Napovedovanje spola slovenskih blogerk in blogerjev

Iza Škrjanec, Nada Lavrač, Senja Pollak

Izvleček

Napovedovanje spola avtorjev je zanimiv raziskovalni problem, izdelava napovednih modelov za razpoznavo spola pa je koristna za uporabo v različnih aplikacijah, npr. na področju trženja in analize kupcev. Cilj naše raziskave je razvoj in evalvacija napovednih modelov za avtomatsko razpoznavo spola avtorjev in avtoric slovenskih blogovskih zapisov. Za to nalogo uporabimo množico blogov 177 blogerjev in 96 blogerk, kot posamezno enoto klasifikacije pa upoštevamo vsa besedila posameznega avtorja v korpusu, združena v eno enoto. V prispevku primerjamo dva tipa modelov za napovedovanje spola avtorja besedil: model z ročno zgrajenimi pravili in model, zgrajen z metodami strojnega učenja. Modeli s pravili upoštevajo rabo slovničnega spola v delih besedila, v katerih se avtor nanaša nase. Za izgradnjo modelov s strojnim učenjem pa smo preizkusili več algoritmov in tipov značilk. Oba tipa modelov dosežeta klasifikacijsko točnost nad 85 %, najuspešnejši pa je model strojnega učenja, naučen na unigramih pojavnic s pomočjo metode podpornih vektorjev. Analiza najbolj informativnih značilk tega modela je pokazala, da se besedila blogerk in blogerjev razlikujejo na slovnični ravni (raba slovničnega spola in zaimkov), izbiri teme besedila (npr. večji poudarek na družini, ljubezni in spolnosti pri blogerkah) in slogovnih značilnostih (npr. raba kletvic v besedilih blogerjev).

Ključne besede: profiliranje avtorjev, spol, družbeni mediji, klasifikacija blogov

1 UVOD

Jezik je družbeni pojav in kot tak podvržen variaciji in spremembam. Med družbenimi dejavniki variacije sociolingvisti preučujejo tudi spol govorcev glede na jezikovne prakse, ki jih uporabljajo ženske in moški. Napredek v obdelavi naravnega jezika raziskovalcem omogoča, da modelirajo jezikovno variacijo v obsežnih besedilnih korpusih in z uporabo avtomatskih pristopov.

V prispevku primerjamo jezik moških in žensk s pristopom gradnje modelov za avtomatsko razpoznavo spola avtorjev besedil. Za to uporabimo podkorpus slovenskih blogovskih zapisov, ki so bili v okviru raziskovalnega projekta JANES zbrani in ročno označeni s podatkom o spolu avtorja.

Profiliranje avtorjev besedil glede na njihov spol je aktualen raziskovalni problem. Ena prvih odmevnih raziskav s tega področja je primerjala jezik govork in govorcev na podlagi Britanskega nacionalnega korpusa (BNC) in pokazala, da je v slogu pisanja žensk več poudarka na medosebni komunikaciji (npr. z rabo zaimkov), medtem ko je za moške med drugim bolj značilna raba členkov, njihov slog pisanja pa je informativne narave (Koppel et al. 2002). Prve raziskave avtomatske razpoznave spola avtorja so analizirale besedila v angleščini, kmalu pa so bili v področje raziskav vključeni tudi drugi jeziki, še posebej na podlagi virov iz družbenih medijev (Schler et al. 2006, Schwartz et al. 2013, Plank in Hovy: 2015, Peersman et al. 2011, Nguyen et al. 2013, Verhoeven et al. 2016, Ljubešić et al. 2017). Profiliranje avtorjev na podlagi spola je tudi ena od kategorij v sklopu vsakoletnega tekmovanja PAN (prim. Rangel et al. 2017).

Med jezike, ki jih pokrivajo raziskave iz računalniške stilometrije in profiliranja avtorjev, spada tudi slovenščina. Zwitter Vitez (2011, 2013) je predstavila prvo raziskavo o ugotavljanju avtorstva, in sicer je identificirala najverjetnejšega avtorja anonimnega besedila, ki je bilo objavljeno na uradni spletni strani ene od slovenskih parlamentarnih strank. Anonimno besedilo je primerjala z besedili potencialnih avtorjev s pomočjo vrste leksikalnih in berljivostnih značilk. Raziskovalni projekt JANES je prinesel nove možnosti za profiliranje avtorjev spletnih uporabniških vsebin, saj je korpus Janes opremljen z metapodatki o spolu in vrsti računa uporabnikov in uporabnic (Erjavec et al. 2018). V podkorpusu tvitov in blogov so ti metapodatki pripisani ročno. Verhoeven et al. (2017) so razvili model za identifikacijo spola avtorjev tvitov iz korpusa Janes in rezultate primerjali s podobnimi modeli, ki so bili naučeni na nemških, nizozemskih, italijanskih, španskih, francoskih in portugalskih tvitih (Verhoeven et al. 2016). Martinc et al. (2017) so razvili klasifikatorje za profiliranje avtorjev besedil v različnih jezikih, razviti modeli za določanje spola avtorjev tvitov pa so, skupaj z drugimi orodji za procesiranje nestandardne slovenščine, na voljo tudi v obliki spletnih delotokov

(Martinc et al. 2018). Z vidika spola avtorja pa sta bila podkorpusa tvitov in blogov analizirana v magistrskem delu Škrjanec (2017), na katerem je osnovano to poglavje.

V pričujočem poglavju uporabimo metode, s katerimi raziščemo, ali lahko na podlagi jezikovne variacije med spoloma avtomatsko razlikujemo med avtoricami in avtorji blogovskih zapisov. V poglavju predstavimo dva tipa napovednih modelov za določanje spola avtorja, in sicer model na podlagi ročno zgrajenih pravili in model, zgrajen z metodami strojnega učenja. Modeli s pravili upoštevajo rabo slovničnega spola v primerih avtorjevega samonanašanja v glagolskih zvezah, torej v sestavljenih glagolskih zvezah, pri katerih je pomožni glagol v prvi osebi ednine, deležnik na -l pa ima žensko ali moško obliko. Ti modeli predstavljajo osnovo za primerjavo s kompleksnejšimi in časovno bolj zahtevnimi modeli strojnega učenja, pri gradnji katerih smo eksperimentirali z različnimi značilkami in algoritmi.

Poglavje vsebuje sledeče razdelke. V drugem razdelku predstavimo korpus blogov, ki smo ga uporabili za analizo jezikovne variacije. Tretji razdelek opisuje metodologijo za gradnjo napovednih modelov. Četrti razdelek poda opis rezultatov klasifikacije, v petem pa se posvetimo napakam modela s pravili ter analiziramo tiste značilke izbranega modela strojnega učenja, ki imajo večjo težo pri klasifikaciji spola avtorja dokumenta. Šesti razdelek opiše sklepne ugotovitve in poda nekaj idej za prihodnje delo.

2 KORPUS BLOGOV

V raziskavi smo uporabili podkorpus blogov Janes-Blog (Erjavec et al. 2018), ki vsebuje bloge s portalov publishwall.si (18.515 besedil 615 uporabnikov) in rtvslo.si (23.515 blogov 243 uporabnikov), objavljene med oktobrom 2006 in januarjem 2016. Kot je podrobno opisano v Erjavec et al. (2018), je korpus bogato jezikoslovno označen, opremljen pa je tudi s številnimi dragocenimi metapodatki o besedilih (stopnja jezikovne in tehnične standardnosti, sentiment in jezik besedila) in njihovih avtorjih (tip uporabniškega računa, spol).

Glede na to, da smo želeli v eksperimentih izvesti binarno klasifikacijo avtorjev in avtoric, smo v analizo vključili le zasebne račune, ki imajo pripisano oznako ženskega ali moškega spola, torej smo izpustili korporativne račune in uporabnike z nedoločenim spolom. Poleg tega smo upoštevali le tiste uporabnike, ki so objavili vsaj 10 blogovskih zapisov v slovenščini. Končni podkorpus za eksperimente vsebuje 28.697 blogovskih zapisov skupno 273 avtorjev (od tega 64,84 % moških in 35,16 % žensk), kot prikazuje Tabela 1. Vsa slovenska besedila posameznega avtorja smo združili v en dokument, s čimer se uvrščamo med pristope profiliranja

na ravni uporabnika (Stamatatos 2009), ki ga uporabljajo tudi pri zasnovi tekmovanja PAN (Rangel et al. 2017); pri alternativnem pristopu pa se določa spol za vsako posamezno besedilo. Za eksperimente smo uporabili tako nelematizirana kot lematizirana besedila.

Tabela 1: Velikost učne množice blogov.

	Uporabniki	Blogovski zapisi	Pojavnice
Ženske	157	9.056	3.393.315
Moški	275	20.105	8.362.668
Ženske (10 ≥ besedil)	96	8.874	3.124.734
Moški (10 ≥ besedil)	177	19.823	6.968.164

3 METODOLOGIJA

Problem napovedovanja spola avtorja smo formulirali kot problem klasifikacije besedil, ki smo ga naslovili z gradnjo dveh tipov napovednih modelov: model napovedovanja s pravili in model z algoritmi strojnega učenja. Z vidika klasifikacije je naloga posameznemu avtorju pripisati razred, to je moški oz. ženski spol. V tem razdelku predstavimo njuno gradnjo in delovanje. Prvi pristop zahteva jezikovno znanje za določanje pravil, vendar pa je pristop z vidika razvoja in uporabe hitrejši. Za strojno učenje je potrebna velika množica ročno označenih dokumentov, ne potrebujemo pa nobenega jezikovnega znanja, saj se algoritmi značilnosti naučijo sami na podlagi primerov. V našem primeru modeli z ročno zgrajenimi pravili temeljijo na izražanju slovnično spola v samonanašanju, v modelih, naučenih s strojnim učenjem, pa so potencialno odločujoče vse uporabljene besede, model pa sam na podlagi učnih primerov pripiše težo posameznim značilkam (besedam, znakom oz. njihovi kombinaciji).

3.1 Modeli s pravili

Modeli s pravili uporabljajo ročno zgrajena klasifikacijska pravila za pripisovanje razreda avtorjem. Klasifikacijska pravila upoštevajo referenčni spol, o katerem sklepamo na podlagi rabe slovničnega spola v glagolskih zvezah. Referenčni spol je odvisen od tega, na koga se določena jezikovna oblika (npr. samostalnik ali zaimek) nanaša izven besedila (Motschenbacher 2010). Modeli s pravili so zgrajeni pod poenostavljeno predpostavko, da raba slovničnega spola odraža referenčni spol in da se slednji ujema s spolom avtorja. Model pri tem upošteva dele

besedila, v katerih se avtor nanaša sam nase. Tako raba ženskega spola v samonanašanju implicira, da gre za avtorico, medtem ko na podlagi moškega spola model predvideva, da gre za avtorja moškega spola.[1]

Modeli s pravili preverjajo rabo spola v samonanašalnih glagolskih zvezah. Upoštevajo rabo spola v glagolskih deležnikih na -l, in sicer pravila poiščejo pojavitve oblik pomožnega glagola: *sem*, *nisem*, *bom* in oblik z nestandardnim črkovanjem *sm* in *nism*. Razviti program v besedilu preveri, ali besede v okolici teh pomožnih glagolov nosijo oznako za spol, pri čemer pregleda okolico v velikosti dva (torej zadnji dve besedi tik pred pomožnim glagolom in prvi dve besedi po pojavitvi pomožnega glagola). Zanima nas končnica okoliških besed. Za vsakega avtorja izračunamo indikator za ženski ali moški spol na podlagi končnic okoliških besed: končnica *-la* signalizira žensko obliko deležnika na -l (npr. *naredi**la** sem*), medtem ko končnice *-al*, *-il* in *-el* nakazujejo moško obliko deležnikov (npr. *bom še vid**el***). Vsaka tovrstna pojavitev doda eno točko indikatorju za ženski oz. moški spol posameznega avtorja. Na koncu primerjamo vrednosti obeh indikatorjev: če večina najdenih indikatorjev (70 % ali več) spada k enemu razredu (ženski ali moški), pravila avtorju pripišejo ta razred. Poleg tega smo določili minimalno število indikatorjev na avtorja in pri tem primerjali uspešnost modela, če nastavimo minimalni prag na tri ali pet indikatorjev. V primeru, da besedila avtorja vsebujejo premalo indikatorjev ali pa je razmerje med obema razredoma preveč enakovredno, model avtorju pripiše razred *nedoločeno*.

Fišer et al. (2016) so opisali metodo, s katero so avtomatsko označili avtorje v korpusu Janes z oznako za spol. Tudi njihova metoda je osnovana na pravilih glede na rabo slovničnega spola, vendar so avtorja za razliko od našega pristopa, ki uporablja besedne oblike, uporabili oblikoskladenjske oznake, s katerimi so poiskali povedi, ki vsebujejo pomožni glagol in deležnik na -l. Podobno kot mi so tudi oni šteli indikatorje za ženski in moški spol. Če je bilo razmerje enih do drugih večje od 0,7, je model pripisal žensko oz. moško oznako, sicer je bil uporabnik uvrščen v razred nedoločenih. Poleg tega so dodali pogoj, da mora vsaj 1 % besedil uporabnika vsebovati indikatorje za spol, sicer je bil uporabnik uvrščen med nedoločene. S to metodo so uspešno označili 78 % blogerk in blogerjev. Na ta način so označili vse blogerje v korpusu, mi pa smo uporabili podmnožico blogerjev in blogerk (glej razdelek 2 in Tabelo 1). Njihov model prav tako ne vključuje pogoja glede števila indikatorjev v vseh besedilih. Zaradi teh razlik rezultati med našim modelom in modelom v Fišer et al. (2016) niso popolnoma primerljivi, vendar nas kljub temu zanima, kateri od obeh modelov z ročno grajenimi pravili je uspešnejši in primernejši za avtomatsko napovedovanje spola blogerjev.

1 Potrebno je poudariti, da za namene gradnje klasifikacijskih modelov predvidevamo, da je spol avtorja binaren razred (ženski ali moški). Vprašanje spolne identitete in samoidentifikacije posameznih avtorjev je relevantno na področju profiliranja avtorjev, vendar presega obseg tega poglavja.

3.2 Pristop z metodami strojnega učenja

V tem razdelku opišemo gradnjo klasifikacijskih modelov za določanje spola avtorja. Predstavimo pripravo besedil in eksperimente s tremi različnimi algoritmi strojnega učenja. Za gradnjo modelov in luščenje značilk smo uporabili knjižnico Scikit-learn (Pedregosa et al. 2011).

3.2.1 Priprava podatkov in učenje modelov

V enoto za klasifikacijo smo združili vse posamezne blogovske zapise enega avtorja v en skupni dokument, če so ti ustrezali kriterijem števila dokumentov na avtorja (glej razdelek 2). Vsako enoto smo predstavili v obliki vektorja značilk. Za model predstavitve smo izbrali vrečo besed (angl. *bag-of-words* ali BOW), v katerih enota besedila (npr. beseda, znak ali n-gram) predstavlja eno značilko. V eksperimentih smo kot značilke preizkusili posamezne besede ter besedne in znakovne n-grame (n-gram je enota n zaporednih besed ali znakov). Za primer vzemimo poved »*To pa je bila top sprostitev.*«, iz katere lahko zgradimo naslednje besedne n-grame dolžine ena (unigrame): »*To*«, »*pa*«, »*je*«, »*bila*«, »*top*«, »*sprostitev*«, ».«. Iz iste povedi dobimo naslednje besedne bigrame: »*To pa*«, »*pa je*«, »*je bila*«, »*bila top*«, »*top sprostitev*«, »*sprostitev.*«. Na podoben način, kot delimo poved na besedne n-grame, lahko besede razdelimo na znakovne n-grame. Iz besede »*bila*« tako lahko dobimo štiri znakovne unigrame (»*b*«, »*i*«, »*l*« in »*a*«) ali tri znakovne bigrame (»*bi*«, »*il*« in »*la*«). V svojih poskusih smo uporabili besedne uni- in bigrame ter znakovne tri- in tetragrame.

Za primerjavo lahko uporabimo različne metode za uteževanje značilk, npr. pogostost besede v posameznem dokumentu (angl. *term frequency*), binarno utež prisotnosti besede v besedilu ali pa mero TF-IDF (angl. *term frequency–inverse document frequency*). Mera TF-IDF upošteva pogostost besede v dokumentu in inverzno pogostost besede v celotnem korpusu, s čimer mera pripiše manjšo utež besedam, ki so pogoste v celotnem korpusu, ter večjo mero besedam, ki so bolj značilne za nekatere, ne pa vse dokumente (Kobayashi 2007). Z mero TF-IDF bi lahko uteži tako besedne kot tudi znakovne n-grame.

Za učenje modela, ki klasificira besedila glede na spol avtorja, smo testirali in med seboj primerjali tri algoritme strojnega učenja: naivni Bayesov klasifikator, logistično regresijo in metodo podpornih vektorjev (angl. *support vector machine*, v nadaljevanju SVM).

Vektorji značilk, ki predstavljajo besedila, so lahko precej veliki. Da bi zmanjšali velikost prostora značilk, prihranili čas za pripravo podatkov in morda izboljšali

točnost klasifikacijskega modela, pogosto uporabimo katero od metod za izbor značilk (Dhillon 2004). Pred učenjem modela smo število značilk najprej zmanjšali glede na število pojavitev in tako ohranili le tiste značilke, ki se pojavijo v besedilih najmanj 5 in največ 218 (80 %) blogerjev. Poleg tega smo uporabili metodo *SelectFromModel*[2] iz knjižnice Scikit-learn. Med učenjem algoritem vsaki značilki pripiše določen koeficient, metoda *SelectFromModel* pa iz prostora značilk odstrani tiste značilke, pri katerih je vrednost koeficienta nižja od določene vrednosti, pri čemer smo za učenje uporabili samodejne nastavitve praga. Pri modelih, naučenih z Bayesovim algoritmom, je ta koeficient logaritem verjetnosti značilke glede na razred. Modeli na podlagi logistične regresije in metode podpornih vektorjev pa koeficient pripišejo s pomočjo odločitvene funkcije.

Zgrajene klasifikacijske modele primerjamo glede na klasifikacijsko točnost. Ko gradimo napovedni model, nas zanima, kako uspešno lahko model napove razred na neznanih primerih, ki jih nismo uporabili v procesu učenja. Uspešnost modela ocenimo z uporabo prečnega preverjanja (Witten in Frank 2005). Pri tem postopku razdelimo učno množico na k delov, od katerih k-1 dele uporabimo za učenje, del, ki smo ga izpustili, pa uporabimo za testiranje. Proces ponovimo k-krat. V poskusih modele ocenimo in primerjamo z 10-kratnim prečnim preverjanjem in poročamo o aritmetični sredini in standardnem odklonu klasifikacijskih točnosti.

4 REZULTATI NAPOVEDOVANJA SPOLA BLOGERJEV

V tem razdelku predstavimo uspešnost za napovedovanje spola blogerjev, in sicer najprej poročamo o točnosti modelov s pravili, nato pa se osredotočimo na klasifikacijske modele strojnega učenja.

4.1 Modeli s pravili

Modeli s klasifikacijskimi pravili na podlagi spola deležnikov na -l klasificirajo blogerje v enega od treh razredov: *ženski*, *moški* ali *nedoločeno*. Tabela 2 predstavi klasifikacijsko točnost modela s pravili glede na to, katere oblike (standardne ali nestandardne skupaj z nestandardnimi) pomožnega glagola smo uporabili, da smo prepoznali glagolsko zvezo. Klasifikacijsko točnost smo izračunali tako, da smo preverili, koliko blogerjev in blogerk je model pravilno klasificiral glede na ročno oznako, avtorje, ki jih je model uvrstil v razred nedoločenih, pa razumemo kot nepravilno klasificirane primere. Tabela 2 vsebuje točnost glede na minimalno

2 Opis metode *SelectFromModel* je opisan na povezavi http://scikit-learn.org/stable/modules/feature_selection.html.

število indikatorjev. Kot lahko razberemo iz tabele, upoštevanje nestandardnega zapisa pomožnega glagola ne spremeni klasifikacijske točnosti. Najbolj točen je model, pri katerem mora imeti besedilo vsaj tri indikatorje za spol, točnost tega modela pa znaša 85,71 %. Za primerjavo tabela vključuje tudi rezultate za večinski klasifikator, torej rezultat, ki ga dobimo, če vsem dokumentom pripišemo večinski moški spol, ki ga najboljši pristop s pravili preseže za 21 %.

Tabela 2: Rezultati klasifikacije modela s pravili.

Minimalno št. indikatorjev	Pomožni glagol	Klasifikacijska točnost (%)
3	sem, nisem, bom	**85,71**
3	sem, nisem, bom, sm, nism	**85,71**
5	sem, nisem, bom	80,95
5	sem, nisem, bom, sm, nism	80,95
Večinski klasifikator		64,84

4.2 Modeli strojnega učenja

V tem razdelku predstavimo rezultate klasifikacije besedil z metodami strojnega učenja. Testirali smo tri različne algoritme (metoda podpornih vektorjev, logistična regresija in naivni Bayesov klasifikator). Preizkusili smo tudi več tipov značilk: besedne uni- in bigrame, znakovne bi-, tri- in tetragrame ter unijo teh značilk. V poskuse smo vključili tako nelematizirano kot lematizirano verzijo korpusa.

Tabela 3 prikazuje povprečje in standardni odklon točnosti pri 10-kratnem prečnem preverjanju. Tabela vključuje rezultate vseh treh algoritmov glede na različne značilke in obliko besedila (pojavnice pomenijo nelematizirano besedilo).

Tabela 3: Povprečna klasifikacijska točnost ± standardni odklon pri 10-kratnem prečnem preverjanju modela za napoved spola z uporabo različnih značilk, oblik besedila in treh algoritmov: metoda podpornih vektorjev (SVM), logistična regresija (LR) in naivnega Bayesov klasifikator (NB).

Značilke	Oblika	SVM (%)	LR (%)	NB (%)
besedni unigrami	pojavnica	**86,85 ± 6,00**	81,67 ± 5,89	64,89 ± 8,21
besedni uni- in birami	pojavnica	85,29 ± 8,34	79,81 ± 6,56	64,84 ± 6,29
znakovni (2-4)-grami	pojavnica	80,58 ± 8,17	78,76 ± 9,77	64,83 ± 6,25
besedni unigrami	lema	83,90 ± 7,82	80,57 ± 9,03	65,26 ± 10,89
besedni uni- in bigrami	lema	82,42 ± 7,12	78,02 ± 7,81	64,18 ± 1,65
Večinski klasifikator		64,85%		

Kot prikazuje Tabela 3, se je metoda podpornih vektorjev (SVM) odrezala bolje kot logistična regresija in Bayesov klasifikator. SVM doseže najvišjo klasifikacijsko točnost (86,85 % ± 6,00 %), ko uporabimo nelematizirane besedne unigrame kot značilke. SVM in logistična regresija presežeta večinski klasifikator v vseh preizkušenih kombinacijah z značilkami. V nasprotju pa Bayesov klasifikator preseže večinski klasifikator le v dveh poskusih, in sicer ko uporabimo nelematizirane ali lematizirane besedne unigrame kot značilke, vendar pa je razlika med rezultati minimalna, standardni odklon pa precej visok. Bayesov klasifikator se je torej v teh poskusih izkazal za manj uspešnega.

Če primerjamo uspešnost modelov glede na značilke, so se nelematizirani besedni n-grami izkazali kot bolj uporabni v primerjavi z lematiziranimi, kar velja tako za SVM kot za logistično regresijo. Oba omenjena algoritma dosežeta najvišjo točnost z nelematiziranimi besednimi unigrami (vrstica 1). Ko poleg besednih unigramov uporabimo še besedne bigrame, se točnost obeh modelov nekoliko zniža (za 1,56 % pri SVM-ju in za 1,86 % pri logistični regresiji), kar lahko pripišemo prevelikemu prileganju učni množici. SVM v poskusih z znakovnimi bi-, tri- in tetragrami doseže nižjo klasifikacijsko točnost kot v poskusih z besednimi n-grami. Znakovni n-grami se niso izkazali kot najboljši tudi za logistično regresijo, vendar pa so razlike v točnosti z besednimi n-grami manjše.

5 ANALIZA IN INTERPRETACIJA REZULTATOV KLASIFIKACIJE

V prejšnjem razdelku smo predstavili rezultate klasifikacije na podlagi pravil in klasifikacije z napovednimi modeli strojnega učenja glede na točnost napovedovanja spola. V razdelku 5.1 se najprej posvetimo modelu s pravili, in sicer analiziramo blogerje, ki jih je model klasificiral v napačni razred. Nato v razdelku 5.2 pregledamo značilke, ki so služile kot najbolj informativne za model strojnega učenja z najvišjo klasifikacijsko točnostjo, pri čemer nas zanimajo razlike in tudi podobnosti v besedilih blogerjev in blogerk.

5.1 Analiza napak modela s pravili

Model s pravili razvršča avtorje glede na rabo ženskega ali moškega spola v sestavljenih glagolskih zvezah v prvi osebi ednine. Kot prikazuje Tabela 2, dosežemo najvišjo točnost, ko je model manj strog in zahteva le tri indikatorje ženskega ali moškega spola v samonanašanju, da avtorja uvrsti v ženski ali moški razred (sicer

je bil avtor uvrščen v razred *nedoločeno*). Ta model uspešno klasificira 85,71 % avtorjev korpusa, Tabela 4 pa prikazuje razvrstitveno tabelo pravilnih in napačnih klasifikacij glede na ročne oznake. Na podlagi tabele lahko izračunamo, da je model pravilno klasificiral 79,17 % blogerk in 89,27 % blogerjev.

Tabela 4: Razvrstitvena tabela za najbolj uspešen model s pravili.

Napovedani razred	Dejanski razred	
	ženski	moški
ženski	76	2
moški	4	158
nedoločeno	16	17
skupaj	96	177

Najprej se osredotočimo na avtorje, ki jih je model uvrstil v razred nasprotnega spola. Kot vidimo v Tabeli 4, je model klasificiral štiri blogerke v moški razred. Pregledali smo njihova besedila; pri dveh od teh štirih blogerk je model našel zelo majhno število oznak za slovnični spol, vendar pa sta obe blogerki vseeno uporabili več moških kot ženskih oblik, saj sta objavili daljše zapise oz. citate v prvi osebi ednine moških avtorjev. Podobno velja za drugi dve blogerki, ki jih je model uvrstil v moški razred, saj je bila izmed več kot 40 oznak za spol večina moškega spola; tudi ti blogerki sta objavili več citatov in predvsem daljših leposlovnih zapisov v prvi osebi moškega spola. Izmed vseh blogerjev moškega spola so pravila razvrstila le dva blogerja v ženski razred. Po pregledu njunih besedil smo ugotovili, da je indikator za ženski spol večji od moškega predvsem zaradi pripovedi v prvi osebi ženskega spola in zaradi citatov ženskih oseb.

Tako za razred blogerjev kot blogerk velja, da je model avtorje največkrat napačno razvrstil v razred nedoločenih, saj je vanj uvrstil skoraj 17 % blogerk in skoraj 10 % blogerjev. Izmed 16 blogerk, ki jih je model klasificiral v razred nedoločeno, jih je šest vključevalo veliko število (nad 30) glagolskih zvez v prvi osebi ednine, vendar pa je število teh zvez v ženskem spolu skoraj izenačeno s številom zvez v moškem spolu. Po pregledu zapisov teh blogerk smo ugotovili, da so v besedila vključile veliko premega govora udeleženk in udeležencev dialoga. Nekateri izmed njihovih blogovskih zapisov pa so v celoti zapisani iz stališča pripovedovalca moškega spola.

Pri preostalih 10 blogerkah, ki jim spol ni bil pripisan (razred nedoločeno), model ni našel zadostnega števila (vsaj treh) indikatorjev za ženski ali moški spol v sestavljenih glagolskih zvezah v prvi osebi. To sicer še ne pomeni, da blogerke niso kako drugače izrazile svojega spola skozi besedne oblike. Tako lahko najdemo primere glagolskih zvez, ki jih naš klasifikator ni upošteval pri izračunu

indikatorja spola, saj med pomožnim glagolom in deležnikom leži več kot ena beseda, npr. v povedi: »Malo sem po naključju sledila.« Med ročnim pregledom smo našli tudi primere, v katerih je spol avtorice izražen v pridevnikih, npr.: »A v to sem prepričana.«

Pravila so v razred nedoločeno uvrstila 17 blogerjev. Štirje od njih so uporabili relativno veliko (40 ali več) tako ženskih kot moških oblik deležnikov na -l, in sicer predvsem v premem govoru. Klasifikator ostalim blogerjem spola ni pripisal in jih je uvrstil v razred nedoločenih, saj model ni našel več kot treh indikatorjev spola. Branje blogovskih zapisov teh avtorjev je pokazalo, da se v besedilih na splošno nase ne nanašajo pogosto, zato nismo našli prvoosebnih oblik niti v glagolih niti v pridevnikih.

5.2 Najbolj informativne značilke modela strojnega učenja

V tem razdelku analiziramo značilke, ki so bile uporabljene v klasifikacijskem modelu z najvišjo točnostjo. Kot smo pokazali v Tabeli 2, se je kot učni algoritem najbolje odrezal SVM z nelematiziranimi besednimi unigrami kot značilkami in tako dosegel klasifikacijsko točnost 86,85 %. Za interpretacijo vzamemo značilke, ki jim je klasifikator pripisal največje uteži in so bile torej najbolj informativne, da je klasifikator avtorju pripisal posamezni razred (spol). Iz klasifikacijskega modela jih izluščimo s funkcijo, ki vrača seznam značilk, ki imajo največje uteži za ženski oz. moški razred. Za ta prispevek smo s funkcijo izluščili 1.000 značilk za vsak razred, značilke smo na seznamu razvrstili po padajoči vrednosti uteži in vsakega od teh dveh seznamov nato analizirali ter primerjali med seboj.

Primerjanje vrednosti uteži nam lahko delno nudi vpogled v razlike v jezikovni rabi med ženskami in moškimi, saj lahko preverimo, katere značilke so bolj pomembne za blogerke in manj za blogerje ter obratno. Interpretacijo oz. pomembnost značilk za posamezni razred pa je treba vzeti z nekaj previdnosti, saj se moramo zavedati, da pri so pri algoritmu SVM značilke med seboj povezane in jih je težko interpretirati neodvisno od ostalih značilk, kar pa zaradi možnosti interpretacije v tem prispevku zanemarimo.

Na vrhu seznama najbolj informativnih značilk najdemo deležnike na -l v ženski (na seznamu ženskega razreda) oz. moški obliki (na seznamu moškega razreda). Za moški razred predstavljajo te oblike kar 15 % najbolj informativnih značilk, največje uteži pa imajo moške oblike bolj splošnih glagolov, npr. *šel, videl, dal*. Na seznamu ženskega razreda prav tako najdemo splošne glagole (npr. *imela, šla, vedela*, *dobila*), med 1.000 značilkami pa je deležnikov v ženski obliki 13 %. Na seznamih značilk se pojavljajo tudi druge besedne oblike, ki nakazujejo spol, in

sicer pridevniki, npr. *vesela* in *ponosna* na seznamu ženskega razreda ter *vesel* in *prepričan* na seznamu moškega razreda.

Poleg deležnikov in pridevnikov se na seznamih značilk pojavijo tudi druge besedne vrste. Pri obeh razredih imajo veliko utež različni prislovi, ki jih lahko razvrstimo med časovne, prostorske, številske ali modalne. V splošnem je na seznamu ženskega razreda več časovnih prislovov, še posebej takih, ki izražajo pogostost (npr. *znova, včasih, pogosto, nikdar*). Na seznamu moškega razreda najdemo več časovnih prislovov, ki so vezani na eno točko v času (npr. *nocoj, sinoči, včeraj*). Zanimive razlike med najbolj informativnimi značilkami obeh razredov se pojavijo v zaimkih. Med značilkami, informativnimi za ženski razred, so predvsem osebni in svojilni zaimki za prvo osebo ednine (npr. *moja/mojega/mojih, mene/zame/menoj*) in dvojine (npr. *naju/nama, najin*). Na seznamu moškega razreda je veliko manj zaimkov, najdemo pa osebne zaimke za prvo osebo množine (*naši/naše*), tretjo osebo ednine (npr. *njej, njegovega/njegov*) ali množine (*njihovi*).

Med prvimi stotimi značilkami z največjimi utežmi so predvsem deležniki ženskega oz. moškega spola, nato pa se na obeh seznamih začnejo pojavljati tudi samostalniki in lastna imena. Čeprav so na seznamu besedni unigrami, torej le posamezne besede brez konteksta, lahko na podlagi seznama sklepamo o temah, ki so bolj značilne za enega od razredov in manj za drugega. Med značilkami ženskega razreda so pogoste besede, ki zaznamujejo družino in družinske člane (npr. *otroci/otroka/otroke*, *mama/mami, očka, starši, sestro, družina, otroštvo*). Poleg tega na tem seznamu najdemo več besed, ki se nanašajo na romantične zveze in spolnost (npr. *spolnost, ljubček, zaljubljenost, seks*). Med značilkami, ki so bolj tipične za ženski razred, je besedišče, povezano s čustvi in občutki, ki so lahko pozitivni (*ljubezen/ljubezni, strasti, nasmeh*), najde pa se več primerov negativnih čustev (*otožnost, jokala, solze, samota, zavist, žalostna, sram, sramota*).

Poleg družinske in ljubezenske tematike lahko s seznama visoko uteženih značilk ženskega razreda sklepamo še o eni temi, ki je bolj priljubljena pri blogerkah kot blogerjih. Na seznamu ženskega razreda namreč najdemo več besed, povezanih s prehrano, kar nakazuje na to, da gre v besedilih za recepte ali nasvete glede prehranjevanja, npr. *cvetače, zelenjave, maslo, kokosovo, testo, penino, cimet*. Med manj vidnimi, vendar še vedno razlikovalnimi temami je tudi zdravje, saj na seznamu ženskega razreda najdemo več besed, ki se nanašajo na zdravstvene zadeve (npr. *zdravljenje, kemoterapija/kemoterapije, rakom/rak, zboleti, cepiv*).

Besedišče, ki je povezano s političnim dogajanjem, se pojavlja med najbolj informativnimi značilkami obeh razredov (npr. na seznamu ženskega razreda najdemo primere *demokratske, socialisti, isis*), vendar pa ta tema veliko bolj opredeljuje seznam značilk moškega razreda in je prevladujoča glede na ostale. Na seznamu moškega razreda najdemo besede, ki se nanašajo na državo, državne organe in

mehanizme (*država, sodišča, volitvah, vlade, referendum*), politične funkcije (*predsednik, državljanov*), Cerkev in različne politične in ekonomske ureditve (*demokracija, kapitalizem*). Seznam vključuje tudi besede, ki so povezane z Jugoslavijo (*udba*, *komunisti*, *jla*) in drugo svetovno vojno (*nob, belogardisti*). Na seznamu moškega razreda se v različnih oblikah pojavita tudi besedi *politika* in *politično* (*politično/političnega/politične/političnega*). Poleg obsežnega besedišča s področja politike je za moški razred zelo značilna športna tematika, saj se to besedišče pojavlja le pri moškem razredu (*ligi, žogo, prvenstvo, tekmo/tekem*).

Razlike med blogerkami in blogerji, o katerih sklepamo na podlagi seznamov najbolj informativnih značilk, se vezane tudi na slog in register pisanja. Na seznamu moškega razreda najdemo denimo primere vulgarizmov (*hudiča, jebo/jebe, scat*) in slabšalnih poimenovanj za manjšinsko skupino (*cigan/ciganov*).

Kot je bilo izpostavljeno, se moramo zavedati, da smo pri interpretaciji rezultatov zanemarili dejstvo, da so pri modelih dejansko značilke med seboj povezane. Tovrstna interpretacija nam omogoča delno razumevanje razlik med besedili, na podlagi besed, ki so bile za klasifikacijski model zelo pomembne. Za preverjanje relevantnosti razlik pa bi bili potrebni dodatni statistični testi.

6 SKLEP

V tem prispevku smo se lotili vprašanja avtomatske napovedi spola avtorja na podlagi besedil, s ciljem napovedati spol slovenskih blogerk in blogerjev, katerih zapisi so zbrani v korpusu spletnih uporabniških vsebin Janes. V prispevku predstavimo dva tipa napovednih modelov: model s pravili (ročno zgrajeni klasifikacijski modeli) in modele zgrajene z metodami strojnega učenja (avtomatsko zgrajeni klasifikacijski modeli).

Klasifikator na podlagi pravil je zasnovan tako, da spol avtorjem pripišemo glede na izražanje slovničnega spola z deležnikom v glagolskih zvezah v prvi osebi ednine. Klasifikator pripiše avtorjem ženski ali moški spol glede na število glagolskih zvez, ki izražajo določen spol oz. jih uvrsti v razred nedoločeno, če je premalo indikatorjev za katerega koli od spolov. V poskusih smo z najboljšim modelom s pravili uspešno klasificiral 85,71 % avtorjev v učni množici (glej Tabelo 2), pri čemer je presegel večinski klasifikator za okoli 21 %. Zanimivo je, da upoštevanje nestandardnih zapisov pomožnega glagola (*sm*, *nism*) ni spremenilo uspešnosti modela. Iz rezultatov lahko sklepamo, da je raba spola v glagolskih zvezah v primerih samonanašanja precej predvidljiva tudi v blogih, saj velika večina avtorjev uporablja dovolj prvoosebnih glagolskih oblik, pri katerih sta pomožni glagol in deležnik relativno blizu, pomožni glagol pa je zapisan na standardni način. Kljub

temu da je napoved spola avtorjev v delu Fišer et al. (2016) zastavljena nekoliko drugače, lahko ugotovimo, da klasifikacijska točnost našega modela s pravili ta model preseže za skoraj 8 %.

V razdelku 4.2 poročamo o uspešnosti modelov za napoved spola, ki smo jih zgradili s tremi različnimi algoritmi strojnega učenja (SVM, logistična regresija in naivni Bayesov klasifikator) ter njihove rezultate primerjali glede na uporabljene značilke. Modele smo testirali z 10-kratnim prečnim preverjanjem. Rezultati so pokazali, da se je najbolje odrezal SVM z uporabo besednih unigramov nelematiziranega besedila, saj je dosegel 86,85-% klasifikacijsko točnost (Tabela 3). Tako SVM preseže večinski klasifikator za 22 %, medtem ko je od modela s pravili boljši le za 1 %.

Avtomatsko zgrajeni model z najvišjo klasifikacijsko točnostjo smo ovrednotili tudi kvalitativno. Analizirali smo tiste značilke, ki jim je model pripisal največje uteži pri odločitvi za uvrstitev v ženski ali moški razred. Analiza je pokazala, da je med najbolj informativnimi značilkami veliko število takih, ki vsebujejo informacijo o spolu (deležniki na -l in pridevniki). Zanimive pa so razlike glede na druge besedne vrste, in sicer je med značilkami ženskega razreda več zaimkov. Podoben trend v besedilih avtoric je bil opažen tudi na primeru angleških sporočil na družbenem omrežju Facebook (Schwartz et al. 2013), angleških blogovskh zapisih (Schler et al. 2006), angleškega pisanega in govorjenega jezika (Newman et al. 2008), kot tudi na pilotni analizi slovenskih tvitov (Verhoeven et al. 2017).

Razlike v najbolj informativnih značilkah so vezane tudi na različne teme, ki so bolj značilne za avtorje enega spola in manj za drug spol. O osredotočanju na teme, kot sta družina in ljubezenski odnosi, torej na teme, ki zaznamujejo socialne procese, v povezavi z besedili avtoric poročajo tudi Schler et al. (2006), Newman et al. (2008) in Schwartz et al. (2013). Schler et al. (2006) in Schmid (2003) so podobno kot mi opazili, da moški v primerjavi z ženskami večkrat pišejo ali govorijo o politiki in športu. Blogovske zapise v korpusu Janes smo preučevali tudi v Škrjanec in Pollak (2016), kjer smo s pomočjo metode gručenja podatkov izdelali ontologije tematik in s tem identificirali prevladujoče teme, o katerih pišejo slovenski blogerji. Rezultati so pokazali, da tako blogerke kot blogerji pišejo o politiki, družini, romantičnih odnosih, okolju in prehrani; blogerji se s primerjavi z blogerkami več posvečajo temam o športu, glasbi, literaturi, Cerkvi, begunski krizi in temam o naravi. V zapisih blogerk pa je več poudarka na religiji, socialni politiki in čustvih.

Raba vulgarizmov in kletvic naj bi bila bolj značilna za jezik moških (Bamman et al. 2014, Newman et al. 2008, Schwartz et al. 2013), kar se je pokazalo tudi na našem seznamu značilk, vendar bi bila potrebna tudi podrobnejša analiza teh besed v kontekstu.

Glede na našo analizo in primerjavo dveh tipov napovednih modelov lahko zaključimo, da je za slovenske bloge pristop s pravili o rabi spola v glagolskih zvezah lahko enako uspešen kot bolj kompleksni modeli, zgrajeni z algoritmi strojnega učenja. Model s pravili ima to prednost, da ne potrebuje označenih podatkov za učenje, vendar pa se je treba zavedati, da lahko avtorji zavestno manipulirajo z rabo slovničnega spola v samonanašanju, da prikrijejo svoj spol. V tem pogledu bi bilo zanimivo in koristno preveriti, kako uspešni so modeli, iz katerih bi izključili besedne značilke, ki imajo referenčni spol (deležniki in pridevniki). Če stopimo korak dalje, Daelemans (2013) trdi, da bi morali biti modeli za profiliranje avtorjev neodvisni od žanra in teme besedila, za kar bi bilo potrebno ohraniti le slogovne značilke, odstraniti pa tiste, ki se nanašajo na razlikovalne teme.

Svojo raziskavo o avtomatski napovedi nameravamo razširiti na več načinov. Model s pravili je možno izboljšati tako, da poleg upoštevanja spola v glagolskih zvezah vključimo tudi pridevnike. Kompleksnejša naloga pa bi bila ugotavljanje citiranih delov besedila, ki so bili, kot je pokazala analiza napak, pogosto razlog za napačno klasifikacijo. Za boljšo uspešnost z metodami strojnega učenja pa bi potrebovali večjo učno množico, v nadaljnjih eksperimentih pa bomo upoštevali nove kombinacije besednih in znakovnih n-gramov ter alternativne metode rangiranja značilk (Guyon et al. 2002). Posebno pozornost bomo posvetili nadaljnji interpretaciji razlik med avtorji blogov glede na spol, kjer bomo izsledke predstavljene raziskave preverili s statističnimi testi, kar smo delno že obravnavali (Škrjanec 2017). Interpretacija informativnih značilk modela, naučenega na lematiziranih besedilih, nam bo omogočila, da več pozornosti namenimo značilkam, ki so neodvisne od izražanja slovničnega spola. Nenazadnje bomo modele preizkusili tudi na drugih žanrih spletnih uporabniških vsebin, ki jih vsebuje korpus Janes (forumi, komentarji novic in uporabniške in pogovorne strani na Wikipediji).

Literatura

Bamman, David, Jacob Eisenstein in Tyler Schnoebelen, 2014: Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 1/2. 135–160.

Daelemans, Walter, 2013: Explanation in computational stylometry. *Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing* (CICLing'13). Vol 2. Berlin, Heidelberg: Springer. 451–462.

Dhillon, Inderjit, Kogan, Jacob in Nicholas, Charles, 2004: Feature selection and document clustering. Berry, Michael (ur.): *A Comprehensive Survey of Text Mining*. New York: Springer. 73–100.

Erjavec, Tomaž, Nikola Ljubešić in Darja Fišer, 2018: Korpus slovenskih spletnih uporabniških vsebin Janes. Fišer, Darja (ur.): *Viri, orodja in metode za analizo spletne slovenščine*. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani. 16–43.

Fišer, Darja, Tomaž Erjavec in Nikola Ljubešić, 2016: Janes v0.4: Korpus slovenskih spletnih uporabniških vsebin. *Slovenščina 2.0* 4/2. 67–99.

Guyon, Isabelle, James Weston, Stephen Barnhill in Vladimir Vapnik, 2002: Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning* 46/1. 389–422.

Jurafsky, Dan in James H.Martin, 2009: *Speech and Language Processing, second edition*. Upper Saddle River, New Jersey: Prentice-Hall, Inc.

Kobayashi, Mei in Aono, Masaki, 2007: Vector space models for search and cluster mining. Berry, Michael W. in Malu Castellanos (ur.): *Survey of Text Mining: Clustering, Classification and Retrieval.* Springer. 109–127.

Koppel, Moshe, Shlomo Argamon in Anat R.Shimoni, 2002: Automatically categorizing written texts by author gender. *Literary and Linguistic Computing* 17/4. 401–412.

Ljubešić, Nikola in Tomaž Erjavec, 2016: Corpus vs. lexicon supervision in morphosyntactic tagging: The case of Slovene. *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016). Portorož, Slovenia: European Language Resources Association (ELRA). 1527–1531.

Ljubešić, Nikola, Fišer, Darja in Tomaž Erjavec, 2017: Language-independent gender prediction on Twitter. *Proceedings of NLP+CSS: Second Workshop on Natural Language Processing and Computational Social Science*. Vancouver, Kanada: ACL. 1–6.

Martinc, Matej, Iza Škrjanec, Katja Zupan in Senja Pollak, 2017: PAN 2017: author profiling - gender and language variety prediction. Cappellato, Linda, Nicola Ferro, Lorraine Goeuriot in Thomas Mandl (ur.): *Working notes papers of CLEF 2017 Conference and Labs of the Evaluation Forum*. CEUR Workshop Proceedings.

Martinc, Matej, Senja Pollak in Ana Zwitter Vitez, 2018: Delotoki za nadaljnje analize nestandardne slovenščine. Fišer, Darja (ur.): *Viri, orodja in metode za analizo spletne slovenščine*. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani. 16–43.

Motschenbacher, Heiko, 2010: *Language, Gender and Sexual Identity: Poststructuralist Perspectives*. Amsterdam: John Benjamins.

Nguyen, Dong, Rilana Gravel, Dolf Trieschnigg in Theo Meder, 2013: TweetGenie: Automatic age prediction from tweets. *ACM SIGWEB Newsletter* 4/4. 1–6.

Osrajnik, Eneja, Darja Fišer in Damjan Popič, , 2015: Primerjava rabe ekspresivnih ločil v tvitih slovenskih uporabnikov in uporabnic. *Zbornik konference Slovenščina na spletu in v novih medijih*. Ljubljana: Znanstvena založba Filozofske fakultete. 50–74.

Peersman, Claudia, Daelemans, Walter in Van Vaerenbergh, Leona, 2011: Predicting age and gender in online social networks. *Proceedings of the Third International Workshop on Search and Mining User-generated Contents*, ACM. 37–44.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay, 2011: Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12. 2825–2830.

Plank, Barbara and Hovy, Dirk, 2015: Personality Traits on Twitter or How to Get 1,500 Personality Tests in a Week. *Proceedings of the Sixth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. 92–98.

Rangel, Francisco, Paolo Rosso, Martin Potthast in Benno Stein, 2017: Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. Cappellato, Linda, Nicola Ferro, Lorraine Goeuriot in Thomas Mandl (ur.): *Working notes papers of CLEF 2017 Conference and Labs of the Evaluation Forum*. CEUR Workshop Proceedings.

Schler, Jonathan, Koppel, Moshe, Argamon, Shlomo, in Pennebaker, James, 2006: Effects of age and gender on blogging. *Proceedings of AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs* 6. 199–205.

Schwartz, H. Andrew, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, Martin E. P. Seligman in Lyle H. Ungar, 2013: Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one* 8(9).

Stamatatos, Efstathios, 2009: A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology* 60/3. 538–556.

Škrjanec, Iza, 2017: *Gender-based analysis of Slovene user-generated content*. Magistrsko delo. Ljubljana: Mednarodna podiplomska šola Jožefa Stefana.

Škrjanec, Iza in Senja Pollak, 2016: Topic ontologies of the Slovene blogosphere: A gender perspective. Fišer, Darja in Michael Beißwenger (ur.): *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities*, 27-28 September 2016, Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia. 62–65.

Verhoeven, Ben, Daelemans, Walter in Plank, Barbara, 2016: TwiSty: A multilingual Twitter stylometry corpus for gender and personality profiling. V *Proceedings of the 10th Language Resources and Evaluation Conference (LREC 2016)*. ELRA, Portorož, Slovenia.

Verhoeven, Ben, Iza Škrjanec in Senja Pollak, 2017: Gender Profiling for Slovene Twitter Communication: The Influence of Gender Marking, Content and Style. *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*. 119–125.

Zwitter Vitez, Ana, 2011: Povej mi karkoli in povem ti, kdo si: Ugotavljanje avtorstva besedil. Kranjc, Simona (ur.): *Obdobja 30: Meddisciplinarnost v slovenistiki*. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete. 565–570.
Zwitter Vitez, Ana, 2013: Le décryptage de l'auteur anonyme : l'affaire des électeurs en survêtements. *Linguistica* 53/1. 91–101.
Witten, Ian H. in Eibe Frank, 2005: *Data Mining: Practical Machine Learning Tools and Techniques.* San Francisco: Morgan Kaufmann.

Zahvala

Avtorice se zahvaljujemo recenzentom za koristne komentarje in predloge.