AUTOMATIZOVANA KLASIFIKACIJA LIČNOSTI
AUTOMATED PERSONALITY CLASSIFICATION

Aleksandar Kartelj[1], Vladimir Filipović[1], Veljko Milutinović[2]
[1]Matematički fakultet, Univerzitet u Beogradu
[2]Elektrotehnički fakultet, Univerzitet u Beogradu

Sadržaj – U ovom radu je dat opis problema automatizovane klasifikacije ličnosti (AKL) i aktuelnih metoda za njegovo rešavanje. Kao naučni doprinos predstavljamo klasifikaciju predloženih metoda za rešavanje problema AKL. Takođe, iznosimo nekoliko ideja za razvoj novih rešenja i unapređenje postojećih. U radu su razmotrene i buduće smernice za rešavanje problema AKL u kontekstu socijalnih mreža.

Abstract – *In this paper we give a description of the problem of the automated personality classification (APC) and the presentation of the existing solutions to this problem. As a scientific contribution, we present a classification of the existing solutions to the problem of the APC. We also expose several ideas for development of new solutions to this problem and for improvement of the existing ones. Future directions for solving the problem of the APC in the context of social networks are also discussed.*

1. UVOD

Klasifikacija ličnosti je jedan od problema koji se razmatra u psihologiji ličnosti, grani psihologije. Fokus ove oblasti je u izučavanju ličnosti i individualnih razlika. Prema tom učenju, ličnost se može definisati kao dinamički i organizovani skup karakteristika osobe koji jedinstveno utiče na spoznaju, motivaciju i ponašanje te osobe. U ovom radu se razmatra problem automatizovane klasifikacije ličnosti neke osobe na osnovu informacija koje proističu iz sledećih sadržaja: tekstualni sadržaj koji je ta osoba napisala, i meta informacija o osobi, dobijenih na zahtev, kroz socijalne mreže ili neke druge načine. Postoje istraživanja koja uključuju i govor, analizu facijalnih karakteristika osobe, gestikulacije, i drugih aspekata ponašanja, ali ona nisu predmet našeg rada. Standardni pristup u rešavanju problema AKL na osnovu pomenutih sadržaja je opisan sledećim koracima: A) pravljenje korpusa, B) određivanje karakteristika ličnosti učesnika i C) kreiranje klasifikacionog modela.

A) Korpus podrazumeva kolekciju sadržaja (tekstualnih i meta) učesnika nad kojima se vrši kreiranje modela. U postojećim istraživanjima ova baza je obično bila sastavljena iz studentskih eseja i propratnih meta informacija koji su prvobitno prikupljeni i opisani u [24] i [25]. Druga istraživanja su kao učesnike razmatrala internet blogere i njihove blogove [10, 14, 18, 21, 22 i 32]. U [23] i [28] su korišćene i-mejl liste i SMS, dok su u najnovijim istraživanjima [3, 26 i 27] iskorišćene informacije dostupne kroz upotrebu socijalnih mreža, Twitter i Facebook.

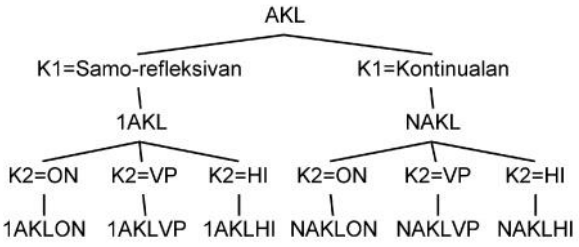
B) Određivanje karakteristika ličnosti se tradicionalno određuje podvrgavanjem učesnika testu ličnosti. Najzastupljenija implementacija testa ličnosti je pet-faktorski model (eng. The Big Five Model of traits) koji vrši klasifikaciju ličnosti na pet zasebnih karakteristika, gde je svaka od njih ocenjena na realnoj skali. Razmatrane karakteristike su: 1) otvorenost ka iskustvu (eng. openness to experience), 2) savesnost (eng. conscientiousness), 3) ekstravertnost (eng. extraversion), 4) saglasnost i prijatnost (eng. agreeableness) i 5) neuroticizam (eng. neuroticism). Neke od implementacija ovog modela, bazirane na upitnicima, opisane su u [2, 8 i 15].

C) Nakon prikupljanja korpusa i određivanja karakteristika ličnosti za svakog od učesnika, potrebno je razviti odgovarajući klasifikacioni model. Ovaj postupak podrazumeva odabir nezavisnih promenljivih, tj. skupa relevantnih atributa korpusa koji dobro određuju zavisnu promenljivu, gde je zavisna promenljiva jedna ili više karakteristika ličnosti. U [6] i [24] su utvrđena dva skupa atributa, respektivno LIWC i MRC za koje se pokazalo da su korelisani sa nekim od karakteristika ličnosti. LIWC (Linguistic Inquiry and Word Count) predstavlja bazu od nekoliko stotina reči, za koje je utvrđeno da doprinose određivanju karakteristika ličnosti, npr. reči *hate* i *kill* su vrlo značajne u određivanju nivoa neurotičnosti. MRC je psiholingvistička baza podataka koja sadrži reči klasifikovane po različitim merama, poput slikovitosti, slogovitosti, konkretnosti, frekvenciji upotrebe reči i slično. Npr. reč *ship* je visoko ocenjena na skali konkretnosti, dok je reč *patience* vrlo nisko ocenjena. Nakon što je skup relevantnih atributa formiran, nad njim se gradi odgovarajući klasifikacioni model koji može biti baziran na različitim lingvističkim, stilističkim i statističkim tehnikama. Detaljnije izlaganje o korišćenim klasifikacionim modelima i algoritmima će biti dato u trećem delu.

U narednom delu ćemo opisati klasifikaciju postojećih rešenja za problem AKL korišćenjem drvolike strukture. Nakon toga, u trećem delu će biti prikazana neka od reprezentativnih istraživanja na temu AKL, grupisana u skladu sa klasifikacijom datom u drugom delu. U četvrtom delu ćemo predstaviti ideje za razvoj novih i poboljšavanje već postojećih rešenja za problem AKL. U petom delu će biti izloženi zaključci našeg istraživanja.

2. KLASIFIKACIJA POSTOJEĆIH REŠENJA

Kako bi što sistematičnije predstavili rešenja problema AKL, uvodimo klasifikaciono drvo koje deli postojeći skup rešenja na relativno balansirane podskupove, kada je u pitanju broj referenci u svakom od njih (Slika 1). Drvo koristi dva nivoa podele, odnosno klasifikaciona kriterijuma. Prvi, označen sa *K1,* odgovara tipu interakcije koji je primenjen u sadržaju, i on može biti samo-refleksivan (eng. self-reflexive) ili kontinualan (eng. continuous conversation). Samo-refleksivan sadržaj je vezan za samo jednu osobu (nalik monologu), a kategorije sadržaja koje potpadaju pod ovu kategoriju su: eseji, blogovi, biografije i dr. Kontinualni tipovi sadržaja su oni koji su nastali unakrsnom komunikacijom dva ili više lica (nalik dijalogu). Ovoj kategoriji pripadaju i-mejl, SMS, diskusije na forumima, i još neki tipovi dijaloga i podataka na zahtev (eng. on-demand data). Neki autori klasifikuju blogove i kao samo-refleksivnu i kao kontinualnu interakciju [7] i [19]. Njihovo objašnjenje je da je to posledica mogućnosti čitalaca da ostavljaju komentare na blogovima. U našem istraživanju blogovi se smatraju za isključivo samo-refleksivni tip sadržaja, jer verujemo da su funkcionalno mnogo bliži biografijama nego forumskim diskusijama. Kriterijum drugog nivoa, *K2*, klasifikuje prema eksperimentalnom pristupu: pristup odozgo na dole (OD, eng. top-down), pristup vođen podacima (VP, eng. data-driven), i hibridni (HI), koji je kombinacija prethodna dva. [9] razmatra sličan kriterijum podele, samo bez hibridnog. [20] kategoriše korelacionu analizu LIWC i MRC atributa i karakteristika ličnosti kao pristupe odozgo na dole. Sa druge strane, stratifikovanu kolokacionu analizu (koristi n-gramsku analizu), individualnu upotrebu kolokacija, kontekstualne metrike, i analizu zasnovanu na frekvencijama reči kategoriše kao podacima vođene metode.



Slika 1. Klasifikacija postojećih rešenja

Predloženo klasifikaciono drvo sadrži šest listova, što znači šest različitih kategorija rešenja za problem AKL. Svakom od listova je pridružena jedinstvena skraćenica koja sugeriše na tip interakcije: 1 za samo-refleksivnu i N za kontinualnu konverzaciju. Prefiksi OD, VP, i HI odgovaraju pristupima odozgo na dole, pristupu vođenom podacima i hibridnom, respektivno. Na osnovu ovog klasifikacionog drveta smo grupisali naučne radove, nama poznate do ovog momenta u dve tabele: Tabela 1, u kojoj su izlistani radovi koji opisuju rešenja za samo-refleksivne podatke i Tabela 2, u kojoj su radovi koji se bave sadržajem kontinualne prirode.

3. PREZENTACIJA POSTOJEĆIH REŠENJA

U ovom delu prezentovana su neka od postojećih rešenja za problem AKL. Odabran je po jedan rad iz svake od šest kategorija, osim u slučaju poslednje kategorije, u kojoj, kao što se vidi iz Tabele 2, ne postoji nijedan rad. Za svaki od radova, pored opisa, predstavljena je i arhitektura sistema pomoću dijagrama.

Tabela 1: Radovi sa samo-refleksivnim sadržajem

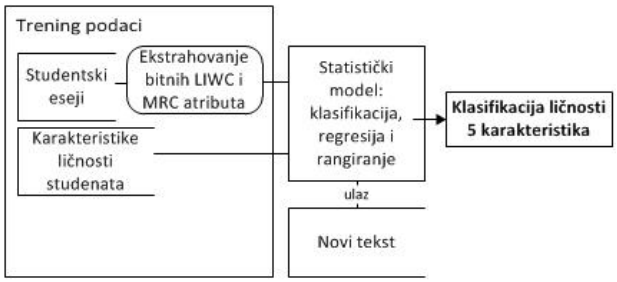
K1	Samo-refleksivan tekst		
Broj radova	12		
K2	ON	VP	HI
Broj radova	5	5	2
Radovi	[10, 16, 17, 18, 35]	[11, 14, 21, 22, 32]	[1, 31]

Tabela 2: Radovi sa kontinualnim sadržajem

K1	Kontinualan tekst		
Broj radova	8		
K2	ON	VP	HI
Broj radova	3	5	0
Radovi	[23, 28, 29]	[3, 4, 5, 26, 27]	

3.1 (1AKLON) Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text

U [17] tri različita statistička modela su primenjena nad korpusom teksta i govora u cilju pravljenja klasifikatora ličnosti prema pet-faktorskom modelu (Slika 2). Pored algoritama za klasifikaciju, autori razmatraju i primenu regresije i modela zasnovanog na rangiranju. Korpus je bio sastavljen od studentskih eseja, prethodno prikupljenih u istraživanju [25]. Korišćenje već postojećeg korpusa je doprinelo transparentnosti i uporedivosti dobijenih rezultata. Autori su takođe upotrebili već postojeće rezultate testa ličnosti studenata koji su napisali pomenute eseje. LIWC i MRC skupovi atributa su korišćeni za pravljenje klasifikatora. Dodatno, autori su za atribute uzeli i svojstva govora, kao što su odnosi između upotrebe naredbi, zahteva, upita i izjava.



Slika 2. Različiti statistički modeli nad LIWC i MRC

Sva tri statistička modela su testirana na već postojećim implementacijama statističkog alata Weka. Svi modeli su prilagođeni tako da izvršavaju binarnu klasifikaciju (nisko ili visoko) po svakoj od pet karakteristika ličnosti zasebno. Pokazano je da sva tri modela klasifikuju sa uspešnošću većom od donje granice 50% (donja granica je rezultat dobijen izvršavanjem slučajnog algoritma). Evaluacija otvorenosti ka iskustvu je bila najpreciznija u

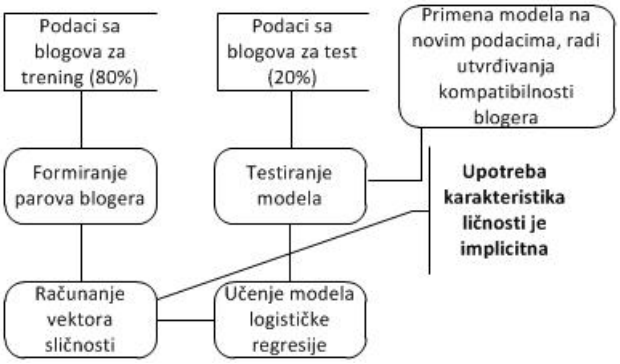
sva tri primenjena modela. Pokazano je i da su MRC atributi korisni u evaluaciji emocionalne stabilnosti (neuroticizam), dok su LIWC atributi bili značajni u pogledu svih pet karakteristika ličnosti.

3.2 (1AKLVP) Personality Based Latent Friendship Mining

U [32] predložen je metod koji rešava problem pronalaženja kompatibilnih prijatelja među blogerima (Slika 3). Za razliku od prethodnih pristupa u rešavanju ovog problema u kojima se pretpostavljalo da kompatibilni blogeri pišu o srodnim temama, autori ovde koriste hipotezu da kompatibilni blogeri imaju slične karakteristike ličnosti. Na ovaj način, problem klasifikacije ličnosti se razmatra kao potproblem, tj. algoritam za rešavanje AKL predstavlja ugnježdeni algoritam u okviru algoritma za pronalaženje kompatibilnih blogera. Algoritam za AKL je zasnovan na treniranju modela logističke regresije, koji se, nakon što je istreniran, koristi za binarnu klasifikaciju. Iz blogova učesnika je ekstrahovano 20 atributa: broj komentara, broj slika, broj interpunkcijskih znakova, čitljivost, broj rečenica, hiper veza, prosečna dužina rečenice, vektor reči itd. Nakon toga je za svaka dva blogera w_i i w_j, definisan vektorski profil sličnosti θ između vektora njihovih atributa. Logistička funkcija je definisana kao $Y=e^A/(1+e^A)$, gde je $A=\lambda_0+\lambda_1X_1+...+\lambda_{20}X_{20}$. X_i je element vektora θ na poziciji i. Vrednost koju vraća logistička funkcija pripada [0,1], tako da je vrednost 0.5 korišćena kao granica između dve klase blogera: nisu potencijalni prijatelji (0), i jesu potencijalni prijatelji (1). Početni skup blogova je pikupljen na slučajan način sa MSN live space internet mreže. Nakon toga, blogovi onih blogera koji su prijatelji blogerima iz početnog skupa su takođe preuzeti i parsirani. Na ovaj način je uspostavljena mreža relacija prijateljstva u okviru korpusa blogova. Nakon toga, prethodno opisani model logističke regresije je primenjen na 80% podataka, dok je preostalih 20% iskorišćeno za testiranje modela. Rezultati primene ovog modela pokazuju preciznost od 80% i odziv od 75%. Međutim, smatramo da je u radu korišćena veoma jaka, a možda i nerealna hipoteza da su osobe koje imaju slične tipove ličnosti često i prijatelji na mreži blogera.

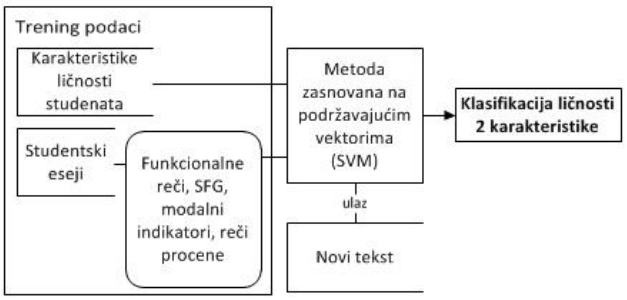
3.3 (1AKLHI) Lexical Predictors of Personality Type

U [31] autori razmatraju problem binarne klasifikacije neuroticizma i ekstravertnosti (Slika 4). Upotrebljena su četiri različita skupa atributa: funkcijske reči, konjuktivne fraze, modalni indikatori, i atributi procene. Metoda zasnovana na podržavajućim vektorima (eng. support vector machine – SVM) je zatim primenjena za binarnu klasifikaciju. Ova studija je zasnovana na psihologiji jezika (eng. language psychology) i računarskoj stilistici (eng. computational stylistic). Računarska stilistika je naučna disciplina koja na značenje teksta posmatra iz aspekata afekta, žanra, funkcije, i ličnosti autora teksta. Ova disciplina sugeriše da se svi ovi aspekti mogu ekstrahovati iz stila pisanja. U tom kontekstu, funkcijske reči su one često upotrebljavane reči poput *and, for, the* itd. Ideja za analizu njihove upotrebe proističe iz verovanja da je upotrebu funkcijskih reči teško svesno kontrolisati, tako da se frekvencija njihove upotrebe može upotrebiti za detektovanje stila autora. Konjuktivne fraze predstavljaju koncept iz teorije sistemskih funkcionalnih gramatika (SFG), jednog od funkcionalnih pristupa lingvističkoj analizi [12] koji pokušava da formalizuje validne upotrebe pisanog teksta. Modalni indikatori kvalifikuju događaje ili entitete u tekstu prema njihovoj mogućnosti, neophodnosti i svojstvenosti (modalni glagoli, priloški dodaci itd.). Konačno, atributi procene opisuju stav i orijentaciju.



Slika 3. Utvrđivanje potencijalnih prijatelja na osnovu ličnosti

U istraživanju je korišćen korpus studentskih eseja [25]. Pokazano je da atributi procene daju najveću preciznost kada je neuroticizam u pitanju. U kontekstu određivanja ekstrovertnosti, samo funkcijske reči su poboljšale preciznost. Istraživanje je pokazalo da ekstrovertne osobe preferiraju upotrebu reči koje prožima sigurnost (*immediate, am, so, being, very* itd.), dok introvertne koriste reči koje prožima nepotpunost ili nesigurnost (*perhaps, nobody, try, hardly* itd.).

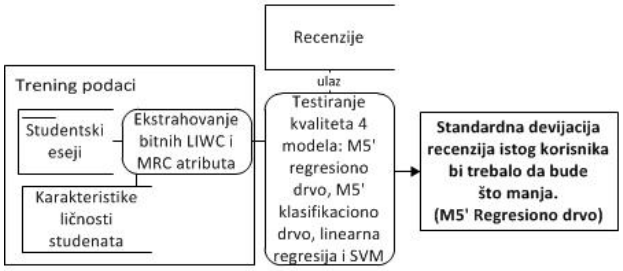


Slika 4. Pristup zasnovan na stilističkoj analizi

3.4 (NAKLON) A Comparative Evaluation of Personality Estimation Algorithms for the TWIN Recommender System

U [29] je izložena komparativna evaluacija nekoliko algoritama za klasifikaciju ličnosti (Slika 5). Cilj istraživanja je bio traženje najadekvatnijeg algoritma za upotrebu u okviru TWIN ("Tell me What I Need") sistema za predlaganje turističkih destinacija koji se koristi na internet lokaciji *tripadvisor.com*. Sistemi za predlaganje (eng. recommender systems) su često bazirani na evolutivnom profilisanju korisnika. Manji deo informacija potrebnih za profilisanje je prikupljen prilikom same registracije. Međutim, ostatak se obično prikuplja implicitno kroz nadzor aktivnosti korisnika sistema. Ključna hipoteza u ovom radu je da grupe ljudi

sa sličnim tipovima ličnosti preferiraju odabir sličnih hotela i destinacija. U okviru TWIN sistema, korisničke recenzije se prikupljaju, na osnovu njih se određuje profil ličnosti korisnika, a potom na osnovu njega određuje pogodna destinacija i hotel.



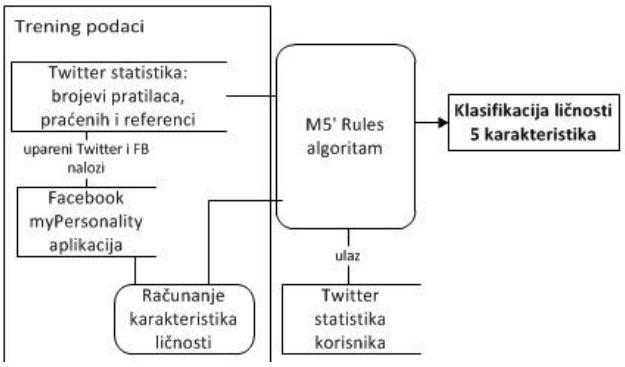
Slika 5. Evaluacija algoritma za AKL u TWIN sistemu

Autori su upotrebili prethodno formirani korpus studentskih eseja i njihove profile ličnosti, bazirane na pet-faktorskom modelu. Za atribute su uzeti LWC i MRC, a potom su nad njima primenjena četiri modela učenja. Iskorišćene su Weka implementacije sledećih algoritma: linearna regresija, M5' klasifikaciono drvo, M5' regresiono drvo, i metoda zasnovana na podržavajućim vektorima (SVM). Testiranje kvaliteta je izvedeno na recenzijama 15 slučajno odabranih korisnika sistema koji su imali bar 30 recenzija. Za svaku recenziju je utvrđen rezultat po svakoj od pet karakteristika ličnosti, korišćenjem sva četiri algoritma. Za meru kvaliteta je uzeta standardna devijacija u rezultatima za recenzije istog korisnika po svakoj karakteristici ličnosti zasebno. Hipoteza je da je algoritam dobar ukoliko daje sličan rezultat svim recenzijama napisanim od strane istog korisnika, tj. ukoliko je standardna devijacija minimalna. M5' regresiono drvo se pokazalo kao najkvalitetnije u skladu sa ovom merom kvaliteta, pa je iz tog razloga i implementirano u okviru TWIN sistema. Uspeh TWIN sistema sugeriše da postoji prostor za indirektnu upotrebu automatizovane klasifikacije ličnosti u cilju rešavanja nekih praktičnih problema koji se pojavljuju u ovim ili srodnim vrstama sistema.

3.5 (NAKLVP) Our Twitter Profiles, Our Selves: Predicting Personality with Twitter

U [26] se analizira veza između karakteristika ličnosti i različitih tipova korisnika na socijalnoj mreži Twitter (Slika 6). Autori su oformili korpus sa podacima o 335 Twitter korisnika, a zatim napravili analizu odnosa pet-faktorskog modela ličnosti sa pet tipova različitih Twitter korisnika: slušaoci (oni koji prate mnoge druge korisnike), popularni, često čitani (eng. highly read), i još dva tipa uticajnih korisnika. Pored toga kreirali su i prediktor ličnosti korisnika zasnovan na tri javno dostupne informacije: broju pratilaca (eng. followers), broju profila koje korisnik prati (eng. following) i broju spominjanja korisnika u listama drugih korisnika (eng. reading lists).

Merenje karakteristika ličnosti je izvršeno indirektno preko Facebook aplikacije *myPersonality*. Naime, ova aplikacija je dostupna Facebook korisnicima, od kojih su neki i korisnici Twitter socijalne mreže. Prikupljeni su rezultati testiranja ličnosti 335 osoba koje su ih objavile paralelno i na svojim Twitter nalozima. Određivanje tipa korisnika je rađeno zasebno za različite tipove, a spomenućemo samo primer detektovanja uticajnih korisnika korišćenjem dve mere. Prva je dobro poznata među analitičarima socijalnih medija, tzv. *Klout* mera (*klout.com*). Ova mera uzima u obzir broj kliknutih „tvitova", onih na koje je odgovoreno, i onih koji su prosleđeni dalje. Druga mera, korišćena od strane TIME magazina za rangiranje javnih ličnosti je jednostavno data sledećim izrazom $(2n_{followers}+n_{facebook})/2$, gde je $n_{followers}$ broj pratilaca na Twitter, a $n_{facebook}$ broj kontakata na Facebook socijalnoj mreži.



Slika 6. AKL u socijalnim mrežama

Korelaciona analiza je pokazala signifikante veze između različitih grupa korisnika i nekih karakteristika ličnosti. Kada je u pitanju problem predikcije ličnosti, autori su ga sveli na ispitavanje tri javno dostupne informacije. Za svaku od pet karakteristika ličnosti je zasebno sprovedena regresiona analiza sa unakrsnom validacijom upotrebom M5' Rules algoritma [34]. Korišćenjem ovog algoritma dobijena je greška predviđanja od 0.88 po RMSE (root-mean-square error) skali, koja uzima vrednosti iz intervala [1,5]. Prema objašnjenju autora ova greška je mala, jer je za, npr. najbolji kolaborativni filter algoritam za predviđanje korisničke ocene filmova, za koji je kompanija Netflix dodelila nagradu od million dolara, RMSE iznosio 0.8567.

4. IDEJE ZA BUDUĆI RAD

U ovom delu iznosimo nekoliko ideja za unapređenje AKL. Radi preglednosti, ideje smo kategorisali na sledeći način: A) novi tipovi korpusa, B) novi načini merenja karakteristika ličnosti, C) novi metodi za klasifikaciju i D) dinamički AKL.

A) Veliki broj istraživanja na temu AKL je razmatrao internet blogove i studentske eseje. U [29] uzete su u obzir i recenzije korisnika iz *tripadvisor.com* sistema za predlaganje turističkih destinacija, ali samo u cilju evaluacije modela koji su formirani na osnovu korpusa studentskih eseja iz [25]. Naše mišljenje je da bi korisničke recenzije i komentari trebali da se iskoriste za kreiranje, a ne samo za testiranje kvaliteta modela. Novinski portali i komentari koje šire stanovništvo ostavlja na njima bi mogli da budu upotrebljeni u rešavanju problema određivanja nekih karakteristika

ličnosti, npr. ekstravertnosti, neuroticizma i otvorenosti ka iskustvu. Na sličan način bi mogli da se upotrebe i komentari sa internet lokacije *youtube.com.* Analiza javnih tokova podataka u socijalnim mrežama je još jedan od resursa koji je trenutno neiskorišćen u cilju rešavanja problema AKL. U narednom paragrafu ćemo razmotriti alternativne, prevashodno indirektne načine merenja karakteristika ličnosti na osnovu ovih novih korpusa.

B) Merenje karakteristika ličnosti učesnika je težak zadatak, jer zahteva kooperaciju osoba koji su autori sadržaja. Naša ideja je da se pojednostavljenjem samog problema AKL pojednostavi i ovaj aspekat istraživanja, koji je obično zasnovan na upotrebi kompleksnih upitnika. U kontekstu novniskih portala i *Youtube*, ovo bi podrazumevalo označavanje (tagovanje) članaka i video sadržaja nekom prožimajućom karakteristikom ličnosti ili kombinacijom karakteristika, npr., naučni sadržaji jasno upućuju na otvorenost ka iskustvu, ekstremni sportovi na kombinaciju ekstrovertnosti i neuroticizma, i slično. Na ovaj način bi se simultano dobio i korpus sadržaja i skup informacija o nekim karakteristikama ličnosti autora tih sadržaja. U kontekstu merenja karakteristika ličnosti na socijalnim mrežama, npr. na Facebook-u, ideja bi bila da se napravi aplikacija koja je zanimljiva korisnicima, a pritom ima prikrivenu funkcionalnost merenja karakteristika ličnosti.

C) Postoji pregršt potencijalnih algoritama koji bi mogli da se primene u rešavanju problema AKL. Kombinovanje, odnosno hibridizacija već postojećih metoda, je jedan od načina za postizanje boljih rezultata, npr. u [30] se razmatra regresioni model, čija je evaluacija zasnovana na rangiranju umesto na standardnom pristupu zasnovanom na rezidualima. Svođenje problema AKL na problem klasterovanja je takođe jedna od alternativa. Ono bi podrazumevalo posmatranje više karakteristika ličnosti istovremeno, a ne svake od njih zasebno. Ovakav pristup bi, po našem mišljenju, bio prirodniji, jer verujemo da su karakteristike međusobno uslovljene, tj. ne mogu se kombinovati na potpuno proizvoljne načine. Pored metode podržavajućih vektora (SVM), ostale metode iz klase *soft computing* algoritama, poput neuralne mreže i metode rasplinutih logika (eng. fuzzy logics) bi mogle dati dobre rezultate.

D) Pod dinamičkim AKL sistemom podrazumevamo sistem koji je u stanju da se prilagođava korisnicima i ulaznim podacima i da „uči“ kroz upotrebu. Pokazano je da psihološke karakteristike prate normalnu raspodelu, međutim, momenti raspodele se razlikuju u pogledu različitih starosnih doba, demografskih karakteristika, stepena obrazovanja i drugih informacija o autoru sadržaja. Dinamički AKL sistem bi takođe učio parametre distribucije kroz korišćenje, a ovo bi moglo da se postigne primenom Bajesovog učenja.

5. ZAKLJUČAK

U ovom radu je napravljen pregled postojećih rešenja za problem automatizovane klasifikacije ličnosti. Predložena je klasifikacija tih rešenja, koja je omogućila sistematičnu analizu istraživanja na temu AKL. Izneli smo nekoliko ideja koje bi mogle da doprinesu poboljšanju kvaliteta rešenja za AKL. Prema našem mišljenju, u skoroj budućnosti će doći do široke, direktne i indirektne upotrebe AKL u okviru sistema za predlaganje, socijalnim mrežama i ekspertskim sistemima.

ZAHVALNOST

Ovo istraživanje je podržano od strane projekata br. 174010 i br. III44006 koje finansira Ministarstvo prosvete i nauke Republike Srbije.

LITERATURA

[1] Argamon, S., Chase, P., Dhawle, S., Raj, S., Navendu, H., and Levitan, G. S., *"Stylistic text classification using functional lexical features"*, Journal of the American Society of Information Science 7, pp 91-109, 2007.

[2] Buchanan, T., Johnson, J. A., and Goldberg, L. R., *"Implementing a five-factor personality inventory for use on the internet"*, European Journal of Psychological Assessment 21, 2, pp 115-127, 2005.

[3] Celli, F., *"Mining user personality in twitter",* Tech. rep., 2011.

[4] Celli, F., *"Unsupervised recognition of personality from linguistic features"*, Tech. rep., 2011.

[5] Celli, F., *"Unsupervised personality recognition for social network sites"*, In Proceedings of the Sixth International Conference on Digital Society ICDS, 2012.

[6] Coltheart, M., *"The MRC psycholinguistic database"*, Quarterly Journal of Experimental Psychology 33A, pp 497-505, 1981.

[7] Efimova, L. and Moor, A. D., *"Beyond personal webpublishing: An exploratory study of conversational blogging practices"*, In Proceedings of HICSS'05 - Track 4 - Volume 04, IEEE Computer Society, Washington, DC, USA, 2005.

[8] Eysenck, H. and Eysenck, S., *"Eysenck Personality Questionnaire-Revised",* Hodder, London, 1991.

[9] Gill, A., *"Personality and language: The projection and perception of personality in computer-mediated communication",* Ph.D. thesis, University of Edinburgh, 2003.

[10] Gill, A. J., Nowson, S., and Oberlander, J., *"What are they blogging about? personality, topic and motivation in blogs"*, ICWSM, 2009.

[11] Gill, A. J. and Oberlander, J., *"Taking care of the linguistic features of extraversion",* In Proceedings of the 24th Annual Conference of the Cognitive Science Society, pp 363-368, 2002.

[12] Halliday, M. A. K., *Introduction to Functional Grammar*, 2 ed. Edward Arnold, 1994.

[13] Heylighen, F. and Marc Dewaele, J., *"Variation in the contextuality of language: An empirical measure in Context"*, Special issue of Foundations of Science, pp 293-340, 1998.

[14] Iacobelli, F., Gill, A. J., Nowson, S., and Oberlander, J., *"Large scale personality classification of bloggers",* In Proceedings of the 4th international conference on Affective computing and intelligent interaction - Volume Part II, ACII'11, Springer-Verlag, Berlin, Heidelberg, pp 568-577, 2011.

[15] John, O. P., Donahue, E. M., and Kentle, R. L., *"The big five inventory: Versions 4a and 5b"*, Tech. rep., Berkeley: University of California, Institute of Personality and Social Research, 1991.

[16] Mairesse, F. and Walker, M. A., *"Words mark the nerds: Computational models of personality recognition through language",* In Proceedings of the 28th Annual Conference of the Cognitive Science Society, 2006.

[17] Mairesse, F., Walker, M. A., Mehl, M. R., and Moore, R. K., *"Using linguistic cues for the automatic recognition of personality in conversation and text",* Journal of Artificial Intelligence Research, Vol 30, pp 457-501, 2007.

[18] Minamikawa, A. and Yokoyama, H., *"Personality estimation based on weblog text classification"*, In Proceedings of the 24th international conference on Industrial engineering and other applications of applied intelligent systems conference on Modern approaches in applied intelligence - Volume Part II. IEA/AIE'11, Springer-Verlag, Berlin, Heidelberg, pp 89-97, 2011.

[19] Nilsson, S., *"The function of language to facilitate and maintain social networks in research weblogs",* Ph.D. thesis, Umea Universitet, Engelska Lingvistik, 2003.

[20] Nowson, S., *"The language of weblogs: A study of genre and individual differences"*, Ph.D. thesis, University of Edinburgh, 2006.

[21] Nowson, S., Oberlander, J., and Gill, A. J., *"Weblogs, genres and individual differences",* In Proceedings of the 27th Annual Conference of the Cognitive Science Society, Cognitive Science Society, pp 1666-1671, 2005.

[22] Oberlander, J., *"Whose thumb is it anyway? classifying author personality from weblog text"*, In Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL), pp 627-634, 2006.

[23] Paul, K., *"Text messaging and personality"*, M.S. thesis, Ball State University, 2011.

[24] Pennebaker, J. W., Francis, M. E., and Booth, R. J., *"Inquiry and Word Count: LIWC"*, Lawrence Erlbaum, Mahwah, NJ, 2001.

[25] Pennebaker, J. W. and King, L. A., *"Linguistic styles: Language use as an individual difference",* Journal of Personality and Social Psychology 77, pp 1296-1312, 1999.

[26] Quercia, D., Kosinski, M., Stillwell, D., and Crowcroft, J., *"Our twitter profiles, our selves: Predicting personality with twitter",* In Proceedings of the 3rd IEEE Conference on Social Computing (SocialCom), 2011.

[27] Quercia, D., Lambiotte, R., Kosinski, M., Stillwell, D., and Crowcroft, J., *"The personality of popular facebook users"*, 2011.

[28] Rigby, P. C. and Hassan, A. E., *"What can oss mailing lists tell us? a preliminary psychometric text analysis of the apache developer mailing list",* In Proceedings of the 4th International Workshop on Mining Software Repositories. MSR '07, IEEE Computer Society, Washington, DC, USA, 2007.

[29] Roshchina, A., Cardiff, J., and Rosso, P., *"A comparative evaluation of personality estimation algorithms for the twin recommender system",* In Proceedings of the 3rd international workshop on Search and mining user-generated contents, SMUC '11. ACM, New York, NY, USA, pp 11-18, 2011.

[30] Rosset, S., Perlich, C., and Zadrozny, B., *"Ranking-based evaluation of regression models",* Knowledge and Information Systems 12, 3, pp 331-353, 2007.

[31] Sushant, S. A., Argamon, S., Dhawle, S., and Pennebaker, J. W., *"Lexical predictors of personality type",* In Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America, 2005.

[32] Wang, F., Hong, Y., Zhang, W., and Agrawal, G., *"Personality based latent friendship mining",* In DMIN 2009, R. Stahlbock, S. F. Crone, and S. Lessmann, Eds. CSREA Press, pp 427-433, 2009.

[33] Wehrli, S., *"Personality on social network sites : An application of the five factor model personality on social network sites"*, Sociology The Journal Of The British Sociological Association 7, pp 1-17, 2008.

[34] Witten, I. H., and Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, 1999.

[35] Yarkoni, T., *"Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers",* Journal of Research in Personality 44, 3 (June), pp 363-373, 2010.