TakeLab

Laboratorij za analizu teksta i inženjerstvo znanja
Text Analysis and Knowledge Engineering Lab
Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva
Unska 3, 10000 Zagreb, Hrvatska

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br.

Profiliranje autora na društvenim mrežama pomoću strojnog učenja

Lukrecija Puljić

Zagreb, lipanj 2017.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA
ODBOR ZA ZAVRŠNI RAD MODULA

Zagreb, 3. ožujka 2017.

ZAVRŠNI ZADATAK br. 5327

Pristupnik: **Lukrecija Puljić (2223070466)**
Studij: Računarstvo
Modul: Računarska znanost

Zadatak: **Profiliranje autora na društvenim mrežama pomoću strojnog učenja**

Opis zadatka:

Profiliranje autora odnosi se na skup računalnih modela i postupaka za utvrđivanje karakteristika autora teksta, poput dobi, spola ili crta ličnosti, na temelju stilometrijskih obilježja teksta. Profiliranje autora svoju primjenu nalazi u računalnoj forenzici, istraživanju tržišta i društva, marketingu, znanosti o književnosti i obrazovanju. U novije vrijeme posebna je pažnja usmjerena na profiliranje autora na društvenim mrežama primjenom metoda statističkoga strojnog učenja.
U okviru završnoga rada potrebno je proučiti postupke za profiliranje autora temeljene na nadziranom i nenadziranom strojnom učenju. Osmisliti računalni postupak za utvrđivanje dobi i spola odnosno za utvrđivanje područja studiranja autora za tekstove na hrvatskome jeziku na temelju statističkih stilometrijskih obilježja teksta. Izgraditi prikladne skupove podataka s društvenih mreža označene dobi i spolom autora odnosno područjem studiranja autora te provesti statističku analizu stilometrijskih značajki. Implementirati postupak za nadziranu klasifikaciju dobi i spola te postupak za grupiranje i vizualizaciju području studiranja autora. Provesti vrednovanje modela, usporedbu s referentnim modelom, statističku obradu rezultata te analizu pogrešaka. Radu priložiti izvorni i izvršni kod razvijenog sustava, označene skupove podataka i potrebnu dokumentaciju te citirati korištenu literaturu.

Zadatak uručen pristupniku: 10. ožujka 2017.
Rok za predaju rada: 9. lipnja 2017.

Mentor:

Izv. prof. dr. sc. Jan Šnajder

Djelovođa:

Doc. dr. sc. Tomislav Hrkać

Predsjednik odbora za završni rad modula:

Prof. dr. sc. Siniša Srbljić

SADRŽAJ

1. Uvod

U posljednja dva desetljeća broj osoba koje imaju pristup internetu se je povećao za više desetaka puta. Tako je 1995. godine pristup internetu imalo svega 1% svjetske populacije, dok je 2016. godine čak 46% svjetske populacije imalo pristup internetu[1]. U razvijenim zemljama ta brojka prelazi 85% populacije. S porastom broja osoba koje imaju pristup internetu u zadnjih deset godina raste i broj korisnika društvenih mreža. Tako je 2016. godine bio 2.34 miljardne aktivnih korisnika društevnih mreža, od kojih je najpopularniji Facebook[2]. Na društvenim mrežama, ali i različitim stranicama za ocjenjivanje hotela, restorana i drugih proizvoda stvara se velika količina korisnički generiranog teksta za kojeg su zainteresirane brojne privatne tvrtke, ali i sigurnosne služe i državne agencije. No za obradu tako velikih količina podataka potrebna je pomoć računala. Međutim, takvi podaci su najčešće u nestrukturiranom obliku te ih je potrebno transformirati u struktruirani oblik kako bih ih računalo moglo analizirati.
Interakcijom računala i prirodnog jezika bavi se obrada prirodnog jezika (engl. *natural language processing*), područje računalne znanosti (engl. *computer science*), umjetne inteligencije (engl. *artificial intelligence*) i računalne lingvistike (engl. *computational linguistics*). Područja primjene obrade prirodnog jezika su brojna te uključuju odgovaranje na pitanja (engl. *question answering*), određivanje teme (engl. *topic detection*), određivanje sentimenta (engl.*sentiment analysis*), analizu autora (engl. *author analysis*).
Analiza autora dijeli se na: određivanje autora (engl. *author verification*), odnosno potvrđivanje autorstva nad djelom, detekciju plagijata (engl. *plagiarism detection*), tj. pronalaženje sličnosti između dva teksta te profiliranje autora (engl. *author profiling*), određivanjem osobina autora kao što su spol, dob, jezik, razina obrazovanja ili osobine ličnosti (Stamatatos, 2009). Profiliranje korisnika društevnih mreža je zanimljivo i stoga što su korisnici društvenih mreža najčešće anonimni, a kako bi se iskoristio korisnički generirani sadržaj s društvenih mreža potrebno je znati osim što je napisano

[1] http://www.internetlivestats.com/internet-users/
[2] https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

i tko je napisao taj tekst. Profiliranjem korisnika društvenih mreža tvrtke mogu prilagoditi svoje reklame i proizvode ciljnim skupinama analiziranjem što pripadnici tih skupina misle o njihovim proizvodima.

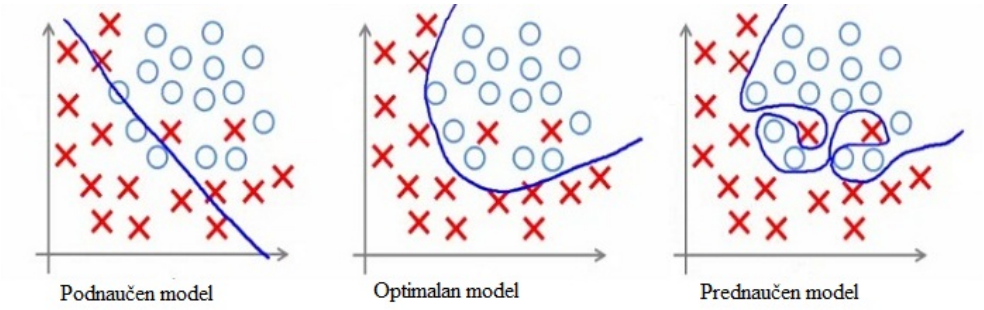
Ovaj se rad bavi profiliranjem korisnika Facebooka i foruma metodama strojnog učenja. Sastoji se od dva odvojena zadataka. Prvi je određivanje spola korisnika foruma i facebooka. Specifičnost ovog rada je u tome što nastoji razviti model profiliranja korisnika društvenih mreža koji bi bio primjenjiv i na drugim društvenim mrežama, a ne samo na onoj društvenoj mreži na kojoj je razvijen. Drugi dio rada se odnosi na određivanje znanstvenog područja kojem pripada fakultet kojega pohađa korisnika foruma.

U nastavku rad je odganiziran tko da su u 2. poglavlju objašnjene metode strojno učenja često korištene pri profiliranja autora te radovi koji su korišteni kao terorijska osnova za izradu ovog rada. Potom je u 3. poglavlju opisan postupak prikupljanja podataka za ovaj rad kao i njihov konačni oblik te model korišten u ovom radu. Na kraju u 4. poglavlju su prikazani glavni rezultati ovoga rada i doneseni zaključci.

2. Profiliranje autora

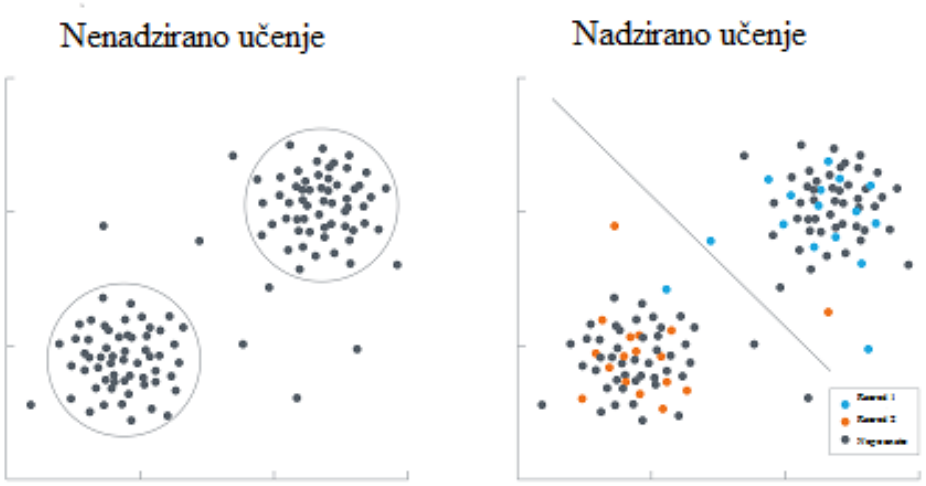
2.1. Strojno učenje

Područje računalne znanosti koje se bavi razvojem i evaluacijom algoritama koji mogu učiti iz podataka te stvarati predikcije o budućim podacima na temelju naučenog naziva se strojno učenje. Ono je podgrana umjetne inteligencije te se je razvilo iz prepoznavanja uzorka i statistike. Prema Mitchell et al. (1997), računalni program uči iz iskustva E s obzirom na neku klasu zadataka T i mjeri učinkovitost P ako se njegova učinkovitost u zadacima u T, mjerena pomoću P, poboljšava s iskustvom E.
Naime, algoritam strojnog učenja stvara model koji je definiran parametrima čije se vrijednosti određuju iz podataka na kojima algoritam uči. Tijekom tog procesa može doći do prenaučenosti i podnaučenosti modela. Do prenaučenosti modela dolazi kad je model prekompleksan, npr. kad ima previše parametara s obzirom na broj podatak na kojima algoritam uči, te tada model opisuju slučajne značajke podataka, a ne temeljne odnose između podataka. Nasuprot prenaučenosti do podnaučenosti dolazi kad je model prejednostavan, npr. kad je model linearn, a podaci su nelinearni, te ne može zahvatiti temeljne odnose između podataka. U oba slučaja prediktivna moć modela je mala.



Slika 2.1: Različiti modeli ovisno o naučenosti

Osnovna podjela strojnog učenja je na nadzirano (engl. *supervised learning*) i nenadzirano (engl. *unsupervised learning*) strojno učenje (Love, 2002).
Kod nadziranog strojnog učenja sa svakim primjerom povezan je njegov izlaz, odnosno ulazni podaci su oblika (x,y), pri čemu je x vektor značajki, a y oznaka klase kojoj pripada, a potrebno je naći preslikavanje y=f(x). Nadzirano strojno učenje se dalje dijeli na klasifikaciju (engl. *classification*) kod koje je y diskretna varijabla i regresiju (engl. *regression*) kod koje je y kontinuirana varijabla.
Za razliku od nadziranog strojnog učenja kod nenadziranog strojnog učenja ulazni podaci nisu povezani s svojim izlazom, odnosno klasom kojoj pripadaju, te je zadatak strojnog učenja pronaći temeljne odnose između podataka. Nenadzirano strojno učenje se dijeli na grupiranje (engl. *clustering*), procjenu gustoće (engl. *density estimation*) i smanjenje dimenzionalnosti (engl. *dimensionality reduction*).

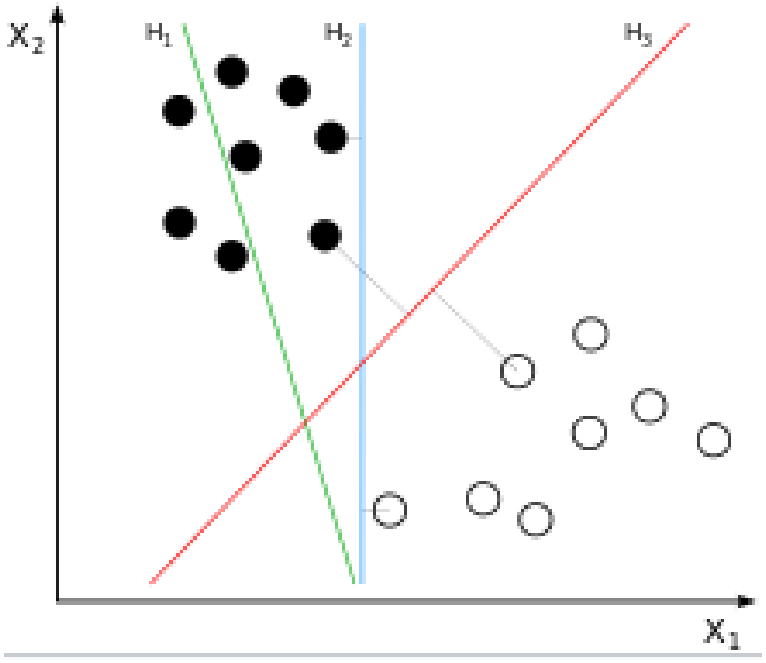


Slika 2.2: Nadzirano i nenadzirano strojno učenje

2.2. Stroj potpornih vektora

Preferirana metoda strojnog učenja kod profiliranja autora teksta, bilo da se radi o profiliranju korisnika društvenih mreža ili profiliranju autora druge vrste teksta kao što su znanstveni članci, je nadzirano strojno učenje, odnosno podvrsta strojnog učenja kasifikacija (Sebastiani, 2002). Pri tome je najčešće korištena metoda za klasifikaciju stroj potpornih vektora (engl. *supprot vector machines, SVM*) (Stamatatos, 2009) te je stoga i u ovom radu korišten SVM-klasifikator.

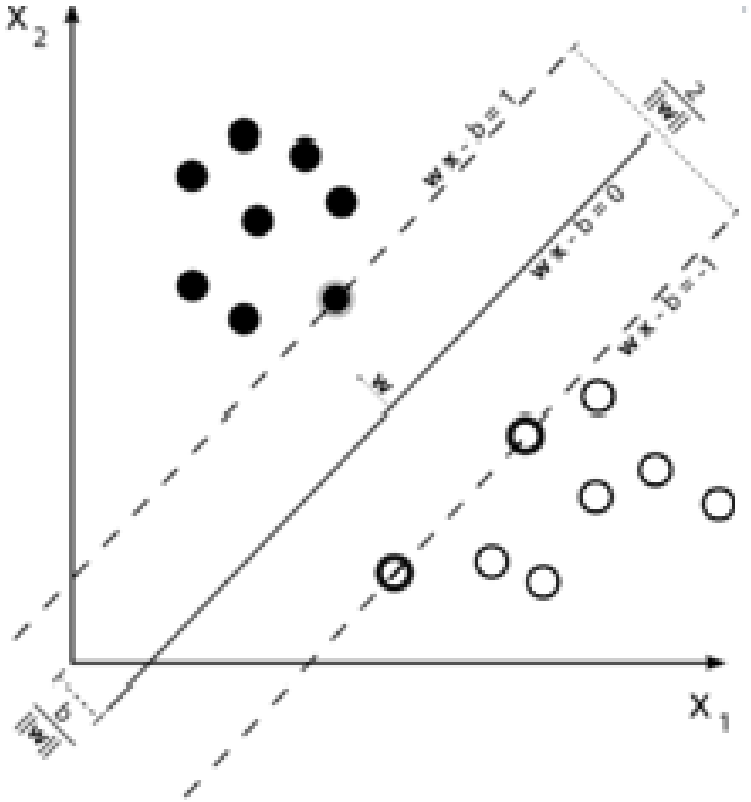
Pošto se radi o metodi nadziranog strojnog učenja, SVM na ulaz dobiva podatke povezane s klasom kojoj pripadaju te ih prikazuje kao točke u prostoru raspoređene na način da su točke koje prestavljuju podatke koji pripadaju različitim klasama međusobno što razmaknutije (Fradkin i Muchnik, 2006). Udaljenost između točaka koje pripadaju različitim klasama naziva se margina razdvajanja klasa te SVM nalazi hiperravninu koja maksimizira marginu oko hiperravnine razdvajanja. Potporni vektori su oni vektori koji se nalaze na rubovima i najbliži su podacima. Na slici 2.3.[1] prikazano je kako SVM odabire optimalnu hiprerravninu razdvajanja. Naime, hiprerravnina H_1 ne razdvaja klasa, hiperavnina H_2 pak razdvaja klase ali s malom marginom razdvajanja, dok hiperravnina H_3 razdvaja kalase s najvećom marginom razdvajanja.



Slika 2.3: Odabir optimalne hiperravnine razdvajanja pomoću SVM

[1] https://en.wikipedia.org/wiki/Support_vector_machine

Linarna klasifikacija strojem potpornih vektora dijeli se na metodu čvrste margine (engl. *hard-margine*) i metodu meke margine (engl. *soft-margine*).
Metoda čvrste margine se koristi ukoliko je podatke na kojima SVM uči moguće linearno radvojiti. Tada se mogu odabrati dvije paralelne hiperravnine koje razdvajaju podatke na način da je razmak između hiperravnina najveći mogući. Prostor između ove dvije hiperravnine se naziva margina te se maksimalna margina razdvajanja nalazi točno između te dvije hiperravnine.

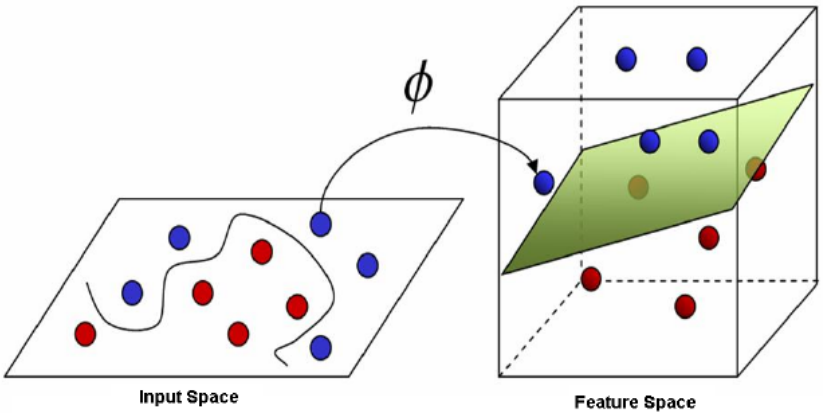


Slika 2.4: Metoda čvrste margine[2]

Metoda meke margine koristi se kada ravninom nije moguće razdvojiti vektore različitih klasa. Pri tome omogućujemo određenu pogrešku prilikom presjecanja prostora hiperravninom, te se uvodi varijabla ξ_i (engl. *slack variable*) koja predstavlja pogrešku klasificiranja nekog vektora x_i prema ravnini podjele. Pogreška raste s udaljenošću vektora od ravnine. Uvodi se i parametar C koji služi kao kazna za krivu klasifikaciju. Kako je cilj minimalizirati broj krivo klasificiranih primjera izračunatoj margini se pridaje $C \sum_{i=1}^{N} \xi_i$, pri čemu je N broj primjera na kojima SVM uči.

[2]https://en.wikipedia.org/wiki/Support_vector_machine#Soft-margin

Nelinearna klasifikacija koristi se u situacijma u kojima nije moguće linarano podijeliti točke koje predstavljaju podatke u prostoru, nego se prelazi u višedimenzionalni prostor kako bi se onda u tom višedimenzionalnom prostoru mogla koristiti linearna klasifikacija. Transformacijska funkcija koja preslikava iz linearno nerazdvojivog prostora u višedimenzionalni, linearno razdvojiv, prostor naziva se jezgrena funkcija (engl. *kernel function*) te ovisi o prametrima C i γ za koje je potrebno provesti optimiranje kako bi klasifikator ima što veću sposobnost generalizacije.



Slika 2.5: Nelinearna klasifikacija[3]

Iako je SVM razvijen za primjenu u binomnoj klasifikaciji (Fradkin i Muchnik, 2006), može ga se koristiti i za rješavanje problema klasificiranja u više klasa. Tada se problem razlaže na više binomnih problema, te razlikujemo dva pristupa: jedan protiv svih (engl. *one versu all*) i jedan protiv jednog (engl. *one versus one*).
Kod pristupa jedan protiv svih izgrađuje se onoliko klasifikatora koliko ima klasa, te se klasifikacija provodi onim klasifikatorom koji ima najveću uspješnost pri klasifikaciji. Jedan protiv svih je najčešće korištena metoda višestruke klasifikacije.
Pristup jedan protiv jednog pak izgrađuje po jedan klasifikator za svaki par klasa te se klasifikacija providi na način da se vektoru pridodaje klasa koja mu je najviše puta dodijeljna.

[3]https://www.researchgate.net/figure/260283043_fig13_Figure-A15-The-non-linear-SVM-classifier-with-the-kernel-trick

2.3. Relevantni radovi

Područje profiliranja autora na društvenim mrežama posljednjih godina je popularno te stoga postoji i veliki broj radova iz ovog područja (Rangel et al., 2014; Zhuang et al., 2015). No, većina radova iz područja profiliranja autora na društvenim mrežama ima ograničenje u smislu da je model razvijan i testiran na podacima s iste društvene mreže. Naime, u takvim uvjetima modeli su se uvelike oslanjali na značajke teksta ovisne o sadržaju. Upravo to želimo izbjeći u ovome radu. Želimo razviti model koji će moći odrediti spol autora neovisno o društvenoj mreži iz koji potječe tekst kojega je napisao autor.

Rangel i Rosso (2013) su, analizirajući veliki broj tekstova na španjolskom i engleskom jeziku s Wikipedije, [4] blogova, forum, Facebooka [5] i Twittera [6], utvrdili kako žene i muškarci na različiti način koriste jezik kako bi se izrazili. Muškarci više od žena koriste prijedloge, dok žene više od muškaraca koriste zamjenice i uzvike tj. riječi koje izražavaju emocije kao sto su "aha", "hum", "bravo" (engl. *interjections*). Također su predložili i stilometrijske značajke teksta kojima bi se trebale zahvatiti temeljne razlike između žena i muškaraca pri korištenu jezika neovisne o temi o kojoj se govori u tekstu. Te značajke su: broj riječi koje počimnju velikim početnim slovom, broj riječi pisan velikim slovima, broj riječi s ponovljenim slovima (npr. Heeeellllllooo), broj interpunkcijskih znakova: točka, zarez, upitnik, uskličnik, dvotočka, točkazarez, navodnici, vrste riječi (engl. *part-of-speech, POS*) tj. učestalost korištenja svake od vrsta riječi te emotikoni, odnosno ukupan broj emotikona i broj emotikona za svaku od emocija: sreća, tuga, ljutnja, gađenje, iznenađenje, zbunjenost.

Na natjecanju PAN 2016[7] jedan od zadataka sudionika u kategoriji profiliranje autora bio je određivanje spola autora teksta, ali s tim da se je skup podataka za razvoj modela sastojao od tvitova dok je potom model testiran na skupu podataka koji se je sastojao od blogova (Rangel et al., 2016). Prvoplasirani (op Vollenbroek et al., 2016) i drugoplasirani (Bilan i Zhekova, 2016) modeli su koristili stilometrijske značajke kao gore opisane (Rangel i Rosso, 2013) te još i oznake na razini rečenica kao što su: započinje li rečenica velikim početnim slovom, završava li rečenica interpunkcijskim znakom, broj riječi u rečenici te prosječna duljina riječi. Oba modela su postigla uspješnost kasifikacije od oko 75%.

Jedini nama poznat rad na nekom od slavenskih jezika, na ruskom jeziku koji se ba-

[4] https://en.wikipedia.org, https://es.wikipedia.org

[5] https://www.facebook.com/

[6] https://twitter.com/

[7] http://pan.webis.de/clef16/pan16-web/author-profiling.html

vio profiliranjem autora (Sboev et al., 2016), također je koristio samo stilometrijska obilježja kako bi odredio spol autora teksta, te je postigao upješnost klasifikacije od 85%. Ovaj je rad ukazao na još jedan problem koji je prisutan kod profiliranja autora. Gotovo svi radovi se bave velikim svjetskim jezicima kao što su engleski i španjolski, te zadnjih godina i arapski (Estival et al., 2007).

U ovome radu korištene su stilometrijske značajke kako bi se izgradio model za određivanje spola autora teksta na društevenim mrežama, te je model testiran na podacima s Facebooka dok je izgrađen na podacima s foruma. Nadalje, koliko nam je poznato ovo je jedini rad koji se bavi određivanjem spola autora na društevenim mrežama na hrvatskom jeziku. Drugi dio ovoga rada odnosi se na određivanje znanstvenog područja kojemu pripada fakultet koji autor teksta na forumu pohada. Koliko nam je poznato prijašnji radovi iz profliranja autora nisu se bavili određivanjem područja kojemu pripada fakultet kojega autor teksta pohađa, bilo na hrvatskom jeziku ili nekom drugom jeziku. Duong et al. (2016) su u svom radu nastojali odrediti zanimanje autora teksta na forumu. Koristili su tri široka područja zanimanja: zdravstvo, tehnika i tehnologija, trgovina i administracija te su uspjeli postići uspješnost klasifikacije od 57% koristeći stilometijska obilježja teksta i obilježja ovisna o sadržaju teksta. Tako smo se u ovom radu ograničili na četiri znanstvena područja: prirodne znanosti, tehničke znanosti, biomedicinske znanosti te društveno-humanističke znanosti.

3. Podaci i model

3.1. Prikupljanje i predobrada podataka

Kako smo u ovome radu izgradili dva modela, jedan za određivanje kojemu području znanosti pripada fakultet kojega korisnik foruma pohađa te drugi model za određivanje spola korisnika foruma koji je još korišten i za određivanje spola korisnika društvene mreže Facebook, bilo je potrebno prikupiti tri skupa podataka: dva za korisnike foruma te jedan za korisnike društvene mreže Facebook.

3.1.1. Podaci o korisnicima foruma

Podatke o korisnicima foruma prikupili smo s društvene mreže forum.hr.[1] Forum.hr je najveći forum u Hrvatskoj te sadrži podforum namijenjen raspravama o različitim fakultetima u Hrvatskoj. Korištenjem internet crawlera[2] prikupljeni su podaci s foruma. Pošto su podaci bili organizirani po temema pri čemu se je svaka tema odnosila na jedan od fakulteta u Hrvatskoj, odlučili smo se ograničiti na petnaest tema odnosno fakulteta. Za područje biomedicinskih znanosti odabrani su fakulteti: Medicinski fakultet Zagreb, Medicinski fakultet Rijeka, Medicinski fakultet Osijek i Farmaceutsko biokemijski fakultet. Fakultet elektrotehnike i računarstva, Fakultet strojarstva i brodogradnje, Građevinski fakultet Zagreb i Tehnički fakultet Rijeka predstavljali su područje tehničkih znanosti. Područje prirodnih znanosti predstavljali su matematički, fizički, kemijski i biološki odsjek Prirodoslovno-matematičkog fakulteta u Zagrebu. Ekonomski fakultet u Zagrebu i Osijeku te Pravni fakultet u Zagrebu i Osijeku odabrani su za društveno-humanističko područje. Pretpostavili smo da oni korisnici foruma koji su pisali na temi samo jednog fakulteta i pohađaju taj fakultet te su samo oni bili uključeni u daljnju obradu podataka, dok su ostali isključeni. Daljnja ograničenja koja smo postavili bila su da svaki korisnik foruma mora imati barem tri komentara na

[1] www.forum.hr

[2] https://bitbucket.org/infomiho/forum.hr-scraper/src

forumu te da ukupno svi njegovi komentari moraju sadržavati barem 400 riječi. Naime, i smanjenje dostupne količine teksta kojom je predstavljen pojedini autor smanjuje se i mogućnost učenja algoritma iz tako malobrojnih podataka (Zhang i Zhang, 2010). Većina istraživanja iz područja profiliranja autora po svakom autoru je imala barem 300 riječi. Svaki autor predstavljen je listom svojih komentara na forum.
Konačan broj korisnika foruma koji su uključeni u daljnju obradu i njihova raspodjela po znanstvenim područjima prikazana je u tablici 3.1.

Tablica 3.1: Raspodjela korisnika foruma po znanstvenim područjima

Znanstveno područje	Broj korisnika
Prirodne znanosti	95
Tehničke znanosti	238
Biomedicinske znanosti	136
Društveno-humanističke znanosti	209
Ukupno	678

Za određivanje spola autora korišteni su isti podaci s društvene mreže forum.hr, samo bez ograničenja kako svaki korisnik može komentirati na samo jednoj temi. Kako bi se odredio spol korisnika foruma izgrađen je crawler koji je s javnog profila svakoga od korisnika foruma uključenog u analizu uzeo podatak o spolu korisnika. Ako korisnik foruma nije na svom profilu ispunio rubriku spol isključen je iz daljnje analize. Preporučeno je da se podaci o dobi, spolu te drugi demografski podaci korisnika društvenih mreža prikupljaju s njihovih javnih profila jer je ručno određivanje tih podataka podložno pristranosti procjenjivača (Nguyen et al., 2013). Raspodjela po spolu i ukupan broj korisnika foruma uključenih u daljnju obradu prikazana je u tablici 3.2.

Tablica 3.2: Raspodjela korisnika foruma po spolu

Spol	Broj korisnika
Ženski	278
Muški	334
Ukupno	612

Predobrada podataka, odnosno tekstova koijma je predstavljen svaki od autora također je napravljeno u ovoj fazi. Naime, dijeljenje teksta u rečenice, tokeniziranje teksta i svođenje riječi na leme te određivanje vrste riječi iziskuje vrlo mnogo vre-

mena, pa su predobrađeni tekstovi zajedno s izvornim tekstovima kojima je predstavljen svaki od autora spremljeni u serijaliziranu JSON-datoteku, po jednu za određivanje spola i jednu za određivanje znanstvenog područja fakulteta kojega pohađa korisnik foruma. Za dijeljenje teksta u rečenice i tokeniziranje rečenica korišten je javno dostupa Python-biblioteka text-sentence,[3] dok je za svođenje riječi na leme i određivanje vrsta riječi korišten je također javno dostupan *tagger* i *lemmatizer* za hrvatski, slovenski i srpski jezik.[4]

Na kraju, svaki korisnik foruma uključen u analizu predstavljen je s:

- korisničkim imenom;
- korisničkim idijem;
- komentarima na forumu (lista);
- komentarima na forumu podijeljeni u tokenizirane rečenice;
- komentarima na forumu, riječi svedene na leme;
- komentarima na forumu, riječi svedene na vrste riječi.

Korisnici foruma koji su bili uključeni u određivanje spola imali su i varijablu spol.

3.1.2. Podaci o korisnicima Facebooka

Podaci o korisnicima Facebooka prikupljeni su iz dvije Facebook grupe "Grupa za rasprave - Društvo za promociju znanosti i kritičkog mišljenja" i "Ateisti i agnostici Hrvatske". Ove dvije grupe su odabrane jer obje grupe imaju veliki broj članova te se na njima često razviju duge rasprave. Za prikupljanje podataka s Facebooka korišten je crawler razvijen za potrebe TakeLabovog projekta CATACX.[5] Opet su prikupljeni komentari bili u obliku liste objava na svakoj od grupa te pridruženih komentara pa ih je trebalo reorganizirati na način da budu grupirani po korisnicima Facebooka. S obzirom da su komentari na Facebook grupama značajno varirali po veličini, u daljnjoj analizi su zadržani oni korisnici Facebooka čiji su komentari i objave imali ukupno više od 400 riječi, a nije postavljeno ograničenje na broj komentara i objava na Facebook grupi. Spol korisnika Facebooka određivala su tri anotatora jer je malen broj korisnika imao javno objavljen podatak o svom spolu, a pristranost anotatora je niska jer su na raspolaganju uz ime i prezime svakog od korisnika imali i njegovu profilnu sliku. Ukoliko se sva tri anotatora nisu slagala oko procjene spola korisnika Facebooka, taj

[3] https://pypi.python.org/pypi/text-sentence

[4] https://github.com/clarinsi/reldi-tagger

[5] http://takelab.fer.hr/catacx/

korisnik je isključen iz daljnje analize. Raspodjela korisnika Facebooka po spolu i njhov ukupan broj vidljiva je iz tablice 3.3.

Tablica 3.3: Raspodjela korisnika Facebooka po spolu

Spol	Broj korisnika
Ženski	43
Muški	72
Ukupno	115

Podaci o korisnicima Facebooka su također pohranjeni u zasebnu serijaliziranu JSON-datoteku u istom obliku kako i podaci o korisnicima foruma za određivanje spola korisnika.

3.2. Model

Za implementaciju je korišten programski jezik Python, verzija 2.7.12. Programski jezik Python je odabran zbog velikog broja dostupnih knjižnica razvijenih za potrebe strojnog učenja. Kako je već navedeno, izgradnja modela za profiliranje autora teksta se najčešće provodi metodom nadziranog strojnog učenja i to klasifikacijom te se je kao najprikladniji klasifikator pokazao stroj potpornih vektora (Zhuang et al., 2015). Tako je i u ovom radu korišten već opisan stroj potpornih vektora implementiran u Python-knjižnici scikit-learn (Pedregosa et al., 2011). Razvijena su dva modela: jedan za određivanje spola autora i drugi za određivanje znanstvenog područja kojemu pripada fakultet kojega pohađa autor teksta.
Za razvoj modela korištena su dva skupa značajki teksta. Prvi skup značajki teksta namijenje je zahvaćanju sadržaja teksta, a drugi skup značajki je namijenjen za zahvaćanje stila autora teksta. Značajke teksta su podijeljene na one usmjerene na sadržaj teksta i one usmjerene na stil autora kako bi se moglo odrediti koje su korisnije za profiliranje autora na društevenim mrežama.

3.2.1. Značajke sadržaja teksta

Skup značajki za zahvaćanje sadržaja teksta temeljio se na frekvencija pojma - inverzna frekvencija dokumenta (engl. *Term Frequency-Inverse Document Frequency*) faktor (TF-IDF faktor), što je učestalo korištena procedura za određivanje informativnosti svake od riječi u korpusu (Sparck Jones, 1972). Za svaku lemu koja se je pojavila u tekstu nekog od autora TF-IDF faktor izračunat je na način da se broj pojavljivanja te riječi pomnoži s logaritnom omjera ukupnog broja autora i broja autora koji su koristili tu lemu. Korištena je TF-IDF implementacija dostupna u Python-knjižnici scikit-learn (Pedregosa et al., 2011). TF-IDF je izračunat na unigramima i bigramina lema tj. na nizu od dvije uzastopne leme te na n-gramima od 1 do 4 znaka, odnon na nizu od 1 do 4 uzastopna znaka. Prije samog izračuna TF-IDF uklonjene su zaustavne riječi. Zaustavne riječi su npr. članci, prijedlozi, veznici koji se pojavljuju u u mnogim tekstovima te se stoga isključuju iz analize.

3.2.2. Značajke stila autora

Skup značajki za zahvaćanje stila autora napravljen je prema uzoru na Rangel i Rosso (2013) i Bilan i Zhekova (2016) te se sastojao od šesnaest značajki i TF-IDF faktora za unigrame i bigrame vrsta riječi. TF-IDF faktora za unigrame i bigrame vrsta riječi pokazao se koristan za određivanje spola autora u (op Vollenbroek et al., 2016), a i Rangel i Rosso (2013) su pokazali kako se muškarci i žene razlikuju prema učestalosti korištenja pojedinih vrsta riječi.
Preostale značajke bile su:

- *Hapax legomenon*, odnosno broj riječi koje su se pojavile samo jednom;
- Prosječan broj riječi u poruci na forumu, odnosno Facebook komentaru;
- Ukupan broj riječi;
- Prosječna duljina riječi;
- Duljina najdulje riječi;
- Udio riječi pisanih velikim slovima;
- Udio riječi koje su brojevi;
- Udio riječi kraćih od 4 znaka;
- Udio riječi duljih od 5 znakava;
- Udio riječi duljih od 7 znakava;
- Prosječan broj rečenica u poruci na forumu, odnosno komentaru na Facebooku;
- Prosječan broj riječi u rečenici;
- Udio rečenica koje završavaju interpunkcijskim znakom;
- Udio rečenica koje završavaju s tri točke;
- Prosječan broj zareza u rečenici;
- Udio složenih rečenica.

Često korištena značajka pri profiliranju autora na društvenim mrežama su i emotikoni (Rangel i Rosso, 2013), ali pošto Facebook i forum.hr koriste različite oblike zapisa emotikone, oni nisu korišteni kao značajka.

4. Rezultati

4.1. Evaluacijske mjere

Prije prikaza samih rezultata klasifikacije korisnika društevenih mreža s obzirom na spol i znanstveno produčje kojemu pripada fakultet kojega pohađa korisnik društvene mreže pomoću razvijenih modela potrebno je ukratko opisati korištene mjere.
Standardne mjere za vrednovanje klasifikatora su točnost (engl. *accuracy*), preciznost (engl. *precision*) i odziv (engl. *recall*) (Joachims, 2002). Ove mjere se temelje na matrici zabune. Matrica zabune je kvadratna matrica veličine [n_class, n_class], pri ćemu stupci predstavljaju klasu kojoj primjer stvarno pripada, a retci predstvljaju klasu koju je klasifikator odredio za dani primjer.
Za svaku klasu se određuje broj: ispravno pozitivnih primjera (engl. *true positive, TP*), lažno pozitivnih primjera (engl. *false positive, FP*), ispravno negativnih primjeran (engl. *true negative, TN*) i lažno negativnih primjera (engl. *false negative, FN*). Ispravno pozitivni primjeri su oni primjeri za koje je klasifikator ispravno odredio kategoriju kojoj primjer pripadaju, u matrici zabune nalaze se u dijagonali. Lažno pozitivni primjeri su oni primjeri kojima je klasifikator neispravno odredio klasu, u matrici zabune nalaze se u nedijagonalnim elementima retka. Ispravno negativni primjeri su oni za koje je klasifikator ispravno odredio da ne pripadaju određenoj klasi, u matrici zabune nalaze se izvan stupca i retka koji je povezan s određenom klasom. Lažno negativni primjeri su oni primjeri za koje je klasifikator pogrešno odredio da ne pripadaju određenoj klasi.

Točnost (engl. *accuracy*) je definirana kao udio točno klasificiranih primjera u skupu svih primjera.

$$Acc = \frac{TP + TN}{TP + FP + FN + TN}$$

Preciznost (engl. *precision*) je definirana kao udio točno klasificiranih primjera u skupu pozitivno klasificiranih primjera.

$$P = \frac{TP}{TP + FP}$$

Odziv (engl. *recall*) je definiran kao udio točno klasificiranih primjera u skupu svih pozitivnih primjera.

$$R = \frac{TP}{TP + FN}$$

F1-mjera mjera računa se kao harmonijska sredina preciznosti i odziva, ali se može izraziti i kao težinski prosjek preciznosti i odziva pri čemu parametar β određuje da li se veća važnost pridaje preciznosti ili odzivu.

$$F1 = \frac{2PR}{P + R} = (1 + \beta)^2 \frac{PR}{\beta^2 P + R}$$

Kod računanja uspješnosti klasifikacije kod višeklasne klasifikacije svaka od mjera uspješnosti klasifikacije se računa zasebno za svaku klasu, a kao mjera cijelog sustava uzima se prosjek mjera svih klasa.

4.2. Određivanje spola autora

Za određivanje spola autora korišten je model SVM za binarnu klasifikaciju. Odabir optimalnih hiperparametara modela proveden je pretragom po rešetci (engl. *grid search*) i unakrsnom validacijom (engl. *cross-validation*).
Skup značajki koji se odnose na sadržaj teksta i skup značajki koji se odnosi na stil autora prvo su primijenjeni zasebno, pa potom i zajedno kako bi se ispitala pretpostavka (Bilan i Zhekova, 2016; op Vollenbroek et al., 2016) kako su značajke koje zahvaćaju stil autora učinkovitije za profiliranje autora od značajki koje zahvaćaju sadržaj teksta što izrazito dolazi do izražaja kad se model testira na skupu podataka iz drugačijeg izvora od onog na kojemu je model izgrađen. Kao što je već rečeno, model za određivanje spola autora izgrađen je pomoću tekstova korisnika foruma, dok je testiran na tekstovima korisnika Facebooka i na tekstovima korisnika foruma.
Raspodjela autora prema spolu u skupovima testnih podataka prikazana je u tablici 4.1.

Tablica 4.1: Raspodjela autora prema spolu u skupovima testnih podataka

Društvena mreža	Žene	Muškarci
Forum	51	71
Facebook	43	72

4.2.1. Određivanje spola autora korištenjem značajki sadržaja teksta

Rezultati klasifikacije autora prema spolu pri korištenju samo značajki teksta koje su namijenjene zahvaćanju sadržaja teksta prikazan je u tablici 4.2.

Tablica 4.2: Uspješnost klasifikacije i optimalni parametri SVM klasifikatora kod određivanja spola autora korištenjem značajki sadržaja teksta

Društvena mreža	Točnost	Preciznost	Odziv	F_1	jezgra	C
Forum	0.58	0.55	0.58	0.49	linear	10
Facebook	0.63	0.39	0.63	0.48	linear	10

Vidljivo je kako klasifikacije autora prema spolu korištenjem samo značajki teksta koje zahvaćaju sadržaj teksta nije uspješna niti kad se testiranje vrši nad podacima s iste društvene mreže kao što je model i izgrađen, kako niti kad se koriste podaci s druge društvene mreže. Rezultati za skup podataka s Facebooka su očekivani, naime

podaci s Facebooka su prikupljeni iz dvije Facebook grupe "Ateisti i agnostici Hrvatske" i "Grupa za rasprave - Društvo za promociju znanosti i kritičkog mišljenja" dok su podaci s foruma prikupljeni s tema o različitim fakultetima u Hrvatskoj, pa nije za očekivati da će korisnici tih Facebook grupa i podforuma o fakultetima pisati o istim temama. Premda se rezultat na skupu podataka s foruma isprva čini neočekivano nizak, uvidom u tablicu 4.3. koja prikazuje deset najvažniji lema za kategoriju "muškarci" i "žene" vidimo da su i muškarci i žene na forumu raspravljali o istoj temi, pa ih stoga model nije mogao razlikovati.

Tablica 4.3: Deset najvažnijih bigram i unigram lema za određivanje spola

Muškarci	Žene
preddiplomski studij	frendica
dio element	seminar
povjerenstvo	prijaviti
razlikovni predmet	čestitati
studij	izvanredan
godina	fax
grupa	zg
tabela	god
student	prijaviti
gimnazija	molekularan

Iz matrica zabune (tablica 4.4.) za skup podataka s forma i Facebooka vidljivo je da model ne razlikuje žene od muškaraca. U skupu podataka s foruma samo je 5 žena ispravno klasificirano, pa je odziv za klasu "Žene" svega 0.1. Rezultat je još lošiji na skupu podataka s Facebooka gdje je niti jedna žena ispravno klasificirana pa je odziv za klasu "Žene" 0.

Tablica 4.4: Matrice zabune za korisnike foruma i Facebooka, značajke sadržaja teksta

Forum	Facebook
$\begin{bmatrix} 5 & 46 \\ 5 & 66 \end{bmatrix}$	$\begin{bmatrix} 0 & 43 \\ 0 & 72 \end{bmatrix}$

4.2.2. Određivanje spola autora korištenjem značajki stila autora

Rezultati klasifikacije autora prema spolu pri korištenju samo značajki teksta koje su namijenjen zahvaćanju stila autora prikazani su u tablici 4.4.

Tablica 4.5: Uspješnost klasifikacije i optimalni parametri SVM-klasifikatora kod određivanja spola autora korištenjem značajki stila autora

Društvena mreža	Točnost	Preciznost	Odziv	F_1	jezgra	C
Forum	0.96	0.96	0.96	0.96	linear	100
Facebook	0.88	0.88	0.88	0.88	linear	100

Odmah je vidljivo kako su značajke usmjerena na zahvaćanje stila autora značajno uspješnije u određivanje spola autora bilo da se model testira na podacima s iste društvene mreže iz koje su uzeti i podaci na kojima je model razvjen ili da su testni podaci preuzeti s neke druge društvene mreže.

Tablica 4.6: Matrice zabune za korisnike foruma i Facebooka, značajke stila autora

Forum	Facebook
$\begin{bmatrix} 49 & 2 \\ 3 & 68 \end{bmatrix}$	$\begin{bmatrix} 32 & 11 \\ 3 & 69 \end{bmatrix}$

Uspješnost klasifikacije korištenjem značajke usmjerena na zahvaćanje stila autora vidljiva je i u tablicama zablude, udio lažno pozitivnih primjera je malen za oba skupa testnih podataka.

4.2.3. Određivanje spola autora korištenjem značajki stila autora i sadržaja teksta

Rezultati klasifikacije autora prema spolu pri korištenju oba skupa značajki: značajki teksta koje su namijenjene zahvaćanju stila autora i značajake koje su namijenjene zahvaćanju sadržaja teksta prikazani su u tablici 4.7.
Vidljivo je da su rezultati klasifikacije korisnika društvenih mreža prema spolu lošiji ako se koriste značajke stila autora i značajke sadržaja teksta te negativan učinak značajki sadržaja teksta više utječe na klasifikaciju kod podataka s Faceboka tj. kod podataka preuzetih s različite društvene mreže nego na klasifikaciju podataka preuzietih s iste društvene mreže kao i podaci na kojima je model razvijen.

Tablica 4.7: Uspješnost klasifikacije i optimalni parametri SVM-klafifikaotra kod određivanja spola autora korištenjem značajki stila autora i sadržaja teksta

Društvena mreža	Točnost	Preciznost	Odziv	F_1	jezgra	C
Forum	0.89	0.89	0.89	0.89	linear	10
Facebook	0.80	0.83	0.79	0.77	linear	10

Tablica 4.8: Matrice zabune za korisnike foruma i Facebooka, značajke stila autora i sadržaja teksta

Forum	Facebook
$\begin{bmatrix} 48 & 3 \\ 11 & 60 \end{bmatrix}$	$\begin{bmatrix} 20 & 23 \\ 1 & 71 \end{bmatrix}$

4.2.4. Zaključak o određivanju spola autora

Cilj ovoga rada bio je izgraditi model za određivanje spola korisnika društvenih mreža koji bi bio primjenjiv na različitim društvenima mrežama na hrvatskom jeziku. Kako bismo to postigli, prikupljeni su podaci s dvije društvene mreže: foruma i Facebooka. Model je razvijena na podacima s foruma. Razvijena su tri klasifiaktora korištenjem različitih značajki teksta: značajki koje se odnose na stil autora, značajke koje se odnose na sadržaj teksta te značajke koje se odnose i na sadržaj teksta i na stil autora, tj. objedinjena dva prethodna skupa značaki. Najboljim modelom se pokazao onaj model koji je koristio samo značajke stila autora, ali ne i značajke sadržaja teksta. Dobivena je uspješnost klasifikacije od 96% na podacima s foruma te 88% na podacima s Facebooka. Ovakvi nalazi su očekivani. Kao što je već navedeno, Rangel i Rosso (2013) su pokazali kako muškarci i žene koriste jezik na drugačiji način te su i Bilan i Zhekova (2016) i op Vollenbroek et al. (2016) dobili slične rezultate kad su razvijali model na podacima s jedne društvene mreže, a testirali na podacima s druge društvene mreže, odnosno rezultat na podacima s društvene mreže koja je korištena samo za testiranje modela bio je niži nego rezultat na podacima za testiranje s društvene mreže iz čijih podataka je model i razvijen.

4.3. Određivanje znanstvenog područja

Drugi cilj ovog rada bio je odrediti kojem znanstvenom području pripada fakultet kojeg pohađa korisnik društvene mreže. U ovom dijelu rada korišteni su samo podaci s društvene mreže forum i četiri znanstvena područja: prirodne znanosti, tehničke znanosti, biomedicinske znanosti i društveno-humanističke znanosti. U ovom dijelu rada korišten je SVM-klasifikator s višestrukom klasifikacijom jedan protiv jednog. Odabir optimalnih parametara modela proveden je pretragom po rešetci (engl. *grid search*) i unakrsnom validacijom (engl. *cross-validation*).
Također i u ovom dijelu rada korištena su ista dva skupa značajki: značajke usmjerene na zahvaćanje sadržaja teksta i značajke usmjerene na zahvaćanje stila autora.
Raspodjela korisnika foruma u testnim podacima prikazana je u tablici 4.9.

Tablica 4.9: Raspodjela korisnika foruma prema znanstvenom području kojemu pripada fakultet kojeg autor pohađa

Znanstveno područje	broj primjera
Prirodne znanosti	30
Tehničke znanosti	30
Biomedicinske znanosti	30
Društveno-humanističke znanosti	30

4.3.1. Određivanje znanstvenog područja korištenjem značajki sadržaja teksta

Rezultati klasifikacije autora prema znanstvenom području kojemu pripada fakultet koji pohađa autor pri korištenju samo značajki teksta koje zahvaćaju sadržaj teksta prikazan je u tablici 4.10.

Tablica 4.10: Uspješnost klasifikacije i optimalni parametri SVM-klasifikatora kod određivanja znanstvenog područja korištenjem značajki sadržaja teksta

Točnost	Preciznost	Odziv	F	jezgra	C	γ_1
0.84	0.90	0.84	0.83	rbf	0.10	0.0001

Kako je vidljivo iz tablice 4.10. i 4.11., uspješnost klasifikacije bila je čak 84% te je model imao problema samo s klasifikacijom korisnika foruma koji pohađaju jedan

Tablica 4.11: Matrice zabune za korisnike foruma, značajke sadržaja teksta

$$\begin{bmatrix} 14 & 15 & 1 & 0 \\ 0 & 30 & 0 & 0 \\ 0 & 1 & 29 & 0 \\ 0 & 2 & 0 & 28 \end{bmatrix}$$

od fakulteta iz područja znanosti i to na način da nije bio siguran treba li ih klasificirati u tehničke ili prirodne znanosti.

Tablica 4.12: Deset najvažnijih bigram i unigram lema za određivanje znanstvenog područja

Prirodne znanosti	Biomedicinske znanosti	Tehničke znanosti	Društveno-humanističke znanosti
pmf	biologija	ects	matura
nastavnicki	raditi	tvz	izborni
program	diplomski	teorija	kolegij
popravni	kemija	program	smjer
smjer	kolegij	upisati	kolega
znanost	dr	digitalan	raditi
biologija	profesor	mat1	matematika
praktikum	matematika	tehnički	semestar
math	ispit	strojarstvo	predmet
fizika	prijemni	elektrotehnika	kolokvij

U tablici 4.12 je prikazano deset najvažnijih lema za svako od znanstvenh područja. Vidljivo je da korisnici foruma raspravljaju o fakultetima koje pohađaju, pa i navode njihova imena. Također učestale su i riječi kao što su matura, upisati, prijemni sto upućuje na to da srednjoškolci kod odabira fakulteta posjećuju ovaj forum i raspituju se o željenom fakultetu.

4.3.2. Određivanje znanstvenog područja korištenjem značajki stila autora

Rezultati klasifikacije autora prema znanstvenom području kojemu pripada fakultet kojeg pohađa autor pri korištenju samo značajki stila autora prikazani su u tablici 4.13.

Klasifikacija korisnika foruma prema znanstvenom području kojemu pripada fakultet kojega pohađa korisnik bila je izrazito neuspješna kada su korištene samo značajke koje zahvaćaju stil autora. Kako je vidljivo iz matrice zabuna (tablica 4.14.), klasifika-

Tablica 4.13: Uspješnost klasifikacije i optimalni parametri SVM-klasifikatora kod određivanja znanstvenog područja korištenjem značajki stila autora

Točnost	Preciznost	Odziv	F	jezgra	C	γ_1
0.35	0.23	0.34	0.26	rbf	10	0.001

tor je gotovo sve primjere svrstao u jednu kategoriju. Teme o kojima su pisali korisnici foruma bile su o fakultetu i o upisima (tablica 4.12) te se stoga može pretpostaviti da je većina korisnika ili na početku studija ili se tek upisuje pa još uvijek nisu stekli stil pisanja i izražavanja svojstven pojedinom znanstvenom području i zanimanju koji se stječe tijekom studija i kasnije na radnom mjestu.

Tablica 4.14: Matrice zabune za korisnike foruma, značajke stila autora

$$\begin{bmatrix} 0 & 22 & 4 & 4 \\ 0 & 26 & 2 & 2 \\ 1 & 19 & 1 & 9 \\ 0 & 13 & 2 & 15 \end{bmatrix}$$

4.3.3. Određivanje znanstvenog područja korištenjem značajki stila autora i sadržaja teksta

Rezultati klasifikacije autora prema znanstvenom području kojemu pripada fakultet koji pohađa autor pri korištenju značajki stila autora i sadržaja teksta prikazani su u tablici 4.12, a u tablici 4.14. je prikazana matrica zabune.

Tablica 4.15: Uspješnost klasifikacije i optimalni parametri SVM klasifikatora kod određivanja znanstvenog područja korištenjem značajki stila autora i sadržaja teksta

Točnost	Preciznost	Odziv	F	jezgra	C	γ_1
0.55	0.76	0.55	0.52	rbf	100	0.01

Model u kojemu su korištena oba skupa značajki; i značajke stila autora i značajke sadržaja teksta, postigao je uspješnost klasifikacije od svega 55% te se iz matrice zabune vidi kako i ovaj model kao i model temeljen samo na značajkama stila autora nije razlikovao korisnike foruma prema znanstvenom području kojemu pripada fakultet koji pohađa korisnik foruma.

Tablica 4.16: Matrice zabune za korisnike foruma, značajke stila autora i sadržaja teksta

$$\begin{bmatrix} 5 & 21 & 0 & 4 \\ 0 & 28 & 0 & 2 \\ 0 & 12 & 12 & 6 \\ 0 & 9 & 0 & 21 \end{bmatrix}$$

4.3.4. Zaključak o određivanju znanstvenog područja

Iako je bilo očekivano kako će najuspješniji model za određivanje znanstvenog područja kojemu pripada fakultet koji pohađa korisnik forum biti model koji koristi i značajke sadržaja teksta i značajke stila autora (Duong et al., 2016), najuspješniji model je koristio samo značajke sadržaja teksta. Već tijekom studija stječe se način izražavanja svojstven određenoj struci, a zajednički interesi koje dijele pripadnici pojedine struke ili studenti sličnih fakulteta trebali bi potaknuti raspravu o temama vezanim uz zajedničke interese. Rezultati ovog rada nisu potvrdili ove pretpostavke, no sumnjamo da je razlog tome u specifičnosti korištenja podataka. U budućim istraživanjima trebalo bi posebnu pažnju posvetiti prikupljanju podataka, npr. koristiti podatke s više foruma namijenjenih pojedinoj struci ili fakultetu.

5. Zaključak

Porastom broja korisnika društvenih mreža raste i zanimanje za sadržaj koji generiraju korisnici društvenih mreža. Jedan od čestih zadataka analiziranja korisnički generiranog sadržaja društvenih mreža kojim se bavi grana računalne znanosti obrada prirodnog jezika je profiliranje autora teksta. Kod zadatka profiliranja autora teksta određuju se pojedine demografske karakteristike autora teksta kako što su spol, dob ili ličnost.
U ovom radu željeli smo odrediti spola autora teksta, ali na način da model za određivanje spola autora teksta bude novisan o društvenoj mreži iz koje potječe tekst, stoga su za izgraju modela korišteni podaci s foruma, a za testiranje podaci s Facebooka i foruma. Izgrađena su tri model koji su redom koristili značajke koje zahvaćaju sadržaj teksta, značajke koje zahvaćaju stil autora teksta, značajke koje zahvaćaju i sadržaj i stil autora teksta. Najuspješnijim se je pokazao model koji je koristio samo značajke koje zahvaćaju stil autora teksta. Postignuta je uspješnost klasifikacije of 96% kad je model testiran na podacima koji potječu s foruma te 88% na podacima s Facebooka. U budućim radovima trebalo bi se usmjeriti na izgradnju modela koji može profilirati autora teksta i u situacijama kad na raspolaganju imamo malu količinu teksta kojom je predstavljen autor. U ovom radu smo se ograničili na situacije u kojima nam je dostupno barem 400 riječi po autoru, a u stvarnim situacijama često raspolažemo s manjim količinama teksta.
Drugi cilj ovog rada bio je odrediti kojem znanstvenom području pripada fakultet kojega pohađa korisnik foruma. Također su izgrađena tri modela koja su koristila značajke koje zahvaćaju sadržaj teksta ili značajke koje zahvaćaju stil autora teksta ili začajke koje zahvaćaju i sadržaj i stil autora teksta. Suprotno očekivanjima (Duong et al., 2016), prema kojima bi najuspješniji model trebao biti onaj koji se temelji i na značajkama sadržaja teksta i na značajkama stila autora, najuspješniji model je koristio samo značajke sadržaja teksta te je postigao uspješnost klasifikacije od 84%. Uvidom u najznačajnije leme svake od klasa pokazalo se je kako korisnici foruma najviše pišu o upisima na fakultet, što ukazuje da korisnici foruma studenti prve godine ili maturanti. Stoga bi u budućim radovima trebalo veću pažnju usmjeriti na prikupljanje podataka

na način da autori tekstova budu ili studenti viših godina kod kojih bi u većoj mjeri došao do izražaja stil izražavanja svojstven pojedinoj struci.

LITERATURA

Ivan Bilan i Desislava Zhekova. Caps: A cross-genre author profiling system. 2016.

Duc Tran Duong, Son Bao Pham, i Hanh Tan. Using content-based features for author profiling of vietnamese forum posts. U *Recent Developments in Intelligent Information and Database Systems*, stranice 287–296. Springer, 2016.

Dominique Estival, Tanja Gaustad, Son Bao Pham, Will Radford, i Ben Hutchinson. Tat: an author profiling tool with application to arabic emails. U *Proceedings of the Australasian Language Technology Workshop*, stranice 21–30, 2007.

Dmitriy Fradkin i Ilya Muchnik. Support vector machines for classification. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 70:13–20, 2006.

Thorsten Joachims. *Learning to classify text using support vector machines: Methods, theory and algorithms*. Kluwer Academic Publishers, 2002.

Bradley C Love. Comparing supervised and unsupervised category learning. *Psychonomic bulletin & review*, 9(4):829–835, 2002.

Tom M Mitchell et al. Machine learning. wcb, 1997.

Dong-Phuong Nguyen, Rilana Gravel, RB Trieschnigg, i Theo Meder. " how old do you think i am?" a study of language and age in twitter. 2013.

Mart Busger op Vollenbroek, Talvany Carlotto, Tim Kreutz, Maria Medvedeva, Chris Pool, Johannes Bjerva, Hessel Haagsma, i Malvina Nissim. Gronup: Groningen user profiling. 2016.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

Francisco Rangel i Paolo Rosso. Use of language and author profiling: Identification of gender and age. *Natural Language Processing and Cognitive Science*, 177, 2013.

Francisco Rangel, Paolo Rosso, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, Walter Daeleman, et al. Overview of the 2nd author profiling task at pan 2014. U *CEUR Workshop Proceedings*, svezak 1180, stranice 898–927. CEUR Workshop Proceedings, 2014.

Francisco Rangel, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, i Benno Stein. Overview of the 4th author profiling task at pan 2016: cross-genre evaluations. *Working Notes Papers of the CLEF*, 2016.

Aleksandr Sboev, Tatiana Litvinova, Dmitry Gudovskikh, Roman Rybka, i Ivan Moloshnikov. Machine learning models of text categorization by author gender using topic-independent features. *Procedia Computer Science*, 101:135–142, 2016.

Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

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

Cathy Zhang i Pengyu Zhang. Predicting gender from blog posts. *University of*, 2010.

G Zhuang, KW Gentle, Patrick Diamond, J Chen, B Rao, L Wang, Kaijun Zhao, Sanghee Han, Yuejiang Shi, YH Ding, et al. Overview of the recent research on the j-text tokamak. *Nuclear Fusion*, 55(10):104003, 2015.

Profiliranje autora na društvenim mrežama pomoću strojnog učenja

Sažetak

Porastom broja korisnika društvenih mreža raste i zanimanje za sadržaj koji generiraju korisnici društvenih mreža. U analiziranje korisnički generiranog sadržaja ubraja se i profiliranje autora teksta, odnosno određivanje demografskh karakteristika, kao što je spol i dob, autora teksta. Ovaj rad bavio se određivanjem spola autora i znanstvenog područja kojem pripada fakultet koji pohađa autor metodama strojnog učenja. Rad se sastoji od cijelog procesa prikupljanja podataka s Facebooka i foruma i preobrade podataka. Potom slijedi razvoj modela određivanja spola i znanstvenog područja kojem pripada fakultet koji pohađa autor teksta. Provedeno je i eksperimentalno vrednovanje modela i statistička analiza dobivenih podataka.

Ključne riječi: obrada prirodnog jezika, profiliranje autora, društvene mreže, hrvatski jezik, strojno učenje

Author Profiling of Social Media Users Using Machine Learning

Abstract

As social networks gain more and more users, interest for the content they're creating is growing also. Analysis of user generated content includes profiling text's authors, and determining demographic characteristics, such as gender and age of the authors. The focus of this thesis was determining author's gender and the scientific area of the author's faculty, using machine learning methods. It consists of the complete process of gathering data off of Facebook and forums and transforming this data. Then there's the development of the model used to determine the gender of the text's author and scientific area of the author's faculty. Experimental model evaluation and statistical analysis of the results was also done.

Keywords: natural language processing, author profiling, social network, Croatian language, machine learning