Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Brina Škoda

Rudarjenje razpoloženja na komentarjih rtvslo.si

diplomsko delo

univerzitetni študijski program prve stopnje
računalništvo in informatika

Mentor: doc. dr. Dejan Lavbič

Ljubljana 2013

Besedilo je oblikovano z urejevalnikom besedil LaTeX.

Univerza *v Ljubljani*
Fakulteta *za računalništvo in informatiko*

Tržaška 25
1001 Ljubljana, Slovenija
telefon: 01 476 84 11
01 476 83 87
faks: 01 426 46 47
01 476 87 11
www.fri.uni-lj.si
e-mail: dekanat@fri.uni-lj.si

Št. naloge: 00133/2013
Datum: 15.04.2013

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Kandidat: **BRINA ŠKODA**

Naslov: **RUDARJENJE RAZPOLOŽENJA NA KOMENTARJIH RTVSLO.SI**
SENTIMENT ANALYSIS ON RTVSLO.SI USER COMMENTS

Vrsta naloge: Diplomsko delo univerzitetnega študija prve stopnje

Tematika naloge:

Analiza mnenja oz. razpoloženja je področje, ki v zadnjem času pridobiva na pomembnosti. Predvsem je bilo veliko napredka opaziti ob pojavu Web 2.0 virov podatkov (blogi, novice, družbena omrežja ipd.), na podlagi katerih je takšno analizo možno izvajati na različnih problemskih domenah. Ker je takšna analiza močno odvisna od jezika vira, ki ga analiziramo, je večina uspešnih implementacij v angleškem jeziku, medtem ko v slovenskem jeziku takšnih rešitev skoraj ni. V okviru diplomske naloge zato na podlagi komentarjev novic uporabnikov na slovenskem spletnem portalu rtvslo.si analizirajte njihovo razpoloženje z različnimi pristopi (strojno učenje, metode na osnovi slovarjev, leksikalni pristopi ipd.) in nato uporabite tistega, ki se bo izkazal za najbolj primernega. Rezultate predstavite v obliki delujočega prototipa v poljubnem programskem jeziku. Podajte tudi smernice za nadaljnji razvoj oz. na katerem področju so lahko rezultati vaše analize koristni.

Mentor:

doc. dr. Dejan Lavbič

Fakulteta za računalništvo in informatiko
2
UNIVERZA V LJUBLJANI

Dekan:

prof. dr. Nikolaj Zimic

Izjava o avtorstvu diplomskega dela

Spodaj podpisana Brina Škoda, z vpisno številko **63090108**, sem avtor diplomskega dela z naslovom:

Rudarjenje razpoloženja na komentarjih rtvslo.si

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelala samostojno pod mentorstvom doc. dr. Dejana Lavbiča,
- so elektronska oblika diplomskega dela, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko diplomskega dela
- soglašam z javno objavo elektronske oblike diplomskega dela v zbirki "Dela FRI".

V Ljubljani, dne 7. septembra 2013 Podpis avtorja:

Zahvaljujem se mentorju doc. dr. Dejanu Lavbiču za vse nasvete, strokovno pomoč in korekten odnos pri izdelavi diplomske naloge. Prav tako se zahvaljujem vsem ostalim, ki ste kakorkoli pripomogli pri dosegu tega cilja.

Kazalo

Povzetek

Glavni namen diplomske naloge je bil bolj podrobno spoznati področje rudarjenja razpoloženja na besedilih v slovenskem jeziku. Pridobljeno znanje smo nato uporabili v povezavi s praktičnim primerom. Za podatke smo vzeli novice in komentarje portala rtvslo.si, ki je eden izmed dveh najbolj popularnih portalov pri nas. Poizkusili smo različne algoritme, s katerimi smo poskušali komentarje s čim večjo stopnjo pravilnosti klasificirati kot pozitivne oziroma negativne, nato pa smo svoje rezultate privzeli za pravilne in nad njimi naredili še nekaj rudarjenj. Na začetku smo si zadali osnovne cilje, ki smo jih nato z pridobljenim znanjem nadgrajevali. Na tem področju v slovenskem jeziku še ni bilo veliko narejenega, zato smo naleteli na nemalo težav. Najprej smo raziskali teorijo in jo povezano predstavili, tako je nadaljno delo bralcu bolj razumljivo. Predstavili smo uporabljene metode in ugotovitve, in prikazali rezultate.

Ključne besede

rudarjenje razpoloženja, rudarjenje mnenj, podarkovno rudarjenje, strojno učenje, obdelava naravnega jezika, metoda podpornih vektorjev, naivni Bayes, maksimalna entropija, klasifikacija, zlati standard, značilke

Abstract

The main purpose of this thesis was getting to know the field of opinion mining on texts in Slovene language in greater detail. We then used the acquired knowledge in connection with a practical example. For the data we used the news and comments from the rtvslo.si news portal, which is one of the two most popular portals in Slovenia. We tested various algorithms trying to classify comments as positive or negative with the greatest possible accuracy, then we assumed our results as correct and did some more opinion mining on them. At the beginning, we have set ourselves the basic goals which we then upgraded with acquired knowledge. There has not been much done in this field in Slovene language, consequently we came across a great deal of problems. Firstly, we studied the theory and presented it, so the further work is more understandable to the reader. We presented the used methods, findings and results.

Keywords

opinion mining, sentiment analysis, natural language processing, data mining, machine learning, SVM, maximum entropy, naive Bayes, classification, gold standard, features

Poglavje 1

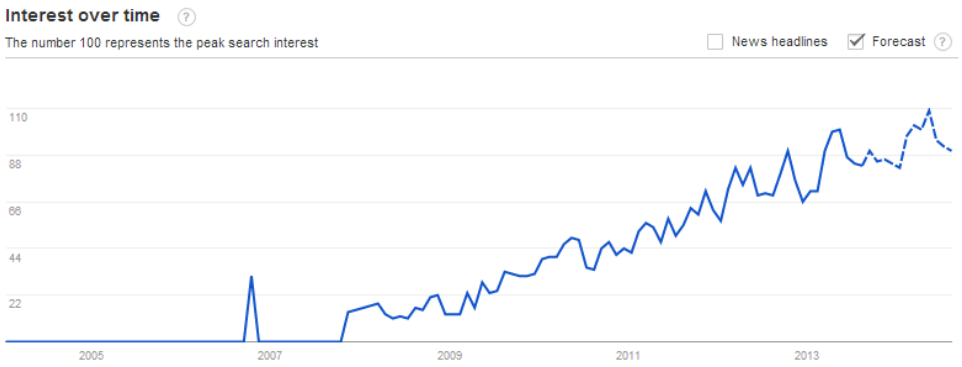
Uvod

Pri vsakodnevnem odločanju se ljudje zelo zanašamo na mnenja drugih. Zanimajo nas priporočila hotelov, da na podlagi le teh izberemo najboljšega, prijatelja vprašamo za najljubšo restavracijo v mestu, ali za fitnes, ki ga obiskuje. Mnenja o stvareh in dogodkih ves čas krožijo, sploh ker jih lahko delimo z milijoni ljudi na svetovnem spletu. In ker si ljudje mnenja tako vztrajno izmenjujemo, jih je ogromno, celo preveč, zato jih želimo znati računalniško opisati, povzeti in znati povedati kaj je splošno mnenje ljudi o neki stvari.

Rudarjenje mnenj (opinion mining) ali analiza mnenja/razpoloženja (sentiment analysis) je področje, ki je trenutno zelo aktualno. Ta veja umetne inteligence je v razmahu že vse odkar uporabljamo Web 2.0. Ena prvih člankov na to temo [3, 4] sta iz leta 2002 in obravnavata polariteto komentarjev filmov in produktov. Na to temo je v nadaljevanju sledilo še ogromno člankov in raziskav. Bolj podrobno je ta tema opisana v knjigi Boa Panga in Lilliana Leea [8]. Danes, ko se skoraj vsak dan razvije kakšno novo družabno omrežje ali mikroblog, pa je ta tema postala še bolj priljubljena kot kdajkoli prej. Rudarjenje mnenj je postalo zelo pomembno za vsako podjetje, ki hoče preživeti na dolgi rok. Ker je konkurenca ogromna, smo lahko v prednosti le, če točno vemo kaj si naše stranke želijo in kaj potrebujejo.

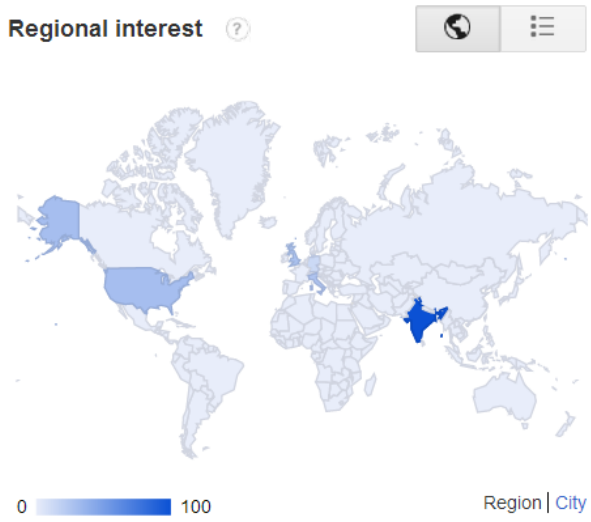
Na spletni strani Google trendi [34] smo si pogledali, kakšen trend ima

pojem *sentiment analysis*, oziroma analiza razpoloženja v svetu, kar lahko vidite na grafu 1.1. Prikazano je, kako narašča iskanje pojma v brskalniku Google.



Slika 1.1: Prikaz naraščanja popularnosti tematike

Na sliki 1.2 je prikaz popularnosti te tematike glede na države, kjer lahko vidimo, da je to področje najbolj popularno v Indiji. V Evropi je poleg Italije in Anglije položaj za zdaj precej slabša.



Slika 1.2: Prikaz naraščanja popularnosti tematike

Na začetku se je analiza razpoloženja osredotočala bolj na eno domeno, na primer oceno filma, danes pa se osredotoča na razvoj metod, ki jih lahko uporabimo za več domen, saj vnaprej pogosto ne vemo, za katero področje gre. Ker je težko pričakovati, da bo algoritem, ustvarjen za identifikacijo čustev o nekem filmu, pravilno deloval tudi za restavracije ali politiko, je to področje velik izziv. Razvijalci si delo poskušajo olajšati, tako lahko na primer pri mikroblogih srečamo dodane znake višaja (#), s katerim uporabniki sami opredelijo domeno in tako rudarjenje poenostavijo. Tudi ostala družabna omrežja so začela prevzemati ta označevanja.

Slika 1.3: Primer komentarja z označbami na mikroblogu Twitter.

Analiza razpoloženja nam lahko koristi tudi pri blogih in novicah. V ta namen smo jo uporabili tudi pri naši diplomski nalogi. To področje lahko razdelimo na analizo razpoloženja in predvidevanje razpoloženja. Del, kjer obravnavamo le subjektivne komentarje, je klasična analiza razpoloženja, če pa zraven vključimo še objektivno novico in jo obravnavamo v povezavi s subjektivnimi komentarji, temu pravimo predvidevanje razpoloženja, ki pa ga pa v naši diplomski nalogi nismo obravnavali.

Do našega cilja lahko pri analizi razpoloženja pridemo na več načinov. Uporabimo lahko samo strojno učenje, v mnogih primerih pa je učinkovito, če uporabimo strojno učenje v povezavi z metodami na osnovi slovarjev ali leksikalnimi pristopi [1]. Te temeljijo na vrednotenju besed v slovarju. Večina

raziskav strojnega učenja je osredotočenih na nadzorovane tehnike strojnega učenja v povezavi s statističnimi pristopi, seveda pa najdemo tudi raziskave, kjer so uporabljene nenadzorovane tehnike učenja. V naslednjem poglavju predstavimo metode, ki smo jih nato uporabili na praktičnem delu diplome, oziroma na kratko predstavimo kaj to področje zajema. V tretjem poglavju opišemo kako smo klasificirali komentarje portala rtvslo.si in kakšni so bili naši rezultati. Nato pa najboljšo metodo uporabimo za zadnji del, kar predstavimo v četrtem poglavju, kjer rudarimo po podatkih in prikažemo nekaj zanimivih povezav. V zadnjem, petem poglavju zaključimo svoje delo z nekaj mislimi o tematiki in pa idejami, kaj se bi še lahko naredilo.

1.1 Opredelitev pojmov

Rudarjenje mnenj [8] kot pojem, se je v povezavi z računalništvom pojavil prvič v dokumentu od Dave et. al. [31] leta 2003, definirano kot: *"Glede na ocenjeno besedilo dokumentov D, ki vsebuje mnenja, oziroma čustva o predmetu (oseba, organizacija, produkt ...), je namen rudarjenja mnenj pridobiti sestavne dele v vsakem od dokumentov D komentiranega predmeta in ugotoviti, ali so komentarji pozitivni, negativni ali nevtralni."* Idealno orodje za rudarjenje mnenj bi glede na Dave-ov opis: *"obdelalo niz rezultatov za določeno postavko, ustvaril seznam lastnosti produkta (kakovost, značilnosti ...) in združil mnenja o vsakem od njih (slabo, mešano, dobro).* Dokument je bil objavljen na konferenci svetovnega spleta (WWW conference) leta 2003, to področje pa je nato postajalo vse bolj priljubljeno. Esuli in Sebastian [32] sta leta 2005 definirala pojem kot: *"Nedavna disciplina na razpotju med iskanjem informacij in računalniškim jezikoslovjem, ki se ne ukvarja s temo o kateri govori dokument, pač pa z mnenjem, ki ga izraža."* Iz računalniškega vidika sta prav tako leta 2005, Kim in Hovy [33] pojem definirala *"kot štirikratnega (tema, nosilec, trditev, občutki,) v katerem nosilec verjame, da je vrednost trditve o temi glede na prepričanje in se v večini primerov povezuje z občutki - dobrimi ali slabimi."* V Kolikor občutki zadevajo

neko temo, jih avtorja opredeljujeta kot čustva. Trdita, da vsako mnenje ne izraža čustev.

Še nekaj definicij pojmov, ki jih je postavil Bing Liu leta 2010:

- "**Mnenje** o značilki f ima pozitivno ali negativno stališče, odnos, čustvo ali presojo na f tistega, ki izraža mnenje."
- "**Lastnik mnenja** je oseba ali organizacija, ki izraža mnenje."
- "**Eksplicitno mnenje** o značilki F, je mnenje eksplicitno izraženo v subjektivnem stavku."
- "**Implicitno mnenje** o značilki F, je mnenje implicitno izraženo v objektivnem stavku."
- "**Stavek, ki izraža mnenje** , je stavek, ki eksplicitno ali implicitno izraža pozitivna ali negativna mnenja. Lahko je objektiven ali subjektiven."
- "**Čustva** so naši subjektivni občutki in naše misli."

V povezavi s to tematiko, se tako v literaturi, kot v naši diplomski nalogi veliko pojavlja še pojem polarnost komentarja, kjer se raziskovalci na tem področju strinjajo z naslednjo definicijo od Pang-a in Lee-a iz leta 2008 [8], in sicer: *"Binarna klasifikacijsko oziroma razvrstitveno naloga označevanja dokumenta, ki v splošnem izraža pozitivno oziroma negativno mnenje, imenujemo polariteta stavka ali polariteta klasifikacije."*, ki smo ji v našem primeru rekli kar polariteta komentarja.

Poglavje 2

Tehnike rudarjenja mnenj

Rudarjenje mnenj [8], ki ga lahko prištevamo k obliki podatkovnega rudarjenja, nam odpira nove možnosti raziskovanja jezika. Pogosto se na tem področju uporabljajo metode strojnega učenja, s katerimi se želi proces analize avtomatizirati. Kljub temu, da je na voljo veliko strukturiranega besedila, primankuje ustrezno velikih korpusov z označenimi primeri, ki bi jih lahko uporabili za nadzorovano učenje pri analizi razpoloženja. Še večji izziv predstavljajo sintaktične in semantične posebnosti posameznih jezikov, saj ima vsak jezik svoja pravila in izjeme. Čeprav nek algoritem na enem jeziku deluje dobro, ni nujno, da bo deloval tudi na drugem. Večina prosto dostopnih korpusov je v angleščini, čeprav se raziskovalci trudijo in jih počasi delajo tudi za druge jezike. Leta 1921 je lingvist Edward Sapir rekel: *"Na žalost, ali na srečo, noben jezik ni tiransko konsistenten. Vse slovnice puščajo."*

In ravno zato, ker imajo vse slovnice toliko posebnosti, je to področje še posebno zanimivo in predstavlja velik izziv. Ne poznamo popolnega algoritma, ki bi ga lahko splošno uporabili na vseh primerih, ampak ima vsak svoje posebnosti, ki se lahko v različnih situacijah izkažejo za dobre.

Pojma strojno učenje in podatkovno rudarjenje sta velikokrat pomešana, saj sta med seboj zelo prepletena. Rudarjenje mnenj je podveja podatkovnega rudarjenja, zato je primerno, da na začetku bralcu predstavimo osnovne algoritme in pristope, ki se uporabljajo pri strojnem učenju in podatkovnem

rudarjenju, saj se ti posledično uporabljajo tudi pri rudarjenju mnenj.

Strojno učenje je področje umetne inteligence. Pomeni pridobivanje znanj na podlagi izkušenj. Leta 1959, je Arthur Samuel definiral strojno učenje [35] kot: *"Študijsko področje, ki daje računalniku možnost učenja, ne da bi bil eksplicitno programiran."* Klasične metode strojnega učenja temeljijo na ekspertnih pravilih, ki jih določijo strokovnjaki za določeno področje. Cena pa je tudi njihova največja ovira, saj so ekspertni sistemi izjemno dragi. Ker je podatkov navadno ogromno, se morajo algortimi strojnega učenja učiti le iz malega števila podatkov, nato pa pridobljeno znanje dobro uporabiti še na ostalih. Tudi podatkovno rudarjenje uporablja podobne metode, le z drugim namenom.

Podatkovno rudarjenje se osredotoči na odkrivanje neznanih lastnosti s pomočjo iskanja vzorcev in asociacij v nekem dokumentu, ki je lahko besedilo, slika, zvok,... Omogoča nam torej avtomatizirano iskanje informacij, ki jih bomo v svojem primeru uporabili na besedilih. Podatkovno rudarjenje uporablja metode strojnega učenja, ki jih razdelimo na nadzorovane metode (supervised methods) in nenadzorovane metode (unsupervised methods).

2.1 Nadzorovane metode

Nadzorovano učenje [2] je princip strojnega učenja za kreiranje funkcije na podlagi učne množice vzorcev. Naloga sistema je, da na pravilen način posploši znanje, ki ga dobi iz učne množice. Pri postopku učenja z učiteljem [14] moramo najprej določiti področje uporabe, s tem pa tudi vrsto podatkov, ki jih bo sistem obdeloval. Nato izberemo podatke, iz katerih se bo sistem učil. Podatki, na katerih se bo sistem učil, morajo biti značilni za končno področje delovanja. Pri nadzorovanih metodah je zelo pomemben del določitev značilk (features), kar je poudarjeno praktično v vseh raziskavah, povezanih s to tematiko. Na značilke se bomo osredotočili v naslednjem poglavju. Izbrati moramo še primeren algoritem, njegove parametre pa postavimo optimalno glede na rezultate testiranj na testni množici.

Nadzorovane metode razdelimo na regresijo, logične relacije, asociacije in klasifikacijo oziroma uvrščanje. V našem primeru bodo zadoščale klasifikacijske metode, zato se bomo osredotočili nanje.

Klasifikacija

je ena izmed najbolj uporabljenih metod na področju strojnega učenja [14]. Naloga klasifikatorja je za objekt, opisan z množico značilk, določiti, kateremu izmed možnih razredov pripada. Zato, da lahko klasifikator določi razred, mora imeti predstavljeno diskretno funkcijo, ki preslika prostor atributov v razred. Lahko je podana vnaprej, ali pa je naučena iz podatkov, ki so rešeni problemi iz preteklosti. V diplomski nalogi smo preizkusili naslednje algoritme, ki jih bomo v nadaljevanju tudi bolj podrobno predstavili: naivni Bayesov klasifikator, klasifikator po metodi najbližjih sosedov, klasifikator po metodi podpornih vektorjev (SVM) in maksimalna entropija.

2.1.1 Naivni Bayes

Naivni Bayesov klasifikator [38] je statistična metoda, ki temelji na Bayesovem izreku s predpostavko pogojne neodvisnosti vrednosti različnih atributov pri danem razredu. Naivni Bayesov klasifikator predpostavlja, da je prisotnost ali odsotnost določene vrednosti atributa neodvisna od prisotnosti ali odsotnosti katerekoli vrednosti drugega atributa.

Kljub njegovi naivnosti lahko rečemo, da klasifikator deluje presenetljivo dobro. Njegova dobra lastnost je, da deluje dobro tako za majhne, kot tudi za velike učne množice. Naivni Bayesov klasifikator lahko neposredno uporabimo pri diskretnih atributih, pri zveznih pa moramo atribute najprej diskretizirati. Dobro se obnaša tudi, kadar predpostavka o pogojni neodvisnosti ne drži popolnoma, če pa imamo med atributi popolne odvisnosti, pa ta klasifikator ni dobra izbira. Metoda predpostavlja pogojno neodvisnost vrednosti različnih atributov pri danem razredu, zato je pri domenah, kjer med razredi obstajajo odvisnosti, manj uspešna.

Osnovna formula naivnega Bayesovega klasifikatorja, izpeljana s pomočjo Bayesovega pravila, je:

$$P(c|v_1, v_2, ...v_n) = P(c) \prod_{i=1}^{n} \frac{P(c|v_i)}{P(c)},$$

kjer c predstavlja razred, v_i pa lastnost. Naivni Bayesov klasifikator izračuna verjetnost za vsak razred, na koncu primer klasificira v razred z najvišjo verjetnostjo [37].

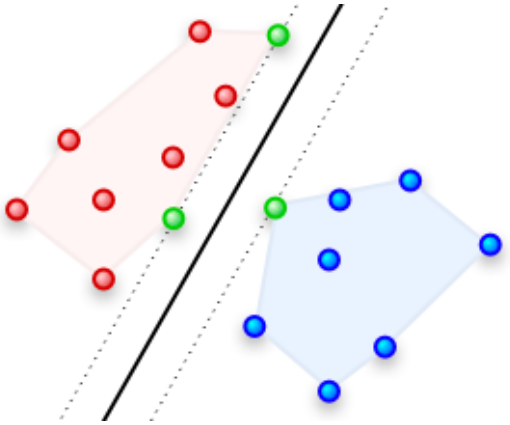
Obstaja več različnih implementacij naivnega Bayesa [15]. Razlikujejo se v predstavitvi dokumenta. Dokument oziroma besedilo lahko predstavimo kot binarni vektor pojavitve besed, kjer vsaka komponenta besedila pripada določeni besedi. Vrednost komponente je ena, če se določena beseda v besedilu pojavi, sicer pa je nič. Takšni predstavitvi rečemo Bernoullijev model naivnega Bayesa [15]. Drugi, ki se za tekstovno uporabo uporablja še pogosteje, pa je multinominal naivni Bayes [15]. Pri tem modelu vsaka komponenta predstavlja število, kolikokrat se posamezna beseda pojavi v besedilu. V večini primerov naj bi dal multinominalni naivni Bayesov model boljše rezultate. To se je izkazalo za resnično tudi v našem primeru, kar predstavimo v nadaljevanju.

2.1.2 Metoda podpornih vektorjev (SVM)

Metoda podpornih vektorjev [5], ki se pogosto uporablja kar pod kratico SVM (Support vector machine), je med najbolj uspešnimi metodami klasifikacije in regresije. V osnovi so klasifikacijski SVM namenjeni razločevanju dveh razredov med seboj. Če imamo razredov več, postopek ponovimo za vsak razred, ki ga želimo ločiti od ostalih. Izhodišče te metode je množica učnih predmetov, za katere vemo, kateremu razredu pripadajo. Vsak predmet predstavimo z vektorjem v vektorskem prostoru, naloga metode pa je poiskati

v tem n-dimenzionalnem prostoru hiperravnino, ki ločuje predmete (točke) iz različnih razredov. Pri postavljanju hiperravnine ni potrebno upoštevati vseh učnih vektorjev. Vektorji, ki so daleč od hiperravnine ne vplivajo na njeno lego. Njena lega je odvisna od njej najbljižnjih vektorjev. Te vektorje imenujemo podporni vektorji in so na sliki 2.1 označeni z zeleno. Razdalji hiperravnine od podpornih vektorjev pravimo rob (margin).

Prednost SVM je točnost in neprilagajanje učni množici, slabost pa računska zahtevnost.



Slika 2.1: Prikaz SVM metode [30].

Vsako hiperravnino lahko zapišemo kot množico točk X, ki zadošča:

$$w \cdot x - b = 0,$$

kjer $\cdot$ pomeni skalarni produkt, w je normala in je pravokotna na hiperravnino, parameter $\frac{b}{||w||}$ pa določa odmik hiperravnine od izhodišča vzdolž normale. Izbrati je potrebno takšen w in b, da maksimizirata rob.

Ta metoda se je tudi v našem primeru izkazala za najbolj uspešno, zato smo njene rezultate na naših komentarjih privzeli za pravilne in z njihovo

pomočjo izvedli nadaljno rudarjenje.

2.1.3 Maksimalna entropija ali logistična regresija

Model maksimalne entropije za procesiranje naravnega jezika [6] je enak modelu logistične regresije v statistiki in strojnem učenju. Metoda je podobna linearni regresiji, saj prav tako išče model, ki se najbolj prilega učnim podatkom, le da se logistična regresija uporablja za diskretne vrednosti. Naloga modela torej je, da iz naučenega modela napove, kolikšna je verjetnost, da vhodnim atributom x priprada na primer vrednost razreda ena. Iščemo model, katerega parametri določajo obliko funkcije, ki se čim bolje prilega učnim podatkom. Funkcija napake meri točnost modela, in če si predstavljamo, da imamo neko verjetnost $P(Y = 1|X = x)$, kot funkcijo z vhodnim parametrom x, je vprašanje, koliko more biti x, da bo verjetnost največja.

2.1.4 Metoda najbližjih sosedov

Metoda najbližjih sosedov (K nearest neighbor - kNN) [7], je še ena od enostavnih metod, ki jih je lahko razumeti in v praksi dajejo dobre rezultate. Algoritem kNN gleda na podatke, ki so testni in učni primeri, kot na točke v prostoru. Učni primeri so že razdeljeni v razrede, ostale primere pa metoda označi glede na bližino ostalih razredov. Metodi moramo določiti, koliko najbližjih sosedov za nek iskan primer upošteva, kar je vhodni parameter. Prav tako moramo določiti kako računa razdaljo, kjer je na voljo več možnosti. Največkrat se uporablja Evlidska razdalja [9], poznamo pa še Hammingovo razdaljo [10], Minkowsko [11] in ostale. Metoda, glede na najbližje razrede, izbere, kateremu razredu pripada iskan podatek.

2.2 Nenadzorovane metode

Nenadzorovano učenje [2] oziroma učenje brez učitelja temelji na podatkih, kjer so podane le značilke vzorcev, ne pa tudi razredi, v katere spadajo.

Število razredov je določeno vnaprej ali pa jih določi učni algoritem sam. Njegova naloga torej je, da določi neko število koherentnih razredov, ki so si med seboj čimbolj podobni. Slaba stran tega principa je, da je v veliko primerih težko ugotoviti, ali je izbor števila razredov res optimalen. V primeru tekstovnih dokumentov lahko dokumente preberemo in tako ocenimo pravilnost. Podobnost med vzorci je odvisna od izbrane vrste razdalje, ki zelo vpliva na rezultat razvrščanja.

Najbolj znane nenadzorovane metode so metode gručenja (clustering). Imajo dolgo zgodovino, uporabljajo pa se skoraj na vseh področjih. V medicini, sociologiji, biologiji in mnogih drugih. Z gručami lahko urejamo slike, veliko pa se uporabljajo pri rudarjenju podatkov, kjer imamo neko zbirko besedil in jih želimo organizirati v skupine. Kakovost gručenja je torej najbolj odvisna od izbranega algoritma in funkcije razdalje.

Gručenje razdelimo na hierarhično (hierarchical clustering) [12] in razdelitveno (partitional clustering) [2]. Hierarhični algoritmi uporabljajo za uspešno gručenje predhodno ustvarjene gruče, razdelitveni algoritmi pa določajo število gruč. Med najbolj znane algoritme gručenja, ki je iz skupine razdelitvenih, spada K-means.

Med nenadzorovane metode poleg metod gručenja spadajo tudi leksikalni pristopi, ki jih bomo opisali v nadaljevanju.

2.2.1 Metoda voditeljev (K-means)

Algoritem K-means [13], ki je najbolj znana metoda gručenja, dobi na vhodu učne primere in želeno število gruč. Najprej naključno izbere k središč in vsak učni primer z uporabo evklidske razdalje pripiše središču, ki mu je najbližje. Primeri, ki so zbrani okrog središča, predstavljajo gručo. Nato v naslednjem koraku algoritem s povprečenjem vrednosti atributov središča izračuna na novo, pripisovanje učnih primerov pa ponovi. Postopek iterativno ponavlja, dokler središča gruč ne ostanejo nespremenjena, oziroma dokler ne doseže neko omejeno število iteracij, če jih omejimo na začetku.

Slabost algoritma je občutljivost na začetne približke, saj lahko ostane v

lokalnem minimumu. Njegova uspešnost je odvisna od izbrane mere podobnosti. Za zmanjšanje občutljivosti na ekstreme je dobro uporabiti manhattansko razdaljo in računati mediano namesto povprečij, to verzijo algoritma imenujemo K-medians [13].

2.2.2 Leksikalni pristop

Leksikalni pristopi prav tako spadajo pod nenadzorovane metode, ampak ne strojnega učenja. Tukaj analiza razpoloženja temelji na analizi posameznih besed oziroma fraz. Imamo tri vrste leksikalnih pristopov [2]. Takšnega, ki temelji na slovarju (dictionary-based approach), ročni pristop (manual approach) in takšnega, ki temelji na korpusu.

Ročni pristop je zelo zamuden, zato se običajno ne uporablja sam, ampak v kombinaciji z avtomatiziranima pristopoma, ki temeljita na slovarju in korpusu. Ker pa avtomatizirani pristopi delajo napake, je ročni pristop za končno preverjanje zelo dobrodošel. Avtomatizirana pristopa bomo tudi podrobno predstavili, tako kot sta opisana v knjigi [2].

Pristop s slovarjem

Najbolj osnoven pristop k tej metodi je opisan v delu [18]. Najprej se vzame manjši nabor besed za katere poznamo polarnost, torej ali so pozitivne ali negativne izberemo ročno. Nato algoritem v že obstoječem slovarju išče sopomenke, oziroma protipomenke, in novo najdene besede avtomatično doda v nov slovar. To naredi vedno znova, dokler novih besed ne more več najti.

Bolj prefinjen pristop je bil predlagan v delu [16], kjer uporabljena metoda temelji na razdalji med besedami. Razdalja med besedama je določena kot dolžina najkrajše poti, ki povezuje ti dve besedi v že obstoječem slovarju. Vrednost za besedo t je bila na primer določena: $SO(t) = \frac{d(t,slabo)-d(t,dobro)}{d(dobro,slabo)}$; t je pozitiven, če je $SO(t) > 0$, sicer pa je negativen.

Leta 2010 sta Hassan in Radev [20] predstavila še en možen pristop v povezavi s slovarjem, in sicer Markov naključni vzorec hoje (Markov random walk model). Najprej sta s pomočjo slovarja zgradila sorodnostni graf

sopomenk in nadpomenk. Vrednost $h(i|S)$ je predstavljala število korakov od vozlišča i do niza vozlišč S. Predstavlja povprečno število korakov, ki ga naključni sprehajalec potrebuje, da pride do prvega $k \in S$. Dan je sklop pozitivnih S^+ in negativnih S^- besed. Torej če je $h(w|S^-) > h(w|S^+)$, potem je beseda negativna, sicer pozitivna.

Obstaja še ogromno različnih izvedb pridobivanja besed, ki so mešanica teh pristopov, nekateri so tudi kombinacija slovarjev in korpusov. Ugotovili so, da je mogoče enostavno in hitro zbrati besede in jih ovrednotiti s pomočjo slovarjev. Čeprav ima lahko naša lista besed veliko napak, lahko z ročnim preverjanjem te napake odpravimo. To je res zamudno delo, ampak je le enkratno opravilo, ki ga lahko materni govorec opravi v parih dneh.

Največja slabost je, da so opredeljene besede neodvisne od konteksta, kar pa se lahko izkaže za slab pristop. Polariteta besede je namreč mnogokrat odvisna od konteksta.

Pristop s korpusom

Za razliko od slovarjev, kjer so besede zložene po abecedi, je besedilni korpus obsežna zbirka besedila, zajeta v določenem obdobju iz množičnih medijev, knjig, interneta ali drugih besedil. Ta besedila so zajeta v strukturirani obliki, z različnimi označbami besed. Prva je osnovna oblika besede ali lema (npr. besede jagode, jagodi, jagodam imajo lemo jagoda), druga je t.i. oblikoskladenjska oznaka. Ta oznaka opisuje, v katero besedno vrsto spada beseda (samostalnik, glagol, pridevnik itd.) in, kakšne so njene lastnosti (npr. spol, število, sklon).

```
<s>
  <w msd="Sozed" lemma="vlada">Vladi</w>
  <S/>
  <w msd="Gp-ste-n" lemma="biti">je</w>
  <S/>
  <w msd="Ggdd-es" lemma="uspeti">uspelo</w>
  <S/>
  <w msd="Ggdn" lemma="rešiti">rešiti</w>
  <S/>
  <w msd="Ppnmeid" lemma="gordijski">gordijski</w>
  <S/>
  <w msd="Somei" lemma="vozel">vozel</w>
  <S/>
  <w msd="Ppnmer" lemma="plačen">plačnega</w>
  <S/>
  <w msd="Somer" lemma="sistem">sistema</w>
  <S/>
  <w msd="Dm" lemma="v">v</w>
  <S/>
  <w msd="Ppnmem" lemma="javen">javnem</w>
  <S/>
  <w msd="Somem" lemma="sektor">sektorju</w>
  <c>.</c>
</s>
```

Slika 2.2: Primer stavka v korpusu Kres.

Korpusi so namenjeni za potrebe raziskovanja jezika, ki pa je v večini primerov povezano s strojnim učenjem. Uporablja se tudi za potrebe strojnega prevajanja v računalniškem jezikuslovju. Lahko so splošni ali pa tematsko orientirani, kar pomeni, da se za izgradnjo le teh uporabi bolj ozko, glede na temo usmerjena besedila.

Nekaj korpusov, ki obstajajo v slovenskem jeziku [21]

GIGAFIDA - Korpus Gigafida je obsežna zbirka slovenskih besedil najrazličnejših zvrsti, od dnevnih časopisov, revij, do knjižnih publikacij vseh vrst (leposlovje, učbeniki, stvarna literatura), spletnih besedil, prepisov parlamentarnih govorov in podobno, vsebuje pa skoraj 1,2 milijarde besed oz. natančneje 1.187.002.502 besedil.

KRES - Korpus Kres je nastal v okviru projekta Sporazumevanje v slovenskem jeziku v letih 2008-2012, vsebuje pa skoraj 100 milijonov besed oz.

natančno 99.831.145 besed. Za razliko od korpusa gigafida, kjer je večina besedil iz časopisja (77 %), je Kres iz Gigafide vzorčeni uravnoteženi podkorpus, saj vsebuje približno isto razmerje besedil iz različnih vrst virov.

GOS - To je korpus govorjene slovenščine. Obsega okol 120 ur posnetkov pogovora v najrazličnejših situacijah iz vsakdanjega življenja.

ŠOLAR - To je korpus besedil, ki so jih učenci slovenskih osnovnih in srednjih šol samostojno tvorili pri pouku. Od večine korpusov se razlikuje v tem, da (a) besedila niso nastala na pobudo projekta, ampak predstavljajo dejansko šolsko produkcijo učencev in dijakov, (b) so jezikovni popravki, ki so označeni v korpusu realni, naredili pa so jih učitelji in ne raziskovalci. Zaradi izrabe učiteljskih popravkov za namene označevanja jezikovnih napak učencev je korpus edinstven ne samo v slovenskem, ampak tudi v evropskem oziroma svetovnem merilu.

2.3 Obdelava naravnega jezika

Do zdaj smo govorili o klasifikaciji, o iskanju polaritet besed oziroma besednih zvez, ničesar pa še nismo povedali o obdelavi jezika, ki je pravzaprav začetna stopnja vsega. Lahko imamo še tako dober algoritem, pa nam ne bo koristil, če jezika prej ne obdelamo. Uporabljene tehnike so odvisne od lastnosti, ki jih ima posamezen jezik, kar pomeni, da moramo upoštevati tudi posebnosti le tega. Za začetek si poglejmo nekaj osnovnih pojmov.

Tokenizacija je razdelitev besedila na besede oziroma pojavnice.

Segmentacija je razdelitev besedila na povedi, kjer večinoma predpostavimo, da so ločniki ločila.

Krnjenje je proces normalizacije besed, kjer besedam odstranimo predpone in pripone.

Lematizacija pomeni, da vsaki besedi pripišemo njeno osnovno obliko, oziroma nedoločnik.

Oblikoslovno označevanje pomeni, da vsaki besedi pripišemo njeno oblikoslovno oznako.

Filtriranje je metoda, ki s pomočjo seznama besed iz dokumentov odstrani besede, ki za nas nimajo vrednosti (angl. stop words).

S tako pripravljenimi dokumenti se lahko lotimo nadaljnje analize besedila. Na prvi pogled bi morda želeli upoštevati vse besede, ki so prisotne v besedilu. To bi si lahko privoščili le, če bi obravnavali le eno besedilo, do česar pa v praksi ne prihaja ravno pogosto. Kakor hitro imamo več dokumentov, moramo dokumente predstaviti drugače. V naslednjih poglavjih bomo predstavili nekaj osnovnih pristopov [22].

2.3.1 Vreča besed

Pri vreči besed [17] so besede združene v seznam s prirejeno stopnjo pomembnosti. Za določanje pomembnosti si pomagamo z vektorsko predstavitvijo, kjer je vsaka beseda predstavljena numerično. Da dobimo vse te besede, uporabimo zgoraj omenjen proces tokenizacije. Z odstranitvijo ločil in zamenjavo nekonsistentnih znakov s praznimi znaki, besedilo razbijemo v niz besed, imenovan slovar, ki ga uporabimo za nadaljnje procesiranje. Sprocesirane besede imenujemo značilke (features).

2.3.2 Določitev praga

Eden najpreprostejših načinov za določitev značilk je, da določimo minimalni in maksimalni prag frekvence [22]. Na ta način čas procesiranja zmanjšamo. Menimo namreč, da je verjetnost, da se redka beseda pojavi v novem, neznanem dokumentu, zanemarljivo majhna, zato lahko takšno besedo kar izpustimo. Prav tako predvidevamo, da so za klasifikacijo besedila neprimerne tudi zelo pogoste besede, saj se praviloma pojavljajo v dokumentih v vseh razredih.

2.3.3 TF-IDF

Drug način, poleg štetja pogostosti besed, je upoštevanje pomembnosti besede. Za to se uporablja merilo TF-IDF [19]. Pri tem merilu je frekvenci dodan še koeficient pomembnosti, ki ga imenujemo obratna pogostost besede (inverse document frequency). Ta v obratnem razmerju upošteva število enot, ki vsebujejo določeno besedo. Na ta način besedi, ki se pojavlja v veliko enotah, zmanjšamo pomen. V primeru, ko se pa beseda pojavi bolj poredko in le v malo besedilih, se ta koeficient poveča, saj predpostavimo, da je taka beseda pomembnejša. Merilo TF-IDF je pozitivno število, ki zamenja klasično frekvenco. Večja kot je vrednost, bolj pomembna je beseda za odkrivanje zakonitosti v dokumentih.

2.3.4 N-grami

Še en uspešen način obdelave značilk je s pomočjo n-gramov, s pomočjo katerih upoštevamo tudi vrstni red terminov v dokumentu. N-gram [22] je zaporedje črk, besed ali zlogov dolžine n. Pristop temelji na predpostavki, da imajo dokumenti v istih razredih podobno porazdelitev frekvenc n-gramov. Prednost te metode v primerjavi z vrečo besed je, da nam ni treba poznati jezika, v katerem je dokument napisan.

Ni dobro, če je izbran n premajhen, saj se lahko majhni zlogi oziroma malo število besed bolj pogosto pojavlja v več razredih. V praksi se najpogosteje uporablja n=4.

2.3.5 Informacijski prispevek

Namesto uteži TF-IDF lahko uporabimo kriterij informacijskega prispevka. Ta upošteva količino informacije, ki jo značilka vsebuje. Informacijski prispevek meri število bitov informacije, ki se pridobi na razred, glede na prisotnost oziroma odsotnost določenega termina v dokumentu. Pristop temelji na entropiji, izbrane pa so značilke, ki imajo informacijski prispevek višji od željenega, ki ga določimo sami.

2.4 Opis rezultatov

Ko se enkrat odločimo, kako bomo obdelali značilke in katere metode bomo uporabili za klasifikacijo, je pomembno še, da izberemo primeren opis rezultatov. Ker ti ne bodo nikoli popolnoma pravilni, je treba določiti, na kakšen način bomo njihovo kakovost merili.

	Dejanski razred	
Pričakovan	TP (Pravi pozitivni)	FP (Lažni pozitivni)
razred	FN (Lažni negativni)	TN (Pravi negativni)

TOČNOST = $\frac{TP+TF}{TP+FP+FN+TN}$; pove kakšen delež predvidevanja je bil pravilen (angl. accuracy)

NATANČNOST = $\frac{TP}{TP+FP}$; pove nam razmerje med pravilnimi izbranimi entitetami in vsemi izbranimi (angl. precision)

PRIKLIC = $\frac{TP}{TP+FN}$; pove nam razmerje med pravilnimi izbranimi in vsemi pravilnimi entitetami(angl. recall)

F VREDNOST = $\frac{2\times\text{PRIKLIC}\times\text{NATANČNOST}}{\text{PRIKLIC}+\text{NATANČNOST}}$; je harmonično povprečje natančnosti in priklica

2.5 Primer analize razpoloženja v slovenskem jeziku

Za razliko od angleškega jezika, kjer je raziskav na temo rudarjenja komentarjev ogromno, v slovenščini to področje še ni zelo razvito.

Angleščina je zelo vpletena v naše vsakdanje življenje, naše znanje angleščine pa je vse bolj kvalitetno. S prihodom podjetja Amazona in e-knjigami, ki so nam na dosegu roke, je slovenščina vse bolj izrinjena in tudi knjige pri nas že beremo v angleščini, saj je izbira veliko večja.

Za angleščino je narejenih ogromno korpusov, raznih slovarjev in knjižnic, v katerih so že kar testni primeri za angleški jezik. Ko smo se lotili tega področja, je bilo takoj jasno, da bi bilo naše delo veliko lažje in bolj učinkovito, če bi obravnavali angleški jezik. Zaradi razširjenosti raziskav besedil v angleškem jeziku in zaradi pomanjkanja oziroma nedostopnosti določenih jezikovnih orodij za slovenski jezik se tudi veliko Slovencev v svojih diplomskih, magistrskih ali doktorskih nalogah pogosto odloči za raziskavo angleškega jezika [22].

Kar nekaj nalog se je že ukvarjalo s procesiranjem slovenskega jezika [24, 25, 26]. Za naš konkretni primer pa je bolj zanimiva raziskava, ki jo je za Pop TV v času volitev 2012 naredilo podjetje System Gama v sodelovanju z Institutom Jože Štefan.

2.5.1 Volitve 2012

Prvi medij pri nas, ki je analizo mnenj v slovenščini uporabil tudi na konkretnem primeru, je bila televizija PoP TV. V času volitev so spremljali družabno omrežje Twitter, kjer so v realnem času procesirali mnenja državljanov [23].

Razvili so platformo, katera je zbirala in analizirala tvite(besedilo maksimalne dolžine 140 znakov, ki izraža mnenje o aktualnem dogodku ali osebi) o treh predsedniških kandidatih, diagrami o trenutnem položaju, pa so bili sproti prikazani na TV zaslovnih med debatami na televiziji in ves čas na portalu 24ur.com.

Rezultati so bili sporni, saj so bili v nasprotju z rezultati različnih agencij. Vse agencije so predvidile, da bo v prvem krogu zmagal takratni predsednik Danilo Türk, medtem ko je sistem na podlagi analize mnenj kazal vodstvo Boruta Pahorja. Seveda je na koncu Borut Pahor res zmagal.

Glavna ideja je bila, da se klasifikacijski model uči iz samodejno označenih podatkov. Ta pristop je drag, saj je časovno zahteven zaradi ročnega označevanja podatkov. Označili so med 20.000 in 30.000 tvitov, nad to številko se delež pravilnosti ni več izboljševal. Označeni so bili politični tviti. V primeru

Bolgarskih volitev (za katere so prav tako naredili takšen sistem), je bilo političnih tvitov pred volitvami še premalo, tako da so klasifikator testirali na splošnih in ga potem sproti prilagodili na politične.

Nevtralne tvite so pri učenju ignorirali. Pri označevanju pa so določili vrednost zanesljivosti, in če ji tvit ni ustrezal, so ga označili kot nevtralnega. S tem pristopom smo tudi mi izboljšali kakovost svojih klasifikacij.

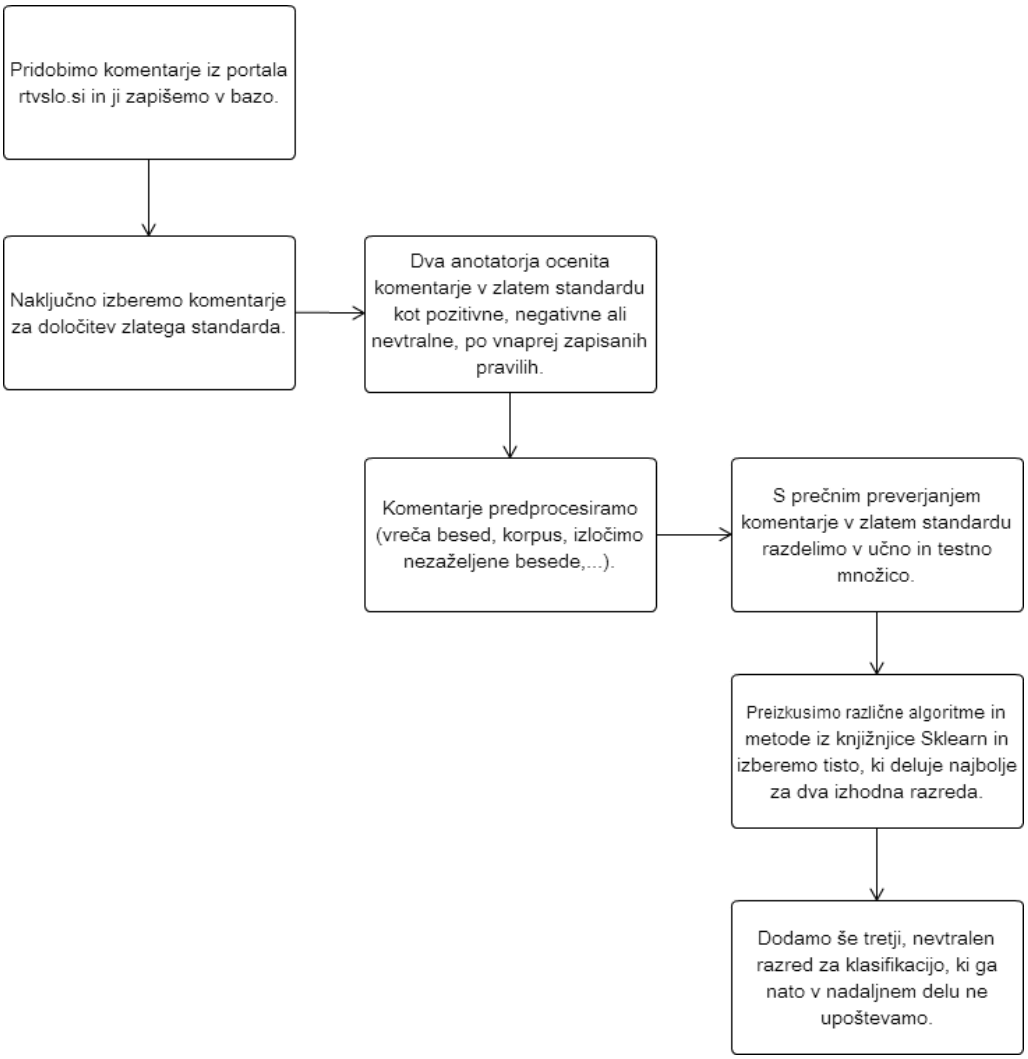
Pri klasifikaciji so se odločili za pristop s strojnim učenjem. Uporabili so metodo podpornih vektorjev, za vrednotenje pa so vzeli 10-kratno prečno preverjanje [36].

V njihovem primeru je bil ključ do dobrega klasifikatorja v kakovosti ocenjenih tvitov uporabljenih za učenje. K dobrim rezultatom je pripomoglo tudi, da so poleg razredov pozitivnih in negativnih tvitov pri učenju naredili še nevtralni razred.

Poglavje 3

Klasifikacija komentarjev portala rtvslo.si

Za implementacijo proučenega znanja iz prejšnjih poglavij smo izbrali komentarje portala rtvslo.si. Za razliko od tvitov je dolžina komentarjev praktično neomejena in ni osredotočena le na eno čustvo, saj lahko en komentar vsebuje tako pozitivno kot negativno čustvo. Tudi ni nujno, da se nanaša le na eno osebo ali dogodek, saj je že v eni novici lahko predstavljenih več ključnih dogodkov, ki so lahko predmeti komentarjev. Kot smo že omenili na začetku, so prednost tudi oznake (tag-i), ki označujejo, na kaj se tvit nanaša, česar pa v komentarjih ni. V nadaljevanju smo predstavili podatke, način kako smo se lotili obravnave, kakšne algoritme smo preizkusili in opišemo, na kakšne težave smo naleteli. Za lažjo predstavo pa prilagamo strnjen diagram poteka.



Slika 3.1: Diagram poteka klasifikacije komentarjev.

3.1 Uporabljeni pristopi

Za področje analize razpoloženja, procesiranja naravnega jezika in strojnega učenja obstaja ogromno knjižnic v vseh mogočih jezikih. Za našo diplomsko nalogo smo uporabili programski jezik Python 2.7 in knjižnico Scikit learn, oziroma sklearn [27]. Knjižnica vsebuje orodja za rudarjenje in analizo podatkov. Zgrajena je na knjižnicah NumPy, SciPy in matplotlib. Sklearn je

odprtokodna knjižnica, ki se jo lahko uporablja tudi za komercialno uporabo z BSD licenco. Ima tudi zelo dobro dokumentacijo. Vsebuje metode klasifikacije, regresije, gručenja, preprocesiranja podatkov in mnoge druge.

Za ta del raziskovanja smo uporabili metode klasifikacije in predprocesiranja podatkov. Pred vsem tem pa smo uredili še značilke in besede lematizirali.

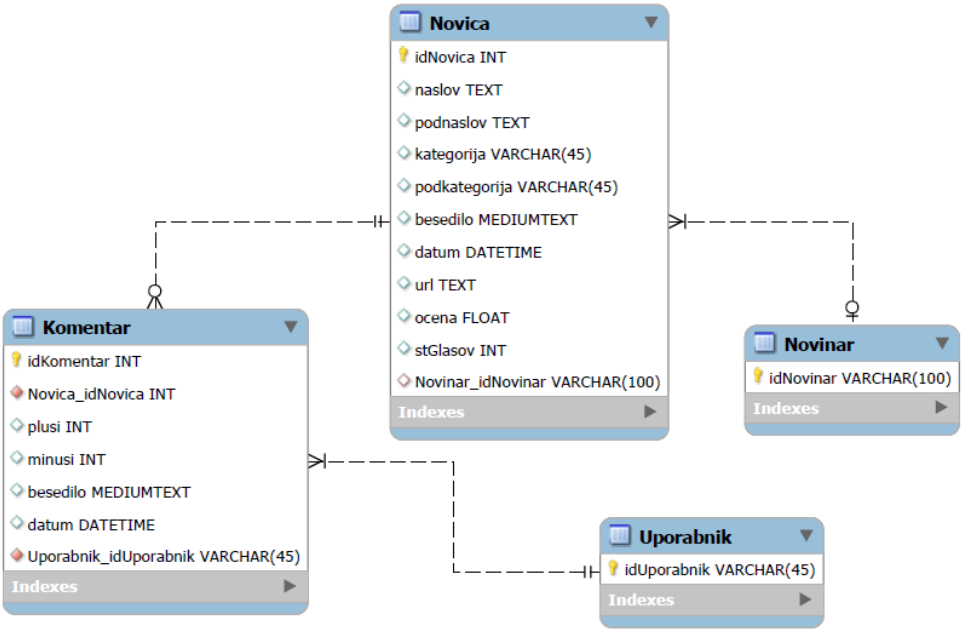
H kvalifikaciji komentarjev smo želeli pristopiti tudi na drug način, s pomočjo slovarja čustev, kjer bi bile besede označene podobno, kot je to narejeno v angleški različici [28]. Besedam v tem slovarju je določen delež pozitivnosti, negativnosti in nevtralnosti, katerih vsota je 1. Vrednosti za različne oblike besede so različne, prav tako je navedeno v kakšnem kontekstu in s katerimi okoliškimi besedami se ta najpogosteje pojavlja, saj imajo iste besede v različnih okoliščinah različne deleže pozitivnosti, negativnosti in nevtralnosti.

Na žalost pa v slovenskem jeziku takšen slovar še ne obstaja, zato smo se morali zadovoljiti le s strojnim učenjem. Izdelava takšnega slovarja je večji projekt, ki ga v sklopu diplomske naloge nismo naredili.

3.2 Pregled podatkov

Iz portala rtvslo.si smo zajeli novice in komentarje za obdobje med 12.9.2011 in 2.7.2013, kar nam je dalo skupaj 42430 novic in 1606226 komentarjev.

Podatke smo zapisali v podatkovno bazo, kjer je naš model bil videti tako:



Slika 3.2: Podatkovni model novic in komentarjev.

Ko izločimo novice brez komentarjev nam ostane še 37784 novic. Povprečno število komentarjev na novico je 42.

Na portalu je v tem času nastalo ogromno kategorij in podkategorij, mi bomo predstavili tiste, ki zajemajo 35545 novic, kar je večina vseh novic, saj smo želeli večina komentarjev združiti v neko smiselno število kategorij. Povprečno število komentarjev le na teh novicah je 38.

V kategoriji Slovenija so novice večinoma v povezavi s politiko v Sloveniji, v kategoriji svet pa je prevladujoča tema politika v tujini. Pri drugih so tematike s Slovenijo in tujino pomešane. Slovenija, svet in gospodarstvo spadata v podkategorijo novice, medtem ko so ostale samostojne kategorije. V tabeli 3.2 lahko vidimo, da je največ komentarjev na temo politike v Sloveniji, za njo pa ne zaostaja dosti kategorija šport.

Kategorija	**Število novic**
Slovenija	4611
Svet	5249
Gospodarstvo	2287
Sport	10401
Zabava	6951
Kultura	6046

Tabela 3.1: Število novic na portalu rtvslo.si glede na kategorijo

Kategorija	**Število komentarjev**
Slovenija	581510
Svet	161718
Gospodarstvo	76425
Sport	425362
Zabava	77621
Kultura	31719

Tabela 3.2: Število komentarjev na portalu rtvslo.si glede na kategorijo

3.2.1 Ustreznost komentarjev

Velika težava na področju analize razpoloženja je sarkazem, ki se v primeru komentiranja na portalih, spletnih dnevnikih, forumih se kar precej uporablja. Čeprav že obstajajo raziskave o zaznavanju sarkazma [29], smo mi v okviru svojega dela to zanemarili. To pa je vplivalo na slabši rezultat.

Na spletnih portalih je pogosto komentiranje komentarjev in ne novic. Debata se lahko hitro preusmeri v medsebojne debate, ki nimajo velike povezave z novico, pod katero so zapisani. Na rtvslo.si ni možnosti komentiranja komentarja in velja nenapisano pravilo, da uporabniki, ki komentirajo komentar določenega uporabnika, v svojem komentarju uporabijo znak afna skupaj z uporabnikovim imenom. Primer:

> @ Kaja, zakaj minus (verjetno vaš), če pa sem opozoril na nonsens v besedilu. Raje preberite in se prepričajte, preden objavite...

Slika 3.3: Primer komentarja na komentar.

Tudi te komentarje smo zanemarili. Komentarjev, ki se začnejo z znakom alfa, torej nismo niti označili, niti na njih nismo izvajali testiranja. Ker pa je to nenapisano pravilo, še vedno obstajajo komentarji, ki se ne nanašajo na novico, ampak jih na ta način ne moremo zaslediti, saj ne uporabljajo značilnih znakov.

3.3 Zlati standard

Kot smo že omenili, smo se odločili, da za učenje klasifikatorja uporabimo nadzorovane tehnike strojnega učenja, za kar smo morali izdelati zlati standard (gold standard).

Zlati standard so naključni primeri komentarjev, ki jih ročno označimo, te pa potem privzamemo za pravilno označene, ter na njih učimo izbran

klasifikator.

Na začetku smo označili 500 komentarjev v razmerju, kot smo ga prikazali v tabeli 3.3. V razmerjih po modulu smo vzeli komentarje iz različnih kategorij, ki sta jih nato ocenila dva anotatorja.

Kategorija	**Število ocenjenih komentarjev**
Slovenija	215
Svet	60
Gospodarstvo	28
Sport	157
Zabava	12
Kultura	29
SKUPAJ	500

Tabela 3.3: Število ročno ocenjenih komentarjev na portalu rtvslo.si glede na kategorijo

Komentar sta anotatorja lahko ocenila kot pozitiven, negativen ali nevtralen. Vsak anotator je ocenil iste naključno izbrane komentarje, nato pa smo za učne primere upoštevali le komentarje, v katerih sta se anotatorja strinjala, in ki niso bili nevtralni. Nevtralne komentarje smo se odločili zanemariti.
Navodila za ocenjevanje komentarjev so bila naslednja:

1. Polariteto komentarja ocenjuj vedno v kontekstu z novico. Če v komentarju zaznaš sarkazem, označi namen komentarja in ne dobesednega pomena.

2. Če je komentar glede na novico pozitiven ga označi s +, torej kot pozitivnega.

3. Če je komentar glede na novico negativen ga označi z -, torej kot negativnega.

4. Če je komentar glede na novico nevtralen; torej vsebuje stavke, ki ne izražajo pozitivnih ali negativnih mnenj, ampak so objektivni; ga označi z n - nevtralno.

5. Če se komentar nanaša na drug komentar in lahko iz njega kljub vsemu razbereš polariteto glede na novico, ga označi pozitivno ali negativno, sicer pa ga označi kot nevtralnega.

Anotatorja sta se strinjala v približno 80% (z nevtralnimi vred), ko smo pa od vsega odšteli še nevtralne komentarje nam je ostalo še 300 komentarjev. Dve tretjini komentarjev je bilo ocenjenih negativno, le ena tretjina pa pozitivno. Rezultati po kategorijah:

Kategorija	**Pozitivni**	**Negativni**	**Skupaj**
Slovenija	30	104	134
Svet	9	24	33
Gospodarstvo	5	16	21
Sport	48	43	91
Zabava	6	10	16
Kultura	2	4	6
SKUPAJ	100	201	301

Tabela 3.4: Delež pozitivnih in negativnih komentarjev na portalu rtvslo.si glede na kategorijo

Kot lahko vidimo v tabeli 3.4, so komentarji v skoraj vseh kategorijah zelo negativno usmerjeni. Tudi ko smo poiskusili, kako uspešni so algoritmi na tej množici podatkov so bili rezultati slabi, okrog 60%.

Vsekakor je bila naša težava, da smo imeli premalo podatkov. Učna množica je majhna, sploh, če jo primerjamo s številko 30.000 pri prej omenjeni raziskavi političnih tvitov v slovenskem jeziku. Mi toliko primerov nismo mogli oceniti, vseeno pa smo se odločili, da se usmerimo le na eno

področje. Glede na rezultate v tabeli 3.4 zgoraj je bila logična izbira šport. Ima negativne in pozitivne komentarje vsaj v testni množici lepo razporejene. Lahko rečemo, da je šport izmed vseh izbranih edina kategorija, ki jo množica doživlja tudi pozitivno.

Naslednji korak je bil naključna izbira 800 primerov, ki pa so bili tokrat le iz kategorije šport, in jih spet ročno označili. Od vseh komentarjev smo jih na koncu, kot učno množico lahko uporabili 511. Vso nadaljno testiranje algoritmov opravljamo na teh primerih v kategoriji šport.

3.4 Predprocesiranje podatkov

Za predprocesiranje komentarjev smo najprej uporabili korpus, iz katerega smo vzeli besede v obliki leme, torej besede v nedoločniku. Za ta namen smo uporabili korpus *ssj500kv.xml* [21], ki vsebuje 78665 različnih besed, od katerih jih v teh 511 ocenjenih komentarjih pri športu z njihovimi lemami zamenjamo 75%. Ostale besede pustimo v prvotni obliki.

```
1 def load_corpus_form_xmlfile(xmlfilename, namespace):
2     corpus = {}
3     tree = et.ElementTree(file=xmlfilename)
4     for elem in tree.iter(tag="{%s}w" % namespace):
5         word  = elem.text.lower()
6         lemma = elem.get("lemma").lower()
7         if (word not in corpus) or corpus[word] == lemma:
8             corpus[word] = lemma
9     return corpus
```

Iz množice besed izločimo še:

1. Množico besed, za katere menimo, da imajo za komentar nevtralen pomen, torej izdelamo seznam stop besed (angl. stop words).

2. Vsa ločila, saj želimo imeti le vrečo besed.

3. Črke v besedah, ki se pojavijo več kot dvakrat, zapišemo le dvakrat. Na primer; besedo toooo spremenimo v too. V tem primeru je mišljen vzklikni izraz veselja, če bi ponavljajočo se črko v tem primeru nadomestili le z eno in bi dobili "to", bi ta beseda predstavljala drug pomen.

4. Vsa števila, saj menimo, da nimajo nobenega pomena pri polarnosti komentarja.

Ko smo dobili želeno obliko svojih komentarjev, smo morali nato besede spremeniti v vektorski prostor, saj knjižnica sklearn operira le s številkami v vektorju. To smo storili s pomočjo metod za vektorizacijo,

```
1 train_vectorizer_array = vectorizer.fit_transform(get_words_set(
    train_data)).toarray()
```

kjer smo iz besed komentarjev dobili array oziroma vektor. Številke v vektorju so bile odvisne od tega, kakšen 'vectorizer' smo izbrali. Knjižnica sklearn vsebuje metodo CountVectorizer(), ki predstavlja navadno vrečo besed, ter TfidfVectorizer(), ki predstavlja uteženo vrečo besed. Obema lahko nastavimo različne parametre. Za nas so bili pomembni naslednji:

max_df , ki je zapisan kot število v zapisu s plavajočo vejico med 0.0 in 1.0 ali kot celo število, pomeni, da se naj ignorirajo izrazi, ki se v besedilih pojavijo večkrat, kot je vrednost max_df. Sklepamo namreč, da ti izrazi ne izražajo polaritete. To so lahko vezniki, predlogi ali druge zelo pogosto uporabljane besede. Mi smo tovrstne besede izločili že s seznamom stop besed, zato te vrednosti ne uporabljamo.

min_df deluje podobno kot max_df, le da ignoriramo besede, ki se pojavijo manjkrat od želene vrednosti. V našem primeru uporabljamo min_df = 2, saj so rezultati za odtenek boljši, kot če upoštevamo vse besede,

saj na ta način izločimo besede, ki se pojavijo le enkrat zanje težko rečemo, da imajo kakšen vpliv na klasifikacijo.

ngram_range uporabimo, če hočemo namesto unigramov uporabljati bigrame, trigrame, ali kakršnokoli poljubno število. Spremenljivka je določena kot interval in če je na primer $(1, 3)$, se preverijo vse kombinacije trigramov, bigramov in unigramov. Mi uporabljamo kombinacijo bigramov in unigramov $(1, 2)$, saj nam to da enak rezultat kot kombinacija s trigrami, vendar boljšega, kot če uporabimo zgolj unigrame$(1, 1)$.

binary v primeru, da je postavljen na True, postavi vse vrednosti v vektorju, ki so večje od 0 na 1. Torej imamo le 1 in 0, ki nam pove le, ali se beseda pojavi ali ne. Naši rezultati so bili v primeru uporabe slabši, zato je ne uporabimo.

3.5 Rezultati z izbranimi klasifikatorji

Izbrali smo pet algoritmov, s pomočjo katerih smo učili klasifikator na naših že označenih komentarjih, oziroma na zlatem standardu. Preizkusili smo metodo podpornih vektorjev (SVM), maksimalno entropijo (MaxEnt), metodo najbližjih sosedov (KNN) in pa multinominalno (MNB) in Bernoulli-jevo (BNB) metodo naivnega Bayesa. Najprej smo naredili program, ki je dal rezultat le z dvema razredoma, ker pa s tem nismo bili zadovoljni, smo rezultat predstavili s tremi razredi. Vse rezultate bomo v nadaljevanju predstavili.

3.5.1 Dva razreda

V tem delu so predstavljeni rezultati, ki so vsebovali razreda s pozitivnimi in negativnimi komentarji.

Kot lahko vidimo v tabeli 3.5, sta metodi najbljižjega soseda in Bernoulli naivnega Bayesa najslabši, medtem ko se druge tri ne razlikujejo veliko.

Poskusili smo tudi, kakšna je razlika med unigrami in bigrami in med TF-IDF utežitvijo in navadno vrečo besed. Razlike med unigrami in bigrami

so praktično zanemarljive. Malo večja razlika je med predstavitvijo z vrečo besed in TF-IDF utežmi, kjer se uteži obnesejo bolje.

Rezultati temeljijo na prečnem preverjanju, kjer se klasifikator uči na 60% primerih, njegovo učinkovitost pa preverimo na 40% primerov. Uporabili smo mešano porazdelitev, ki nam glede na izbrano velikost učne množice naključno izbere komentarje, vse ostale pa porabi za testiranje. Vsak komentar je uporabljen le enkrat, in je lahko bodisi v učni ali v testni množici.

		točnost	**natančnost**	**priklic**	**f1**
TF-IDF	UNIGRAMI				
	SVM	73%	72%	70%	71%
	MaxEnt	72%	74%	66%	69%
	NKK	54%	54%	77%	59%
	MNB	71%	76%	58%	66%
	BNB	60%	55%	83%	66%
	BIGRAMI				
	SVM	74%	71%	72%	71%
	MaxEnt	74%	73%	67%	70%
	NKK	53%	52%	76%	57%
	MNB	73%	75%	61%	67%
	BNB	60%	53%	88%	66%
Vreča besed	UNIGRAMI				
	SVM	69%	65%	78%	71%
	MaxEnt	69%	65%	77%	71%
	NKK	58%	54%	88%	66%
	MNB	71%	77%	58%	66%
	BNB	57%	53%	93%	67%
	BIGRAMI				
	SVM	67%	61%	79%	69%
	MaxEnt	69%	63%	80%	70%
	NKK	55%	52%	89%	63%
	MNB	72%	77%	57%	65%
	BNB	51%	48%	97%	64%

Tabela 3.5: Uspešnost posameznih metod.

Prečno previrjanje smo naredili dvajsetkrat, in vzeli povprečje, rezultate pa smo prikazali v zgornji tabeli 3.5. Najbolje se je obnesla metoda podpornih vektorjev, ki jo od tukaj dalje tudi edino uporabljamo. Z rezultati še

vedno nismo bili zadovoljni. Ko smo z izbrano metodo označili do tedaj še neoznačene komentarje smo ugotovili, da imamo težavo.

Nato smo naredili zlati standard, iz katerega smo izločili vse nevtralne komentarje, klasifikator pa smo učili le na pozitivnih in negativnih, kjer je bilo pri kategoriji šport v učni množici razmerje približno 50% - 50%. Nasprotno pa je na realnih podatkih. Naša metoda mora označiti tako pozitivne in negativne, kot tudi nevtralne. Za določitev pa ima na voljo le dva razreda, pozitivni in negativni razred. Rezultati so bili naslednji:

+	119616	29,88%
-	280697	70,12%
skupaj	400313	

Tabela 3.6: Razmerje pozitivnih in negativnih komentarjev, ki nam jih da SVM metoda.

Metoda je označila 70% komentarjev za negativne, po čemer sklepamo, da nevtralnih komentarjev ni enakomerno porazdelila med pozitivne in negativne, ampak so negativni prevladali. Do tega je lahko prišlo iz več razlogov.

Zelo verjetno je razlog razlika v dolžini pozitivnih in negativnih komentarjev. Iz nabora zlatega standarda vidimo, da je dolžina pozitivnih komentarjev veliko krajša od dolžine negativnih, iz tega sledi, da je večja verjetnost, da najde metoda iskane besede v negativnih komentarjih, kot v pozitivnih.

+	4004	36,24%
-	7045	63,76%
skupaj	11049	

Tabela 3.7: Število in delež besed v pozitivnih in negativnih komentarjih.

3.5.2 Trije razredi

Zaradi ugotovitev, ki smo jih že omenili zgoraj, smo se odločili, da zgradimo še klasifikator, ki za komentar ne določi le pozitivnega ali negativnega razreda, ampak mora biti za to tudi dovolj velika verjetnost, sicer komentar označi kot nevtralen. Podoben pristop je uporabljen za klasifikacijo tvitov v slovenskem jeziku, ki smo ga opisali zgoraj [23]. Želeli smo izboljšati svoje rezultate in predvsem izločiti komentarje, ki niso dovolj prepričljivi, da bi spadali v določen razred.

Tokrat nam klasifikator ni vrnil le + ali -, ampak nam je dal za vsak komentar dve vrednosti. Ugotovili smo, da se razen metode podpornih vektorjev, pri vseh ostalih vrednosti vrtijo med 40% in 60%, večina okrog 50%, medtem, ko so bile številke pri metodi podpornih vektorjev veliko bolj raznolike. To nam je dalo še eno potrditev, da smo se pravilno odločili pri izbiri algoritma. Prejšnji način je deloval tako, da je komentarju priredil razred, ki mu je pripadal višji delež, tokrat pa smo postavili mejo. Če nobena izmed vrednosti ni bila večja od 70%, je komentar ostal nevtralen. Naši novi rezultati so bili naslednji:

	točnost	**natančnost**	**priklic**	**f1**
SVM	82%	83%	77%	80%

Tabela 3.8: Rezultat metode SVM na testnih podatkih.

Rezultat se je pričakovano precej izboljšal. Ko smo našo novo metodo pognali nad vsemi komentarji na temo šport, je bilo razmerje takšno:

+	95512	23,86%
-	178566	44,61%
Nevtralni	126235	31,53%
skupaj	400313	

Tabela 3.9: Rezultat metode SVM na testnih podatkih, z dodanim nevtralnim razredom.

Skoraj tretjino vseh komentarjev je naša metoda označila za nevtralne. Še vedno je med pozitivnimi in negativnimi precejšnja razlika, ampak je ta vsekakor manjša kot prej. V nadaljevanju bomo komentarje, ki so označeni nevtralno, zanemarili.

Poglavje 4

Rudarjenje po komentarjih portala rtvslo.si

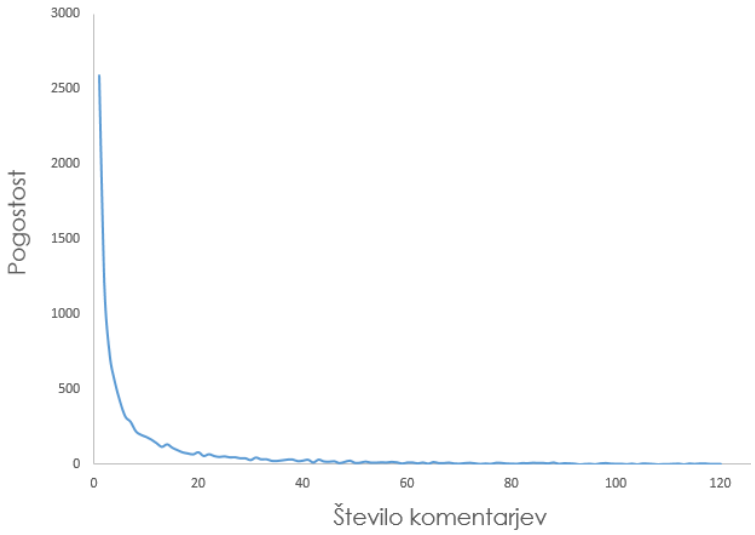
Cilj rudarjenja razpoloženja je, da znamo pridobljene klasifikacije tudi uporabiti in prikazati kakšno zanimivo povezavo, kar je bil tudi cilj v drugem delu praktičnega dela. V tem delu smo uporabili orodje Orange, kjer smo uporabili metodo k-means, ki spada med metode gručenja. Uporabili smo jo v prvem delu pri iskanju podobnosti uporabnikov, naprej pa smo delali s poizvedbami in MySQL 5.6 bazo, grafe pa smo prikazovali s pomočjo Pythona in knjižnice matplotlib.

4.1 Iskanje podobnosti uporabnikov

Do zdaj smo se ukvarjali le s klasifikacijo komentarjev, torej ali so pozitivni ali negativni. To smo storili za kategorijo šport, kjer smo našo metodo privzeli za pravilno in našim komentarjem v kategoriji šport v bazi podatkov zapisali še polarnost, ali so pozitivni, negativni ali nevtralni. Za nadaljnje delo pa nevtralnih nismo upoštevali.

V tej sekciji bomo opisali, na kakšen način smo poskušali razdeliti uporabnike v skupine in kakšni so bili rezultati.

Za začetek si oglejmo, kako uporabniki komentirajo:



Slika 4.1: Prikaz pogostosti števila komentarjev v kategoriji šport.

Iz zgornje slike lahko vidimo, da ima zelo majhno število uporabnikov veliko število komentarjev. Mi smo za svojo analizo vzeli število uporabnikov, ki so napisali 70% vseh komentarjev. Glede na to merilo smo obravnavali 892 uporabnikov od 9931 vseh, kar je približno 9% vseh. Največ uporabnikov ima samo en komentar, teh je kar 7343, torej 74% vseh. Ne smemo seveda pozabiti, da govorimo o številkah samo na področju športa. Sklepamo pa lahko, da so deleži tudi drugje zelo podobni.

Lastnosti, ki opredeljujejo vsakega uporabnika, smo razdelili v naslednje tri kategorije:

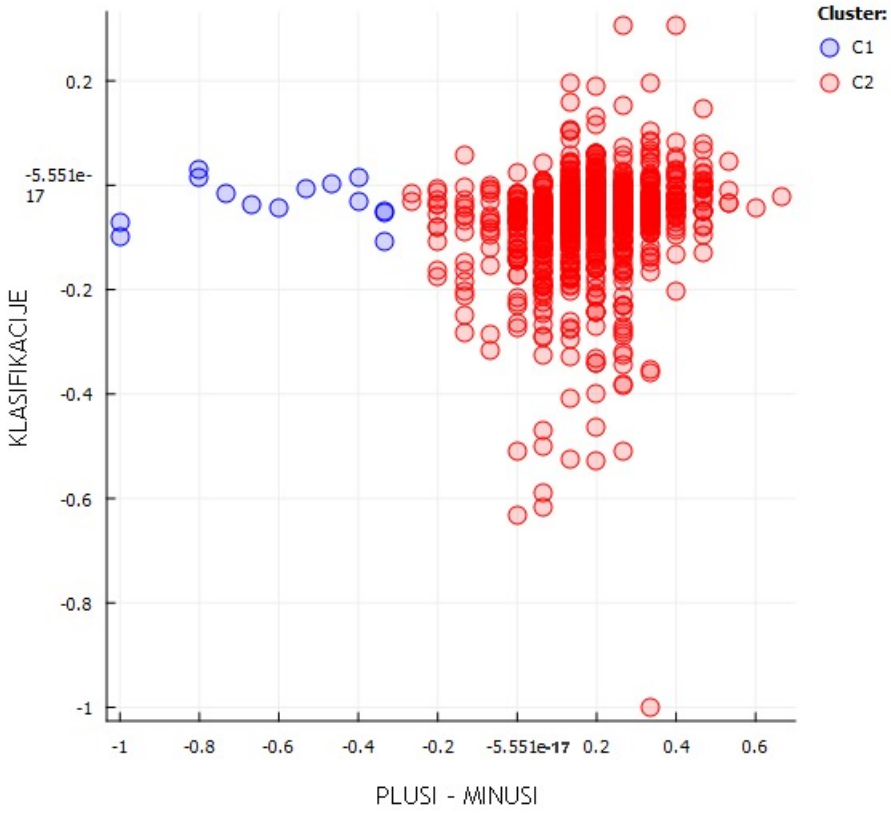
1. Kolikšen je delež pozitivnih komentarjev (naša klasifikacija).

2. Kakšna je všečnost komentarjev (koliko je razmerje med vsemi plusi in minusi glede na njegove komentarje).

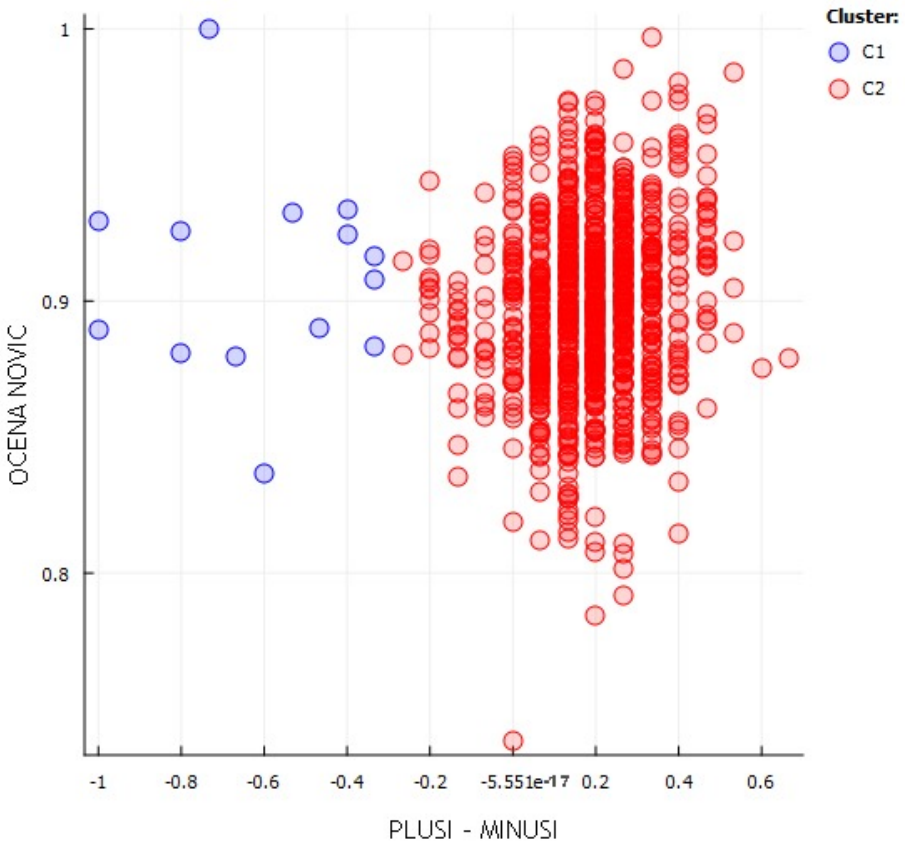
3. Kakšne novice komentira (povprečna ocena novic).

Najprej smo vsako kategorijo normalizirali, nato pa smo v orodju Orange, s pomočjo metode gručenja k-means, točke, ki so predstavljale posamezne uporabnike, tudi prikazali. Ugotovili smo, da se večina točk, razen tistih nekaj redkih izjem, nahaja v sredini pri vseh kategorijah. Vendar kljub vsemu je odnosu med prvo in zadnjo kategorijo moč zaznati neko korelacijo.

Za lažjo predstavo prilagamo 2D slike v vseh treh kombinacijah.

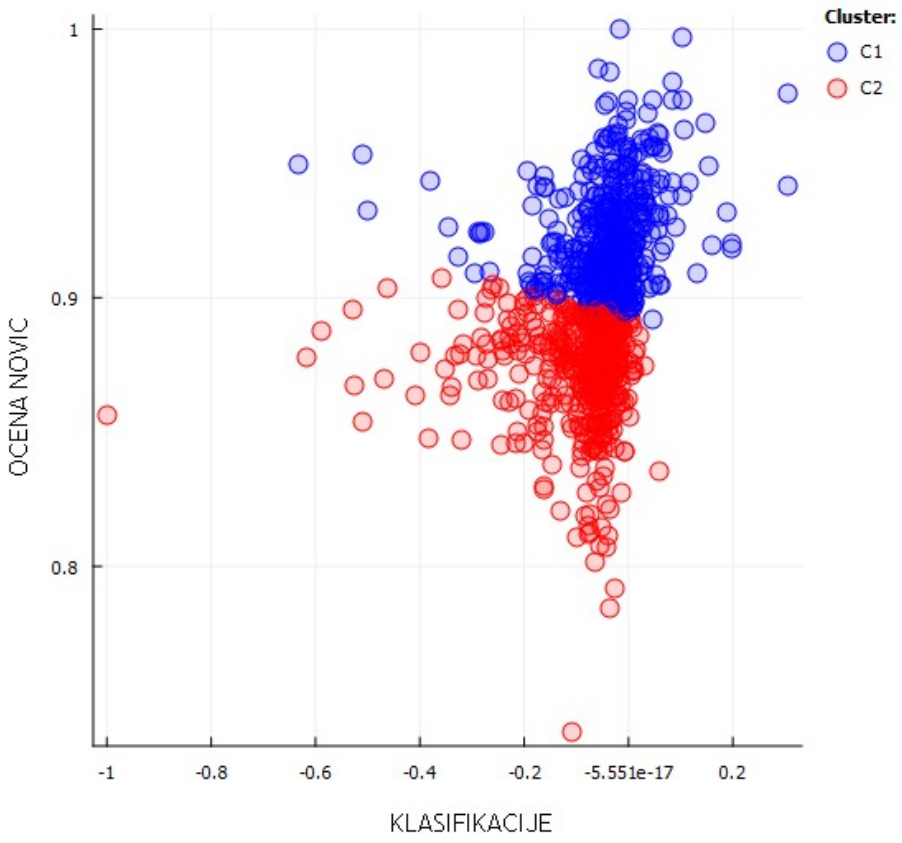


Slika 4.2: Prikaz ne korelacije plusov in minusov s klasifikacijo komentarjev.



Slika 4.3: Prikaz ne korelacije plusov in minusov z oceno novice.

Zgornji sliki 4.2 in 4.3 tipično prikazujeta neodvisnost med kategorijama. Vsaka točka predstavlja uporabnika, ki ima dve lastnosti; x in y. Rečemo lahko, da to, kakšna je polarnost uporabnikovih komentarjev, nima povezave z njihovo všečnostjo. Prav tako z všečnostjo nima povezave to, kako ocenjene novice uporabnik komentira.

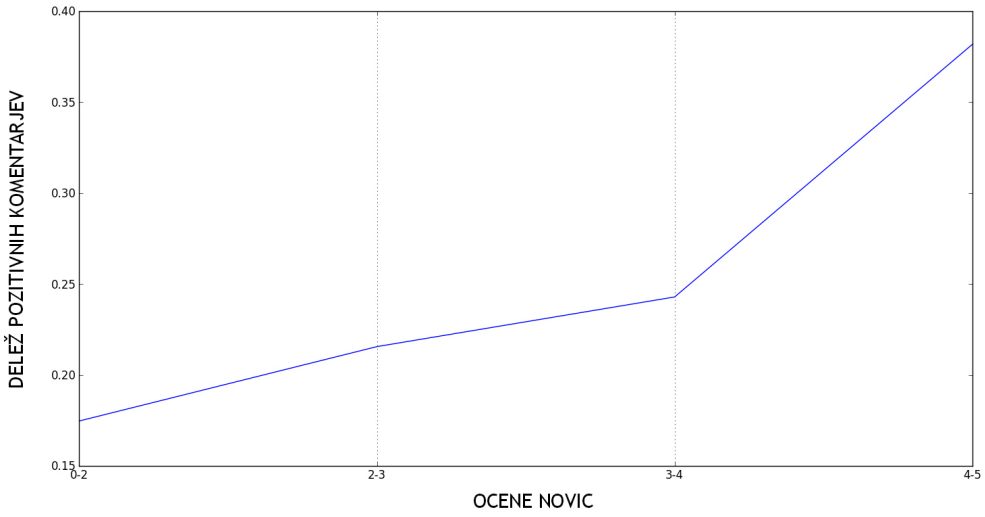


Slika 4.4: Prikaz korelacije med povprečno oceno novic, katere komentira uporabnik, in med njegovimi povprečno klasificiranimi komentarji.

V povezavi med našo klasifikacijo uporabnikovih komentarjev in ocenami novic, katere uporabnik komentira, pa lahko na grafu 4.4 vidimo rahlo korelacijo. Boljša(večja) kot je klasifikacija komentarjev, boljšim novicam uporabnik daje komentarje. Lahko rečemo, da imamo dve skupini uporabnikov. Tisti, ki v povprečju komentirajo slabše novice in imajo bolj negativno klasificirane komentarje, in pa tiste, ki komentirajo boljše ocenjene novice in imajo bolj pozitivne komentarje.

Tudi spodnji graf 4.5, nam na podlagi vseh komentarjev in novic na področju športa pokaže, da imajo bolje ocenjene novice bolj pozitivne komentarje. V grafu lahko vidimo, kako odstotek pozitivnih komentarjev narašča

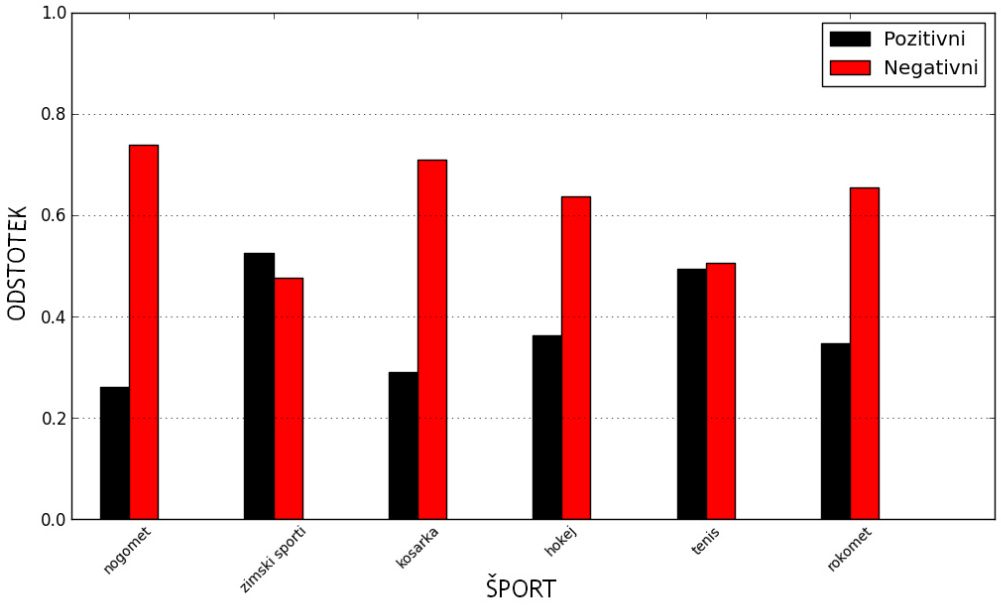
glede na interval ocene novice. Iz tega naša korelacija med uporabniki tudi izvira.



Slika 4.5: Naraščanje števila pozitivno klasificiranih komentarjev z naraščanjem ocene novice.

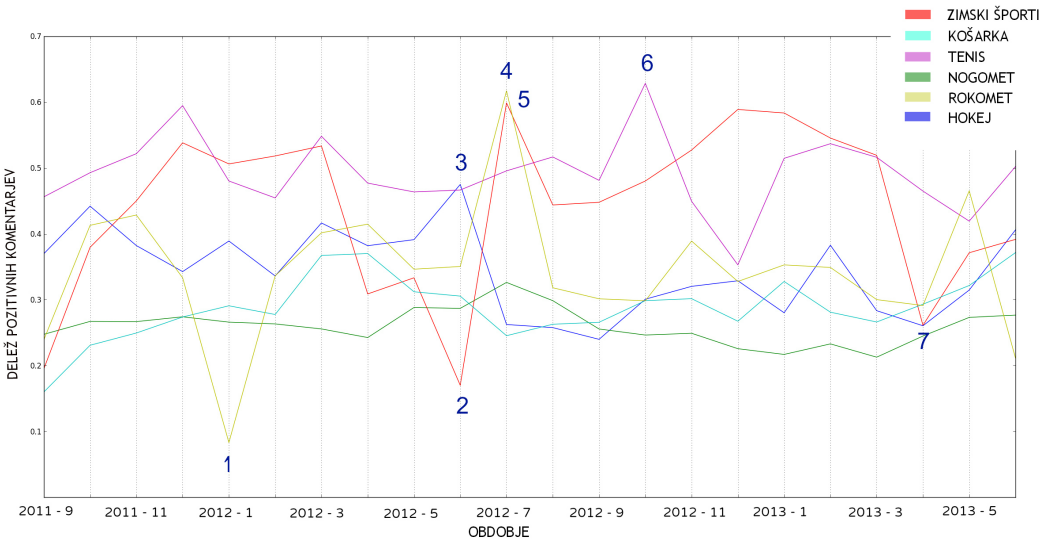
4.2 Športne kategorije skozi čas

Zanimalo nas je tudi, kako se komentarji za različne športe spreminjajo čez čas. Izbrali smo šest najbolj komentiranih podkategorij v kategoriji šport, in sicer nogomet, košarko, zimske športe, hokej, tenis in rokomet.



Slika 4.6: Razmerje med pozitivno in negativno klasificiranimi komentarji glede na posamezni šport.

V spodnjem diagramu 4.6 si lahko ogledamo, kako se spreminja odstotek pozitivnih komentarjev čez čas. Za časovno obdobje smo vzeli 22 mesecev, od septembra 2011 do junija 2013, saj so to podatki, ki smo jih imeli na voljo.



Slika 4.7: Časovni prikaz pozitivnosti komentarjev glede na šest najbolj komentiranih športov, z obrazložitvijo točk spodaj.

1. Prvi večji negativni vrh imamo januarja 2012, in sicer pri rokometu, ki pa ima ta mesec le dve novici z zelo malo komentarjev, zato težko govorimo o kakšnem splošnem negativnem dogodku.

2. Junija 2012 so imeli zimski športi negativno špico, saj je bila takrat afera med predsednikom Olimpijskega komiteja Slovenije Janezom Kocijančičem in predsednikom Smučarske zveze Slovenije Tomažem Lovšetom.

3. Junija 2012 ima hokej pozitiven vrh, saj je Anže Kopitar z Los Angelesom osvojil Stanleyev pokal.

4. Julija 2012, kjer ima rokomet pozitivno špico, je bilo evropsko prvenstvo za mlade do dvajset let, kjer so bili naši reprezentanti zelo uspešni.

5. Prav tako julija 2012 imajo pozitivno špico tudi zimski športi, saj so v tem obdobju potekali poletni smučarski skoki, kjer so na ekipni tekmi naši skakalci zmagali.

6. Oktobra 2012, kjer ima pozitiven vrh tenis, imajo največ komentarjev novice, ki govorijo o Žemlji, saj je bil prvi Slovenec v zgodovini, ki se je uvrstil v finale turnirja za ATP.

7. Aprila 2012 imajo negativni vrh zimski športi, saj je bila prvoaprilska šala, da bo Jakov Fak tekmoval za Nemčijo zelo slabo sprejeta.

Na sliki 4.6 je prikazano razmerje za vseh šest najbolj komentiranih športov po vrsti po številu komentarjev. Vidimo, da so komentarji večinoma negativni, le zimski športi imajo več pozitivnih komentarjev, sledi jim tenis. Največ negativnih komentarjev ima nogomet, ki je tudi najbolj komentiran, sledi mu košarka. Vsak izmed športov ima v obdobju 22 mesecev kakšen padec ali dvig, ki odstopa od povprečja. Zanimalo nas je, ali lahko iz teh ekstremov kaj vidimo, in ugotovili smo, da se zelo povezujejo z aktualnimi dogodki tistega časa. Zaključimo lahko, da so bile naše metode, ki jih opisujemo v prejšnjem poglavju o klasifikacijah, dobre, saj se ujemajo z naravnimi dogodki.

Poglavje 5

Sklepne ugotovitve

V okviru diplomske naloge smo spoznali področje analize razpoloženja, teoretično smo raziskali kaj se v praksi uporablja in to s pomočjo knjižnice sklearn implementirali ter preizkusili na našem naboru podatkov.

Našo kodo smo se odločili objaviti na spletu, za kar smo uporabili spletni servis GitHub [39], koda pa je dostopna na naslovu:

https://github.com/brina123/sentianalysis

Ugotovili smo, da je področje analize razpoloženja na splošno zelo razvito, še posebno področje angleškega jezika, kjer si je težko zamisliti raziskavo, ki še ne bi obstajala. Obstaja tudi ogromno slovarjev, kot tudi že kar označenih besed v sklopu knjižnic, zato so tu smeri raziskovanja brezmejne, kar pa povzroči, da se veliko ljudi, tudi če angleščina ni njihov materni jezik, odloči za to možnost, in preučujejo kar angleški jezik.

Popolnoma druga zgodba pa so jeziki, ki niso množično uporabljeni. Tak primer je slovenski jezik. Če kakšne raziskave že obstajajo, niso v večji meri javno dostopne, podobno je s slovarji. Z gotovostjo lahko trdimo, da se pri nas to področje šele razvija. Ugotovili smo, da še ne obstaja slovar čustev za slovenski jezik, zato bi bilo na mestu, da bi začeli kar z grajenjem tega slovarja.

Vprašati pa se je seveda potrebno, koliko je dejansko smiselno raziskovanje slovenščine na takšen način. Ali je to sploh dovolj velik trg? Veliko ljudi

namreč že na družabnih omrežjih piše kar v angleščini. Veliko družabnih omrežij je celo zastavljeno globalno in uporaba slovenščine ne bi bila smiselna.

Še vedno mora biti v interesu spodbujati slovenščino, prav je, da poskušamo biti čim bolj v trendu z angleškim jezikom in njegovimi zmožnostmi, zato je nujno, da tehnike za slovenski jezik čimbolj razvijamo. Na tem področju se še vedno lahko delajo zanimive stvari, četudi je trg le 2-milijonski.

Mi smo v naši diplomski predstavili zelo preproste pristope, ki so nam pravzaprav dali že precej dobre rezultate. Seveda pa je tukaj še precej lukenj in možnosti izboljšav. Za začetek bi bilo dobro povečati naš zlati standard, saj menimo, da bi lahko bili tako rezultati veliko boljši.

5.1 Motiv za nadaljno raziskovanje in uporabo

Kljub temu, da smo v diplomski nalogi le v osnovi preučili polarnost komentarjev, si izbrali samo kategorijo šport in na njej naredili še nekaj raziskav, pa bi lahko, če bi našo tehniko še izboljšali in jo naredili bolj splošno, iz nje naredili še veliko zanimivih stvari.

Ideja je, da bi na portalu imel uporabnik možnost izbire, kakšne novice in komentarji se naj mu prikazujejo. V Sloveniji, in najbrž tudi v tujini, portalom zelo pada raven kakovosti novic, saj je vedno bolj pomembno narediti novico zanimivo, da dobi čimveč "klikov", zato postaja kakovost drugotnega pomena. Zaradi tega razloga bi lahko naredili, da bi si lahko uporabnik izbral prikazovanje le pozitivnih novic. Lahko bi določil, da hoče videti le novice, ki imajo oceno višjo od določene vrednosti. Naredili bi lahko, da bi se program sam učil na uporabnikovih všečkih in ocenah novic in komentarjev in bi mu prikazoval le takšne, ki so mu všeč.

Druga ideja, kjer bi lahko uporabili polarnost komentarjev, je, da bi na portalu imeli prilagojene oglase. Naredili bi raziskavo, kaj imajo raje bolj pozitivno in kaj bolj negativno usmerjeni ljudje. Spletni portali, kot sta na primer 24ur.com in rtvslo.si, katerih lastnika sta televiziji, bi lahko oglaševali

prilagojene oglase za različne skupine uporabnikov. Bolj pozitivni skupini uporabnikov bi lahko oglaševali drugačne filme in oddaje kot negativni skupini.

To sta le dve ideji, ki nam prideta na misel, vsekakor pa se tukaj možnosti ne končajo. Zaključimo lahko, da to področje ponuja ogromno možnosti raziskovanja, vse je odvisno le od naše domišljije in volje.

Literatura

[1] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, B. Liu. (2011). *Combining Lexicon-based and Learning-based Methodsfor Twitter Sentiment Analysis*. Dostopno na: http://www.hpl.hp.com/techreports/2011/HPL-2011-89.pdf.

[2] L. Bing. (2012). *Sentiment analysis and opinion mining*. Dostopno na: http://www.cs.uic.edu/ liub/FBS/SentimentAnalysis-and-OpinionMining.pdf.

[3] Peter D. Turney. (2002). *Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews*. Dostopno na: http://acl.ldc.upenn.edu/P/P02/P02-1053.pdf.

[4] B. Pang, L. Lee.(2002). *Thumbs up?: sentiment classification using machine learning techniques* Dostopno na: http://www.cs.cornell.edu/home/llee/papers/sentiment.pdf.

[5] A. Ng. (2008). Stanford University Video on SVM. Dostopno na: http://www.youtube.com/watch?v=qyyJKd-zXRE.

[6] A. Berger, V. Della, S. Pietra. (1996). *A Maximum Entropy Approach to Natural Language Processing*.Dostopno na: http://dl.acm.org/citation.cfm?id=234289 .

[7] K nearest neighbour. Dostopno na: $http://en.wikipedia.org/wiki/K-nearest\_neighbours\_algorithm$.

[8] B. Pang, L. Lee. (2008). *Opinion Mining and Sentiment Analysis*. Dostopno na: http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf.

[9] Euclidean distance. Dostopno na: *http* : *//en.wikipedia.org/wiki/ Euclidean_distance*.

[10] Hamming distance. Dostopno na: *http* : *//en.wikipedia.org/wiki/ Hamming_distance*.

[11] Minkovski distance. Dostopno na: *http* : *//en.wikipedia.org/wiki/ Minkowski_distance*

[12] Hierarchical clustering. Dostopno na: *http* : *//en.wikipedia.org/wiki/ Hierarchical_clustering*.

[13] T. Kanungo, N. S. Netanyahu, A. Y. Wu. (julij, 2002). Efficient k-Means Clustering Algorithm: Analysis and Implementation. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 24, NO. 7, JULY 2002.

[14] K. Polanec. (2006). Strojno učenje. Dostopno na: http://dat.si/publikacije/Article/Strojno-u–269-enje/66.

[15] A. McCallum, K. Nigam. (1998). *A comparison of Event Models for Naive Bayes Text Classification*. Dostopno na: *http* : *//citeseerx.ist.psu.edu/viewdoc/download?doi* = 10.1.1.65.9324&*rep* = *rep*1&*type* = *pdf*.

[16] J. Kamps, M. Marx, R. J. Mokken, M. Rijke. (2004). *Using WordNet to measure semantic orientation of adjectives*. In Proceedings of LREC-04", 4th International Conference on Language Resources and Evaluation, pages 1115–1118, Lisbon, PT. Dostopno na: http://dare.uva.nl/document/154122.

[17] Bag of words. Dostopno na: *http* : *//en.wikipedia.org/wiki/Bag* − *of* − *words_model*.

[18] M. Hu, B. Liu. (2004). *Mining and Summarizing Customer Reviews.* Dostopno na: *http* : *//citeseerx.ist.psu.edu/viewdoc/download?doi* = 10.1.1.76.2378&*rep* = *rep*1&*type* = *pdf*.

[19] Tf-idf. Dostopno na: http://en.wikipedia.org/wiki/Tf-idf.

[20] A. Hassan, D. R. Radev. (2010). *Identifying text polarity using random walks.* In Proceedings of ACL Uppsala, Sweden. Dostopno na: http://clair.si.umich.edu/ radev/papers/ACL2010-ahmed.pdf.

[21] Korpusi. Dostopno na http://www.slovenscina.eu/korpusi.

[22] M. Verlic (2009). Hibridni pristop za zaznavo elementov subjektivnosti v besedilnih tokovih. Doktorsko delo, Maribor: Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko.

[23] J. Smailovic, J. Kranjc, M. Juršič, M. Grčar, M. Gačnik, I. Mozetič. (2013). *Monitoring the Twitter sentiment during the Bulgarian elections.*

[24] J. Brezovnik. (2009). Programsko orodje za procesiranje besedil v naravnem jeziku, Magistrsko delo, Maribor: Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko.

[25] A. Bregant. (2011). Razreševanje večpomenkosti in odkrivanje ustreznih predlog v sistemu za priklic informacij v naravnem jeziku. Magistrsko delo, Maribor: Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko.

[26] M. Sandić (2011). Klasifikacija spletnih novic s pomočjo metod podatkovnega rudarjenja v besedilu. Diplomsko delo, Maribor: Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko.

[27] Sklearn knjižnica. Dostopno na: http://scikit-learn.org/stable/.

[28] Sentiwordnet slovar. dostopno na: http://sentiwordnet.isti.cnr.it/.

[29] O. Tsur, D. Davidov, A. Rappoport. (2010) *A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews*. 4th Int'l AAAI Conference on Weblogs and Social Media, Washington. Dostopno na: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10 /paper/viewFile/1495/1851

[30] SVM slika. Dostopno na: http://www.m8j.net/(All)Geometry%20of%20Support%20Vector%20Machines

[31] K. Dave, S. Lawrence, D.M. Pennock. (2003). *Mining the peanut gallery: opinion extraction and semantic Classification of product reviews*. Proceedings of the 12th International Conference on World Wide Web, pg. 519 – 528. Dostopno na: $http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.13.2424\&rep=rep1\&type=pdf$.

[32] Andrea Esuli and Fabrizio Sebastiani. Determining the semantic orientation of terms through gloss analysis. In Proceedings of the ACM SIGIR Conference on Information and Knowledge Management (CIKM), 2005.

[33] S. Kim,E. Hovy. (2005). *Automatic identification of pro and con reasons in online reviews*. Dostopno na: http://www.isi.edu/natural-language/people/hovy/papers/06ACL-ProCon-opinions-short.pdf.

[34] Google trends. (2013). Dostopno na: http://www.google.com/trends/explore?q=sentiment+analysis

[35] P. Simon. (March 18, 2013). *Too Big to Ignore: The Business Case for Big Data*.

[36] Cross validation. Dostopno na: $http://en.wikipedia.org/wiki/Cross-validation\_(statistics)$.

[37] Naivni Bayes. Dostopno na: http://kt.ijs.si/PetraKralj/ DataMining0809/NaiveBayes.pdf.

[38] Naivni Bayes classifier. Dostopno na: *http* : *//www.princeton.edu/ achaney/tmve/wiki*100*k/docs/Naive_Bayes _classifier.html*.

[39] (2013). Wikipedia. GitHub. Dostopno na: http://en.wikipedia.org/wiki/GitHub.