Primjena mašinskog učenja u procesu klasifikacije oglašenih radnih mjesta

Branislava Cvijetić
Agencija za statistiku Bosne i Hercegovine
Sarajevo, Bosna i Hercegovina
branislava.cvijetic@bhas.gov.ba

Zaharije Radivojević
Elektrotehnički fakultet, Univerzitet u Beogradu
Beograd, Republika Srbija
zaki@etf.bg.ac.rs

Sažetak—**Institucije koje proizvode službene statistike teže da što više, pored redovnih statističkih istraživanja, primjenjuju eksterne izvore podataka, kao što su administrativni izvori podataka. Osim navedenih izvora podataka, veliki podaci su takođe prepoznati kao novi izvori podataka. Jedan od osnovnih zadataka u statistici, bez obzira o kom izvoru podataka je riječ, jeste automatska klasifikacija tekstualnih podataka. U ovom radu korišćeni su algoritmi mašinskog učenja nad podacima prikupljenim sa četiri veb sajta, a sve u cilju da se provjeri uspješnost primjene algoritama mašinskog učenja prilikom automatske klasifikacije tekstualnih nestrukturiranih podataka.**

Ključne riječi- veliki podaci; mašinsko nadgledano učenje; klasifikacija tekstualnih podataka; obrada govornog jezika

I. Uvod

Veliki podaci (engl. *Big data*) su tehnologija koja služi za prikupljanje, obradu i analizu velike količine podataka. Ovi podaci mogu biti strukturirani, polu-strukturirani i nestrukturirani, a generišu se i pristižu velikom brzinom i to u različitim intervalima (ponekad i u realnom vremenu). Sve navedeno ove podatke čini vrlo složenim za analizu. Međutim, prikupljanje i skladištenje velikih količina podataka nije ono što samo obilježava tehnologiju velikih podataka. Zapravo, mogućnost obrade i analiza tih prikupljenih podataka za dalju upotrebu, kao i izdvajanje informacija iz datih podataka, je ono što ovu tehnologiju čini značajnom. Bez mogućnosti analize i potrebnih alata za izdvajanje informacija, bila bi to samo gomila prikupljenih podataka. Mogućnost primjene zahvatila je mnoga područja, počev od analize sadržaja neke objave na socijalnim mrežama, pa čak i do politike, odnosno analize podataka o ponašanju i javno dostupnim mišljenjima birača s ciljem kreiranja efikasnih izbornih kampanja [1].

Evropski statistički sistem (engl. *European Statistical System - ESS*) sa svojim strateškim partnerima, 2010. godine [2], prepoznao je značaj korišćenja velikih podataka prilikom proizvodnje zvanične statistike. U posljednjih desetak godina kreirane su strategije, pokrenuti projekti, organizovani seminari, a sve sa ciljem da statističke institucije budu spremne da odgovore izazovu zvanom veliki podaci, tj. da se pripreme za integraciju ovog izvora podataka u proizvodnju zvanične statistike. Jedna od komponenti projekta ESSnet Big Data [3] odnosi se na statističke procjene vezane za oglašavanje radnih mjesta na Internetu. Predstavnici zvaničnih statističkih institucija Bosne i Hercegovine nisu uključeni u ovaj projekat, pa je većina informacija o tome šta je urađeno, šta se trenutno radi i šta se planira uraditi, dostupna putem Interneta.

Jedan od izazovnih zadataka u institucijama koje proizvode službene statistike, bez obzira da li su izvor podataka redovna statistička istraživanja, administrativni izvori ili veliki podaci, jeste automatska klasifikacija tekstualnih podataka [4], [5].

Automatska klasifikacija teksta, koja se još naziva i kategorizacija teksta, ima istoriju koja datira sa početka 1960-ih. No, značajan porast dostupnih dokumenata u posljednje dvije decenije, razvoj algoritama mašinskog učenja, razvoj algoritama dubokog učenja (engl. *deep learning*) inspirisanih radom ljudskog mozga, obnovio je interes za automatizovanu klasifikaciju teksta. U početku je klasifikacija teksta bila usmjerena na rješavanje zadatka primjenom skupa pravila temeljenih na stručnom znanju domenskih stručnjaka [6], a danas je fokus na metodama automatske klasifikacije teksta pomoću algoritama mašinskog učenja [7], [8], [9], kao i algoritama dubokog učenja [10].

Literatura koja se odnosi na klasifikaciju radnih mjesta snažno je ukorijenjena u američkoj tradiciji, tj. pretežno u klasifikaciji naziva radnih mjesta korišćenjem O*NET (engl. *Occupational Information Network*) klasfikacijskog sistema, koji potiče iz Ministarstva rada Sjedinjenih američkih država (SAD). Sa druge strane u zemljama Evropske unije, od 2009. godine, koristi se međunarodni standard za klasifikaciju zanimanja (engl. *International Standard Classification of Occupations ISCO-08*) [11]. Osim toga, veliki broj veb stranica za traženje posla ima sjedište u SAD, kao na primjer Monster i CareerBuilder, dok je nekolicina, poput StepStone-a, sa sjedištem u Evropi. Kao posljedica toga u radovima [12], [13] može se pronaći opis sistema za klasifikaciju radnih mjesta koji se koristi u CareerBuilder veb sajtu. Radovi relevantni za evropske zemlje [14], [15], su uglavnom oni gdje se klasifikacija radnih mjesta vrši u skladu sa ISCO-08.

Svrha ovog rada jeste da se nad nestrukturiranim podacima prikupljanim sa Interneta, odnosno nad podacima o oglašenim radnim mjestima u domenu informaciono-komunikacionih tehnologija, primijene algoritmi mašinskog učenja u cilju automatske klasifikacije teksta u više klasa po ISCO-08. U prikupljenom setu podataka opisi radnih mjesta pisani su na engleskom, njemačkom, srpskom i grčkom jeziku. Postoji veliki broj algoritama mašinskog učenja koji se mogu koristiti u procesu klasifikacije teksta, a u ovom radu biće primijenjena

dva algoritma mašinskog učenja: Multinominalni Naive Bajes i metoda potpornih vektora, na višejezičnom setu podataka. Algoritmi su odabrani u skladu sa rezultatima prikazanim u [8] i [9], kao i zbog malog seta prikupljenih podataka nad kojima će se raditi eksperiment. Takođe ovi algoritmi su odabrani jer ne zahtijevaju mnogo računarskih resursa. Ovaj rad predstavlja početak istraživanja sa ciljem da se utvrdi kako se klasifikacija višejezičkih tekstualnih opisa izvršena sa algoritama mašinskog učenja, može koristiti u proizvodnji zvanične statistike.

Ostatak rada organizovan je na način da je u narednom poglavlju dat kratak teorijski osvrt na automatsku klasifikaciju teksta i algoritme mašinskog učenja koji su primijenjeni u eksperimentalnom dijelu. U trećem poglavlje opisan je način prikupljanja nestrukturiranih podataka sa Interneta, te postupak kreiranja seta podataka na osnovu javno dostupnih podataka, koji će se koristiti u eksperimentalnom dijelu. Četvrto poglavlje odnosi se na eksperimentalni dio, tj. proces klasifikacije teksta u više klasa postupkom nadgledanog mašinskog učenja. Na kraju je dat zaključak i navedena lista korišćene literature.

II. KLASIFIKACIJA TEKSTA I ALGORITMI MAŠINSKOG UČENJA

A. *Klasifikacija teksta*

Klasifikacija teksta je postupak razvrstavanja tekstualnih nizova različite dužine u različite klase u zavisnosti od sadržaja nizova. Klasifikacija teksta ima različitu primjenu, a neki od primjera su: razdvajanje kritika o filmu na pozitivne ili negativne, razdvajanje elektronske pošte na neželjenu poštu (spam) ili legitimnu elektronsku poštu, razdvajanje novinskih članaka na politiku, sport, kulturu i sl. U navedenim primjerima svakoj instanci (filmska kritika, elektronska pošta, novinski članak) može se dodijeliti oznaka klase – ciljna vrijednost, kojoj ta instanca pripada. Ako skup ciljnih vrijednosti sadrži tačno dvije klase onda se radi o binarnoj klasifikaciji teksta. Primjer binarne klasifikacije teksta bilo bi razdvajanje filmskih kritika na pozitivne ili negativne. Klasifikacija u više klasa znači klasifikacijski zadatak s više od dvije klase. Primjer klasifikacije sa više klasa je podjela žanra filma po raspoloženju na: akcija, avantura, komedija, drama, fantazija, horor, triler itd. Klasifikacija u više klasa pretpostavlja da je svaka instanca dodijeljena jednoj i samo jednoj klasi. To znači da žanr filma po raspoloženju može biti drama ili horor, ali ne istovremeno i drama i horor. U slučaju kada tekst može spadati istovremeno u više klasa, onda je riječ o više-kategorijskoj klasifikaciji (engl. *Multi-label classification*).

Za potrebe ovog rada klasifikacija teksta se može definisati kao postupak razvrstavanja tekstualnih nizova različite dužine (opisa oglašenih radnih mjesta iz oblasti informacionih tehnologija) u različite kategorije (klase zanimanja prema međunarodnom standardu za klasifikaciju zanimanja ISCO - 08) u zavisnosti od sadržaja nizova, koristeći algoritme mašinskog učenja.

B. *Algoritmi za klasifikaciju teksta*

Najvažniji korak kod rješavanja problema klasifikacije teksta jeste izabrati najbolji algoritam. Postoji veliki broj algoritama mašinskog učenja, a u eksperimentalnom dijelu ovog radu su primijenjena sljedeća dva algoritma:

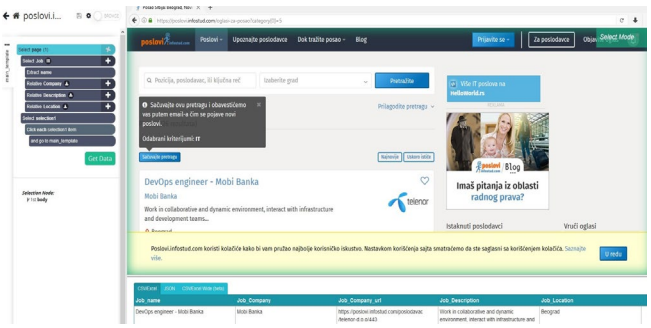
- Multinominalni Naive Bajes (engl. *Multinomial Naive Bayes MNB*) spada u skup Naive Bajes (Naive Bayes) algoritama klasifikacije. Jedna od prednosti ovog algoritma je da daje veoma dobre rezultate kada nema dovoljno podataka i zahtjeva malo računarskih resursa [8]. Ovaj algoritam baziran je na teoremi engleskog statističara i filozofa Tomas Bajesa (Thomas Bayes), a prema kojoj je moguće izračunati vjerovatnoću pojave nekog događaja na osnovu vjerovatnoće pojave svakog pojedinog događaja. To znači da će svaki vektor koji predstavlja tekst morati sadržavati informacije o vjerovatnoći pojavljivanja riječi teksta u tekstovima određene klase, da bi algoritam mogao izračunati vjerovatnoću da taj tekst pripada nekoj klasi.
- Metoda potpornih vektora (engl. *Support Vector Machine SVM*) je još jedna od popularnih tehnika koja se može primijeniti za klasifikaciju teksta. Kao i Naive Bajes, SVM ne zahtjeva mnogo podataka za obuku (engl. *training data*) da bi dao dobre rezultate [8].

III. KREIRANJE SET PODATAKA

Da bi se obučio sistem u režimu nadgledanog učenja, potreban je označen ulazni set podataka. Označeni podaci, za potrebe ovog rada, treba da sadrže šifru zanimanja po ISCO-08 i naziv oglašenog radnog mjesta. Na osnovu podataka koji su prikupljeni putem Interneta i razvijenog poluautomatskog načina dodjeljivanja šifre zanimanja, kreiran je potrebni set podataka. U tekstu niže opisan je proces kreiranja seta podataka koji je kasnije korišćen u eksperimentalnom dijelu rada.

A. *Prikupljanje podataka*

Podaci koji se nalaze na vanjskim izvorima, tj. veb stranicama, prikupljaju se alatima za učitavanje nestrukturiranih podataka (engl. *Web-scraping tools*). U ovom radu korišćen je specijalizovan alat ParseHub (Sl. 1). Naime, u ParseHub aplikaciji kreirana su četiri projekta koja prikupljaju podatke sa četiri različita veb sajta (www.stepstone.at, www.stepstone.de, www.jobfind.gr, poslovi.infostud.com), na četiri različita jezika (engleskom, njemačkom, srpskom i grčkom), a prikupljani su podaci o oglašenim radnim mjestima u oblasti informacionih tehnologija. Podaci su prikupljani pet puta, a prikupljani su sljedeći atributi: naziv radnog mjesta, grad, datum objave oglasa, naziv kompanije i url za otvoreni oglas.



Slika 1. Primjer jednog projekta u ParseHub alatu

B. *Kreiranje seta podataka*

U Tabeli I prikazana je statistika broja oglašenih IT radnih mjesta na četiri veb sajta po datumima prikupljanja podataka. Iz navedenog seta podataka izdvojena su samo oglašena IT radna mjesta koja su jedinstvena po svom nazivu.

TABELA I. BROJ OGLAŠENIH RADNIH MJESTA

	Datum prikupljanja podataka					**Ukupno**
	14.12. 2019.	***22.12. 2019.***	***29.12. 2019.***	***05.01. 2020.***	***12.01. 2020.***	
www.stepstone.at	868	884	822	764	734	**4072**
www.stepstone.de	837	817	768	777	801	**4000**
www.jobfind.gr	149	140	120	110	128	**647**
poslovi.infostud.com	305	302	271	248	238	**1364**
	2159	**2143**	**1981**	**1899**	**1901**	**10083**

Nakon što su uklonjeni svi dupli zapisi, dobijeno je 2984 zapisa. Do dupliranja je došlo zbog činjenice da je veliki broj oglasa, tačnije 883, bio aktivan tokom vremenskog perioda prikupljanja podataka (5 nedelja). Takođe isti naziv radnog mjesta mogao se pojaviti na različitim veb stranicama, kao i na istoj veb stranici.

Za potrebe eksperimentalnog dijela, tj. za proces nadgledanog mašinskog učenja, svakom oglašenom radnom mjestu iz seta podataka, bilo je potrebno dodijeliti odgovarajuću šifru prema međunarodnoj klasifikaciji zanimanja ISCO08. Kao polazna osnova za ovaj postupak iskorišćen je set podataka od 363 zapisa kojima je ručno dodijeljena šifra zanimanja. Nakon detaljne analize seta podataka koji su ručno prešifrirani, započeo je postupak kreiranje seta podataka za potrebe eksperimenta, na sljedeći način:

1. Kreirana je tabela koja sadrži skraćenice i sinonime. Tabela sadrži ukupno 98 zapisa. (Sl. 2)

	abr_in	abr_out
1	Datenbank	database
2	Daten-bank	database
3	Netzwerk	network
4	sap basis	database
5	dba	database
6	baza podataka	database
7	hardware	network

Slika 2. Dio tabele skraćenica i sinonima

2. Kreirana je kontrolna tabela sa pravilima u vidu kombinacija jedne ili dvije ključne riječi. Svako pravilo ima dva nivoa sa kojima je definisan redoslijed izvršavanja pravila. Ova tabela sastoji se od 73 zapisa. (Sl. 3)

	Level1	Level2	Code	Word1	Word2
1	1	1	2521	administrator	database
2	1	2	2521	analyst	database
3	1	3	2521	architect	database
4	1	4	2514	developer	database
5	1	4	2521	specialist	database
6	6	1	2522	administrator	network
7	10	5	2522	engineer	network
8	10	10	2522	developer	network

Slika 3. Dio kontrolne tabele

3. Kreiranim SQL procedurama, koje konsultuju tabele iz prethodne dvije tačke, automatski su dodijeljene šifre zanimanja svim opisima radnih mjesta iz seta podatka. Ovaj set podataka imao je ručno dodjeljene šifre zanimanja.
4. Ovako dobijene šifre su upoređene sa šiframa koje su ručno dodjeljene. Utvrđeno je poklapanje na 325 zapisa ili 89.53 %.
5. Pošto je procenat poklapanja dodjeljenih šifara zanimanja sa šiframa koje su ručno dodjeljene bio zadovoljavajući, na isti način je urađena dodjela šifara zanimanja i setu podataka koji je prikupljen sa Interneta. Na ovaj način dodijeljena je šifra na 2054 zapisa od ukupno 2984 zapisa, tj. za 68.83 % zapisa automatski je dodjeljena šifra zanimanja.
6. Kreiran je set podataka kao skup podataka iz tačke 4 (325) i seta podataka iz tačke 5 (2054). Pošto je bilo preklapanja u zapisima izdvojeni su samo jedinstveni zapisi, pa konačni set podataka koji će dalje koristiti u eksperimentalnom dijelu, sastoji se od 2323 zapisa. (Sl. 4)

	JobDescription	isco08
1	(Junior) Datenbankadministrator (m/w)	2521
2	Administrator baza podataka i aplikativnih platfor...	2521
3	Administrator baza podataka i aplikativnih platfor...	2521
4	Database Administrator	2521
5	Database Administrator (DBA)	2521
6	Database Administrator (m/f/d)	2521
7	Database Administrator *	2521
8	Database Analyst	2521
9	Database Architect / Developer (f/m/d)	2521
10	Database Architect/Engineer	2521
11	Database Engineer/Architect PostgreSQL (m/f/d)	2521
12	Database Specialist (f/m)	2521
13	Datenbank Spezialist (m/w)	2521

Slika 4. Dio finalnog seta podataka koji će se koristiti za eksperiment

IV. EKSPERIMENT

Za potrebe eksperimentalnog dijela korišćen je programski jezik *Python* kroz *Jupyter Notebook*. *Scikit-Learn* je korišćen kao komponenta za mašinsko učenje, kao i *Natural Language Toolkit – NLTK* koji se primjenjuje kod obrade govornog jezika (engl. *Natural Language Processing - NLP*). Kao pomoćni alat korišćen je softver Orange.

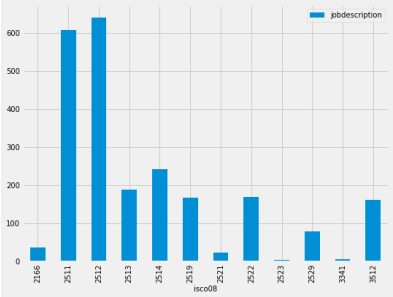
A. *Analiza podataka*

Analizom trening seta podataka utvrđeno je sljedeće:

- Set podataka sadrži 12 klasa prema ISCO-08.
- Podaci nisu ravnomjerno raspoređeni po klasama, odnosno utvrđeno je da se radi o neizbalansiranom setu podataka (engl. *imbalanced dataset*). (Sl. 5)
- Set podataka se sastoji od 11772 riječi. (Sl. 6)

Kada se uoči problem neizbalansirane raspodjele klasa, on se teško može riješiti standardnih algoritmima mašinskog učenja. Konvencionalni algoritmi često su pristrasni prema klasi koja je najviše zastupljena i ne uzimaju u obzir distribuciju podataka. U najgorem slučaju, manjinske klase se tretiraju kao izuzeci i zanemaruju se. U nekim slučajevima, kao na primjer kada se u medicini vrši predviđanje da li se radi o

nekoj rijetkoj bolesti, potrebno je vještački uravnotežiti skup podataka. Postoji više tehnika kako se to može uraditi, a u ovom radu o tome neće biti riječi. U ovom eksperimentu, a uzimajući u obzir da se radi o potencijalnom statističkom izvoru podataka, može se prihvatiti rizik da će algoritmi dati bolje rezultate kod većinskih klasa i da se možda zanemari preciznost kod manjinskih klasa, kao na primjer za klase 2523 ili 3341.



Slika 5. Distribucija seta podataka – broj oglašenih radnih mjesta prema isco08

Slika 6. Word Cloud seta podataka generisan u Orange softveru

B. Prethodna obrada teksta

Prethodna obrada teksta (engl. *Text preprocessing*) ima zadatak da tekst pretvori u oblik koji bi olakšao i poboljšao rad algoritma mašinskog učenja.

Kada se govori o prethodnoj obradi teksta može se reći da se sastoji od sljedećih komponenti:

- Tokenizacija (engl. *Tokenization*) predstavlja razdvajanje nizova teksta na manje dijelove ili „tokene“. Paragraf se može razdvojiti u rečenice, a rečenice se dalje mogu razdvojiti u riječi.
- Normalizacija (engl. *Normalization*) ima za cilj da cjelokupan tekst stavi u neki jedinstven oblik, na primjer sve znakove da pretvori u mala slova. U normalizaciju teksta spada i morfološka normalizacija teksta, kao što je na primjer korjenovanje riječi (engl. *Stemming*).
- Filtriranje (engl. *Filtering*) ima za cilj dobijanje „čistijeg“ teksta. To se postiže na primjer uklanjanjem praznih stringova, specijalnih karaktera, brojeva i sl. U filtriranje se ubraja i uklanjanje iz teksta nekih riječi koje imaju mali ili nemaju nikakav značaj u tekstu, tzv. stop riječi (engl. *Stop Words*). To su na primjer, veznici ili neke druge specifične riječi koje se mogu pronaći u tekstu.

Prethodna obrada teksta, tj. prikupljenih naziva radnog mjesta, obuhvatila je uklanjanje: specijalnih znakova, brojeva, praznih znakova, kao i stop riječi za engleski, grčki i njemački jezik. Takođe su uklonjene i neke specifične stop riječi koje nisu relevantne za klasifikaciju zanimanja, kao na primjer riječi koje su vezane za poziciju (*senior*, *junior*, *head*, *lead*, *expert*, *assistant*, *student*), potrebne uslove rada (*experience*, *knowledge*, *professional*, *german*), grad ili zemlja (*athens*, *stockholm*, *greece*, *germany*, *wien*, *berlin*, *thessaloniki*, *belgium*), radno vrijeme (*vollzeit*, *teilzeit*, *part*), rodnu pripadnost i sl.

Tokenizacija je urađena na nivo riječi i korištena su mala slova.

Nakon završene obrade broj riječi je sa 11772 smanjen na 8064 riječi. Na Sl. 7 prikazan je Word Cloud za obrađen tekst. Ukoliko se ovaj rezultat uporedi sa rezultatom prikazanim na Sl. 6 mogu se primijetiti efekti primijenjene obrade teksta.

Slika 7. Word Cloud obrađenog seta podataka generisan u Orange softveru

C. Izdvajanje karakteristika

Algoritmi mašinskog učenja ne mogu direktno obrađivati tekst kakav je u izvornom obliku. Većina algoritama očekuje da radi sa numeričkim vektorima karakteristika (engl. *feature vectors*) fiksne dužine, a ne sa izvornim tekstom promjenljive dužine. Stoga se, prije izgradnje modela, trebaju izdvojiti karakteristike iz teksta.

Za ovaj postupak korišten je model skupa riječi (engl. *Bag of Words*). U Scikit-Learn modulu izdvajanje odlika, tj. postupak pretvaranja teksta u numeričke vektore odlika, se naziva vektorizacija. U modelu skupa riječi se za svaki dokument, a u ovom eksperimentu za opis radnog mjesta, uzima u obzir prisustvo, kao i frekvencija svake riječi, ali ne i redoslijed kako se te riječi pojavljuju. Prilikom kreiranja ovog modela, tj. izdvajanja karakteristika, postavljeno je da se neka riječ treba pojaviti najmanje 5 puta u čitavom setu podataka da bi se uzela u obzir. Ukupan broj izdvojenih karakteristika u slučaju ovog seta podataka je 335.

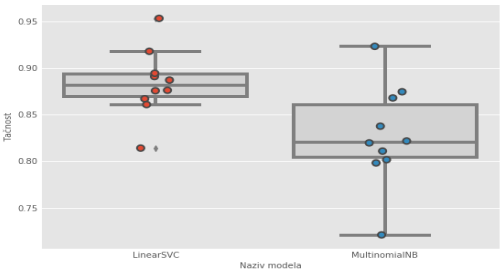
D. Kreiranje modela klasifikovanja teksta

Implementacija klasifikatora u Python programskom jeziku je vrlo jednostavna ako se koristi Scikit-Learn skup alata koji sadrži brojne algoritme za klasifikovanje, kao i multinominalni Naive Bajes i metoda potpornih vektora. Dalje u tekstu će se

ova dva algoritma označavati sa MultinominalNB i LinearSVC.

Za izradu modela klasifikacije teksta korišten je skup od 2323 opisa radnih mjesta sa opcijom desetostrukog unakrsnog testiranja (engl. *10-fold cross-validation*).

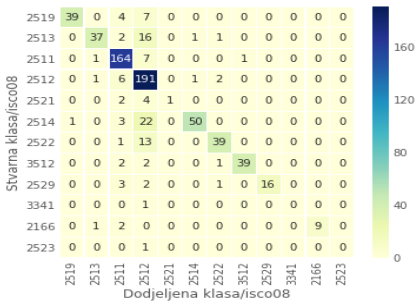
Na Sl. 8 prikazani su rezultati kreiranja modela klasifikacije sa desetostrukom unakrsnom validacijom.



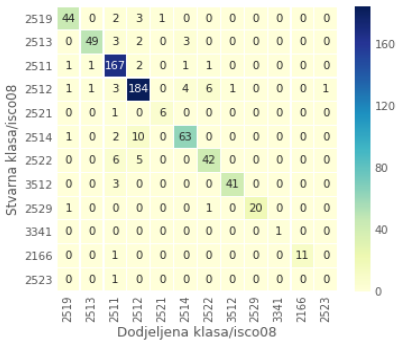
Slika 8. Rezultat obučavanja algoritama LinearSVC i MultinominalNB

E. *Testiranje modela*

Kod klasifikacije u više klasa za procjenu modela može se koristiti matrica zabune (engl. *confusion matrix*). Prema [16] matrica zabune je tabela u kojoj svaka ćelija [*i,j*] sadrži broj koliko je puta oznaka klase *j*, bila jednaka oznaci klase *i*, koja je dodijeljena ručno svakom zapisu testnog skupa podataka. Dijagonala matrice predstavlja ispravne klasifikacije, dok svi ostali elementi matrice predstavljaju grešku u klasifikaciji. Za MultinominalNB klasifikator matrica zabune prikazana je na Sl. 9, dok Sl. 10 prikazuje matricu zabune za LinearSVC klasifikator.



Slika 9. Matrica zabune za MultinominalNB



Slika 10. Matrica zabune za LinearSVC

model_name	Srednja tačnost	Standardna devijacija
LinearSVC	0.88362	0.036363
MultinomialNB	0.82766	0.054005

Slika 11. Parametri procjene

Prvi parametri koji su dobijeni prilikom testiranja algoritama mašinskog učenja, prikazani su na Sl. 11. Algoritam LinearSVC ima bolju srednju tačnost. Tačnost predstavlja procenat slučajeva (instanci) koji su uspješno (korektno) klasifikovani. Na Sl. 9 i Sl. 10 to je prikazano kao dijagonala kod koje je u svakoj ćeliji upisan brojem uspješnih klasifikacija.

U slučaju neravnomjerne raspodjele instanci između klasa, ova mjera je nepouzdana. Kao dodatni parametri procjene koriste se preciznost (engl. *Precision*) i opoziv (engl. *Recall*). Ukoliko su visoke vrijednosti procjene, kako za preciznost tako i opoziv, to ukazuje da klasifikator vraća tačne rezultate (preciznost), kao i da vraća većinu svih pozitivnih rezultata (opoziv). Idealan sistem visoke preciznosti i velikog opoziva vratiće mnoge rezultate, a svi će rezultati biti pravilno označeni. Pošto algoritam LinearSVC (Sl. 13) ima bolju sličnost između preciznosti i opoziva, može se reći da on daje bolje rezultate u odnosu na algoritam MultinominalNB (Sl. 12).

	precision	recall	f1-score	support
2519	0.97	0.78	0.87	50
2513	0.93	0.65	0.76	57
2511	0.87	0.95	0.91	173
2512	0.72	0.95	0.82	201
2521	1.00	0.14	0.25	7
2514	0.96	0.66	0.78	76
2522	0.89	0.74	0.80	53
3512	0.97	0.89	0.93	44
2529	1.00	0.73	0.84	22
3341	0.00	0.00	0.00	1
2166	1.00	0.75	0.86	12
2523	0.00	0.00	0.00	1
micro avg	0.84	0.84	0.84	697
macro avg	0.78	0.60	0.65	697
weighted avg	0.86	0.84	0.83	697

Slika 12. Parametri procjene za MultinominalNB

	precision	recall	f1-score	support
2519	0.92	0.88	0.90	50
2513	0.96	0.86	0.91	57
2511	0.88	0.97	0.92	173
2512	0.89	0.92	0.90	201
2521	0.86	0.86	0.86	7
2514	0.89	0.83	0.86	76
2522	0.84	0.79	0.82	53
3512	0.98	0.93	0.95	44
2529	1.00	0.91	0.95	22
3341	1.00	1.00	1.00	1
2166	1.00	0.92	0.96	12
2523	0.00	0.00	0.00	1
micro avg	0.90	0.90	0.90	697
macro avg	0.85	0.82	0.84	697
weighted avg	0.90	0.90	0.90	697

Slika 13. Parametri procjene za LinearSVC

F. *Testiranje modela I*

Da bi se testirao uticaj primjene stop riječi na rezultate klasifikacije, ponovljen je eksperiment u dijelu kreiranja modela, s tim da nisu filtrirane stop riječi. Osnovni parametri procjene prikazani su na Sl. 14. Rezultati su pokazali da postoji

smanjenje srednje tačnosti za oba algoritma kad se ne koriste stop riječi u procesu pripreme seta podataka za postupak klasifikacije.

model_name	Srednja tačnost	Standardna devijacija
LinearSVC	0.868063	0.045942
MultinomialNB	0.813427	0.058623

Slika 14. Parametri procjene za model bez primjenjenih stop riječi

ZAKLJUČAK

U ovom radu istražena je primjena algoritama mašinskog učenja u procesu klasifikacije nestrukturiranih podataka (oglašenih radnih mjesta) koji su prikupljeni sa Interneta. Za potrebe eksperimenta kreiran je set podataka koji sadrži oglašena IT radna mjesta kojima je dodijeljena šifra zanimanja u skladu sa međunarodnim standardom za klasifikaciju zanimanja ISCO-08. Za ovaj set je specifično da sadrži većinu podataka na engleskom jeziku, ali su takođe prisutni podaci napisani na srpskom, grčkom i njemačkom jeziku. Zbog ove specifičnosti kreirana je specijalizovana lista stop riječi.

Klasifikacija je urađena korišćenjem dva algoritma mašinskog učenja, MultinominalNB i LinearSVC. Za set podataka koji je kreiran, utvrđeno je da su oba algoritma dala dobre rezultate. Međutim algoritam LinearSVC, po parametrima procjene, pokazao se kao bolji. Najveći broj odstupanja, tj. u slučaju gdje se prethodno dodijeljena šifra razlikuje od one koje su dodijelili algoritmi, pojavio se u dvije najzastupljenije klase. Analizom rezultata u nekim slučajevima utvrđeno je da je bitan redoslijed pojavljivanja riječi, kao i da bi se isti sadržaj mogao klasifikovati u dvije različite klase. U budućem radu potrebno je poboljšati model za slučaj gdje je bitan redoslijed riječi. Takođe se pokazala tačna pretpostavka da konvencionalni algoritmi mogu zanemariti klase koje su manje zastupljene zbog neizbalansiranog seta podataka.

U fazi eksperimenta izvršena je i provjera uticaja kreirane liste stop riječi na proces klasifikacije. Uočeno je da se primjenom stop riječi povećava procenat srednje tačnosti za oba algoritma mašinskog učenja.

Pokazano je da se algoritmi mašinskog učenja mogu uspješno koristiti za klasifikaciju oglašenih radnih mjesta. Stoga, u budućem radu je potrebno primijeniti stečena znanja i prilikom klasifikacije velikog broja različitih zanimanja koja nisu iz oblasti informacionih tehnologija. U tom slučaju, pored algoritama koji su primijenjeni u ovom radu, trebali bi se koristiti i algoritmi dubokog učenja

ZAHVALNICA

Rad na ovom projektu djelimično je finansiran od strane Ministarstva prosvete i nauke Republike Srbije (III44009). Autori se zahvaljuju na finansijskoj podršci.

LITERATURA

[1] S. Sudhahar, G. Veltri and N. Cristianini, "Automated analysis of the US presidential elections using Big Data and network analysis, " Big Data & Society, vol. 2, 2015.

[2] https://ec.europa.eu/eurostat/cros/content/big-data_en [pristupljeno 15.01.2020.]

[3] https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/ESSnet_Big_Data [pristupljeno 15.01.2020.]

[4] H. Gweon, M. Schonlau, L. Kaczmirek, M. Blohm and S. Steiner, "Three Methods for Occupation Coding Based on Statistical Learning, " Journal of Official Statistics, vol. 33, 2017.

[5] A. Bethmann, M. Schierholz, K.Wenzig and M. Zielonka, "Automatic Coding of Occupations. Using Machine Learning Algorithms for Occupation Coding in Several German Panel Surveys," WAPOR 67th Annual Conference, 2014.

[6] F. Thabtah, P. Cowling and Y. Peng, "MCAR: Multi-class Classification based on Association Rule," International Conference on Computer Systems and Applications, AICCSA, IEEE, 2005.

[7] S. Fatima and B. Srinivasu, "Text Document categorization using support vector machine," International Research Journal of Engineering and Technology, IRJET, vol. 4, 2017.

[8] K. Kowsari, K. Meimandi. M. Heidarysafa, S. Mendu, L. Barnes and D. Brown, "Text Classification Algorithms: A Survey," Information, 2019.

[9] V. Vijayan, Bindu and L. Parameswaran, "A comprehensive study of text classification algorithms," International Conference on Advances in Computing, Communications and Informatics, ICACCI, 2017.

[10] A. Kadhim, "Survey on supervised machine learning techniques for automatic text classification," Springer, 2019.

[11] https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1396369855234&uri=CELEX:32009H0824 [pristupljeno 01.03.2020.]

[12] F. Javed, Q. Luo, M. McNair, F. Jacob, M. Zhao and T.S. Kang, "Carotene: A Job Title Classification System for the Online Recruitment Domain," IEEE First International Conference on Big Data Computing Service and Applications, 2015.

[13] F. Javed, M. McNair, F. Jacob and M. Zhao, "Towards a Job Title Classification System," ArXiv, 2016.

[14] R. Boselli, M. Cesarini, F. Mercorio and M. Mezzanzanica, "Classifying online Job Advertisements through Machine Learning," Future Generation Computer Systems, 2018.

[15] K. Tijdens and C. Kaandorp, "Classifying job titles from job vacancies into ISCO-08 and related job features - the Netherlands," Technical Report, 2019.

[16] S. Bird, E. Klein and E. Loper, "Natural Language Processing with Python," O'Reilly Media, 2009, p. 240

ABSTRACT

In addition to regular statistical surveys National Statistical Institutions tend to use as much as possible external data source such as administrative data sources. In addition to these data sources, Big Data has also been recognized as a new data source. One of the basic tasks in Statistics, no matter what source of data it is, is the automatic text classification. In this paper, machine learning algorithms were used over data collected from four websites, all in order to verify the success of machine learning algorithms used for text classification unstructured data.

Application of machine learning in the process of classification of advertised jobs

Branislava Cvijetic and Zaharije Radivojevic