Analiza mnenj s pomočjo strojnega učenja in slovenskega leksikona sentimenta

Klemen Kadunc, Marko Robnik-Šikonja

Fakulteta za računalništvo in informatiko, Univerza v Ljubljani
Večna pot 113, 1000 Ljubljana
klemzimeister@gmail.com, marko.robnik@fri.uni-lj.si

Povzetek

V prispevku obravnavamo področje tekstovnega rudarjenja, ki se primarno osredotoča na identifikacijo mnenj v besedilih ter njihovo opredeljevanje kot pozitivna ali negativna. V našem delu je bil uporabljen pristop z metodami strojnega učenja, ki smo ga nadgradili z ročno zgrajenim leksikonom sentimenta. Z leksikonom smo v statističnih modelih želeli zaobjeti leksikalno znanje in s tem izboljšati uspešnost klasifikacije. Z evalvacijo rezultatov na primeru spletnih komentarjev nekaterih bolj obiskanih slovenskih spletnih portalov pokažemo, da je pristop uspešen, saj smo klasifikacijo zaznavno izboljšali.

Opinion Mining Using Machine Learning and Slovene Sentiment Lexicon

The article deals with text mining and focuses on identifying opinions in texts and determining their positive or negative character. The research was based on machine learning methods that were upgraded with a manually compiled sentiment lexicon. Its purpose was to encompass lexical knowledge in statistical models and thus improve the classification. Furthermore, the analysis of online comments found on some of the most visited portals indicated that the approach is successful because it managed to significantly improve the classification.

1 Uvod

Analiza mnenj v besedilih (angl. opinion mining), v ožjem kontekstu poimenovana tudi analiza sentimenta (angl. sentiment analysis), je eno od področij tekstovnega rudarjenja (Liu in Zhang, 2012). Ukvarja se z odkrivanjem piščevega mnenja o predmetu pisanja. Večinoma se uporablja polarna analiza, torej določamo pozitivno, negativno in nevtralno mnenje. Naloga za avtomatske sisteme ni enostavna, saj je potrebno iz besedila izluščiti bistveno semantično informacijo, pri tem pa težavo predstavljajo (večkratno) zanikanje, sarkazem, dvoumnost besedila, kontekstna odvisnost rabe besed itd. Pravilno določanje sentimenta je koristno za številne namene, npr. za napovedovanje uspešnosti izdelkov, izidov volitev ali v sociooloških raziskavah. Zaradi vse širšega izražanja mnenj na internetu, preko spletnih forumov, komentarjev novic, tvitov ter recenzij izdelkov in storitev postaja avtomatska analiza mnenj nujen raziskovalni pripomoček.

Analizo sentimenta lahko izvajamo na različnih nivojih, od najbolj splošnega, tj. na nivoju celotnega besedila, do posameznih vidikov oz. značilnosti, ki se v besedilu pojavijo (npr. piščevo mnenje o porabi goriva testnega avtomobila). Pri slednjem nivoju je velika težava že identifikacija entitet in vidikov. V našem delu smo se osredotočili na nivo celotnega besedila. Takšni analizi sentimenta pravimo tudi klasifikacija sentimenta, saj podobno kot pri drugih nalogah klasifikacije besedil (npr. klasifikacija dnevnih novic pod gospodarstvo, šport ipd.), besedilo klasificiramo v eno od sentimentnih kategorij.

Pri analizi sentimenta sta se v grobem uveljavila dva pristopa, leksikalni in pristop s strojnim učenjem. Pri leksikalni metodi potrebujemo enega ali več leksikonov sentimenta, ki vključujejo besede in fraze s pozitivno in negativno konotacijo. Pristop prešteje te besede in fraze v besedilu, ki ga želimo klasificirati. Če prevladujejo besede z negativno konotacijo, besedilo označi kot negativno, sicer kot pozitivno. Težava pristopa je, da je potrebno izdelati leksikon, poleg tega pa se izrazoslovje s časom spreminja, nekatere besede pa imajo različen sentiment v različnih kontekstih (npr. beseda majhen se pri opisu zaslona mobilnega telefona šteje kot negativna oznaka, pri opisu pomnilniškega ključka pa kot pozitivna), zato je ta pristop v praksi v zadnjem času skoraj vedno združen s strojnim učenjem. Pri pristopu s strojnim učenjem s pomočjo učne množice besedil, ki jim priredimo eno od sentimentnih kategorij (npr. pozitivno ali negativno), tvorimo učno množico, na podlagi katere zgradimo klasifikacijski model. Kljub temu, da sentimentni leksikon ni nujno potreben za analizo s strojnim učenjem, so raziskave pokazale, da je dober leksikon koristen in izboljša klasifikacijsko točnost (Hu in Liu, 2004).

Za angleščino obstajajo številna orodja in besedilni korpusi namenjeni analizi sentimenta. Analiza sentimenta v slovenščini je še v povojih in manjkajo še številni viri, npr. slovenski SentWordNet - sentimentno označen WordNet (Baccianella et al., 2010), orodja za pridobivanje in označevanje spletnih virov kot so tviti in spletni komentarji, korpusi sentimentno označenih besedil pa tudi slovenski leksikon polarnih besed potrebuje dodelavo.

Po vzoru leksikona Hu in Liu (2004) smo v tem delu sestavili prosto dostopen leksikon sentimentnih besed. Predstavljamo njegovo evalvacijo v okviru klasifikacije spletnih komentarjev z metodami obdelave naravnega jezika in strojnega učenja. Širša analiza leksikona in označene podatkovne baze, ki jo uporabljamo, je predstavljena v (Kadunc, 2016).

V 2. razdelku pripravimo kratek pregled obstoječih leksikonov sentimenta v slovenskem jeziku ter podrobneje predstavimo izdelan leksikon, v 3. se osredotočimo na zbirko uporabniških komentarjev, ki smo jo sestavili za potrebe evalvacij klasifikacijskih modelov, v 4. si ogledamo dosežene rezultate, jih kritično ovrednotimo ter jih primer-

jamo z rezultati sorodnih raziskav v drugih jezikih. Prispevek zaključimo z omembo glavnih sklepov ter navedemo nekaj možnih izboljšav.

2 Sentimentni leksikon

Sentimentni leksikoni predstavljajo osnovo za sentimentno analizo z leksikalnim pristopom. V strojnem učenju leksikalno znanje običajno vključujemo v fazi priprave značilk. Leksikoni v najosnovnejši obliki sestojijo iz seznama pozitivnih in seznama negativnih besed ali fraz. Naprednejši poleg informacije o polariteti vsebujejo še uteži (npr. beseda je močno pozitivna), oblikoslovne oznake ipd. Gradnja leksikonov poteka na različne načine, od povsem ročnega (drago in zamudno), do polavtomatskega in avtomatskega. Pri naprednejših pristopih se določi začetno množico besed, ki predstavljajo semena za avtomatsko ekspanzijo s pomočjo Wordneta in drugih strukturiranih leksikalnih baz. Podrobneje so avtomatski pristopi opisani v (Potts, 2011). Predvsem za angleščino obstaja veliko število prosto dostopnih leksikonov sentimenta, od splošnih do bolj specializiranih.

Za slovenščino že obstaja nekaj manjših sentimentnih leksikonov. Martinc (2013) je na podlagi seznama AFINN-111 (Nielsen, 2011), ki vsebuje 2477 besed, sestavil seznam polarnih besed ter vsaki priredil vrednost z razponom od -5 (skrajno negativno) do +5 (skrajno pozitivno). Leksikon je uporabil za izdelavo orodja za analizo sentimenta na družbenem omrežju Twitter. Za razliko od Martinca se je Volčanšek (2015) osredotočila na sentimentno analizo bolj formalnih besedil. Za analizo novic je Volčanšek (2015) uporabila leksikalni pristop, njen leksikon pa je osnovan na angleškem slovarju General Inquirer (Stone, 1997), ki poleg seznamov besed vsebuje še dodatne metapodatke, kot je denimo označba vseh kategorij, v katerih se beseda nahaja. Prevedeni slovar teh podatkov ne zajema. Iz kategorij *Positiv* in *Negativ* je z uporabo avtomatskega in ročnega preverjanja sestavila slovenski slovar sentimenta, ki šteje 1669 pozitivnih in 1912 negativnih besed. Rezultati analize na podlagi klasifikacije 5000 novic so bili po mnenju avtorice pod pričakovanji. Kot sentimentni leksikon lahko uporabimo tudi angleški SentiWordNet (Baccianella et al., 2010), ki temelji na leksikalni bazi WordNet. SentiWordNet je povezan z WordNetom in vsakemu vnosu v WordNetu priredi tri numerične vrednosti, s katerimi meri pozitivnost, negativnost ter objektivnost oziroma nevtralnost pojmov. Za razliko od prejšnjih dveh slovarjev, SentiWordNet hrani ocene za različne pomene iste besede. Ker bi bilo brez opisa nemogoče ločevati med posameznimi pomeni, so vnosi oplemeniteni z gloso oz. kratkim opisom pomena za lažje pomensko razdvoumljanje besed. Ker obstaja leksikalna baza WordNet tudi v slovenskem jeziku pod imenom SloWNet (Fišer, 2008) in ker so vnosi povezani z WordNetom, je mogoče sentimentno informacijo iz angleščine prenesti v slovenščino. V naših poskusih smo uporabili tudi to možnost.

Naš primarni leksikon temelji na angleškem leksikonu (Hu in Liu, 2004), ki ga trenutno sestavljata seznama 2006 pozitivnih in 4783 negativnih besed. Za Hujev leksikon smo se odločili zato, ker je bil uporabljen že v vrsti raziskav iz obravnavanega področja ter se stalno posodablja. Osnova sicer izvira iz prve polovice prejšnjega desetletja, ko so se raziskovalci osredotočali predvsem na analizo sentimenta opisov filmov in raznih produktov, kar je razvidno tudi iz samih vnosov. V seznamu so namenoma vključene tudi napačno črkovane in žargonske besede. Izdelava našega leksikona je potekala tako, da smo za osnovo vzeli slovar sentimentnih besed v angleškem jeziku ter ga ročno prevedli v slovenščino s pomočjo spletnih prevajalskih orodij. Vključili smo tudi polarno obarvane sinonime in nekaj različnih oblik iste besede. V slovenščino neprevedljive besede smo izpustili. Zaradi bogate pregibnosti slovenskega jezika je priporočljiva uporaba lematizacije. Naš leksikon trenutno sestavljata seznama 2646 pozitivnih in 6689 negativnih besed. Slovar je prosto dosegljiv v obliki kompresirane datoteke ZIP, ki vsebuje seznam pozitivnih (*positive_words.txt*) ter seznam negativnih (*negative_words.txt*) besed. Posamezne besede so med seboj ločene z znakom za novo vrstico. Vsebina je shranjena v kodnem naboru UTF-8 (Unicode), tako da je uporaba mogoča na večini računalniških platform.

Tako izdelan leksikon ni brez slabosti. Poleg neupoštevanja konteksta uporabe, ki je ena od splošnih slabosti tovrstnih slovarjev, je težava tudi v tem, da smo s prevajanjem iz angleščine izpustili nekatere pogosteje uporabljane slovenske izraze, predvsem tiste, ki se pogosto uporabljajo v neformalni komunikaciji na spletnih omrežjih in za katere ne obstaja neposredni prevod. Leksikon sicer predstavlja dovolj dobro osnovo za nadaljne posodabljanje z novim izrazoslovjem. Omenili smo, da se v slovarju nahaja nekaj različnih oblik iste besede, ki smo jih pri prevajanju dodajali. Postavi se vprašanje o smotrnosti tega početja z ozirom na dejstvo, da za posamezno besedo lahko obstaja tudi več deset različnih oblik. Za določene rabe bi bilo smiselno vse besedne oblike dodati v leksikon, za druge pa vključiti le lemo vsake besede, saj je lematizator za slovenski jezik prosto dostopen z že narejeno podporo za integracijo z več programskimi jeziki (Juršič, 2007). Za uporabnika leksikona tako ne bi smelo biti težav s pretvarjanjem v osnovne oblike besed. Morda bi bilo smiselno v leksikon vključiti še določene metapodatke, denimo podatek o intenziteti pozitivnosti oziroma negativnosti vnosa (npr. z besedo *odličen* lahko boljše identificiramo besedilo kot pozitivno, kot z besedo *dober*).

Kvalitete izdelanega leksikona nismo neposredno ocenjevali ampak smo zgolj merili njegov doprinos k uspešnosti same klasifikacije (več v 4 razdelku). Pri tem smo uporabili večinsko glasovanje. Besede v besedilu smo primerjali z vnosi v leksikonu. Celotno besedilo je bilo pozitivno, v kolikor so večinsko prevladovale besede s pozitivno konotacijo in obratno, besedilo je bilo negativno, v kolikor so večinsko zastopane besede z negativno konotacijo. V kolikor s pomočjo slovarja ni bilo moč pridobiti informacije o polariteti (npr. število pozitivnih besed je enako številu negativnih besed) smo tudi to informacijo uporabili pri klasifikaciji, saj se je izkazalo, da blagodejno vpliva na uspešnost klasifikacije. Dodatnih možnosti, kot je denimo vpeljava praga, s katerim bi nastavili, koliko pozitivnih oz. negativnih besed je potrebno, da ima besedilo s stališča leksikona polariteto, nismo preizkušali.

3 Zbirka spletnih komentarjev

Analiza sentimenta je z razvojem Spleta 2.0 doživela precejšen razmah, saj so raziskovalci dobili praktično neomejene možnosti analiziranja misli uporabnikov, ki jih dnevno delijo preko objav na socialnih omrežjih, (mikro)blogih ipd. Medtem, ko so bili uporabniki na začetku predvsem pregledovalci vsebin, so z razvojem spleta postali tudi njihovi aktivni ustvarjalci. Govorimo o pojmu uporabniško generiranih vsebin (UGV). Z razmahom Spleta 2.0 se je tako fokus raziskav sentimenta iz opisov filmov, produktov ter analize novic prestavil na UGV. Za razliko od formalnih besedil, kjer je pričakovana čistost, slovnična pravilnost ter malo pravopisnih napak, gre pri UGV predvsem za neformalna besedila s samosvojimi zakonitostmi, ki analizo sentimenta dodatno otežijo. Takšna besedila pogosto vsebujejo sarkazem, ironijo, slovnične napake, okrajšave, sleng ter emotikone, s katerimi avtorji še poudarijo svoja občutja o določeni temi. Besedila so pogosto krajša, kar je po eni strani zaradi jedrnatosti prednost, po drugi pa lahko že ena beseda identificira piščevo mnenje.

Raziskovalci se zadnje čase osredotočajo predvsem na analizo sentimenta UGV. Zelo popularna platforma za tovrstne raziskave je družbeno omrežje Twitter, kjer so objave oz. tviti omejeni na 140 znakov, kar uporabnike sili, da svoje misli posredujejo v neposredni, jedrnati obliki. Ker uporabniki običajno tvite oplemenitijo z oznakami (hashtagi), s katerimi primarno označijo temo, na katero se tvit sklicuje, je poenostavljeno tudi pridobivanje tvitov glede na željeno temo (npr. tviti z oznako #SLOprivatizacija zelo verjetno vsebujejo mnenja uporabnikov o slovenski privatizaciji). Vse to in preprost uporabniški vmesnik za dostop do zbirke tvitov so botrovali temu, da je na voljo precejšnje število prosto dostopnih korpusov za analizo sentimenta. Žal to velja le za večje svetovne jezike, kjer prevladuje angleščina. Prosto dostopnega korpusa za slovenski jezik nismo našli, zato smo se odločili, da za ovrednotenje leksikona sentimenta izdelamo svojega. Ena od zahtev je bila, da korpus vsebuje neformalna besedila. Poleg platforme Twitter so bili naravna izbira uporabniški komentarji slovenskih novičarskih portalov. Glede na to, da na platformi Twitter objavljajo tudi uradne entitete, kot so denimo podjetja, smo menili, da bo identifikacija subjektivnega besedila lažja. Hkrati pa smo se zavedali, da komentarji pogosto niso v skladu z novico (angl. offtopic) ter da je identifikacija mnenja težavnejša zaradi tega, ker se lahko v enem komentarju prepleta več različnih mnenj. Nasploh je analiza uporabniških komentarjev ena težjih nalog s tega področja.

Pri izdelavi korpusa smo uporabili spletne komentarje iz portalov 24ur, Finance, Reporter in RtvSlo. S pomočjo ključnih besed (npr. "trg nepremičnin raste", "begunska kriza"ipd.) in prilagojenih Googlovih iskalnih pogonov, s katerimi smo iskali izključno po naštetih spletnih portalih, smo avtomatizirano pridobili spletne povezave na novice ter iz njih izluščili komentarje. Popoln nabor uporabljenih iskalnih pogojev pri gradnju korpusa je objavljen v (Kadunc, 2016). Naj omenimo, da nekateri slovenski novičarski portali vodijo vse bolj restriktivno politiko komentiranja. V času izdelave korpusa tako portal Dnevnik ni več omogočal neposrednih komentarjev uporabnikov, SiolNet je po drugi strani hranil komentarje le za zadnjih 7 dni ter tako funkcionalnost komentiranja naredil povsem neuporabno za naš eksperiment. Skupaj smo pridobili 5087 uporabniških komentarjev iz 427 različnih strani omenjenih spletnih virov, v povprečju z vsake strani 11 komentarjev, od tega je bilo 4777 uporabnih (ostali so bili npr. v tujem jeziku ali pa so vsebovali le sliko). Komentarje smo uvrstili v eno od tematik: šport, politika, gospodarstvo in drugo.

Ko smo pridobili željene komentarje, jih je bilo potrebno označiti s sentimentnimi ocenami, z namenom pridobitve zadostnega števila primerov za učenje in testiranje klasifikatorjev. Za označevanje primerov je na voljo več tehnik, od povsem ročnih, do avtomatskih. V našem delu so bili vsi primeri označeni ročno, s strani človeških označevalcev. Komentarje so označevali trije označevalci (angl. annotators). Za označevanje smo poleg osnovnih sentimentnih kategorij *pozitivno*, *negativno* ter *nevtralno*, določili še *irelevantno*. S slednjo kategorijo smo želeli označiti ter tako iz končne verzije korpusa izločiti komentarje, ki za samo analizo niso relevantni, npr. vsebujejo le sliko, povezavo, so napisani v tujem jeziku ipd. Vsi označevalci so videli isti nabor komentarjev. Vsak komentar je bil označen natanko trikrat. Stopnjo strinjanja dveh označevalcev smo izmerili s statistično mero Cohen Kappa, ki upošteva tudi morebitno naključno strinjanje. Tako $\kappa = 1$ pomeni popolno strinjanje med označevalcema, $\kappa \leq 0$ pa ujemanje, ki ni večje od naključnega. Pri nas strinjanje med prvim in drugim označevalcem znaša $\kappa = 0,33$, med prvim in tretjim $\kappa = 0,31$ ter med drugim in tretjim $\kappa = 0,54$. Stopnjo strinjanja med vsemi tremi označevalci smo izračunali z mero Fleiss Kappa. Z doseženo vrednostjo $\kappa = 0,38$ smo po interpretaciji iz Landis in Koch (1977) dosegli ustrezno ujemanje, tako da bi moral biti korpus dovolj zanesljiv za uporabo. Vpliva zanesljivosti korpusa na uspešnost klasifikacije sicer nismo merili. To bi lahko naredili tako, da bi za učenje in testiranje klasifikatorjev vzeli samo tiste komentarje, pri katerih je bilo med označevalci popolno strinjanje . Omeniti je potrebno, da gre pri označevanju sentimenta za močno subjektivno nalogo, posledično je visoko stopnjo strinjanja med označevalci težko pričakovati.

Korpus je na voljo v uravnoteženi in neuravnoteženi obliki. Končna verzija korpusa brez uravnoteženja vsebuje 898 pozitivnih, 3291 negativnih ter 588 nevtralnih komentarjev, določenih z večinskim strinjanjem označevalcev. Podrobnejši razpored primerov po posameznih kategorijah novic in spletnih virih je prikazan v tabeli 1. Vidimo lahko, da so komentarji pretežno negativno nastrojeni, razen za kategorijo *šport*, kjer je razmerje med pozitivnimi in negativnimi približno uravnoteženo. Za učenje smo uporabljali tudi uravnoteženi korpus, pri katerem smo naključno izbrali po 580 komentarjev vsake vrste sentimenta. Korpus je prosto dostopen v obliki kompresirane datoteke ZIP, ki vsebuje korpus v formatu XML ter opis strukture dokumenta XML. Za XML smo se odločili zaradi velike razširjenosti uporabe tega formata, ki predstavlja de facto standard za izmenjavo podatkov preko interneta.

4 Evalvacija z metodami strojnega učenja

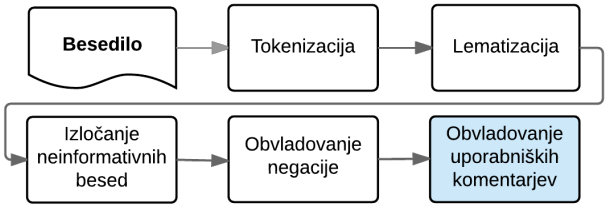
Sentimentni leksikon smo preizkusili v kontekstu strojnega učenja in predobdelave besedil. Izvedli smo vrsto ek-

	gospodarstvo	politika	šport	drugo	RtvSlo	24ur	Finance	Reporter	skupaj
pozitivno	129	26	679	64	566	255	54	23	898
nevtralno	262	33	240	53	441	48	75	24	588
negativno	1420	351	882	638	1614	584	554	539	3291
skupaj	1811	410	1801	755	2621	887	683	586	4777

Tabela 1: Razporeditev komentarjev po kategorijah novic (levo) in po spletnih virih (desno).

sperimentov, s katerimi smo merili vpliv predobdelave in izbire atributov na uspešnost klasifikacije. Zlati standard za naše eksperimentiranje je predstavljal uravnoteženi korpus z enakomerno zastopanostjo komentarjev vsake izmed treh sentimentnih kategorij. Za preverjanje uspešnosti klasifikacije smo uporabili 10-kratno prečno preverjanje. Vsi navedeni rezultati predstavljajo povprečje 10-ih iteracij. Kot klasifikatorje smo izbrali logistično regresijo (LR), metodo podpornih vektorjev (SVM), večvrednostni naivni Bayesov klasifikator (MNB) in binarni naivni Bayesov klasifikator (BNB). Navedeni klasifikatorji se pri analizi sentimenta tudi sicer največ uporabljajo. Kot mere uspešnosti smo izbrali klasifikacijsko točnost (CA) ter za vsako vrsto sentimenta posebej še mero F_1. Naj omenimo, da smo izvajali izključno trirazredno klasifikacijo, tj. klasifikacijo v eno izmed treh sentimentnih kategorij. Veliko sorodnih raziskav se namreč osredotoča le na kategoriji *pozitivno* in *negativno*. Z ozirom na uravnoteženi korpus vrednost CA pri 33,33% predstavlja spodnjo mejo še sprejemljive klasifikacije točnosti. Kot osnovo smo določili konfiguracijo, pri kateri smo za atribute izbrali posamezne besede (unigrami), skupaj jih je bilo 20.388. Z osnovno konfiguracijo smo dosegli vrednost CA 54,5%, kar je 21,2% nad vsakokratno izbiro večinskega razreda.

Najprej smo primerjali različne načine predobdelave besedil (diagram na sliki 1). S predobdelavo želimo besedilo pripraviti in ga očistiti. Pri uporabniških komentarjih lahko pričakujemo veliko šuma, zato je predobdelava še posebej pomembna. Preizkusili smo več načinov predobdelave, od bolj splošnih, ki so široko uporabljeni pri obdelavi naravnega jezika, kot so denimo tokenizacija in lematizacija, do bolj specifičnih, npr. obvladovanje negacije.



Slika 1: Koraki pri predobdelavi vhodnega besedila.

Najprej smo merili vpliv različnih načinov tokenizacije. Primerjali smo tokenizacijo s presledki, Treebank tokenizacijo in Pottsovo tokenizacijo, najboljše se je povečini izkazala slednja. Lematizacija je izboljšala, izločanje neinformativnih besed in enostavno obravnavanje negacije pa sta poslabšali delovanje vseh klasifikatorjev. Za uporabniške komentarje specifični načini predobdelave (npr. zamenjava spletnih povezav s pojavnico *URL* ali zamenjava emotikonov s sentimentno oceno) so povečini izboljšali klasifikacijo. Iz tabele 1 je razvidno, da smo s predobdelavo, v primerjavi z osnovno konfiguracijo, pridobili 8,7% pri CA. Naj omenimo, da smo poleg izboljšanja klasifikacije uspeli zmanjšati tudi število atributov, na 9.198, kar je več kot 50% manj kot pri osnovni konfiguraciji.

Klasifikator	CA	mera F_1			
		pos	*neg*	*neu*	povp.
osnova (LR)	54,5	57,6	51,5	54,3	54,5
LR	61,8	66,8	58,6	60,1	61,8
SVM	59,7	64,6	55,9	58,4	59,6
MNB	**63,2**	67,3	64,2	57,6	**63,0**
BNB	48,9	57,5	44,4	36,4	46,1

Tabela 2: Stanje konfiguracij po predobdelavi besedila. Podani so rezultati klasifikacijske točnosti ter mer F_1 za različne klasifikacijske metode. Najboljša konfiguracija je predstavljena z odebeljeno pisavo.

Nadalje smo preizkusili različne tehnike za izločanje, izbiro in uteževanje značilk. Z višjimi N-grami smo želeli preizkusiti, ali bi zajetje širšega konteksta, kot so npr. fraze, lahko pripomoglo k izboljšanju rezultatov. Če smo kot atribute dodali še bigrame, se je uspešnost nekoliko izboljšala le pri metodi SVM, kar je razvidno iz tabele 3. Dodajanje trigramov je poslabšalo rezultate vseh metod. Tudi z omejevanjem števila značilk glede na pogostost pojavitve v korpusu pri višjih N-gramih nismo dosegli večjih razlik. Da se unigrami zelo dobro obnesejo pri analizi sentimenta so pokazali že v raziskavi (Pang et al., 2002), kjer so analizirali opise filmov.

Model	LR	SVM	MNB	BNB
unigrami	**61,8**	59,7	**63,2**	**48,9**
unigrami + bigrami	61,2	**60,1**	59,0	43,3
bigrami	51,0	49,5	51,6	38,3
uni + bi + trigrami	60,6	59,4	56,4	39,4

Tabela 3: Vpliv različno visokih N-gramov na uspešnost klasifikacije. Odebeljeni so najboljši modeli za posamezno metodo.

Uteževanje značilk (pogostost, frekvenca, tf-idf) je dalo mešane rezultate (tabela 4); tako je pogostost koristila klasifikatorju MNB, uteževanje s tf-idf pa SVM, ki se je zelo približal ostalima dvema metodama. Naj navedemo, da je Smailović (2014) v sorodni raziskavi pri klasifikaciji tvitov s klasifikatorjem SVM prišla do zaključka, da se preprostejša utež tf obnese bolje od tf-idf.

Izbira podmnožice pomembnih atributov z metodo *Hi-kvadrat* je koristila klasifikatorju BNB, kjer smo dosegli izboljšanje CA za 3%. Pri ostalih večjega uspeha nismo zabeležili.

Vektorizacija značilk	LR	SVM	MNB
prisotnost (dvojiška vrednost)	**62,9**	60,2	62,1
pogostost (štetje)	61,8	59,7	**63,2**
utež TF	58,7	61,0	56,6
utež TF-IDF	61,7	**62,6**	61,3

Tabela 4: Primerjava načinov vektorizacije značilk. BNB smo izpustili zaradi definicije Bernoullijevega modela.

V klasifikacijskih modelih smo leksikalno znanje uporabili na način, da smo primerjali besede v besedilu z vnosi v leksikonu ter dodali ustrezno značilko. V kolikor so prevladovale besede s pozitivno konotacijo smo dodali značilko *oznakaSlovarja_POS*, v kolikor so prevladovale negativne besede značilko *oznakaSlovarja_NEG* in *oznakaSlovarja_NEU* za primere, ko iz leksikona ni bilo mogoče ugotoviti polaritete besedila. Rezultate združevanja različnih leksikonov in klasifikatorjev prikazuje tabela 5. Najboljše se je obnesel naš leksikon (KSS). Z njim smo dosegli zaznavno izboljšanje klasifikacije pri vseh metodah strojnega učenja, v primeru klasifikatorja MNB za 1,3%. Z leksikonom General Inquirer (GIS) smo dobili mešane rezultate. Najslabše se je odrezal leksikon avtomatsko pridobljen iz kombinacije SentiWordNeta in SloWNeta (SWN), razloge za to gre lahko iskati v tem, da SWN različne pomene iste besede točkuje z različnimi sentimentnimi ocenami. Za učinkovito izrabo leksikona SWN bi bilo potrebno vključiti sistem razdvoumljanja večpomenskih besed, kar pa bi znalo predstavljati težavo, saj v SloWNetu manjkajo glose in primeri uporabe za precejšnje število vnosov. Tudi z vključitvijo vseh leksikonov skupaj v povprečju nismo uspeli bistveno izboljšati rezultatov leksikona KSS.

Model	LR	SVM	MNB	BNB
unigrami	61,8	59,7	63,2	48,9
unigrami + KSS	**62,9**	**60,6**	64,5	49,8
unigrami + GIS	61,5	59,5	64,4	49,4
unigrami + SWN	61,0	59,8	63,4	49,2
vsi skupaj	62,2	60,5	**65,2**	**50,3**

Tabela 5: Vpliv leksikonov sentimenta na klasifikacijo. Podani so rezultati klasifikacijske točnosti za različne klasifikacijske metode pri različnih sentimentnih leksikonih. Odebeljeni so najboljši rezultati za vsako metodo.

Najboljša konfiguracija, ki vključuje vse koristne predobdelave, leksikone sentimenta in uporablja večvrednostni naivni Bayesov klasifikator, doseže na uravnoteženem besedilu 65,5% klasifikacijsko točnost, kar predstavlja 11% izboljšanje glede na osnovno konfiguracijo in 32,2% izboljšanje v primerjavi z vsakokratno izbiro večinskega razreda. Poglejmo si še primerjavo med najboljšima klasifikatorjema. Razlika v CA znaša slaba 2%, standardni odklon ali standardna deviacija (σ) za logistično regresijo znaša 3,79%, za večvrednostni naivni Bayesov klasifikator je bil izmerjen σ = 3,75%. Glede na to, da smo delali na splošnem klasifikacijskem modelu, smo z rezultati zadovoljni. Za klasifikatorja MNB in LR smo z uporabo Wilcoxonovega testa (Demšar, 2006) izračunali še statistično značilnost razlik. Pri stopnji značilnosti α = 0,05 ničelne hipoteze, da med klasifikatorjema ni razlik, ne moremo zavrniti.

Klasifikator	CA	mera F_1			
		pos	*neg*	*neu*	povp.
osnova	54,5	57,6	51,5	54,3	54,5
LR	63,6	68,1	61,3	61,6	63,7
SVM	63,2	69,0	62,1	58,6	63,2
MNB	**65,5**	68,6	66,8	60,6	**65,3**
BNB	60,1	65,0	56,7	58,4	60,0

Tabela 6: Izbira najboljše metode strojnega učenja za klasifikacijo komentarjev na uravnoteženem korpusu.

Najboljšo konfiguracijo smo preizkusili še v kontekstu neuravnoteženega korpusa. Porazdelitev komentarjev po razredih v tem primeru je: 588 nevtralnih, 898 pozitivnih ter 3291 negativnih komentarjev. Z vsakokratno izbiro večinskega razreda bi dobili 68,9% CA. Model, ki smo ga zgradili na neuravnoteženem korpusu doseže 76,2% klasifikacijsko točnost, mera F_1 na pozitivnih primerih daje vrednost 60,0%, na negativnih pa 85,4%.

4.1 Primerjava z rezultati raziskav za druge jezike

V preteklih letih je bilo narejenih veliko raziskav za večje svetovne jezike, predvsem angleščino. Rezultate raziskav, celo znotraj istega jezika, je medsebojno težko neposredno primerjati. Upoštevati je potrebno več faktorjev, ki lahko bistveno vplivajo na interpretacijo in primerjavo rezultatov posameznih raziskav. Med njimi so vrsta besedila, razrednost klasifikacije, uporabljen korpus ipd. Za angleški jezik, v obliki spletnih storitev, obstaja nekaj javno dostopnih splošnih klasifikatorjev sentimenta, kot je denimo AlchemyAPI[1]. Če bi delali na analizi sentimenta angleških besedil, bi lahko na testnih podatkih klasifikatorje preizkusili in okvirno ocenili, kako se naš klasifikator primerja z drugimi. Žal za slovenski jezik ni na voljo tovrstnih, prosto dostopnih storitev.

Vrsta besedila predstavlja pomemben dejavnik pri vrednotenju rezultatov. V času pred ekspanzijo (mikro)blogov so bili med raziskovalci priljubljeni opisi filmov in raznih produktov. Pri opisih filmov so raziskovalci dosegli rezultate, primerljive z rezultati kategorizacije novic (šport, gospodarstvo ipd.). Abbasi et al. (2008) so v raziskavi nad korpusom opisov filmov (Pang et al., 2002), ki je bil predmet številnih raziskav, pri dvorazredni klasifikaciji dosegli 91,7% CA. Takšnih rezultatov pri bolj neformalnih besedilih (tviti, uporabniški komentarji ipd.) klasifikatorji ne dosegajo. Prav tako uspešnost klasifikacije upade, če poleg kategorij pozitivno in negativno, rešujemo še problematiko nevtralnosti besedila. Smailović (2014) je primerjala nekaj prosto dostopnih klasifikatorjev na množici ročno označenih testnih besedil in pri nekaterih v primeru trirazredne klasifikacije zmogljivosti precej upadejo. Poglejmo nekaj sorodnih raziskav v angleškem jeziku, ki se

[1] http://www.alchemyapi.com/.

osredotočajo na trirazredno analizo sentimenta neformalnih besedil.

Agarwal et al. (2011) so analizo sentimenta izvajali nad 1709 ročno označenimi tviti, enakomerno razporejenimi po vseh treh kategorijah. Z najboljšo konfiguracijo so dosegli 60,83% CA, kar je nekaj slabše kot v primeru našega klasifikatorja. Kot zanimivost naj omenimo, da je tudi pri njih kot osnova služila konfiguracija z unigrami, s katero so dosegli 56,58% CA. V primerjavi z našo raziskavo jim je torej uspel manjši dvig uspešnosti najboljše konfiguracije glede na osnovo.

Prav tako so nad označenimi tviti analizo sentimenta izvajali Hamdan et al. (2013). Z najboljšo konfiguracijo so uspeli doseči 58,87% CA. Raziskava je zanimiva, ker so preizkusili sentimentni leksikon SentiWordNet in, kot mi, prišli do zaključka, da lahko leksikalni viri izboljšajo klasifikacijo.

V okviru tekmovanja SemEval-2013 so v kategoriji analize sentimenta sporočil v sistemu Twitter najboljši klasifikator izdelali Kiritchenko et al. (2014). Dosegli so povprečje mer F_1 pri 69,02% (uporabljen je bil neuravnotežen korpus tvitov, za primerjavo klasifikatorjev so uporabili povprečje mer F_1, klasifikacijske točnosti niso merili). Zmagoviti klasifikator temelji na kombinaciji metod strojnega učenja in intenzivni rabi različnih leksikonov, med drugim so vključili tudi Hujev leksikon sentimenta, ki je tudi nam služil kot osnova za izdelavo slovenskega leksikona sentimenta.

V zgoraj naštetih in mnogih drugih raziskavah se CA pri analizi sentimenta neformalnih angleških besedil giblje med 60% in 70%. Tudi naši rezultati na slovenskih besedilih so na tem intervalu, zato menimo, da smo lahko z njimi zadovoljni.

5 Zaključki

V prispevku smo predstavili rezultate analize sentimenta uporabniških komentarjev v slovenskem jeziku z uporabo leksikalnih virov v kontekstu nadzorovanega učenja. Sklenemo lahko, da leksikalni viri pozitivno vplivajo na analizo sentimenta z uporabo metod strojnega učenja. Najboljša konfiguracija, ki smo jo preizkusili, bistveno preseže klasifikacijo v večinski razred in vse osnovne konfiguracije. Leksikon, ki smo ga izdelali, je javno dostopen[2] in lahko koristno služi pri nadaljnih analizah mnenj v slovenskem jeziku. Možne so še številne izboljšave, predvsem v povezavi s kakovostnimi viri, kot so enojezični in večjezični slovarji ter avtomatsko določanje sentimentnih besed za posamezne kontekste.

Klasifikacijske modele bi lahko še dodatno izboljšali z uvedbo naprednih tehnik, kot je denimo uporaba bolj ali manj zahtevnih lingvističnih pravil za obravnavanje negacije (npr. negacija besede z negativno konotacijo sentiment spremeni v pozitiven). Običajno tovrstne raziskave vključujejo tudi oblikoslovno označevanje. V našem delu smo ga v celoti prezrli, tako da je to dobra iztočnica za nadaljne delo. Z oblikoslovnim označevanjem se odpira vrsta dodatnih možnosti pri pripravi značilk, npr. omejitev unigramov na pridevnike. Za izboljšanje leksikona SWN bi potrebovali sistem za razdvoumljanje večpomenskih besed. Prav tako so odprte možnosti pri izboljšavah korpusa uporabniških komentarjev. Korpus bi lahko razširili z novimi primeri, nekateri tuji tovrstni korpusi vsebujejo tudi več 10 tisoč označenih primerov. Kot je razvidno iz tabele 1, smo zadovoljivo pokrili predvsem kategorijo *šport*. Uporabniške komentarje bi lahko črpali iz raznovrstnejših spletnih virov, morda v korpus dodali tudi objave iz socialnih omrežij ipd.

[2]Korpus označenih uporabniških komentarjev ter slovenski leksikon sentimenta sta dostopna na naslovu http://lkm.fri.uni-lj.si/rmarko/repozitorij/opinionLexicon.

6 Literatura

Ahmed Abbasi, Hsinchun Chen in Arab Salem. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions On Information Systems*, 26(3).

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow in Rebecca Passonneau. 2011. Sentiment analysis of twitter data. V: *Proceedings of the Workshop on Languages in Social Media*, LSM '11, str. 30–38, Stroudsburg, PA, USA. Association for Computational Linguistics.

Stefano Baccianella, Andrea Esuli in Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. V: *Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, zvezek 10, str. 2200–2204.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30, December.

Darja Fišer. 2008. Using multilingual resources for building SloWNet faster. V: *The Fourth Global WordNet Conference*, str. 185–193.

Hussam Hamdan, Frederic Béchet in Patrice Bellot. 2013. Experiments with DBpedia, WordNet and SentiWordNet as resources for sentiment analysis in micro-blogging. V: *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, str. 455–459, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Minqing Hu in Bing Liu. 2004. Mining opinion features in customer reviews. V: *Proceedings of AAAI Conference on Artificial Intelligence*, zvezek 4, str. 755–760.

Matjaž Juršič. 2007. Implementacija učinkovitega sistema za gradnjo, uporabo in evaluacijo lematizatorjev tipa RDR. Univerza v Ljubljani, Fakulteta za računalništvo in informatiko. Diplomsko delo.

Klemen Kadunc. 2016. Določanje sentimenta slovenskim spletnim komentarjem s pomočjo strojnega učenja. Univerza v Ljubljani, Fakulteta za računalništvo in informatiko. Diplomsko delo.

Svetlana Kiritchenko, Xiaodan Zhu in Saif M. Mohammad. 2014. Sentiment analysis of short informal texts. *J. Artif. Int. Res.*, 50(1):723–762, May.

J. Richard Landis in Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1).

Bing Liu in Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. V: *Mining text data*, str. 415–463. Springer.

Rok Martinc. 2013. Merjenje sentimenta na družabnem omrežju Twitter: izdelava orodja ter evaluacija. Univerza v Ljubljani, Fakulteta za družbene vede. Magistrsko delo.

Finn Årup Nielsen. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. V: *Proceedings of the 8th European Semantic Web Conference Workshop on 'Making Sense of Microposts': Big things come in small packagess*, str. 93–98. http://arxiv.org/abs/1103.2903.

Bo Pang, Lillian Lee in Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. V: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, str. 79–86.

Christopher Potts. 2011. Sentiment symposium tutorial: Lexicons. http://sentiment.christopherpotts.net/lexicons.html.

Jasmina Smailović. 2014. *Sentiment analysis in streams of microblogging posts*. Doktorsko delo, International postgraduate school Jožef Stefan, Ljubljana, Slovenia.

Philip J Stone. 1997. Thematic text analysis: New agendas for analyzing text content. V: Carl Roberts, ur., *Text Analysis for the Social Sciences*, str. 33–54. Lawrence Erlbaum Associates Publishers, Mahwah, NJ.

Mateja Volčanšek. 2015. Leksikalna analiza razpoloženja za slovenska besedila. Univerza v Ljubljani, Fakulteta za računalništvo in informatiko. Diplomsko delo.