Seminarski rad u okviru kursa

Istraživanje podataka 1

Matematički fakultet

Analiza skupa podataka Twitter User Gender Classification

Metod klasifikacija

Kristina Pantelić 91/2016

mi16091@alas.matf.bg.ac.rs

19. jun 2019. godine

1 Uvod u podatke

Skup koji je korišćen u ovom seminarskom radu je skup podataka o tvitovima (*engl. tweet*) društvene mreže Twitter, preuzet sa web-sajta Kaggle datasets (https://www.kaggle.com/crowdflower/twitter-user-gender-classification). Ovaj skup podataka bio je korišćen u okviru projekta čiji je cilj bio treniranje prediktora pola *CrowdFlower AI*. Saradnici na tom projektu su prikupili podatke tako što su uprostili pogled na Twitter profile, izdvajajući određene informacije o profilima, i na osnovu tih podataka procenili su da li je korisnik profila bio muškog, ženskog pola, ili je u pitanju bio brend (nije bila individua); ili se na osnovu prikupljenih podataka nije mogla doneti procena. Dakle, ciljni atribut njihovog istraživanja je bila procenjena vrednost za pol korisnika profila.

2 Analiza i pretprocesiranje

Podaci se nalaze u tabeli koja sadrži 20050 instanci, gde svaka instanca predstavlja zapis o jednom Twitter profilu. Svaki profil je opisan pomoću 26 atributa. Sledi lista atributa, kao i njihovi opisi:

unit_id: jedinstven identifikator korisnika	**golden:** indikator koji ukazuje na to da li je korisnik bio uključen u zlatni standard modela; TRUE ili FALSE
unit_state: stanje observacije; finalized (za one koji su procenjeni od strane saradnika na projektu) ili golden (za zlatne standardne observacije)	**trusted_judgments:** broj procena od povenja(int); uvek je 3 za one koji nisu golden, i može biti jedinstven identifikator za standardne observacije
last_judgment_at: datum i vreme poslednje procene profila; prazno polje za zlatne standardne observacije	**gender:** male, female ili brand (za profile koji ne predstavljaju lične profile)
gender:confidence: broj u pokretnom zarezu koji predstavlja pozdanost procenjenog pola	**profile_yn:** "no" znači da je profil trebalo da bude deo skupa za istraživanje, ali nije bio dostupan u trenutku procene
profile_yn:confidence: pouzdanost postojanja/nepostojanja profila	**created:** datum i vreme kada je profil napravljen
description: opis korisničkog profila	**fav_number:** broj tvitova koje je korisnik označio da su mu omiljeni
gender_gold: ako je profil bio golden, koji je u tom slučaju bio pol korisnika profila	**link_color:** boja linka ka profilu, heksadekadna vrednost
name: ime korisnika	**profile_yn_gold:** indikator koji govori o tome da li je korisnički profil bio golden [y/n]
profileimage: link ka slici profila	**retweet_count:** broj puta koliko je korisnik bio retvitovan, tj. koliko se ljudi pozivalo na njegove tvitove
sidebar_color: boja korisničkog profila, heksadekadna vrednost	**text:** tekst jednog od korisničkih tvitova
tweet_coord: ako je korisnikova navigacija bila uključena, koordinate u string formatu: "[latituda, longituda]"	**tweet_count:** broj ukupnih tvitova koje je korisnik postavio
tweet_created: vreme kada je odabrani tvit napravljen	**tweet_id:** identifikator odabranog tvita
tweet_location: lokacija tvita	**user_timezone:** vremenska zona korisnika

Tabela 1. Opis atributa skupa podataka

Na osnovu prethodno opisanih atributa i imajući u vidu da želimo da izvršimo klasifikaciju pola korisnika na osnovu teksta tvita, većina atributa nije korišćena u formiranju modela.

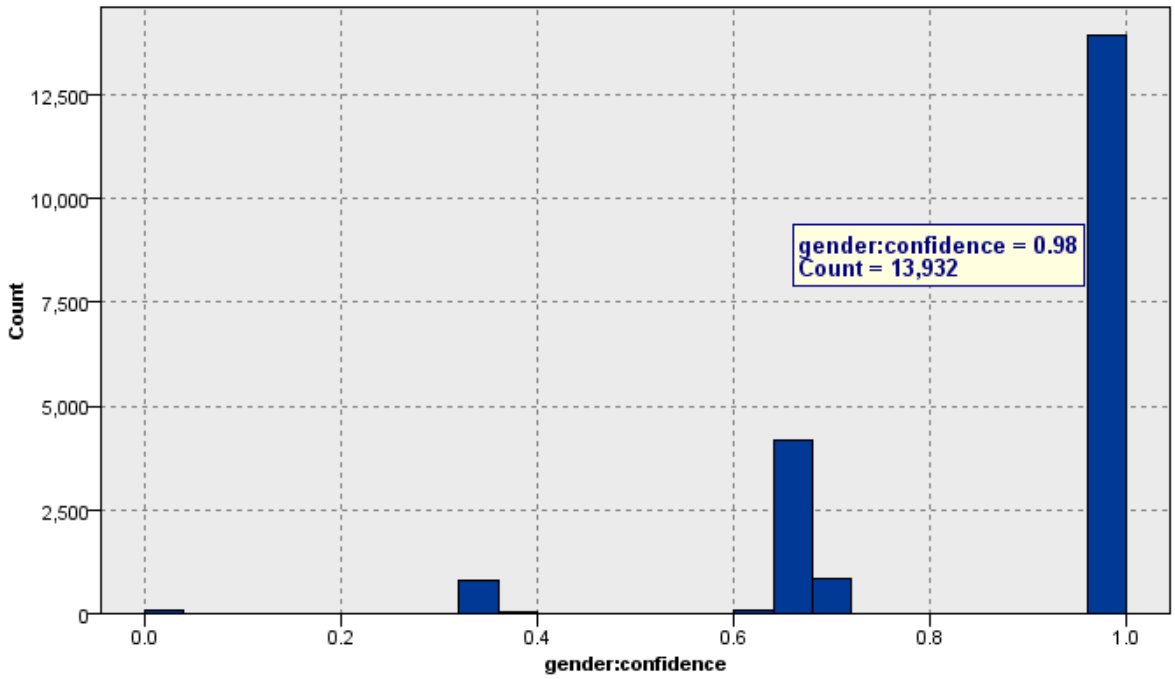
Analiza i pretprocesiranje skupa su rađeni u programskom jeziku Python uz korišćenje odgovarajućih biblioteka za mašinsko učenje, kao i u SPSS modeleru. Fokus istraživanja podataka ovog skupa je određivanje pola korisnika Twitter profila na osnovu teksta tvita. Prvo su analizirani podaci, kako bismo se bolje upoznali sa našim skupom. Analizu i pretprocesiranje skupa podataka počinjemo učitavanjem podataka pomoću Pandas biblioteke u programskom jeziku Python, odnosno upotrebom čvora Var u SPSS-u.

Na početku rada su analizirani tipovi atributa i njihove statistike.

Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev
_unit_id		Continuous	6	915756269	465359445.681	287361795.648
_golden		Flag	--	--	--	--
_unit_state		Flag	--	--	--	--
_trusted_judgments		Continuous	3	274	3.616	12.332
_last_judgment_at		Nominal	--	--	--	--
gender		Nominal	--	--	--	--
gender:confidence		Continuous	0.000	1.000	0.883	0.191

Slika 2.1. Tipovi atributa i njihove statistike

Da bi se eliminisali nedostajući podaci, prvo je provereno da li ih ima u skupu i uklonjeni su svi slogovi iz skupa koji sadrže nedostajuće vrednosti, nakon čega je ostalo 18836 slogova. Kako je fokus na atributima teksta tvita i atributu *gender* koji je ciljni atribut, eliminacija slogova se odnosila na nedostajuće vrednosti koje su se nalazile u ova dva atributa.



Slika 2.2. Histogram vrednosti za poverenje procenjenog pola korisnika profila

S obzirom da je ciljni atribut dobijen procenom profila, odnosno da to nisu sa sigurnošću prave vrednosti ciljnog atributa, cilj je bio uzeti u obzir slogove koje imaju visoku pouzdanost procene radi što bolje klasifikacije slogova skupa. Dobijena je raspodela poverenja procene pola (Slika 2.2.), na osnovu koje je doneta odluka da se eliminišu svi slogovi koji imaju pouzdanost procene pola manju od 0.8.

Zanimljiv atribut za fazu pretprocesiranja je atribut *profile_yn* : *"no"* znači da je profil trebalo da bude deo skupa za istraživanje, ali nije bio dostupan u trenutku procene i izbacujemo sve takve slogove iz skupa podataka. Za oba prethodno navedena uslova (*gender:confidence* > 0.8 i *profile_yn* \='no') za izbacivanje slogova iz skupa su u Python-u eksplicitno navedeni uslovi, dok je u SPSS modeleru korišćen čvor Select.

Statistikom broja slogova po polu korisnika profila, uočeno je da se pored vrednosti ciljnog atributa *male*, *female* i *brand* nalazi i vrednost *nepoznato (engl. unknown)*, i svi takvi slogovi eliminisani su iz skupa (Slika 2.3).

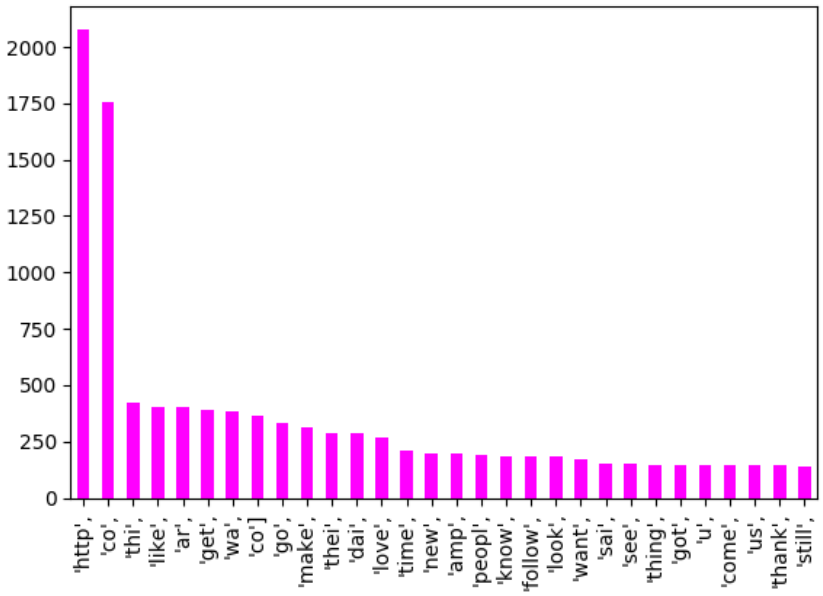
Value	Proportion	%	Count
		0.48	97
brand		29.64	5942
female		33.42	6700
male		30.89	6194
unknown		5.57	1117

Slika 2.3. Statistika atributa *gender*

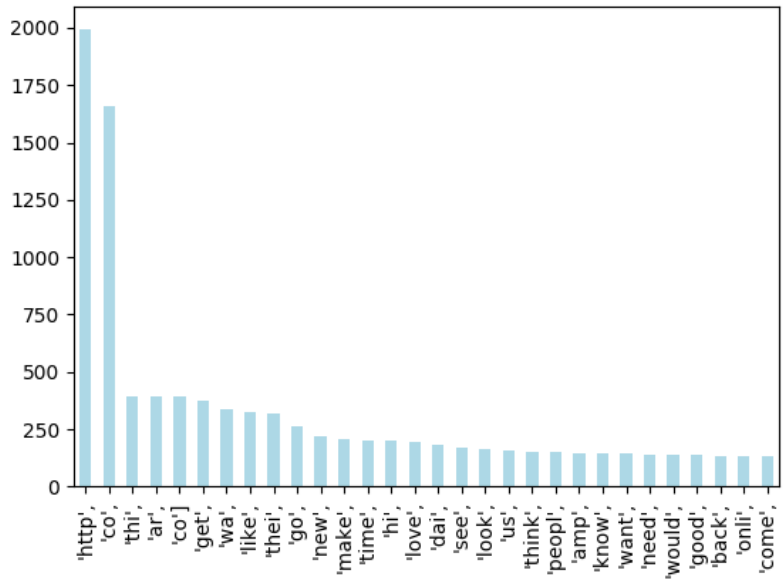
Kako se radi o istraživanju tekstualnih podataka napisanih na engleskom jeziku, pretprocesiranje uključuje eliminaciju engleskih STOP reči koristeći Python-ovu biblioteku NLTK (Natural Language Toolkit) i pronalaženje korena reči za svaku reč u tvitovima, kako bi se reči koje su slične i imaju slično značenje svele na jednu reč. Time je postignuto da se reči poput *love*, *loved, loving*, posmatraju kao jedna reč *lov* i time poboljša proces klasifikacije tvitova. Svođenje reči na koren reči rađeno je pomoću Porterovog stemera (iz uputstva za izradu seminarskog rada sa vežbi).

Pretprocesiranje teksta tvita je rađeno na dva načina. U tekstovima tvita, osim čistih reči postoje pojavljivanja raznih znakova, brojeva, linkova i izražavanja emocija korišćenjem simbola koji su zbog svoje prirode kodiranja u tekstovima tvita preuzetih sa Twitter profila predstavljeni heksadekadnim vrednostima. Prvi način pretprocesiranja podrazumevao je izdvajanje čistih reči iz tvitova (dodatak slika 5.1.1), dok je u drugom pristupu bilo dozvoljeno pojavljivanje svih mogućih reči, uključujući i simbole izražavanja emocija (dodatak slika 5.1.2). Ovakva dva načina pretprocesiranja podataka su korišćena za klasifikaciju i upoređeno je koji od pristupa daje bolje rezultate klasifikacije.

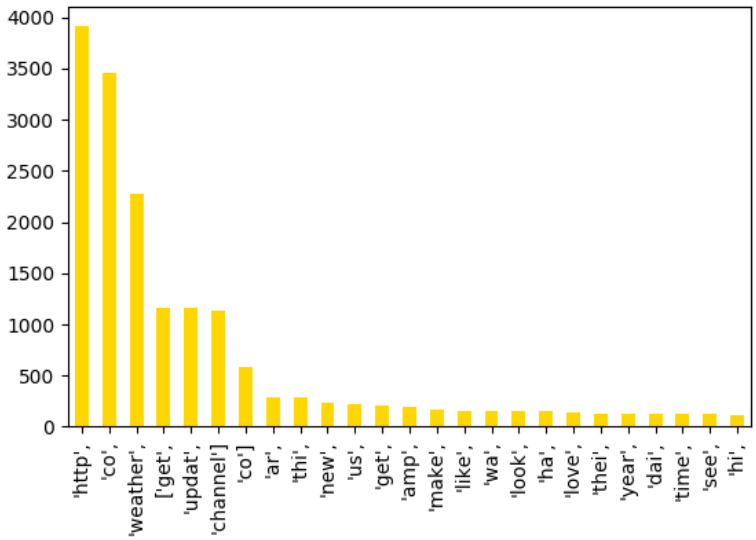
Za izdvajanje termi iz teksa korišćena je klasa CountVectorizer programskog jezika Python koja pretvara kolekciju tekst dokumenata u term-matricu sa brojem pojavljivanja termi u dokumentu. Radi performansi algoritama, odabrano je da se za klasifikaciju koriste 500 najfrekventnijih reči iz celog skupa. Pored pomenute klase, korišćena je klasa TfidfVectorizer() koja pretvara kolekciju tekst dokumenata u matricu sa tf-idf atributima. Ukoliko je pojavljivanje određene reči veoma frekventno u svim dokumentima, onda ona verovatno ne nosi puno informacija. Kako bi se taj problem rešio korišćena je term matrica inverznih frekvencija koja je redukovala pojavljivanja najfrekventnijih reči u svim tvitovima.



Slika 2.4. Najzastupljenije reči u tvitovima koji pripadaju korisnicima profila ženskog pola, prvi pristup pretprocesiranja podataka, pre normalizacije.



Slika 2.5. Najzastupljenije reči u tvitovima koji pripadaju korisnicima profila muškog pola, prvi pristup pretprocesiranja podataka, pre normalizacije.



Slika 2.6. Najzastupljenije reči u tvitovima koji pripadaju korisnicima profila brenda, prvi pristup pretprocesiranja podataka, pre normalizacije.

3 Klasifikacija

Pre procesa klasifikacije, skup je podeljen na test i na trening, gde je podešeno da trening skup iznosi 70% ukupnog skupa instanci, a test skup 30%. Korišćen je prvi pristup pretprocesiranja.

Prvi algoritam za klasifikaciju koji je primenjen na skup je Naivni Bajesov algoritam u Python-u korišćenjem klase MultinomialNB() sa podrazumevano podešenim vrednostima klasifikatora, tj. parametar uglađivanja *alpha=1*, *fit_priori* parametar je *True* jer želimo da klasifikator uči verovatnoće klasa i *class_prior=None*, jer ne želimo da zadajemo početne verovatnoće klasa. Sledi izveštaj klasifikacije trening i test skupa:

```
Trening skup
Matrica konfuzije
        brand  female  male
brand    1823     462   367
female    550    2474   735
male      611    1356  1293


Preciznost za trening skup: 0.5780167511115707


Izvestaj klasifikacije
              precision    recall  f1-score   support

       brand       0.61      0.69      0.65      2652
      female       0.58      0.66      0.61      3759
        male       0.54      0.40      0.46      3260

    accuracy                           0.58      9671
   macro avg       0.58      0.58      0.57      9671
weighted avg       0.57      0.58      0.57      9671
```

Slika 3.1. Izveštaj klasifikacije za trening skup Naivnog Bajesovog klasifikatora

```
Test skup
Matrica konfuzije
        brand  female  male
brand     760     227   149
female    218     991   403
male      286     697   415


Preciznost za test skup: 0.5224312590448625

Izvestaj klasifikacije
              precision    recall  f1-score   support

       brand       0.60      0.67      0.63      1136
      female       0.52      0.61      0.56      1612
        male       0.43      0.30      0.35      1398

    accuracy                           0.52      4146
   macro avg       0.52      0.53      0.52      4146
weighted avg       0.51      0.52      0.51      4146
```

Slika 3.2. Izveštaj klasifikacije za test skup Naivnog Bajesovog klasifikatora

Vidimo da se preciznost trening i test skupa ne razlikuju mnogo, dakle nije došlo do preprilagođavanja modela, ali je preciznost oba skupa relativno mala, što je prihvatljivo s obzirom da za klasifikaciju koristimo samo tekst tvita kao prediktore klasifikacije.

Naredni klasifikator korišćen za klasifikaciju je Drvo odlučivanja (Python klasa DecisionTreeClassifier). Korišćenjem metoda unakrsne validacije određeni su najbolji parametri

modela drveta odlučivanja. Za meru nečistoće su razmatrane entropija i Ginijev indeks, a za maksimalnu dubinu drveta odlučivanja razmatrane su vrednosti 5, 15, 25 i 50. Najbolji model dobijen je korišćenjem Ginijevog indeksa kao mere nečistoće maksimalne dubine drveta 25.

```
Trening skup
Matrica konfuzije
        brand  female  male
brand    2213     426    13
female    686    2900   173
male      755    1819   686


Preciznost 0.5996277530762072


Izvestaj klasifikacije
              precision    recall  f1-score   support

       brand       0.61      0.83      0.70      2652
      female       0.56      0.77      0.65      3759
        male       0.79      0.21      0.33      3260

    accuracy                           0.60      9671
   macro avg       0.65      0.61      0.56      9671
weighted avg       0.65      0.60      0.56      9671
```

Slika 3.3. Izveštaj klasifikacije za trening skup Drveta odlučivanja kao klasifikatora

```
Test skup
Matrica konfuzije
        brand  female  male
brand     839     255    42
female    315    1222    75
male      340     946   112


Preciznost 0.5241196333815726


Izvestaj klasifikacije
              precision    recall  f1-score   support

       brand       0.56      0.74      0.64      1136
      female       0.50      0.76      0.61      1612
        male       0.49      0.08      0.14      1398

    accuracy                           0.52      4146
   macro avg       0.52      0.53      0.46      4146
weighted avg       0.51      0.52      0.46      4146
```

Slika 3.4. Izveštaj klasifikacije za test skup Drveta odlučivanja kao klasifikatora

Uočavamo da klasifikator Drvo odlučivanja daje bolje rezultate klasifikacije od Naivnog Bajesovog klasifikatora, na osnovu poređenja preciznosti klasifikacije nad test podacima.

Naredni algoritam koji primenjujemo je k najbližih suseda (KNN). Korišćenjem metoda unakrsne validacije određeni su najbolji parametri modela k najbližih suseda. Za broj razmatranih suseda algoritmom predložene vrednosti su od 3 do 9 suseda; za parametar rastojanja Minkovskog razmatrani su p=1 (Menhetn rastojanje) i p=2 (Euklidsko rastojanje); za težine suseda su rasmatrene opcije uniformnog rastojanja *'uniform'* (svi susedi imaju podjednak uticaj) i *'distance'* (gde bliži susedi imaju veći uticaj na određivanje klase). Dobijeno je da najbolji model tj. onaj koji daje najveću preciznost uzima 8 suseda, koristi rastojanje Minkovskog (p=1) i gde bliži susedi imaju veći uticaj na određivanje klase.

```
Trening skup
Matrica konfuzije
        brand  female  male
brand    2566      37    49
female    143    3427   189
male      172     163  2925


Preciznost 0.922138351773343


Izvestaj klasifikacije
              precision    recall  f1-score   support

       brand       0.89      0.97      0.93      2652
      female       0.94      0.91      0.93      3759
        male       0.92      0.90      0.91      3260

    accuracy                           0.92      9671
   macro avg       0.92      0.93      0.92      9671
weighted avg       0.92      0.92      0.92      9671
```

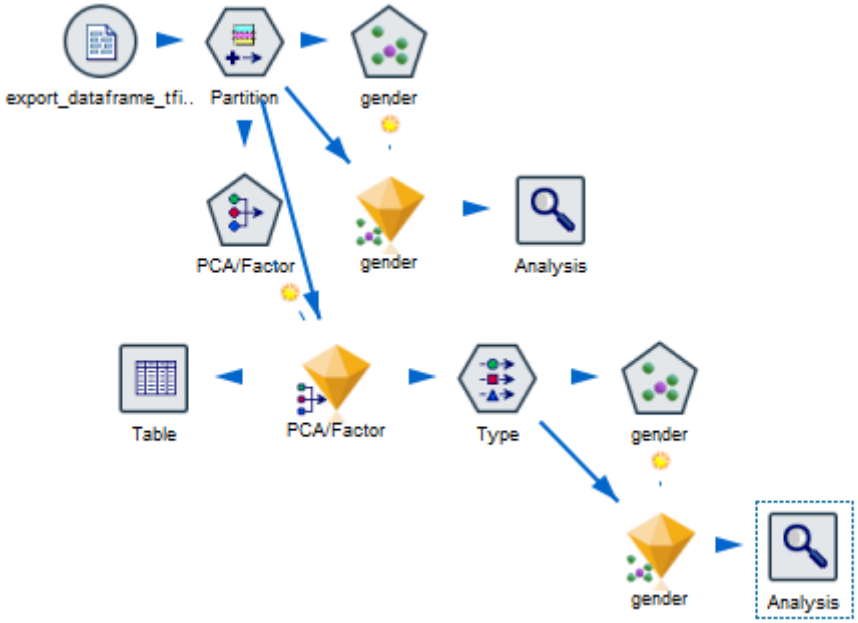
Slika 3.5. Izveštaj klasifikacije za trening skup KNN klasifikatora

```
Test skup
Matrica konfuzije
        brand  female  male
brand     772     207   157
female    285     832   495
male      298     623   477


Preciznost 0.5019295706705258


Izvestaj klasifikacije
              precision    recall  f1-score   support

       brand       0.57      0.68      0.62      1136
      female       0.50      0.52      0.51      1612
        male       0.42      0.34      0.38      1398

    accuracy                           0.50      4146
   macro avg       0.50      0.51      0.50      4146
weighted avg       0.49      0.50      0.49      4146
```

Slika 3.6. Izveštaj klasifikacije za test skup KNN klasifikatora

S obzirom na visoku preciznost trening skupa, a nisku preciznost test skupa, zaključujemo da je došlo do preprilagođavanja modela. Ovom može biti uzrok nenormalizovanost podataka i korišćenje term matrice sa brojem pojavljivanja termi u dokumentima. Naredni korak bio je normalizovati podatke, međutim zbog ograničenja memorije i vremena, a bez odgovarajuće optimizacije, rad KNN algoritma u Python-u nije bio moguć, zato je dalje modelovanje nastavljeno u SPSS-modeleru. Formirana je matrica sa tf-idf atributima sa 500 najfrekventnijih reči iz skupa svih reči i izvezena u CSV format koji je dalje učitan u SPSS modeler preko čvora Var.

U čvoru Var su učitane vrednosti atributa, atributi su neprekidnog tipa, dok je ciljni atribut numeričkog tipa. Na čvor Var primenjen je čvor Partition gde je specifikovano da se skup podeli tako što će se iz celog skupa odabrati 70% instanci za trening skup i 30% instanci za test skup.



Slika 3.7. SPSS tok sa čvorovima za klasifikaciju pomoću KNN algoritma

Nad podacima je primenjena redukcija atributa pomoću rotacije osa (PCA), kako bi se od inicijalnih 500 atributa dobio manji broj nezavisnih atributa na osnovu kojih se mogu klasifikovati podaci. Algoritmu PCA je dodeljen maksimalan broj atributa 100. Napravljena su 2 modela KNN klasifikatora, prvi korišćenjem svih 500 atributa, a drugi korišćenjem rezultata PCA algoritma, dobijenih 100 atributa. Na PCA čvor primenjen je čvor Type u kome je specifikovano da će se za klasifikaciju koristiti samo redukovani atributi. Na dobijeni model primenjen je čvor Analysis pomoću kojeg su dobijene statistike rada klasifikatora nad trening i test podacima, kao i matrica konfuzije.

Results for output field gender

Comparing $KNN-gender with gender

'Partition'	1_Training		2_Testing	
Correct	8,219	62.19%	2,638	46.93%
Wrong	4,996	37.81%	2,983	53.07%
Total	13,215		5,621	

Coincidence Matrix for $KNN-gender (rows show actuals)

'Partition' = 1_Training	0	1	2
0	2,297	1,029	852
1	423	3,593	719
2	434	1,539	2,329

'Partition' = 2_Testing	0	1	2
0	837	551	376
1	247	1,177	541
2	267	1,001	624

Slika 3.8. Rezultat rada SPSS čvora Analysis koji prikazuje statistike modela KNN algoritma nad svim podacima (500 atributa)

Results for output field gender

Comparing $KNN-gender with gender

'Partition'	1_Training		2_Testing	
Correct	8,106	61.34%	3,473	61.79%
Wrong	5,109	38.66%	2,148	38.21%
Total	13,215		5,621	

Coincidence Matrix for $KNN-gender (rows show actuals)

'Partition' = 1_Training	0	1	2
0	2,371	1,057	750
1	430	3,348	957
2	461	1,454	2,387

'Partition' = 2_Testing	0	1	2
0	1,034	423	307
1	168	1,356	441
2	188	621	1,083

Slika 3.9. Rezultat rada SPSS čvora Analysis koji prikazuje statistike modela KNN algoritma nad podacima koji su rezultat rada PCA algoritma

Uočavamo da je KNN klasifikator nad svim atributima pokazao bolje performanse nad trening podacima u odnosu na klasifikator sa PCA atributima, međutim drugi model je bolji jer je visoka i približna preciznost i trening i test skupa. Bolji model je za klasifikaciju koristio 5 suseda, Euklidsko rastojanje kao meru rastojanja i svi susedi su imali podjednaki uticaj u određivanju klase.

Naredni algoritam za klasifikaciju su Neuronske mreže. Klasifikacija je rađena nad inverznom term matricom dokumenata dobijenom prvim pristupom pretprocesiranja.

Results for output field gender

Comparing $N-gender with gender

'Partition'	1_Training		2_Testing	
Correct	6,144	46.49%	2,517	44.78%
Wrong	7,071	53.51%	3,104	55.22%
Total	13,215		5,621	

Coincidence Matrix for $N-gender (rows show actuals)

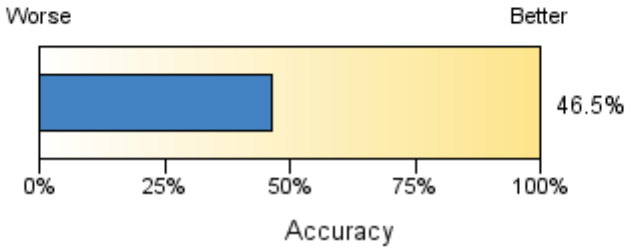
'Partition' = 1_Training	0	1	2
0	2,420	967	791
1	1,039	2,312	1,384
2	1,077	1,813	1,412

'Partition' = 2_Testing	0	1	2
0	998	426	340
1	444	947	574
2	466	854	572

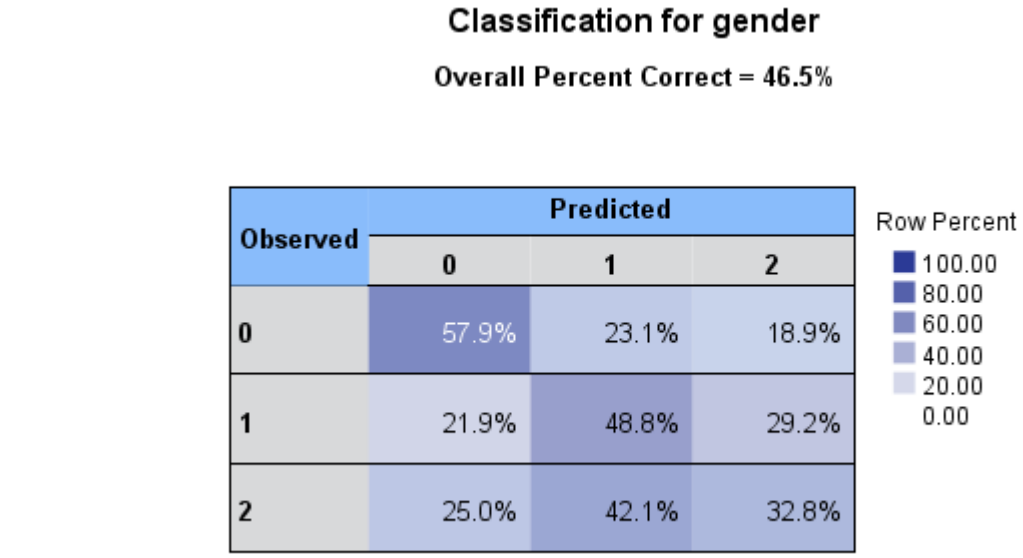
Slika 3.10. Rezultat rada SPSS čvora Analysis koji prikazuje statistike modela Neuronskih mreža

Model Summary

Target	gender
Model	Multilayer Perceptron
Stopping Rule Used	Error cannot be further decreased
Hidden Layer 1 Neurons	9

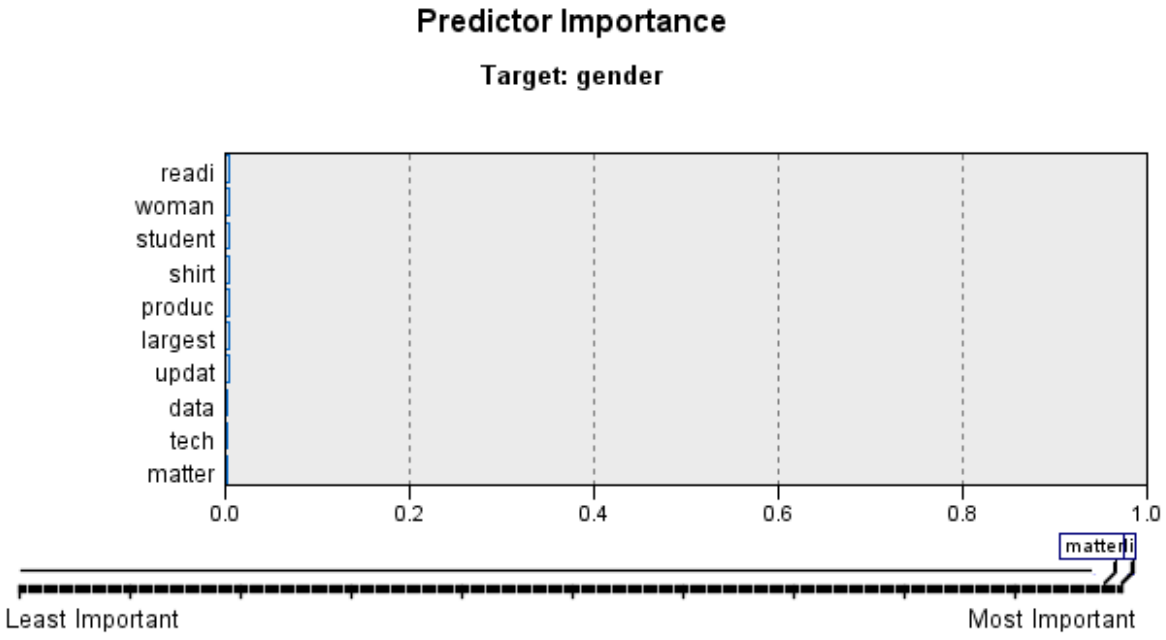


Slika 3.11. Rezultat rada SPSS čvora Analysis koji prikazuje statistike modela Neuronskih mreža, rezime modela



Slika 3.12. Rezultat rada SPSS čvora Analysis koji prikazuje statistike modela Neuronskih mreža, matrica konfuzije

---



Slika 3.13. Rezultat rada SPSS čvora Analysis koji prikazuje statistike modela Neuronskih mreža, značajnost prediktora

Klasifikator Neuronske mreže nije pokazao dobre performanse nad klasifikovanim podacima, s obzirom da je preciznost klasifikacije trening i test skupa manja od 50%. Klasifikator

je za klasifikaciju koristio 9 skrivenih slojeva. Na slici 3.13. su prikazani termi koji su imali najveći procenat značajnosti u procesu klasifikacije.

Naredni klasifikator je metod potpornih vektora (SVM). Korišćen je istoimeni čvor u SPSS-u sa podrazumevanim opcijama klasifikatora. Klasifikator je primenjen nad svim atributima, a zatim nad redukovanim atributima dobijenim kao rezultat rada PCA algoritma. Klasifikacija je rađena nad inverznom term matricom dokumenata dobijenom prvim pristupom pretprocesiranja.

Results for output field gender

Comparing $S-gender with gender

'Partition'	1_Training		2_Testing	
Correct	4,777	36.15%	1,979	35.21%
Wrong	8,438	63.85%	3,642	64.79%
Total	13,215		5,621	

Coincidence Matrix for $S-gender (rows show actuals)

'Partition' = 1_Training	0	1
0	134	4,044
1	92	4,643
2	77	4,225

'Partition' = 2_Testing	0	1
0	50	1,714
1	36	1,929
2	38	1,854

Slika 3.14. Rezultat rada SPSS čvora Analysis koji prikazuje statistike modela SVM klasifikatora, klasifikacija nad svim podacima (500 atributa)

Results for output field gender

Comparing $S-gender with gender

'Partition'	1_Training		2_Testing	
Correct	9,207	69.67%	2,943	52.36%
Wrong	4,008	30.33%	2,678	47.64%
Total	13,215		5,621	

Coincidence Matrix for $S-gender (rows show actuals)

'Partition' = 1_Training	0	1	2
0	3,179	495	504
1	607	3,280	848
2	619	935	2,748

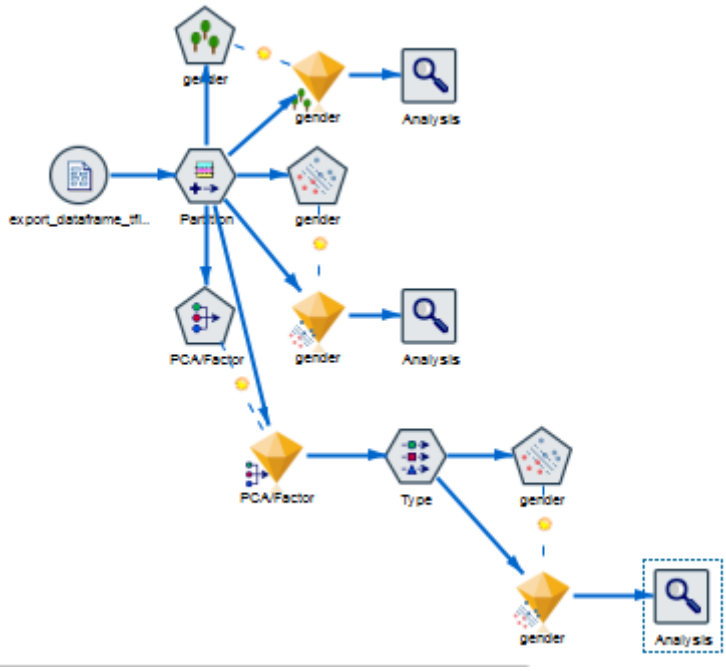
'Partition' = 2_Testing	0	1	2
0	1,173	271	320
1	333	978	654
2	383	717	792

Slika 3.15. Rezultat rada SPSS čvora Analysis koji prikazuje statistike modela SVM klasifikatora, klasifikacija nad redukovanih 100 atributa (PCA)

Uočava se velika razlika u preciznosti SVM klasifikacije nad svim atributima i nad redukovanim atributima (PCA). Klasifikacija nad svim atributima je manja od 40% za oba skupa, i trening i test, dok je klasifikacija nad redukovanim atributima dala preciznost klasifikacije od

skoro 70% za trening i oko 52% nad test skupom. Zaključujemo da je SVM klasifikator nad podacima bolji kada se prethodno izvrši redukcija broja atributa.

Naredni klasifikator je RandomForestTree. Korišćen je istoimeni čvor u SPSS-u sa podrazumevanim opcijama klasifikatora i korišćenim podacima iz prvog pristupa pretprocesiranja.



Slika 3.16. SPSS tok sa čvorovima za klasifikaciju pomoću RandomForestTree klasifikatora nad svim podacima; SVM klasifikatora nad svim podacima i SVM klasifikatora nad redukovanim podacima

Results for output field gender

Comparing $R-gender with gender

'Partition'	1_Training		2_Testing	
Correct	5,354	40.51%	2,163	38.48%
Wrong	7,861	59.49%	3,458	61.52%
Total	13,215		5,621	

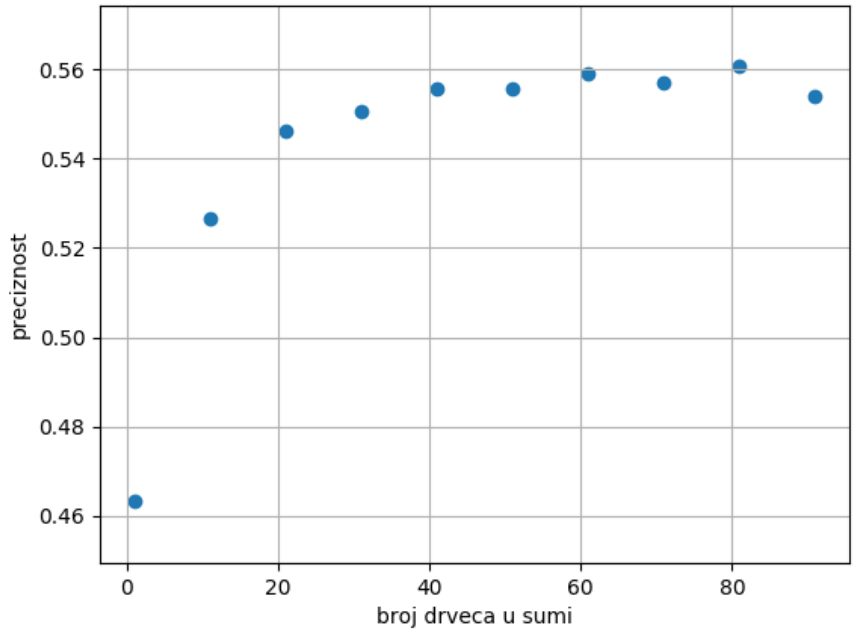
Coincidence Matrix for $R-gender (rows show actuals)

'Partition' = 1_Training	0	1	2
0	3,784	98	296
1	3,259	475	1,001
2	2,991	216	1,095
'Partition' = 2_Testing	**0**	**1**	**2**
0	1,560	53	151
1	1,326	177	462
2	1,356	110	426

Slika 3.17. Rezultat rada SPSS čvora Analysis koji prikazuje statistike modela RandomForestTree klasifikatora, klasifikacija nad svim atributima

Na osnovu dobijenih statistika klasifikacije, zaključuje se da klasifikator nije pokazao dobre performanse u slučaju naših podataka, jer je preciznost i trening i test skupa oko 40%.

Isti klasifikator, RandomForestTree, primenjen je nad podacima koji su pretprocesirani tako što je dozvoljeno da se u termima nalaze brojevi i simboli koji predstavljaju izražavanje emocija korisnika. Klasifikatoru su prosleđene moguće vrednosti za broj drveća u šumi od 1 do 100 sa korakom 10 i dobijeno je da klasifikator koji daje najveću preciznost koristi 90 drveća u šumi za klasifikaciju (*n_estimators* = 80). Izveštaj klasifikacije klasifikatora sa 80 drveta u šumi i Ginijevim indeksom kao kriterijumom za meru nečistoće.



Slika 3.18. Preciznost test skupa u zavisnosti od broja drveća u šumi RandomForestTree klasifikatora, klasifikacija nad svim podacima pretprocesiranim drugim pristupom pretprocesiranja.

```
Trening skup

Matrica konfuzije:
 [[2539   48   65]
 [  82 3556  121]
 [ 110  100 3050]]

Izvestaj klasifikacije:
              precision    recall  f1-score   support

           0       0.93      0.96      0.94      2652
           1       0.96      0.95      0.95      3759
           2       0.94      0.94      0.94      3260

    accuracy                           0.95      9671
   macro avg       0.94      0.95      0.95      9671
weighted avg       0.95      0.95      0.95      9671

Preciznost trening skupa: 0.9456105883569434
```

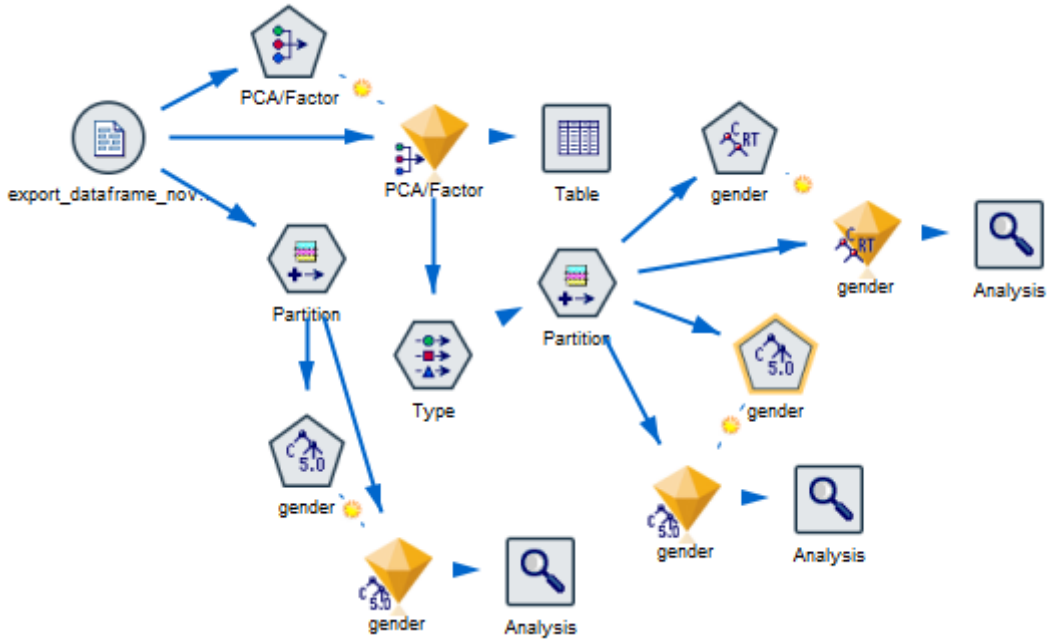
Slika 3.19. Izveštaj klasifikacije za trening skup RandomForestTree klasifikatora

```
Test skup

Matrica konfuzije:
 [[776 185 175]
 [223 956 433]
 [245 595 558]]

Preciznost test skupa: 0.5523396044380126

Izvestaj klasifikacije:
              precision    recall  f1-score   support

           0       0.62      0.68      0.65      1136
           1       0.55      0.59      0.57      1612
           2       0.48      0.40      0.44      1398

    accuracy                           0.55      4146
   macro avg       0.55      0.56      0.55      4146
weighted avg       0.55      0.55      0.55      4146
```

Slika 3.20. Izveštaj klasifikacije za test skup RandomForestTree klasifikatora

Zaključak rada ovog algoritma je da je došlo do preprilagođavanja modela nad trening podacima. Međutim, imajući u obzir rezultate klasifikacije prethodnih klasifikatora nad istim podacima, preciznost test skupa ne odstupa od preciznosti klasifikacije test skupa drugih klasifikatora, ali bez obzira na to, zbog preprilagođenosti, model nije adekvatan.

Naredni algoritmi klasifikacije korišćeni nad podacima o tvitovima su CART i C5.0. Klasifikator C5.0 je primenjen nad svim atributima, kao i nad redukovanim atributima koji su rezultat rada PCA algoritma. Korišćeni podaci su rezultat drugog pristupa pretprocesiranja, dakle dozvoljeni su brojevi, znakovi i simboli izražavanja emocija.



Slika 3.21. SPSS tok sa čvorovima za klasifikaciju pomoću C5.0 i CART klasifikatora.

Results for output field gender

Comparing $R-gender with gender

'Partition'	1_Training		2_Testing	
Correct	5,146	53.56%	2,180	51.95%
Wrong	4,462	46.44%	2,016	48.05%
Total	9,608		4,196	

Coincidence Matrix for $R-gender (rows show actuals)

'Partition' = 1_Training	0	1	2
0	1,819	134	674
1	599	1,470	1,675
2	648	732	1,857

'Partition' = 2_Testing	0	1	2
0	789	75	293
1	288	602	733
2	276	351	789

Slika 3.22. Rezultat rada SPSS čvora Analysis koji prikazuje statistike modela CART klasifikatora, klasifikacija nad redukovanih 100 atributa (PCA)

Results for output field gender

Comparing $C-gender with gender

'Partition'	1_Training		2_Testing	
Correct	5,310	55.27%	2,202	52.48%
Wrong	4,298	44.73%	1,994	47.52%
Total	9,608		4,196	

Coincidence Matrix for $C-gender (rows show actuals)

'Partition' = 1_Training	0	1	2
0	2,019	161	447
1	658	1,334	1,752
2	772	508	1,957

'Partition' = 2_Testing	0	1	2
0	871	88	198
1	326	511	786
2	341	255	820

Slika 3.23. Rezultat rada SPSS čvora Analysis koji prikazuje statistike modela C5.0 klasifikatora, klasifikacija nad redukovanih 100 atributa (PCA)

Results for output field gender

Comparing $C-gender with gender

'Partition'	1_Training		2_Testing	
Correct	5,753	59.88%	2,135	50.88%
Wrong	3,855	40.12%	2,061	49.12%
Total	9,608		4,196	

Coincidence Matrix for $C-gender (rows show actuals)

'Partition' = 1_Training	0	1	2
0	2,010	256	361
1	511	1,872	1,361
2	616	750	1,871

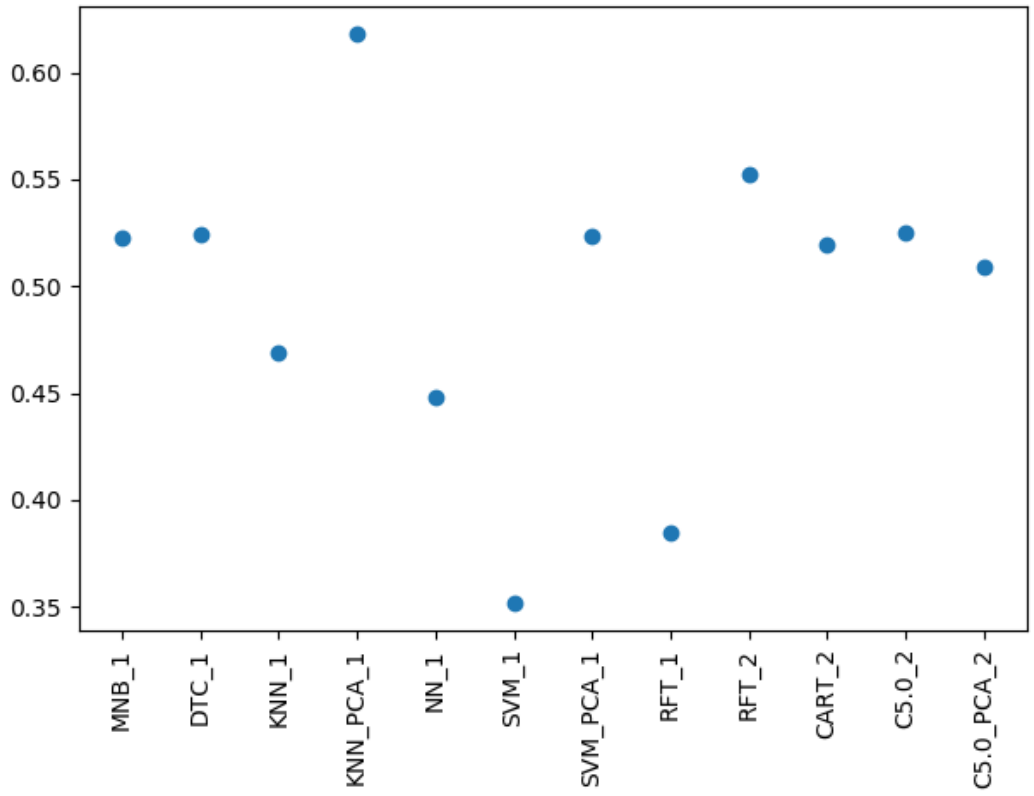
'Partition' = 2_Testing	0	1	2
0	783	179	195
1	279	672	672
2	325	411	680

Slika 3.24. Rezultat rada SPSS čvora Analysis koji prikazuje statistike modela C5.0 klasifikatora, klasifikacija nad svim atributima

Klasifikator C5.0 je pokazao bolje rezultate klasifikacije od klasifikatora CART, gde su klasifikovani podaci uključivali redukovane atribute (PCA), 100 atributa. Dodatno, klasifikatoru C5.0 kao boljem klasifikatoru od ova dva klasifikatora prosleđeni su i svi podaci. Na osnovu rezultata klasifikacije zaključuje se da je klasifikator C5.0 dobar u klasifikaciji svih atributa, kao i redukovanih atributa, ali veći značaj se pridaje klasifikaciji redukovanih atributa zbog veće preciznosti test skupa u odnosu na preciznost test skupa klasifikacije nad svim atributima.

4 Zaključak

Cilj istraživanja podataka skupa podataka o profilima Twitter profila je bio određivanje pola korisnika na osnovu teksta tvita. Tekst tvita je pretprocesiran na dva načina. Prvi način bio je izdvajanje čistih reči iz teksta, a drugi način je bio izdvajanje svih reči uključujući reči sa brojevima i simbolima koji predstavljaju izražena osećanja korisnika. Primenjeni su različiti algoritmi nad prethodna dva skupa podataka koji su sadržali 500 najfrekventnijih reči u obliku inverzne term matrice dokumenata i određene su njihove preciznosti za trening i za test skup. Kao najbolji klasifikator izdvaja se K najbližih suseda (KNN) sa preciznošću nad test podacima od oko 62%. Korišćen je prvi pristup pretprocesiranja, dakle čiste reči i redukcija broja atributa pomoću PCA metode.

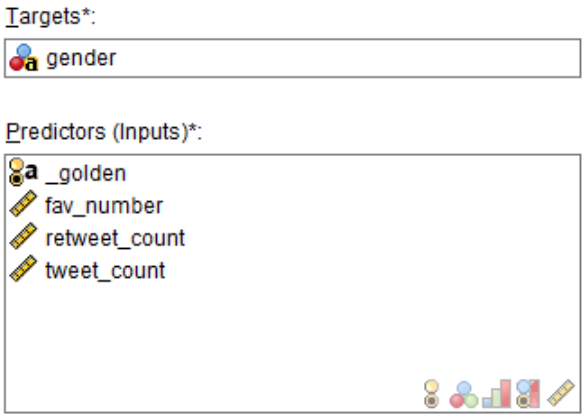


Slika 4.1. Preciznost test skupa u zavisnosti od različitih klasifikatora, broja atributa i načina pretprocesiranja skupa

Na dijagramu (Slika 4.1.) se mogu videti preciznosti klasifikacije nad test podacima različitih klasifikatora, oznaka PCA u imenu klasifikatora označava da je za dati algoritam rađena redukcija atributa pomoću PCA metode, dok je poslednji broj u imenu klasifikatora oznaka vrste pretprocesiranja nad podacima.

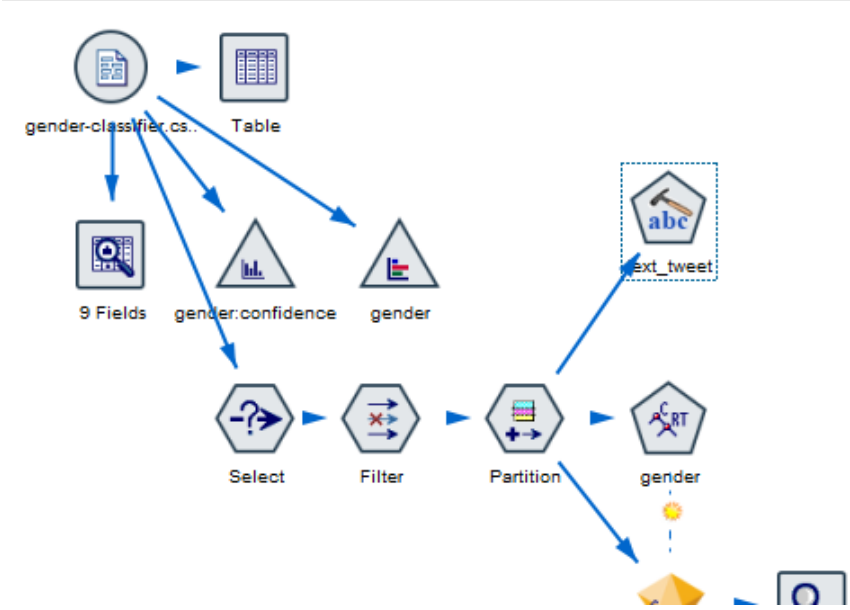
5 Dodatak

Kako je skup podataka pored teksta tvita sadržao i druge atribute kao informacije o profilu korisnika, klasifikovani su podaci koristeći atribute sa slike 5.1.



Slika 5.1. Atributi korišćeni za proces klasifikacije

Odabrani su slogovi koji zadovoljavaju uslove poverenja procene pola koji nije manji od 0.8 i uslov da se među slogovima nalaze samo oni korisnici čiji je profil bio dostupan u trenutku procene pola. Za navedene uslove korišćen je čvor Select. Nadovezan je čvor Filter kojim su odabrani atributi koji će učestvovati u procesu klasifikacije i skup je particionisan na trening i test skup pomoću čvora Partition. Nad podacima je primenjen algoritam CART.



Slika 5.2. SPSS tok sa čvorovima za analizu i klasifikaciju skupa pomoću alterna tivnih atributa.

Comparing $R-gender with gender

'Partition'	1_Training		2_Testing	
Correct	5,030	51.49%	2,159	50.75%
Wrong	4,739	48.51%	2,095	49.25%
Total	9,769		4,254	

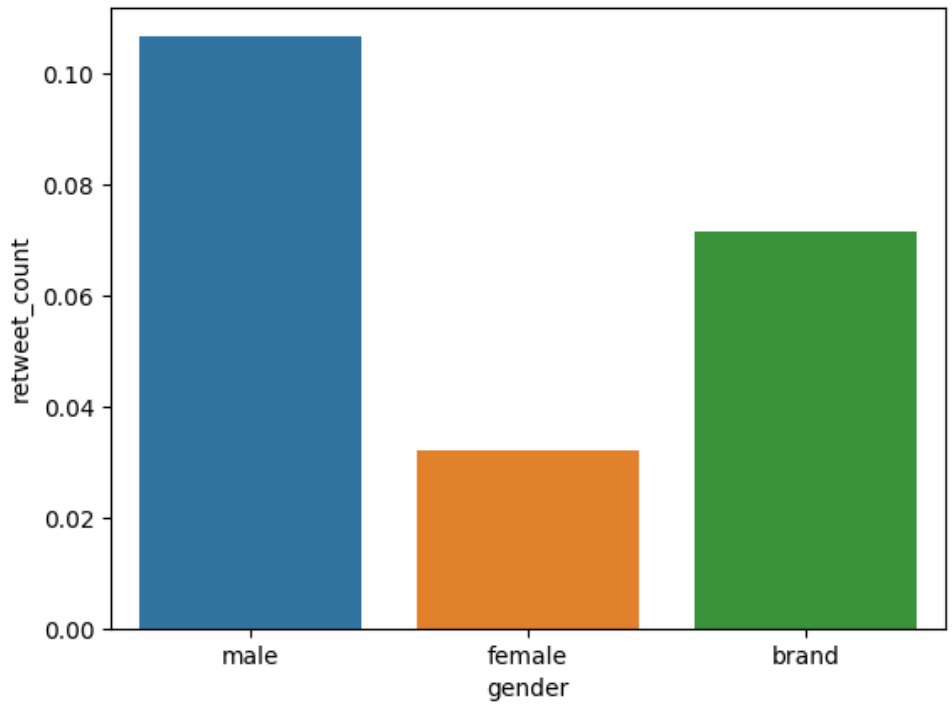
Coincidence Matrix for $R-gender (rows show actuals)

'Partition' = 1_Training	brand	female	male
	12	40	21
brand	1,452	603	576
female	349	2,746	632
male	360	2,059	832
unknown	27	43	17

'Partition' = 2_Testing	brand	female	male
	4	14	6
brand	618	268	267
female	141	1,185	314
male	167	879	356
unknown	20	8	7

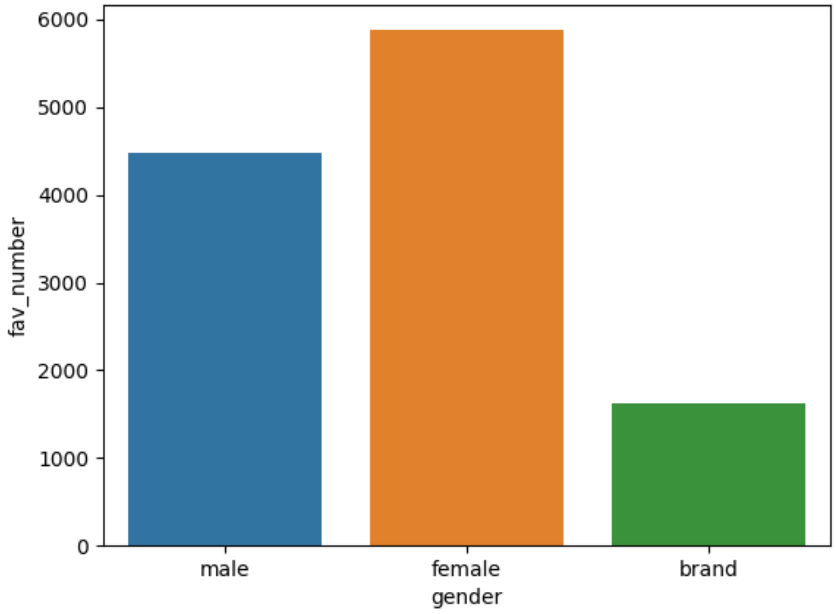
Slika 5.3. Rezultat rada SPSS čvora Analysis koji prikazuje statistike modela CART klasifikatora, klasifikacija nad svim atributima početkog skupa podataka.

Određene su zavisnosti broja retvitova (koliko ljudi je podelilo tvit korisnika) od procenjenog pola korisnika, kao i zavisnosti tvitova koje je korisnik označio da su mu omiljeni u zavisnosti od procenjenog pola korisnika.



Slika 5.4. Zavisnost broja retvitova od pola korisnika profila

Uočava se da su tvitovi postavljani od strane korisnika muškog pola najviše bili deljeni među korisnicima Twitter profila, a najmanje su deljeni tvitovi korisnica ženskog profila.



Slika 5.5. Zavisnost broja favorizovanih tvitova od pola korisnika profila

Uočavamo da korisnice ženskog pola više označavaju određene tvitove kao omiljene u odnosu na korisnike muškog pola i brenda.

5.1. Dodatak

```
account, act, actual, ad, ag, agre, ain, album, alreadi, alwai, ama, amaz, american, amp, ani, announc,
anoth, answer, anyon, anyth, app, appl, appli, ar, art, artist, artistoftheyear, ask, ass, avail, awai,
awesom, babi, bacon, bad, bc, beat, beauti, becaus, becom, bed, befor, believ, best, better, big, biggest,
birthdai, bit, bitch, black, bless, blog, blue, boi, bond, book, break, bring, brother, bui, build, busi,
came, cancer, car, care, cat, catch, caus, celebr, chanc, chang, channel, charg, check, citi, class, click,
close, club, coffe, colleg, come, compani, complet, cool, costum, couldn, countri, coupl, cover, credit, cup,
current, cut, cute, dad, dai, damn, danc, data, date, dead, deal, definit, delai, deserv, design, desk, di,
did, didn, differ, digit, doe, doesn, dog, don, dont, dream, dress, drink, drive, drop, dude, dure, earli,
eat, end, enjoi, enter, episod, event, everi, everydayilovey, everyon, everyth, excit, ey, face, facebook,
fact, fall, famili, fan, favorit, feel, fight, film, final, follow, food, footbal, forc, forevermoreand,
forget, forward, free, fridai, friend, fruit, fuck, fun, funni, futur, game, gener, girl, goal, god, goe,
gone, gonna, good, got, gotta, great, group, grow, gt, gui, ha, hair, half, halloween, hand, happen, happi,
hard, harri, hate, haven, head, health, hear, heard, heart, hei, hell, hello, help, hi, high, histori, hit,
hold, home, hope, host, hot, hour, hous, http, human, idea, im, import, inspir, internet, isn, issu, jame,
job, john, join, just, kid, kill, kind, know, ladi, largest, late, latest, lead, learn, leav, left, let,
life, light, like, line, link, list, listen, liter, littl, live, ll, lmao, lol, long, look, lord, lose, lost,
lot, love, low, mai, make, man, manag, mani, market, matter, mayb, mean, meat, media, meet, men, met, mind,
minut, miss, mom, moment, mondai, monei, month, movi, music, nation, need, new, nice, nigga, night, noth,
number, octob, offer, offic, oh, ok, old, onc, onedirect, onli, onlin, open, order, organ, pai, parent, park,
parti, past, peopl, perfect, person, phone, photo, pic, pick, pictur, pink, place, plai, plan, player, pleas,
pm, point, post, power, ppl, pretti, price, probabl, problem, process, produc, product, project, public,
pull, pumpkin, question, read, readi, real, realiz, realli, reason, red, rememb, report, rest, result,
retweet, review, right, rock, room, round, run, sad, sai, said, sale, saturdai, save, saw, school, season,
second, seen, sell, send, servic, set, share, shirt, shit, shop, sign, sinc, sit, sleep, smile, smoke,
social, someon, someth, sometim, song, soon, sorri, sound, space, speak, special, spectr, st, stai, stand,
star, start, stat, state, step, stop, storag, stori, street, student, stuff, style, success, super, support,
sure, talk, tax, team, tech, tell, test, text, th, thank, thei, thi, thing, think, thought, ticket, time,
tip, todai, togeth, told, tomorrow, tonight, took, tour, train, transform, tri, true, trust, truth, try,
turn, tv, tweet, twitter, uk, understand, unfollow, unit, updat, ur, ve, veri, video, visit, voic, vote, wa,
wai, wait, walk, wanna, want, watch, water, wear, weather, wednesdai, week, weekend, went, white, win, wish,
woman, women, won, wonder, word, work, workbench, world, worst, worth, write, written, wrong, ya, ye, yeah,
year, yesterdai, young, youtub
```

Slika 5.1.1. Najfrekventnije reči dobijene prvim pristupom pretprocesiranja, čiste reči

```
0, 00, 000, 05, 1, 10, 100, 11, 12, 15, 2, 20, 2015, 3, 30, 39, 4, 40, 5, 6, 7, 8, 9, :), _, __ù, _ù, _ùó,
_ùô, _ùõ, account, act, actual, ago, agre, ain, album, alreadi, alway, ama, amaz, american, amp, ani,
announc, anoth, answer, anyon, anyth, app, appl, appli, art, artist, artistoftheyear, ask, ass, avail, away,
awesom, b, babi, bacon, bad, bc, beauti, becaus, becom, bed, befor, believ, best, better, big, birthday, bit,
bitch, black, bless, blog, blue, book, boy, break, bring, brother, build, busi, buy, c, came, cancer, car,
care, catch, caus, celebr, chanc, chang, channel, charg, check, chill, citi, class, click, close, club, come,
complet, cool, costum, couldn, cover, cri, cup, cut, cute, d, dad, damn, danc, date, day, dead, deal,
definit, desk, did, didn, die, differ, digit, doe, doesn, dog, don, dream, dress, drink, drive, drop, dude,
dure, earli, eat, end, enjoy, enter, episod, event, everi, everydayilovey, everyon, everyth, excit, eye,
face, facebook, fact, fall, famili, fan, favorit, feel, fight, film, final, follow, food, footbal,
forevermoreand, forget, forward, free, friday, friend, fuck, fun, funni, futur, game, girl, goal, god, goe,
gone, gonna, good, got, great, group, grow, gt, guy, ha, hair, half, halloween, hand, happen, happi, hard,
hate, haven, head, health, hear, heard, heart, hell, hello, help, hey, hi, high, hire, hit, hold, home, hope,
hot, hour, hous, http, human, idea, im, isn, issu, job, join, just, kid, kill, kind, know, ladi, late,
latest, learn, leav, left, let, lie, life, light, like, line, link, listen, liter, littl, live, ll, lmao,
lol, long, look, lord, lose, lost, lot, love, m, make, man, manag, mani, market, matter, mayb, mean, meat,
media, meet, men, mind, minut, miss, mom, moment, monday, money, month, morn, movi, music, n, nation, need,
new, news, nice, nigga, night, noth, number, o, octob, offer, offic, oh, ok, old, onc, onedirect, onli,
onlin, open, order, p, parent, park, parti, past, pay, peopl, perfect, person, phone, photo, pic, pick,
pictur, place, plan, play, player, pleas, point, post, power, pretti, price, probabl, problem, process,
produc, product, pushawardslizquen, question, quot, read, readi, real, realiz, realli, reason, red, rememb,
report, rest, retweet, right, room, round, run, s, sad, said, sale, saturday, save, saw, say, school, season,
second, seen, sell, send, servic, set, share, shirt, shit, shop, sign, sinc, sit, sleep, smile, social,
someon, someth, sometim, song, soon, sorri, sound, space, special, spectr, stand, star, start, stat, state,
stay, step, stop, storag, stori, street, student, stuff, success, support, sure, t, talk, team, tech, tell,
test, thank, thi, thing, think, thought, ticket, time, today, togeth, told, tomorrow, tonight, took, tour,
train, transform, tri, true, truth, turn, tv, tweet, twitter, u, uk, unfollow, unit, updat, ur, use, ve,
veri, video, visit, voic, vote, w, wa, wait, walk, wanna, want, watch, water, way, wear, weather, week,
weekend, went, whi, white, win, wish, woman, women, won, wonder, word, work, workbench, world, worst, write,
wrong, x, y, ya, ye, yeah, year, yesterday, youtub, ¼, å, å ù, å, õ, û, û , ûª, ûªt, ûó, ü,
```

Slika 5.1.2. Najfrekventnije reči dobijene drugim pristupom pretprocesiranja