Tablica 4.1: Preuzeti podaci

4.2. Označavanje komentara

Da bismo istrenirali naš klasifikator temeljen na metodama *nadziranog* strojnog učenja, najprije moramo pripremiti skup primjera za treniranje. Na tom je skupu ukupno radilo osmero označivača čiji je zadatak bio odrediti stavove iz personaliziranog skupa komentara.

Prije samog označavanja, s označivačima je održan kratak sastanak s ciljem boljeg upoznavanja sa zadatkom te su im podijeljene upute za označavanje iz *dodatka A*. Također, prije prelaska na pravi skup podataka, označivači su morali proći kroz kalibracijski skup podataka s ciljem usklađivanja standarda pri raspoznavanju stavova te boljeg razumijevanja problema.

4.2.1. Kalibracija

U procesu kalibracije, svaki je označivač označio kalibracijski skup komentara koji se sastojao od ukupno 40 komentara. Svaki je događaj bio jednako zastupljen u kalibracijskom skupu, odnosno, kalibracijski skup sadrži po 10 komentara vezanih uz svaki događaj. Osim bolje označenog skupa za treniranje, proces kalibracije nam daje do znanja koliko je zapravo identifikacija stavova u korisničkim komentarima za ljude težak problem, a također i pomaže označivačima da se bolje upoznaju s tematikom odabranih događaja i opisom posla koji je pred njima. Postignemo li automatskom klasifikacijom točnost blisku onoj slaganju označivača, možemo zaključiti da klasifikator radi optimalno.

Tablica 4.2 za svaki par označivača prikazuje broj komentara iz kalibracijskog skupa oko čijih se stavova te dvije osobe *ne slažu*. Iz tih podataka možemo zaključiti da, odaberemo li nasumično dva označivača, vjerojatnost da će se oko nekog komentara složiti iznosi 70.1%. Pregledavši glavne razloge pogrešne identifikacije stavova, identificirali smo sljedeća dva, vrlo lako ispravljiva problema:

- Označivač je pogrešno shvatio značenja oznaka ”ZA!” i ”PROTIV!” na način da su komentari koji emitiraju stav ”ZA!” označeni kao ”PROTIV!” i

obrnuto.

– Označivač nije primijetio da tekst komentara unutar `<ref>` oznaka predstavlja citat iz nekog prijašnjeg komentara na koji se ovaj referencira.

Po završetku kalibracijskog procesa, označivači su upoznati sa spomenutim pogreškama i postotkom neslaganja. Na sva je ostala neslaganja u kalibracijskom skupu teško utjecati. Ponekad je tanka linija između neutralnog komentara i onog koji skreće s teme pa je određen postotak neslaganja razuman. Također, valja naglasiti kako ovaj tip pogreške ne bi smio imati velik utjecaj na konačne rezultate budući da analizu argumenata provodimo nad komentarima koji su klasificirani isključivo kao 'ZA!" ili "PROTIV!".

	Antea	Filip	Frano	IvanP	IvanS	Ivna	Kaja	Paula
Antea	0	15	12	15	9	14	12	12
Filip	15	0	13	14	14	10	14	14
Frano	12	13	0	13	9	13	10	11
IvanP	15	14	13	0	10	12	10	13
IvanS	9	14	9	10	0	12	7	8
Ivna	14	10	13	12	12	0	12	14
Kaja	12	14	10	10	7	12	0	8
Paula	12	14	12	13	8	14	8	0

Tablica 4.2: Matrica neslaganja u kalibracijskom skupu

U tablici 4.3 vidljiva je raspodjela izrečenih stavova na skupu označenih podataka. Napominjemo da ovi podaci najvjerojatnije ne prikazuju pravu raspodijelu stavova nad cijelim skupom komentara. Razlog tome je što su komentari birani ručno kako bismo stavili naglasak na one komentare u kojima izrečeni stav i nije najjasniji. Tablica nam također daje grubu naznaku o donjoj teoretskoj granici uspješnosti klasifikatora, odnosno, zahtjevamo da naš klasifikator radi bolje od primitivnog klasifikatora (engl. *baseline*) koji će sve komentare svrstati u većinsku klasu.

	skretanje s teme	neutralno	ZA!	PROTIV!
ukupno komentara	12	6	7	15
udio u skupu	30%	15%	17.5%	37.5%

Tablica 4.3: Razdioba izrečenih stavova iz kalibracijskog skupa

4.2.2. Označavanje skupova za treniranje i testiranje

Unija skupa za treniranje i skupa za testiranje sastoji se od ukupno 2920 korisničkih komentara pri čemu je svaki od događaja zastupljen uzorkom od 730 komentara. Skup je ravnomjerno podjeljen na osmero označivača s time da presjek personaliziranih skupova svih označivača iznosi 40 komentara, veličine koja odgovara kalibracijskom skupu. Temeljem tog presjeka vrednovat ćemo uspješnost kalibracijskog procesa te odrediti gornju teoretsku granicu uspješnosti automatskog klasifikatora. U tablici 4.4 prikazana su neslaganja među označivačima na jednak način kao u prethodnom poglavlju.

Iz tablice slijedi da vjerojatnost slaganja dvaju nasumično odabranih označivača oko izrečenog stava nekog komentara iznosi 75%. Ova brojka predstavlja gornju teoretsku granicu uspješnosti klasifikatora, odnosno, ne možemo očekivati da će automatska klasifikacija biti preciznije. Prema očekivanjima, zabilježeno je poboljšanje u postoku slaganja između označivača u iznosu od 4.9%.

	Antea	Filip	Frano	IvanP	IvanS	Ivna	Kaja	Paula
Antea	0	7	12	13	11	9	10	11
Filip	7	0	11	11	7	9	11	9
Frano	12	11	0	8	7	11	9	12
IvanP	13	11	8	0	8	12	12	12
IvanS	11	7	7	8	0	9	10	10
Ivna	9	9	11	12	9	0	11	10
Kaja	10	11	9	12	10	11	0	8
Paula	11	9	12	12	10	10	8	0

Tablica 4.4: Matrica neslaganja u zajedničkom presjeku skupa za testiranje

U tablicama 4.5 i 4.6 vidljive su razdiobe izrečenih stavova u označenom skupu komentara te zajedničkom presjeku personaliziranih skupova osmero označivača. Prema očekivanju, razdiobe prikazane tablicama znatno se razlikuju od razdiobe stavova u kalibracijskom setu prikazane u tablici 4.3. Zajednički presjek svih označivača generiran je nasumično pa su razdiobe u donjim tablicama dosta slične. Naravno, zbog razlike iznosa dva reda veličine između broja komentara u skupovima ovakva su odstupanja razumna. Razdioba izrečenih stavova iz čitavog označenog skupa komentara vrlo je relevantan pokazatalj razdiobe stavova među svim preuzetim komentarima. Vidljivo je, i pomalo iznenađujuće, da pribpribližno 60% komentara s web–portala *index.hr* nije vezan uz temu o kojoj se raspravlja.

	skretanje s teme	neutralno	ZA!	PROTIV!
ukupno komentara	27	4	2	7
udio u skupu	67.5%	10%	5%	17.5%

Tablica 4.5: Razdioba izrečenih stavova iz zajedničkog presjeka.

	skretanje s teme	neutralno	ZA!	PROTIV!
ukupno komentara	1683	476	227	534
udio u skupu	57.64%	16.27%	7.80%	18.29%

Tablica 4.6: Razdioba izrečenih stavova iz označenog skupa komentara.

Skup označenih komentara nadalje je podjeljen na skup za treniranje i skup za testiranje u omjeru $7:3$. Skup za treniranje, kao što sama riječ kaže, služi da bismo istrenirali naš model, dok pomoću skupa za testiranje određujemo ispravnost rada našeg modela. Da bismo osigurali podjednaka svojsva obaju skupova, raspodjela je napravljena potpuno nasumično te su na oba skupa osigurane podjednake distribucije događaja na koje se odnose komentari i distribucije stavova koje emitiraju. Konačno, broj komentara u skupu za treniranje iznosi 2044, dok broj komentara u skupu za testiranje iznosi 876.

4.3. Klasifikacija stavova u korisničkim komentarima na internetu

4.3.1. Treniranje i odabir modela

Kao što je prije navedeno, izgradnji modela pristupili smo na dva načina:

1. Četveroklasnim klasificiranjem
2. Kaskadom binarnog klasifikatora između *off–topic* komentara i ostalih klasa te troklasnog klasifikatora između preostalih klasa

U te je svrhe potrebno izgraditi tri temeljna klasifikatora

1. Binarni klasifikator između klase *off–topic* i preostalih triju klasa
2. Troklasni klasifikator između klase neutralnih komentara, klase komentara koji emitiraju stav *ZA!* i klase komentara koji emitiraju stav *PROTIV!*

3. Četveroklasni klasifikator između svih klasa.

Za svaki temeljni klasifikator optimalno su određeni parametri za linearni, odnosno nelinearni klasifikator te je točnost tako istreniranih klasifikatora provjerena na skupu za treniranje, odnosno skupu za testiranje. Za svaki je klasifikator određena i donja granica točnosti koja odgovara točnosti naivnog klasifikatorskog algoritma koji sve komentare stavlja u većinsku klasu. Točnost klasifikatora na nekom skupu primjera modelirana je udjelom ispravno klasificiranih primjera u ukupnom skupu primjera.

	1. klasifikator	2. klasifikator	3. klasifikator
donja granica točnosti (engl. *baseline*)	57.64%	44.40%	57.64%
parametar linearnog klasifikatora (C)	0.5	0.5	2
parametar nelinearnog klasifikatora (C)	2048	8192	2048
parametar nelinearnog klasifikatora (γ)	$1.22 \cdot 10^{-4}$	$3.05 \cdot 10^{-5}$	$4.8 \cdot 10^{-4}$
točnost linearnog klasifikator (trening)	87.17%	87.99%	93.96%
točnost nelinearnog klasifikator (trening)	87.43%	96.73%	94.39%
točnost linearnog klasifikator (test)	75.49%	45.27%	58.75%
točnost nelinearnog klasifikator (test)	74.16%	41.94%	58.86%

Tablica 4.7: Razdioba izrečenih stavova iz označenog skupa komentara.

Pogled na tablicu 4.7 daje nam do znanja da je klasifikacijski problem najbolje odraditi četveroklasnim klasifikatorom s nelinearnom jezgrom. Unatoč tome, u svim se ostalim slučajevima linearni klasifikator pokazao boljim, no ne i znatno boljim odabirom od nelinearnog klasifikatora. Primijetit ćemo također kako je točnost svih elementarnih klasifikatora, izuzev troklasnog nelinearnog, iznad donje granice točnosti. Posebno je impresivan rezultat linearnog binarnog klasifikatora čija je točnost na testnom skupu za čak 17.85% iznad donje granice točnosti. Razlog tome jest taj što binarni klasifikator klasificira komentare s obzirom na tematiku što je dosta lakše od klasifikacije s obzirom na sentiment. Također, vektor značajki bogatiji je značajkama koje upućuju na tematske karakteristike komentara od značajki koje otkrivaju sentiment. Konačno, rad nastavljamo četveroklasnim klasifikatorom s nelinearnom jezgrom koji je na testnom skupu pokazao poboljšanje od 2.22% u odnosu na naivan algoritam.

Primijetimo također dosta veliku razliku u točnostima na skupu za treniranje u odnosu na skup za testiranje. Ovo je pokazatelj da se naš model previše

prilagodio skupu za treniranje (engl. *overfitting*) i nije uspio izvući dovoljno generalne zaključke za klasificiranje. Neke od metoda kojima bismo mogli smanjiti ovu razliku su:

- Korištenje manjeg broja značajki
- Povećavanjem skupa za treniranje
- Povećavanjem regularizacije

Modifikacije značajki i povećavanje skupa za treniranje bile su vremenski prezahtjevne radnje u okviru ovog rada, ali su sasvim sigurno među najvažnijim metodama poboljšanja koje bismo mogli provesti u idućim iteracijama. Povećanje regularizacije svodi se na promjenu parametra C kojeg smo propisno odabrali koristeći odgovarajuću skriptu iz `libSVM` biblioteke pa smatramo da velika razlika u točnosti najvjerojatnije nije posljedica lošeg izbora parametara.

4.3.2. Evaluacijske mjere

Vrednovanje klasifikatora započinjemo izgradnjom matrice zabune (engl. *confusion matrix*) u kojoj element i–tog retka i j–tog stupca odgovara broju komentara koji su klasificirani klasom i, a u stvarnosti pripadaju klasi j. Pretpostavimo najprije da se radi o binarnom klasifikatoru koji klasificira primjere na pozitivne i negativne. Matrica zabune je tada

$$M_z = \begin{bmatrix} T_p & F_p \\ F_n & T_n \end{bmatrix},$$

a njeni elementi poprimaju sljedeća značenja:

- T_p – broj točno klasificiranih pozitivnih primjera (engl. *true positive*)
- F_p – broj pogrešno klasificiranih pozitivnih primjera (engl. *false positive*)
- F_n – broj pogrešno klasificiranih negativnih primjera (engl. *false negative*)
- T_n – broj točno klasificiranih negativnih primjera (engl. *true negative*)

Mjeru točnosti iz prethodnog poglavlja sada možemo formalno izraziti kao

$$Acc = \frac{T_p + T_n}{T_p + T_n + F_p + F_n}$$

Uz točnost, definirat ćemo i evaluacijsku mjeru preciznosti (engl. *precision*) kao udio točno klasificiranih primjera u skupu pozitivno klasificiranih primjera. Formalno, preciznost ćemo izraziti kao

$$P = \frac{T_p}{T_p + F_p}$$

Osim točnosti i preciznosti, koristimo i mjeru odziva (engl. *recall*) koja označava udio točno klasificiranih primjera u skupu *svih* pozitivnih primjera. Formula za izračunavanje odziva glasi

$$R = \frac{T_p}{T_p + T_n}$$

Konačno, odnos preciznosti i odziva mjerimo F–mjerom (engl. *F–measure*) koju najčešće računamo kao harmonijsku sredinu preciznosti i odziva. Odnosno,

$$F = \frac{2}{\frac{1}{P} + \frac{1}{R}} = \frac{2PR}{P + R}$$

U općem slučaju možemo parametrom β pridodati veću važnost preciznosti ili odzivu na način da F–mjeru definiramo kao

$$F_\beta = \frac{(1 + \beta^2)PR}{\beta^2 P + R}$$

Dakako, preciznost naglašavamo uz $\beta < 1$, dok odziv naglašavamo uz $\beta > 1$. U idealnom slučaju (savršen klasifikator), sve bi navedene mjere poprimile vrijednost 1, odnosno 100%.

Budući da smo se odlučili na višeklasno klasificiranje, valjalo bi sve navedene mjere definirati u slučaju n–klasnog klasifikatora. Matrica zabune dimenzija je $n \times n$, a definicija njenih elemenata ostaje ista. Za i–tu klasu C_i definiramo vrijednosti T_{p_i}, F_{p_i}, F_{n_i}, T_{n_i} na sljedeći način:

- T_{p_i} := i–ti element glavne dijagonale (broj točno klasificiranih primjera koji pripadaju klasi i)
- F_{p_i} := zbroj nedijagonalnih elemenata i–tog retka (broj pogrešno klasificiranih primjera u klasu i)
- F_{n_i} := zbroj nedijagonalnih elemenata i–tog stupca (broj pogrešno klasificiranih primjera koji u stvarnosti pripadaju klasi i)
- $T_{n_i} := N - T_{p_i} - F_{p_i} - F_{n_i}$ (preostali točno klasificirani primjeri)

Preciznost, odziv i F–mjeru klase C_i računamo po analogiji na binarni klasifikator. Dakle, vrijede formule:

$$P_i = \frac{T_{p_i}}{T_{p_i} + F_{p_i}}$$

,

$$R_i = \frac{T_{p_i}}{T_{p_i} + T_{n_i}}$$

,

$$F_i = \frac{2}{\frac{1}{P_i} + \frac{1}{R_i}} = \frac{2P_iR_i}{P_i + R_i}$$

Konačno, preciznost i odziv cijelog sustava računamo kao

$$P = \frac{\sum_{i=1}^{n} T_{p_i}}{\sum_{i=1}^{n} T_{p_i} + \sum_{i=1}^{n} F_{p_i}}$$

$$R = \frac{\sum_{i=1}^{n} T_{p_i}}{\sum_{i=1}^{n} T_{p_i} + \sum_{i=1}^{n} F_{n_i}}$$

dok F–mjera ostaje ista kao kod binarnog klasifikatora.

Matrica zabune na skupu za testiranje našeg odabranog četveroklasnog klasifikatora je

$$M_z = \begin{bmatrix} 429 & 23 & 7 & 41 \\ 59 & 31 & 18 & 25 \\ 27 & 13 & 10 & 11 \\ 87 & 40 & 9 & 46 \end{bmatrix}$$

U tablici 4.8 nalaze se dobivene vrijednosti evaluacijskih mjera za svaku od klasa kao i za cjelokupni sustav. Klase su numerirane brojevima od 1 do 4 i to redom off–topic, neutralan komentar, ZA! i PROTIV!.

	klasa 1	klasa 2	klasa 3	klasa 4	ukupno
Preciznost (P)	85.8%	23.3%	16.4%	25.3%	58.9%
Odziv (R)	67.9%	4.4%	1.3%	6.9%	18.5%
F–mjera (F)	75.8%	7.5%	2.3%	10.9%	28.2%

Tablica 4.8: Vrijednosti evaluacijskih mjera za svaku klasu

Da bismo uspješno interpretirali mjere preciznosti i odziva neke klase, moramo se najprije zapitati koje klase smatramo relevantnima za naš problem. Odnosno, je li nam bitnija uspješna klasifikacija komentara koji skreće s teme ili, primjerice, komentara koji emitira stav *ZA!*. Budući da je konačan cilj klasifikacije zapravo isfiltrirati one komentare koji emitiraju stav, najvažniji podaci iz tablice 4.8 odnose se na mjere preciznosti i odziva klasa 3 i 4. Preciznosti na tim mjestima nam govore da će od svih komentara koje smo klasificirali kao *ZA!* svega 16.4% biti ispravno, dok će od svih komentara koje smo klasificirali kao *PROTIV!* ispravno biti klasificirano 25.3% komentara. Još nam lošije vijesti nosi odziv na temelju kojeg zaključujemo da će od svih komentara iz testnog skupa za koje označivači smatraju da emitira stav *ZA!* svega 1.3% biti točno klasificirano, a od komentara za koje označivači smatraju da emitira stav *PROTIV!* točno će biti klasificirano 6.9% komentara. Ovi su rezultati doista poražavajući jer uvelike

narušavaju svojstva dobre klasifikacije na kojima smo temeljili ideju iza metode grupiranja. Naizgled velika preciznost cijelog sustava posljedica je izuzetno visoke preciznosti klase 1, no taj je podatak mnogo manje relevantan od prethodnih.

4.3.3. Poboljšanja

Sudeći prema dobivenim rezultatima i interpretaciji evaluacijskih mjera, naš model vapi za poboljšanjima. Već smo napomenuli kako bi korištenje manjeg broja značajki uz treniranje modela na većem broju primjera moglo poboljšati preveliku prilagođenost modela skupu za treniranje. Ova su poboljšanja izvediva i ne zahtjevaju veće zahvate izuzev dodatno utrošenog vremena. Svejedno, ne možemo se oteti dojmu da nam za znatno bolje rezultate treba nešto više.

Glavni je problem vjerojatno niska kvaliteta značajki. Odnosno, korištene značajke previše su tematski orijentirane i ne pružaju mnogo informacija glede sentimenta. Neke od značajki koje bi mogle pomoći vezane su uz izgradnju stabla komentara pa bismo mogli, primjerice, kao značajke koristiti

- dubinu komentara u lancu diskursa
- veličinu podstabla komentara (količina diskrusa koju generira komentar)
- stav koji emitira komentar roditeljskog čvora
- broj diskursnih lanaca koje generira odgovarajući događaj

Mogli bismo također pokušati sastaviti niz jezičnih fraza hrvatskoga jezika koje se često koriste pri argumentaciji pa značajkama pridodati i vektor argumentacijskih fraza. Nažalost, pri odabiru značajki dosta smo ograničeni jezičnom kvalitetom komentara pa su značajke temeljene na parsiranju teksta gotovo nezamislive.

Unatoč lošim rezultatima smatramo da je odabrani pristup problemu, u pogledu nadziranog strojnog učenja metodom potpornih vektora, sasvim opravdan i ispravan.

4.4. Analiza argumenata nad korisničkim komentarima na internetu

Ovaj smo problem, prisjetimo se, odlučili riješiti metodom Markovljevog grupiranja. Pretpostavlja se da su podaci dobro klasificirani te da će većina korisnika svoje stavove argumentirano braniti. Markovljevim grupiranjem na grupe smo

podijelili ukupno osam grafova, po jedan za svaki od dva moguća stava u svakoj od četiri teme. Broj grupa koje je algoritam izbacio vidljiv je u tablici 4.9, a događaji su navedeni kronološki kao i u poglavlju o klasifikaciji stavova.

	događaj 1	događaj 2	događaj 3	događaj 4
ZA!	35	31	5	38
PROTIV!	69	75	10	135

Tablica 4.9: Veličine grupa dobivene Markovljevim grupiranjem

Odmah je vidljivo da broj grupa sasvim sigurno ne odgovara broju argumenata koji brane taj stav, no izbacimo li grupe koje sadrže vrlo malen broj komentara u odnosu populaciju dobivamo realniju sliku prikazanu u tablici 4.10

	događaj 1	događaj 2	događaj 3	događaj 4
ZA!	7	15	2	23
PROTIV!	7	4	1	4

Tablica 4.10: Veličine grupa nakon izbacivanja svih grupa koje sadrže manje od 10 komentara

Bacimo li pogled na neku grupu komentara, prvo što ćemo primijetiti su pogreške pri klasificiranju. Primjerice, grupa koja odgovara najčešćem argumentu koji koriste pristaše inicijative „U ime obitelji" sadrži komentare koji očito zastupaju suprotan stav poput

- „*jadna li si srednjevjekovna rvatska*" (nakon pobjede na referendumu)
- „*uvjerljivo najviše pedofila, dakle ljudi koji spolno zlostavljaju djecu, ima u obitelji; uglavnom se radi o zlostavljanim djevojčicama od strane očeva, očuha, ujaka, stričeva i slično..dakle, točno unutar onih idiličnih obitelji koje vi jedino i priznajete, djeca i njihovi heteroseksualni roditelji..*"
- „*Moj ludi profesor iz srednje je jednom rekao da tko prizna da je p**** ima 5 za kraj godine(nitko nije htio priznat), i da svi trebamo podržavat p**** i samo neka ih bude što više*"
- „*Radije bih bio siroče nego dijete u obitelji Markić!i*"

Uz ove nedostatke, ipak većina komentara zastupa očekivani stav. Unatoč tome, mišljenja smo da dobivene grupe ne predstavljaju komentare koji jednakom argumentacijom brane odgovarajući stav. Štoviše, jako malen broj komentara

grupe uopće sadrži neki oblik argumentacije što nam predstavlja velik problem jer ruši važnu pretpostavku s kojom smo krenuli u rješavanje ovog problema. Ipak, vidljivo je da sam postupak Markovljevog grupiranja radi očekivano, odnosno grupira komentare u kojima se pojavljuju određene ključne riječi ili fraze. Primjerice, jedna grupa komentara vezana uz skandiranje Josipa Šimunića pretežito sadrži komentare koji u sebi imaju sporan usklik „za dom, spremni". Ostale grupe taj usklik ne sadrže.

U konačnici, smatramo da Markovljevo grupiranje nije dovoljno dobar alat za uspješnu analizu argumenata nad korisničkim komentarima, što potvrđuje i činjenica da se u te svrhe najčešće koriste neke druge metode koje, zbog prirode problema, nismo mogli primijeniti na naš skup podataka.

4.5. Poboljšanja

U okvirima Markovljeva grupiranja, jedino na što možemo utjecati jest mjera sličnosti između komentara koju smo definirali kao kosinus kuta između vektora riječi dvaju komentara. Nadali smo se da će, uz pretpostavku da je većina stavova argumentirana, ova mjera sličnosti biti dovoljna da se grupe formiraju oko najčešće korištenih argumenata. S obzirom na prirodu skupa podataka i algoritma Markovljevog grupiranja, mišljenja smo da nije moguće promjenom mjere sličnosti postići znatno bolje rezultate. Odnosno, da bismo postigli željeni rezultat, nužno je problemu pristupiti koristeći drugačiji model.

Treba istaknuti da, zbog kroničnog nedostatka argumentacije u komentarima koji emitiraju stav, ovaj problem postaje vrlo zahtjevan. Za rješavanje problema vrlo bismo se vjerojatno trebali upustiti u identifikaciju argumentiranih komentara koristeći neka saznanja o načinima strukturiranja argumenata. Za to nam je potrebno opsežno znanje o često korištenim strukturama argumenata u tekstovima pisanim hrvatskim jezikom i potrebna nam je sposobnost parsiranja rečenica u korisničkim komentarima. Kada bismo u tome uspjeli, i dalje radimo pod pretpostavkom da korisnici svoje argumente propisno strukturiraju, što je najvjerojatnije daleko od istine.

5. Zaključak

Cilj ovoga rada bio je klasificirati korisničke komentare na internetu prema stavovima koje emitiraju te analizom komentara koji zastupaju pojedini stav odrediti glavne argumente kojima korisnici taj stav podupiru.

Problemu klasifikacije korisničkih komentara pristupili smo izgradnjom klasifikatora koristeći metodu potpornih vektora. Model koji je davao najbolje rezultate ostvario je poboljšanje u odnosu na primitivni algoritam u iznosu od 2.22%. Unatoč siromašnom poboljšanju, mišljenja smo da je pristup problemu dobar te da će opisana poboljšanja s naglaskom na kvalitetniji izbor značajki uvelike poboljšati konačan rezultat. Zavidan uspjeh u obliku poboljšanja od 17.85% u odnosu na primitivni algoritam od strane binarnog klasifikatora koji filtrira komentare čiji je sadržaj u okviru teme, daje nam do znanja da su odabrane značajke pogodne za klasifikaciju s obzirom na tematiku.

Analizu najčešćih argumenata kojima korisnici brane neki stav proveli smo metodom Markovljevog grupiranja. Nažalost, grupe formirane ovom metodom sasvim sigurno ne odgovaraju grupiranju komentara s obzirom na njihovu argumentaciju. Smatramo da poboljšanja ne treba tražiti u okviru Markovljevog grupiranja, već da je pravi put ka boljim rezultatima promjena modela. Iz srodne literature možemo naslutiti kako promjena modela iziskuje pogodniji skup podataka pa je možda odabir alternativnog izvora korisničkih stavova prvi pravi korak prema uspjehu.

Na kraju, ističemo kako je ovaj, i inače vrlo zahtjevan problem, dodatno otežan niskom jezičnom kvalitetom komentara zajedno s činjenicom da su komentari pisani hrvatskim jezikom. U svrhu uspješnijeg rješavanja srodnih problema u domeni hrvatskoga jezika, smatramo da naglasak treba staviti na razvoj alata koji će nam, koristeći niz jezičnih specifičnosti, omogućiti bezbolniju analizu tekstova pisanih hrvatskim jezikom.

Literatura

Sylvain Brohee i Jacques Van Helden. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC bioinformatics*, 7(1):488, 2006.

Chih-Chung Chang i Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Corinna Cortes i Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Kai-Bo Duan i S Sathiya Keerthi. Which is the best multiclass svm method? an empirical study. U *Multiple Classifier Systems*, stranice 278–285. Springer, 2005.

David Meyer i FH Technikum Wien. Support vector machines. *The Interface to libsvm in package e1071*, 2014.

Raquel Mochales Palau i Marie-Francine Moens. Argumentation mining: the detection, classification and structure of arguments in text. U *Proceedings of the 12th international conference on artificial intelligence and law*, stranice 98–107. ACM, 2009.

Bo Pang, Lillian Lee, i Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. U *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, stranice 79–86. Association for Computational Linguistics, 2002.

Stijin van Dongen. Graph clustering by flow simulation. U *PhD thesis*, stranice 1–96. University of Utrechtm, 2009.

Dodatak A
Upute za označivače

A.1. Motivacija

Korisnički komentari na internetu vrijedan su izvor informacija za analizu stavova i mišljenja ljudi o događajima i njihovim protagonistima, političkim odlukama i političkim subjektima, ideološkim pitanjima, kontroverznim temama itd. Računalna analiza stavova razmjerno je novo područje u okviru analize prirodnog jezika koje se bavi automatskom klasifikacijom i analizom stavova izraženih u tekstu, primjerice korisničkih komentara na internetu. Riječ je o posebno izazovnom zadatku, dodatno otežanom zbog vrlo niske jezične kvalitete korisničkih komentara.
Kako bismo uspješno istrenirali naš klasifikator stavova potrebno je dio komentara *ručno* označiti. Ovaj dokument namijenjen je označivačima kao vodilja kroz posao označavanja komentara.

A.2. Opis posla

Svaki će označivač označiti 400 komentara na članke vezane uz teme koje su prodrmale hrvatsku javnost. Svi su komentari preuzeti s web–portala Index, a trebali bi se osvrtati na jednu od sljedećih tema:

- Monetizacija autocesta u RH
- Josip Šimunić – Za dom spreman!
- Ustavna zabrana istospolnih brakova u RH
- *Šatoraši* iz Savske ulice

Zadatak svakog označivača jest svakom komentaru pridodati jednu od sljedećih oznaka:

- *off–topic* komentar
- neutralan komentar
- ZA!
- PROTIV!

Temeljem iskustvene analize, očekivano vrijeme izvršavanja zadatka nalazi se negdje u intervalu od $[2, 5]$ sati. Važno je napomenuti da nije nužnno sve komentare označiti odjednom. Također, nužno je da svaki označivač svoj dio posla obavi **samostalno**.

A.3. Oznake

Odlomci koji slijede sadrže neke opće smjernice vezane uz pojedinu oznaku. Značenja oznaka koja su usko vezana uz pojedinu temu nalaze se u nastavku teksta, zasad se držimo generalne ideje iza pojedine oznake.

A.3.1. *Off–topic* komentar

Svaki komentar koji u sebi **ne sadrži nikakve naznake** dane teme smatramo *off–topic komentarom*. Komentari ovakvog tipa najčešće su *ad hominem* svađe između korisnika, pokušaji ispravljanja jezičnih grešaka u tekstu članka ili nekog drugog komentara, promocije raznih proizvoda ili usluga, itd.
Valja istaknuti da neke komentare, koji su inače usko vezani uz članak, po definiciji svrstavamo u ovu skupinu. Primjerice, direktne kritike autoru članka, mišeljenja vezana uz osobe ili lokacije koje se spominju u članku i reference na srodne članke smatramo *off–topic* komentarima.

A.3.2. Neutralan komentar

Svaki komentar koji u sebi **ne sadrži nikakve naznake** korisnikovog stava glede zadane teme smatramo *neutralnim komentarom*. Ovakvi komentari najčešće su informativnog tipa, primjerice, definicije nekih pojmova usko vezanih uz temu. Primijetite kako je korisnikov stav nužan sastojak komentara koji nije neutralan. Stoga, komentari koji sadrže analize stavova autora članka ili nekog drugog korisnika, ali ne sadrže stavove autora komentara, neutralni su komentari.

A.3.3. ZA! i PROTIV!

Odabrane teme nam dopuštaju binarnu klasifikaciju korisničkih stavova na one koji podupiru, odnosno protive se, temeljnom stajalištu koje nameće pojedina tema. Komentari koji većinskim dijelom, ili u potpunosti, podupiru temeljno stajalište neke teme, označavamo kao *ZA!*, dok komentare koji se većinskim dijelom, ili u potpunosti, protive temeljnom stajalištu neke teme, označavamo kao *PROTIV!*.

A.4. Teme

Odlomci koji slijede sadrže detaljniji opis svake od navednih tema, značenje oznaka u okviru pojedine teme te dodatne smjernice i primjere koji pomažu pri usklađivanju procjena označivača. Također, za svaku temu definirali smo njenu temeljno stajalište, odnosno stajalište koje zastupaju komentari koje valja označiti oznakom *ZA!*. Dakako, polaritet temeljnog stajališta dogovorno je određen te ni na koji način ne odražava stavove autora ovog teksta.
Važna napomena: Od označivača se očekuje da poslu pristupe kao ljudi, a ne kao računalni sustav. Tokom označavanja u obzir uzmite jezične figure (poput ironije ili sarkazma), stil pisanja, ton komentara, Vaš opći dojam, i sl. Odnosno, eksplicitno zastupanje nekih stavova u komentaru je dovoljan, ali *nije nužan* uvjet da biste taj komentar označili oznakama *ZA!* ili *PROTIV!*.

A.4.1. Monetizacija autocesta

U sklopu akcije *„NE damo naše AUTOCESTE!"* skupljali su se potpisi za raspisivanje referenduma. Mnogi su građani, putem internet komentara, javnosti dali do znanja jesu li *ZA* ili *PROTIV* monetizacije hrvatskih autocesta.
Dakle, u sklopu ove teme željeli bismo klasificirati i analizirati korisničke stavove o monetizaciji hrvatskih autocesta pri čemu temeljno stajalište teme definiramo kao: **"ZA monetizaciju autocesta u RH"**.

A.4.2. Josip Šimunić – Za dom spreman!

Nakon trijumfa nad reprezentacijom Islanda, hrvatski nogometaš australskog podrijetla, Josip Šimunić, uzeo je mikrofon i skandirao – *„Za dom spremni!"*. Budući da je ovaj usklik usko vezan uz Ustaški pokret za vrijeme drugog svjetskog rata,

Šimunićev postupak izazvao je zavidnu reakciju hrvatske javnosti. Temeljno stajalište ove teme definirali smo kao: **"Šimunićevo skandiranje isključivo je odraz domoljublja"**.

A.4.3. Ustavna zabrana istospolnih brakova u RH

Građanska udruga – *„U ime obitelji"* pokrenula je inicijativu koja je u konačnici rezultirala unošenjem odredbe u Ustav Republike Hrvatske kojom se brak definira kao zajednica muškarca i žene. Odnosno, time je Ustavom zabranjeno sklapanje istospolnih brakova. Temeljno stajalište ove teme definirali smo kao: **"Brak je isključivo životna zajednica između muškarca i žene"**.

A.4.4. *Šatoraši* iz Savske ulice

Neprazni podskup hrvatskih branitelja dane provodi u šatoru koji se nalazi u Savskoj ulici 66. Ovim prosvjedom branitelji žele ispraviti nanešene nepravde te se izboriti za niz zakonskih povlastica. Jasno, ovaj je prosvjed izazvao svakakve reakcije hrvatske javnosti. Dio nacije se slaže sa postupcima branitelja, dok se dio nacije tomu protivi (prosvjed protiv prosvjeda). Temeljno stajalište ove teme definirali smo kao: **"Prosvjed hrvatskih branitelja razuman je i opravdan."**. Preciznije, komentari koji se slažu sa postupcima i zahtjevima bratnitelja potrebno je označiti kao *ZA!*, dok je komentare koji se tomu protive potrebno označiti kao *PROTIV!*.

A.5. Primjeri

Upozorenje: Neki od primjera koji slijede sadrže eksplicitne i neprimjerene izraze. Ovakvi primjeri prvenstveno su uvršteni zbog njihove učestalosti u skupu preuzetih komentara.

A.5.1. Monetizacija autocesta

Komentar: „Inicijativa je dobra i protivim se prodaji autocesta, ali ne mogu potpisati jer ju podržava Teodor Celakoski."
Oznaka: *PROTIV!*
Obrazloženje: Korisnik se eksplicitno protivi monetizaciji autocesta u RH.

Komentar: „AC su sam samo jedna gigantska betonska crna rupa za usisavanje teško zarađenog novca svih nas i što ih prije uvalimo drugome to bolje."
Oznaka: *ZA!*
Obrazloženje: Korisnik eksplicitno podržava monetizaciju autocesta u RH.

A.5.2. Josip Šimunić – Za dom spreman!

Komentar: „Ustasa,pod hitno mu zabraniti odlazak na svetsko prvenstvo!"
Oznaka: *PROTIV!*
Obrazloženje: Korisnik se očevidno ne slaže sa tvrdnjom da je Šimunićev postupak isključio odraz domoljublja.

Komentar: „Utakmica gotova, Joe pita publiku jeste spremni za polazak kući?, a navijači kažu da su spremni.. u čemu je problem?"
Oznaka: *neutralan komentar*
Obrazloženje: Donekle duhovit komentar u kojem korisnik nije ostavio svoje prave stavove.

Komentar: „Volim Moderatora03 :)"
Oznaka: *off–topic komentar*
Obrazloženje: Korisnik iskazuje privrženost prema *Moderatoru03*.

A.5.3. Ustavna zabrana istospolnih brakova u RH

Komentar: „Većinu ljudi boli kurac i to je OK. Ja sam mišljenja da sve treba bolit kurac za ovo."
Oznaka: *neutralan komentar*
Obrazloženje: Očevidno je da je komentar neutralan.

Komentar: „Nesto malo tijekom kuhanja slusala to zenu i ukratko moj dojam: klimatkerica, isfrustrirana, ruzna, gluba i nedojebana baba"
Oznaka: *PROTIV!*
Obrazloženje: Iako nije eksplicitno navedeno, ostavlja se dojam kako se komentar odnosi na Željku Markić. Zbog navedenih epiteta, autor ovog teksta je pod dojmom kako se autorica teksta protivi stavovima Željke Markić.

Klasifikacija i analiza stavova u korisničkim komentarima na internetu

Sažetak

Korisnički komentari na internetu vrijedan su izvor informacija za analizu stavova i mišljenja ljudi o događajima i njihovim protagonistima, političkim odlukama i političkim subjektima, ideološkim pitanjima, kontroverznim temama itd. Računalna analiza stavova razmjerno je novo područje u okviru analize prirodnog jezika koje se bavi automatskom klasifikacijom i analizom stavova izraženih u tekstu, primjerice korisničkih komentara na internetu. Riječ je o posebno izazovnom zadatku, dodatno otežanom zbog vrlo niske jezične kvalitete korisničkih komentara. Problemu klasifikacije u okviru završnog rada pristupili smo metodom potpornih vektora, dok smo analizi korisničkih komentara, odnosno, rudarenju argumenata pristupili algoritmom Markovljevog grupiranja.

Ključne riječi: obrada prirodnog jezika, strojno učenje, umjetna inteligencija, metoda potpornih vektora, Markovljevo grupiranje, hrvatski jezik, internet komentar

Stance classification and analysis in online user comments

Abstract

On–line user comments are a valuable source of information on public's opinions regarding certain events and their protagonists, political decisions and subjects, ideological questions, controversial subjects etc. Computational stance analysis is a relatively new field of natural language processing which deals with automatic classification and analysis of stance expressed in texts, such as user generated on–line comments. In this paper, the classification part of the problem was solved using support vector machines, whereas the analysis part was tackled by Markov clustering algorithm.

Keywords: natural language processing, machine learning, artificial intelligence, support vector machine, Markov clustering algorithm, Croatian language, on–line comment