UTJECAJ I VAŽNOST PSIHOMETRIJSKOG PROFILIRANJA NA WEBU

Bekić, Tomislav

VISOKO UČILIŠTE ALGEBRA

DIPLOMSKI RAD

UTJECAJ I VAŽNOST PSIHOMETRIJSKOG PROFILIRANJA NA WEBU

Tomislav Bekić

Zagreb, veljača 2018

Predgovor

Pisanom izjavom posvećujem ovaj rad svojim roditeljima te im ujedno zahvaljujem što su ulagali u moje obrazovanje kroz cijeli život.

Kroz svoj diplomski studij digitalnog marketinga stekao sam znanja koje bih osobno svrstao na prvo mjesto po važnosti s obzirom na izazove koje pred nas stavlja digitalizacija industrije i razvoj tehnologija. Stoga, bila mi je izrazita čast svoj diplomski studij digitalnog marketinga završiti upravo na Algebri, upoznati visoke stručnjake iz područja podatkovne znanosti iz područja telekomunikacija, IT sigurnosti, stručnjaka za analitiku zahvaljujući Tomislavu Krištofu, bez kojeg ovo sve ne bi bilo moguće.

Posebice se zahvaljujem svom mentoru, Sandru Skansi koji me svojim pristupom i metodama učenja motivirao za odabranu temu u kontekstu iskorištavanja podataka s Web-a te mi ukazao kako strateški razmišljati, kritički prosuditi i primijeniti tehnike strojnog učenja na suvremene poslove probleme.

Sažetak

Hrvatski

Proces ekstrakcije informacija omogućava dobivanje uvida u obrasce i trendove u vezi s podacima. U ovom je radu, u realnom vremenu, provedena analiza teksta modela Myers-Briggs (MBTI) tip ličnosti i analiza objava na društvenim medijima. U nastavku slijedi prikaz i interpretacija konfuzijske matrice prediktivnog modela tipa osobnosti. Rezultati rada ukazuju na mogućnost predviđanja MBTI ličnosti analizom objava na društvenim medijima.

Ključne riječi: Myers-Briggs (MBTI) tip ličnosti, objave na društvenim medijima, tekstualna analiza, prediktivna analiza

English:

The information extraction process provides insight into data patterns and trends. This paper carried out the text analysis of the Myers-Briggs (MBTI) model type of personality and analysis of social media releases in real-time. In continuation below is a description and interpretation of the confusional matrix of the predictive model of personality type. The results indicate the possibility of predicting MBTI personality by analyzing social media releases.

Keywords: Myers-Briggs (MBTI) personality type, social media posts, text analysis, predictive analysis

Sadržaj

1. Uvod 1

2. Prilagodba ili personalizacija 3

3. Tipovi učenja s obzirom na tip podataka 4

4. Automatska personalizacija i rudarenje podacima 6

5. Pristup personalizaciji 7

5.1 Sustav temeljen na pravilima 7

5.1.1 Prednosti sustava temeljenog na pravilima 7

5.1.2 Nedostaci sustava temeljenog na pravilima 8

5.2 Filtriranje na temelju sadržaja 9

5.2.1 Prednosti sustava filtriranja temeljnog na sadržaju 10

5.2.2 Nedostaci sustava filtriranja temeljnog na sadržaju 11

5.3 Značajke sustava kolaborativnog filtriranja 12

6. Brojke na društvenim mrežama 14

7. Razumijevanje ličnosti; Model pet dimenzija ličnosti 17

7.1 Analiza ličnosti kroz MBTI model 18

7.2 Osobine i Myers- Briggs 19

7.3 Tipovi ili osobine 20

8. MBTI 21

8.1 Pouzdanost i valjanost MBTI-a 24

9. Tekstualna analiza 26

10. Metodologija istraživanja 29

10.1 Ciljevi istraživanja 29

10.2 Hipoteze istraživanja 29

10.3 Metoda istraživanja 29

11. Metode pret- procesiranja 31

11.1 Metoda Bag of Words 31

11.2 Tokenizacija 32

11.3 Frekvencija riječi u postu 34

11.4 Sentiment analiza 35
12. Modeliranje podataka 38
12.1 Klasifikacija algoritma 38
12.2 Stablo odlučivanja 39
12.3 Random Forest modeliranje 41
12.4 Interpretacija utreniranog seta podataka 42
13. Predviđanje točnosti modela 45
13.1 Rezultati modela IE 45
13.2 Rezultati modela NS 47
13.3 Rezultati modela PJ 49
13.4 Rezultat modela TF 51
14. Evaluacija klasifikacijskog algoritma Random Forest na modelima 53
15. Zaključak 56
Popis slika 58
Popis grafikona 59
Popis tablica 60
Literatura 61

1. Uvod

Društveni mediji predstavljaju mjesto u kojem se korisnici prikazuju svijetu, otkrivaju mišljenja, osobne detalje i uvide u svoje živote. Neke od tih informacija mogu se iskoristiti za poboljšanje iskustava korisnika na internetu. Ultimativni cilj svakog korisničkog adaptivnog sustava jest omogućavanje korisnicima da dobiju što žele bez da ih se eksplicitno pita [1].

Društveno umrežavanje na webu dramatično je raslo tijekom posljednjeg desetljeća, te je samo u 2017. godini, prema podacima Statiste " 71% korisnika interneta ujedno su bili i korisnici društvenih mreža te se očekuje da će brojke korisnika rasti [2] ". Kroz samooblikovanje društvenih mreža, ažuriranje statusa, objava fotografija i pokazivanje raznih područja interesa, veliki dio osobnosti korisnika izlazi iz profila.

Istraživanja u psihologiji na sustavnom razumijevanju ličnosti provode se desetljećima. Opsežnim radom na razvoju i potvrđivanju široko prihvaćenih modela osobnosti, takva istraživanja pokazala su povezanost između osobnosti i psiholoških poremećaja [3] , zadovoljstva [4] pa čak i u romantičnim odnosima [5]. Ovaj rad nastoji premostiti jaz između društvenih medija i istraživanja osobnosti pomoću informacija koje ljudi otkrivaju. Na temelju najpoznatijeg modela osobnosti, Myers- Briggs indikatora analizirane su posljednje objave na raznim društvenim mrežama imajući uvid u tip osobnosti kojem ti isti pojedinci pripadaju. Naše temeljno istraživanje postavlja pitanje može li se analizom postova na forumu predvidjeti MBTI ličnost. Ukoliko potvrdimo navedeno, postoji prilika za integraciju mnogih rezultata o implikacijama čimbenika osobnosti, ponašanju korisnika (korisničko iskustvo) i korištenju profila društvenih medija kao izvor informacija s ciljem shvaćanja individualaca i unaprjeđenja sustava kroz pristupe i tehnike koji služe za razvoj postojećih sustava personalizacije.

Prikupljeni su podaci o tipu osobnosti po Myers-Briggs modelu te njihove posljednje objave na društvenim mrežama. Podaci su prikupljeni s platforme Kaggle, kvantificirani i analizirani kroz alat koji ujedno i vrši analizu teksta kako bi se dobio skup značajki potrebnih za precizniju analizu. Nadalje, nad podacima su korištene tehnike pret-procesiranja, usredotočene na identifikaciju i ekstrakciju tekstualnih podataka koje se potom koriste za transformaciju u strukturirani set te njegovu daljnju analizu. Alati društvenih medija mogu imati koristi od uvida

u osobnosti korisnika s obzirom da se osobine predviđene primjenom društvenih medija mogu koristiti za izgradnju boljeg i dinamičnijeg sustava preporuke. Također, može se koristiti za sustav preporuka prijatelja društvenih web stranica [6], prijedlozi upoznavanja ljudi [7], zdravstvenu zaštitu [8] i mnogih drugih.

Započinjemo s predstavljanjem pozadine sustava personalizacije i opisom njihovih funkcionalnosti, prednosti i nedostataka, izazova u pristupu te primjena u praksi. Nakon toga predstavljena je penetracija društvenih mreža i prikazana su potencijalna kretanja izražena brojkama. Kako bi smo razumjeli odnose između osobnosti i objava na društvenim medijima, opisana je osobnost i njezine značajke, primijenjena analiza ličnosti kroz model za mjerenje osobnosti (MBTI) te su predstavljeni rezultati provedene tekstualne analize. Opisane su tehnike korištene prilikom klasifikacije seta podataka nad kojima su zatim provedeni prediktivni modeli.

2. Prilagodba ili personalizacija

P. Brusilovsky, A. Kobsa, W.Nejdl u svom radu predstavili su glavnu razliku "automatske personalizacije" koja se najčešće odnosi na "prilagodbu" [8] . U tom pogledu, prilagodba i personalizacija imaju za cilj isporučiti sadržaj prilagođen specifičnom korisniku. Glavna značajka, koja razlikuje prilagodbu i personalizaciju jest tko upravlja stvaranju korisničkog profila kao i prezentacija elemenata sučelja korisnika. U prilagodbi, korisnik ima kontrolu specifikacije preferencija i želja, na temelju kojeg su elementi sučelja kreirani. Neki od primjera prilagođenog weba jest MyYahoo [9] koji omogućuje ručnu konfiguraciju sustava ili usluga prije same provedbe željenog cilja. Za razliku od prilagodbe, automatska personalizacija polazi od činjenice kako nakon korisničkog kreiranja profila, potencijalnog *update-a* automatska obrada vrši se uz minimalnu eksplicitnu kontrolu korisnika. Primjer automatske personalizacije je *Amazon.com* koji ima personaliziranu preporuku poput onih koje koristi za preporučivanje knjiga dok pritom ne zahtijeva od korisnika da unese sve informacije o sebi. Nadalje, glazbena aplikacija *Spotify.com*. [10] zapravo ne koristi jedinstveni revolucionarni model preporuka, umjesto toga, kombinira neke od najboljih strategija koje koriste druge usluge kako bi stvorio jedinstveni moćni sustav. Spotify koristi tri glavna modela preporuka: kolaborativni sustav, NLP (engl. *Natural Language Processing*) modele koji analiziraju tekst te audio modele (analiza audio zapisa).

3. Tipovi učenja s obzirom na tip podataka

Svaki od sljedećih predstavljenih pristupa učenja razlikuje se s obzirom na tip podataka prikupljenih za izradu korisničkog profila i određenog tipa algoritamskog pristupa koji se koristi za pružanje personalizirane ponude ili sadržaja. Generalno, proces personalizacije sastoji se od faze prikupljanja informacija usko vezane uz interese korisnika i faze učenja u kojoj se izrađuju korisnički profili od prikupljenih podataka. Prema B. Mobasher [11] učenje na temelju podataka može se klasificirati kao učenje temeljeno na memoriji (engl. *memory based*) ili na temelju modela (engl. *model based*) ovisno o tome je li učenje učinjeno *online* dok sustav obavlja personalizaciju ili *offline* koristeći utrenirane podatke.

Sustavi učenja koji se temelje na memoriji "pamte" sve podatke i generaliziraju iz njega u vrijeme generiranja preporuka, no osjetljivi su kada je u pitanju njihova skalabilnost. Za razliku od njih, sustav učenja temeljen na modelu, kao što je već spomenuto, provodi fazu učenja izvan mreže te obično imaju veću skalabilnost u odnosu na učenje temeljeno na memoriji koji se provodi kao *online* učenje. Nasuprot tome, ukoliko se prikuplja velika količina podataka, sustavno učenje temeljeno na memoriji predstavlja model koji se bolje prilagođava promjenama korisničkih interesa u usporedbi s tehnikama temeljenim na modelu, u kojima model mora biti prilagođen postupnom povećanju ili izgrađen iz početka. Prednosti i nedostaci ovih sustava doveli su do potrebe za opsežnijim istraživanjem i primjenom u praksi. Predstavljeni su načini rudarenja podacima s ciljem personalizacije, kvalitetnijeg targetiranja te prikaza kako se web profiliranje može koristiti i u koju svrhu. Nadalje, Kohavi [12] je u svom opusu rada predložio temelje za uspješno rudarenje podacima s naglaskom na cilj- personalizaciju:

- Opisni podaci koji omogućuju traženje uzoraka izvan jednostavnih korelacija
- Obujam podataka koji će omogućiti izgradnju i implementaciju pouzdanih modela
- Kontrolirano i pouzdano (automatsko) prikupljanje podataka
- Mogućnost evaluacije rezultata
- Jednostavnost u integraciji s postojećim poslovnim procesima (izgradnja sustava koji mogu učinkovito iskoristiti rudarenje podacima).

Kako bi se procesu personalizacije na temelju rudarenja podataka pristupilo na valjani način, važno je prilagoditi personalizaciju svim fazama u rudarenju podacima, uključujući

prikupljanje podataka, pret-procesiranje, otkrivanje uzoraka i evaluacija u *off-line* modelima i na kraju raspoređivanje znanja u realnom vremenu posredovanja između korisnika i weba.

Prednost i fleksibilnost koja se postiže rudarenjem podataka sagledava se kroz činjenicu kako je personalizacija holistički proces više nego li je individualni algoritam na specifičnom tipu podataka. [12]

U sljedećem djelu predstavljeni su načini pristupa obradi podataka temeljem rudarenja podacima. Prvenstveno, cilj je iskoristiti prikupljene podatke uslijed interakcije korisnika s webom kako bi se predstavio odgovarajući model za personalizaciju korisnika koji služi za daljnje profiliranje, ovisno o industriji. Predstavljeni su načini ekstrakcije informacija na temelju rudarenja podacima, pret-procesiranje i integracija podataka iz više izvora, tehnike otkrivanja zajedničkih uzoraka koji se primjenjuju na ove podatke kako bi smo dobili skupni korisnički model i algoritme preporuka za koje možemo, na temelju otkrivenih znanja sa trenutnim statusom aktivnosti korisnika na web mjestu predstaviti preporuke personaliziranog sadržaja.

4. Automatska personalizacija i rudarenje podacima

Sposobnost sustava da prilagodi, personalizira sadržaj i preporučuje stavke podrazumijeva da mora biti u stanju razlučiti što korisnik zahtijeva na temelju prethodnih ili trenutačnih interakcija s korisnikom. Svrha personalizacije može se promatrati kao problem predviđanja: sustav mora predvidjeti korisničku razinu interesa ili korisnost specifičnih sadržaja, stranica ili stavki te ih rangirati prema predviđenim vrijednostima. Često je sustav personalizacije uokviren, u smislu preporuka zadataka u kojoj sustav vrši preporuku s najviše predviđenim vrijednostima interesa korisnika. Općenito, sustav personalizacije može se promatrati kao mapiranje korisnika i stavki na temelju "vrijednosti interesa". Pogled funkcije personalizacije kao zadatka predviđanja proizlazi iz činjenice da ovo mapiranje nije, generalno definirano na cijeloj domeni korisničkih stavki i stoga zahtijeva da sustav procjenjuje interese za samo neke domene parova. Automatski personalizacijski sustavi razlikuju se u vrsti podataka i načinu korištenja za stvaranje korisničkog profila te u vrsti algoritamskih pristupa koji se koriste za predviđanje. [12]

5. Pristup personalizaciji

Polazeći s arhitektonskim i algoritamskim razmatranjem, Mobasher [11] je sustav personalizacije kategorizirao kroz pristupe i tehnike koji služe za razvoj postojećih sustava personalizacije u tri grupe: sustav temeljen na pravilima (engl. *Rule- based system*), sustav filtriranja na temelju sadržaja (engl. *Content based filtering system*) te sustav kolaborativnog filtriranja (engl. *Collaborative filtering system*). Kako bi se upoznali s funkcionalnostima sustava, u nastavku su detaljno opisani navedeni sustavi.

5.1 Sustav temeljen na pravilima

Prema Mobasheru, sustav temeljen na pravilima (eng. *Memory-based system*) oslanja se na ručno definiranim ili automatskim pravilima koji imaju za cilj preporučiti predmete krajnjim korisnicima. Mnoštvo postojećih web stranica koje u svom opusu usluga sadrže web trgovinu (engl. *e-commerce*) primjenjuju sustav temeljen na pravilima. Takav sustav dozvoljava administratoru specificiranje pravila, najčešće prema obilježjima potrebnih za evaluaciju različitih segmenata poput: demografije, psihografske segmentacije, ili nekim drugim obilježjima specifičnih za skupinu. U nekim slučajevima, pravila su ovisna o domeni u mjeri u kojoj se reflektiraju na ciljeve web stranice. Svrha pravila je utjecati na krajnjeg korisnika kroz predstavljeni sadržaj prilikom kojeg je za sustav potrebno da zadovoljava jedan ili više definiranih pravila. Kao i većina sustava temeljenih na pravilima, ovaj tip personalizacije se oslanja na znanje stručnjaka i dizajnera koji konstruiraju pravila baziranih na specifičnim karakteristikama domene kao i na temelju provedenih marketing istraživanja. Profil korisnika, generalno je specificiran kroz interakciju s korisnicima. Istraživanja su pokazala kako su tehnike strojnog učenja (engl. *Machine Learning*) korisne za klasificiranje korisnika u nekoliko kategorija, kao prema demografskim atributima prema kojima se automatski vrše pravila koja se koriste za personalizaciju.

5.1.1 Prednosti sustava temeljenog na pravilima

Sustavi temeljeni na pravilima imaju veliki stupanj korištenosti u praksi, pogotovo u e trgovini, stoga će u ovom djelu, prema tehnološkim ekspertima biti predstavljene njihove prednosti i nedostatke u korištenju s ciljem shvaćanja funkcioniranja samog sustava personalizacije [14].

Prednosti:

1. Dostupnost – sustav je korisniku uvijek dostupan, bez obzira na pravila koja se postavljaju.
2. Troškovna učinkovitost – sustav je troškovno učinkovit i precizan s obzirom na njegov krajnji rezultat
3. Brzina – optimizacija sustava je brza ukoliko poznajete dijelove sustava, pružanje *outputa* odvija se u samo nekoliko sekundi
4. Točnost i mala stopa pogreške – stopa pogreške je vrlo mala zbog unaprijed definiranih pravila
5. Smanjen rizik – rizik je smanjen na minimum s obzirom na preciznost sustava
6. Postojanost odgovora – *output* koji je generirao sustav ovisi o pravilima, stoga je krajnji cilj prema korisniku stabilan, što ukazuje da sustav nije neizvjestan u svom cilju.
7. Modularnost sustava – modularnost i arhitektura sustava temeljenog na pravilima uvelike pomaže tehničkom timu prilikom održavanja, smanjujući vrijeme i uloženi trud.

5.1.2 Nedostaci sustava temeljenog na pravilima

Nedostaci sustava temeljnog na pravila ukazuju na problematiku:

1. Neophodan ručni rad – sustav zahtijeva široko poznavanje domene
2. Vremenski zahtjevno – sustav je vrlo složen zbog pravila koja se definiraju, što je vremenski ograničeno i zahtjevno
3. Manji kapacitet učenja – sustav će generirati rezultate prema pravilima, stoga je kapacitet učenja sustava manji
4. Kompleksnost domene – ukoliko je aplikacija koju želite kreirati kompleksna, gradnja sustava temeljenog na pravilima (definiranje pravila) zahtijeva puno uloženog vremena i analize.

Predstavljene prednosti i nedostaci sustava temeljenih na pravilima donose izazove u samom pristupu, poput:

6. Bihevioralni pristup- izazov prilikom prepoznavanja uzorka i oponašanja čovjeka
7. Kreiranje i dizajn arhitekture sustava zahtijeva kritičko razmišljanje, prema kojem se može povući paralela s bihevioralnim pristupom
8. Sustav zahtijeva ekspertno poznavanje specifične domene i NLP-a (engl. *Natural language procesing*) koji kreira pravila prema uputama stručnjaka.
9. NLP je izazovna domena s obzirom na iznimke koja pokrivaju to područje, pogotovo u slučaju velike količine pravila.
10. Vremenska komponenta kao ključni izazov NLP-a
11. Kompleksnost prepoznavanja uzorka predstavlja izazov u pristupu sustava temeljenog na pravilima.

Tehnološki napredak i digitalna transformacija utjecala su na razvoj i poboljšanje već postojećih sustava, kao i predstavljanje novih, bržih, "pametnijih" u područjima kao što su medicina, proizvodnja, transport, zrakoplovstvo, računovodstvo itd… Sustavi temeljeni na pravilima dokazali su uspješnost na svim dobro uspostavljenim i novim područjima. Pravila i ekspertni sustavi upravljaju znanjima, činjenicama, empirijskim teoremima i sl. Iako postoje ograničenja, sustavi omogućuju preciznost prilikom predviđanja rezultata i odlučivanja.

U nastavku je opisan jedan od najčešće korištenih sustava profiliranja korisnika, koji kao i svi navedeni koriste se u personalizaciji usluga/ proizvoda/ sadržaja krajnjem korisniku.

5.2 Filtriranje na temelju sadržaja

Filtriranje na temelju sadržaja (engl. *Content based filtering system*) prilikom profiliranja, uzima u obzir karakteristike korisnika, poput onih što korisnik prati ili označi sa "sviđa mi se", te na temelju tih karakteristika preporučuje slične stvari. Salton i McGill [15] opisali su ih kao opisi sadržaja stavki koje predstavljaju skup značajki ili atributa koji karakteriziraju tu stavku. Zadatak preporuka u takvim sustavima obično uključuje usporedbu izdvojenih značajki ili neocijenjenih stavki s opisima sadržaja u korisničkom profilu. Upravo na temelju takve usporedbe, korisniku se preporučuju stavke koje su dovoljno slične njegovom profilu.

U većini sustava temeljenih na filtriranju sadržaja, posebice onima koji se koriste na webu ili aplikacijama koje u svom portfelju imaju web- trgovinu, opisi sadržaja su tekstualne značajke unutar web stranica ili u opisu proizvoda. Sustavi, kao takvi, oslanjaju se na poznate tehnike modeliranja s korijenjem u traženju informacija i filtriranju istih. Oba korisnička profila, kao i

stavke vrednovani su kao vektori s pripadajućim vrijednostima (temelj je model TF-IDF. [16]). Predviđanja interesa korisnika u stavkama izvode se na temelju izračuna vektorskih sličnosti (engl. *Cosine similarity measure*) ili na temelju pristupa Bayesian klasifikacije koja prikazuje zajedničku raspodjelu vjerojatnosti za definirani skup varijabli. U ovakvim sustavima, profili su individualni po svojoj prirodi korisnika, izgrađeni od značajki povezanih s stavkama koje je aktivni korisnik vidio ili ocijenio. Primjer ove vrste profiliranja je tvrtka Amazon, koja je u svom portfelju usluga u Sjedinjenim Američkim Državama ima preko 232 milijuna stavki [17]. Amazon je osmislio svoj sustav preporuka (personalizacije) na način u kojem svaki korisnik vidi personaliziranu verziju stranice.

5.2.1 Prednosti sustava filtriranja temeljnog na sadržaju

Kao što je već spomenuto, ključni dio sustava za preporuku na temelju sadržaja jest korisnički profil. Korisnički profili sastoje se od objekata s atributima na temelju kojih je korisnik ostvario interakciju. Atributi koji se pojavljuju uz objekt imaju veću vrijednost (engl. *weight*) od onih koji se pojavljuju rjeđe. Također, napredni algoritmi ne uzimaju u obzir samo stavke koje je korisnik gledao, slušao, čitao, već i one koje je pregledao u prošlosti, kao i preporuke koje su mu već ponuđene. U određivanju važnosti različitih stavki, povratna informacija korisnika ključna je te predstavlja stavku s najvećom vrijednošću sustavu preporuka temeljenog na sadržaju. Prema težinskoj vrijednosti atributa i povijesti pretraživanja, sustav proizvodi jedinstveni model prema sklonostima svakog korisnika korištenjem tehnika strojnog učenja. Atributi koje korisnici definiraju svojom pretragom (engl. *like, dislike*), sustav ponderira prema važnosti. Modeli se zatim uspoređuju sa svim objektima u bazi podataka te im se dodjeljuju bodovi temeljeni na sličnostima s korisničkim profilom. No, kao i svaki sustav, i ovaj model pružanja personalizacije, odnosno preporuke krajnjem korisniku ima određene prednosti u odnosu na sustav temeljen na pravilima, kao i nedostatke u odnosu na kolaborativni sustav, koje će detaljnije biti opisane u nastavku [18].

Prednosti sustava temeljenog na sadržaju:

1. Važnost rezultata- budući da su preporuke temeljene na sadržaju i oslanjaju se na karakteristike objekata, oni ukazuju na relevantnost kod interesa korisnika. Ovo je uvelike važno u organizacijama koje posjeduju veliki broj tipova sadržaja, poput onih koji nude medijske usluge ili pretplatu na sadržaj.

2. Transparentnost preporuka – u procesu u kojem se generiraju, preporuke utječu na povećanje povjerenja od strane korisnika u transparentnost preporuka. U procesu sa kolaborativnim filtriranjem korisnici ne razumiju pojavljivanje određenih stavki, no one su temeljene na sustavnom algoritmu.
3. Brži početak – filtriranje temeljeno na sadržaju izbjegava startne probleme koje često uzrokuju tehnike filtriranja. Iako sustav zahtijeva početne inpute, za razliku od robusnih modela sustava, kvaliteta ranih preporuka puno je veća.
4. Tehnički laka implementacija – podatkovna znanost iza sustava temeljenog na sadržaju relativno je jednostavna.

5.2.2 Nedostaci sustava filtriranja temeljnog na sadržaju

Nedostaci sustava temeljenog na sadržaju koji ujedno predstavljaju izazove:

1. Nedostatak noviteta i raznolikosti – relevantnost u sustavima personalizacije od velikog je značaja, no ne i ključan aspekt. Primjerice, ukoliko korisnik voli i pozitivno ocjeni film *Star Wars,* zasigurno ne treba sustav preporuke da pogleda *The Empire Strikes Back.* Ono što je važno za sustav jest da pronađe filmove koji su raznoliki (širok izbor njihovih interesa) i neočekivani.
2. Skalabilnost kao izazov –u potrazi za shvaćanjima ključni zahtjev koji predstavlja izazov sustava temeljenog na sadržaju je ekspertno znanje stručnjaka koje mnoge tvrtke teško pronalaze. Također, ručno označavanje atributa mora se nastaviti s novim dodanim sadržajem.
3. Atributi mogu biti netočni ili ne dosljedno primijenjeni – s obzirom na količinu (tisuću ili milijun) stavki, izazov je sustava da dosljedno ili precizno vrši preporuku krajnjem korisniku.

Primarni nedostatak sustava za filtriranje na temelju sadržaja jest njegova tendencija prevelike specijalizacije odabira stavki budući da se profili korisnika temelje isključivo na prethodnoj ocjeni korisnika stavke. Daljnja istraživanja [19] pokazala su da korisnici online preporuke smatraju više korisnije ukoliko je preporuka ponudi neočekivane stavke, što sugerira da sama upotreba sličnih sadržaja može rezultirati nedostatkom važnih "pragmatičnih" odnosa među web objektima kao što su njihova zajednička ili komplementarna korisnost u kontekstu određenog zadatka. Nadalje, filtriranje na temelju sadržaja zahtijeva da se stavke učinkovito

prikažu pomoću ekstrahiranih tekstualnih značajki koje nisu uvijek praktične s obzirom na heterogenost web podataka.

5.3 Značajke sustava kolaborativnog filtriranja

Prilikom dizajniranja sustava preporuke sustavi poput onih temeljenih na sadržaju ukazuju na problematiku kada je u pitanju sadržaj koji je preveliki ili raznolik da bi ručno primijenili atribute. Upravo iz tog razloga, kolaborativno filtriranje (engl. *Collaborative filtering*) rješava navedenu problematiku. Sustav kolaborativnog filtriranja ima daleko veću primjenu u društvenim medijima, maloprodaji i *streaming*[1] uslugama. U ovom djelu biti će predstavljeno kako sustav funkcionira, gdje se upotrebljava te koje su mu prednosti i nedostaci u odnosu na izazove budućnosti [20].

Koncept kolaborativnog filtriranja temelji se na ideji da ljudi koji dijele zanimanje za određene stvari dijele i sličan ukus prema drugim sličnim stvarima. Primjer kolaborativnog filtriranja može se lako uočiti u praksi. Kada kao korisnik posjetite određenu mrežu, možete uočiti: *"Kupci koji su kupili ovu stvar također su kupili i..."* kao i *"Korisnicima poput Vas također se svidjelo.."*. Za razliku od sustava filtriranja temeljenog na sadržaju koji filtriranje temelji na atributima određenih objekata, kolaborativni sustav filtriranja oslanja se na ponašanje korisnika. Cjelokupni sustav ima razvijenije performanse u odnosu na sustav filtriranja temeljen na sadržaju:

1) Veća korisnost temeljena na velikoj korisničkoj bazi – velika upotrebljivost usluga koje koriste kolaborativno filtriranje utječe na broj preporuka putem kojih sustav dobiva velik broj, relevantnih podataka, oslanjajući se, bez, dodatnih razvojnih aktivnosti na ekspertizu stručnjaka.
2) Fleksibilnost u različitim domenama – kolaborativni pristup personalizaciji pogodan je za vrlo različite skupove predmeta. Dok se filtriranje temeljeno na sadržaju oslanja na meta podatke, kolaborativno filtriranje temelji se na trenutnoj/ stvarnoj aktivnosti korisnika. Primjerice, kolaborativnim filtriranjem odvija se povezivanje naizgled različitih stavki, poput primjerice, van brodski motor i štap za ribolov, ukazuju sustavu na poveznicu i relevantnost za skup korisnika koji vole ribu.

[1] Streaming (Streaming media ili Internet Streaming); vrsta multimedijskog sadržaja, tehnologija u kojoj korisnik istovremeno prima podatke i reproducira ih, označava način dostavljanja multimedijskog sadržaja.

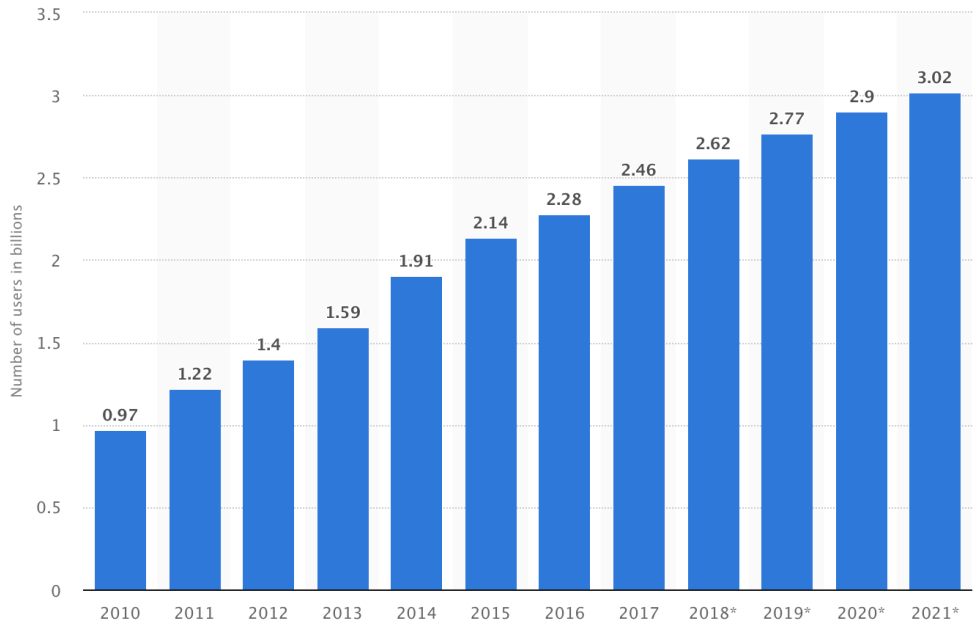
3) Proizvodnja nasumičnih preporuka – kod preporuka, točnost podataka nije uvijek najbitniji element. U tom kontekstu, polazeći od činjenice točnosti podataka, sustav filtriranja temeljen na sadržaju obične prikazuje stavke korisniku sličnim onome koje su pregledane od strane korisnika. Nasuprot tome, kolaborativni sustav razumije da većina korisnika ima interese koji obuhvaćaju različite podskupine, što u teoriji ukazuje na raznovrsnost preporuka.
4) Zahvaćanje većih nijansi oko stavaka – oslanjajući se na razumijevanje sustava kada je u pitanju ponašanje korisnika, sustav predlaže stavke koje imaju potencijalno veći afinitet jedan s drugim nego kako to radi sustav temeljen na sadržaju koji predlaže prema strogoj usporedbi atributa.

Ukoliko pobliže pristupimo kolaborativnom filtriranju, filtriranje možemo podijeliti na temelju stavki i na temelju korisnika. Izvorno, kolaborativno filtriranje razvila je tvrtka Amazon koja predikcije temelji na odnosu različitih stavki artikala koji se kupuju zajedno. Često se primjerice artikli poput maslac od kikirikija i žele nađu zajedno u košarici kupca, stoga sustav filtrira *inpute* koji obavještavaju zajedničku pojavnost artikla. Stoga, ukoliko algoritam prepozna povezanost između artikla, predložiti će artikle od tih sastojaka ili artikle koji idu uz navedene. Nasuprot tome, filtriranje korisničkih stavki zauzima drugačiji pristup. Umjesto da sustav izračunava udaljenost između stavki, izračunava se udaljenost između korisnika temeljem njihovih ocjena (s definiram mjernim algoritmom). Na taj način, prilikom donošenja preporuka korisnicima, sustav pregledava korisnike, uspoređuje ih i povezuju u cjelinu. Pregledavanjem najbližih korisnika predlaže stavke koje su se svidjele drugim, sličnim korisnicima te ih predlaže korisniku koji još nije imao interakciju sa stavkom. Ukoliko ste kao korisnik Facebooka gledali određeni broj videozapisa, Facebook može pregledati druge korisnike kojima su se svidjeli ti videozapisi i preporučiti ih ukoliko ih niste još vidjeli. Iako u navedenim primjerima sustav ne razlučuje zašto se pojedina stavka pojavljuje s drugom, koji je njihov odnos, razumije da ih treba povezati u cjelinu. U određenoj branši, poput *on-line* trgovine ili društvenih medija, upravo ova vrsta filtriranja može se sagledati kao odlika sustava više nego kao nedostatak, pogotovo kada se radi o stavkama koje su heterogene naravi.

6. Brojke na društvenim mrežama

Penetracija društvenih mreža u svijetu sve je veća. Prema posljednjem istraživanju statističkog portala Statiste [2], u 2017. godini: " 71 % korisnika interneta bilo je na društvenim mrežama, a predviđanja ukazuju na daljnji rast u narednim godinama." Društveno umrežavanje jedna je od najpopularnijih aktivnosti na internetu s visokim angažmanom korisnika i daljnjim napredovanjem mobilnih mogućnosti. Ukoliko sagledamo rasprostranjenost korištenja društvenih mreža, Sjeverna Amerika nalazi se na prvom mjestu među regijama u kojima su društvene mreže najpopularnije sa stopom penetracije društvenih mreža od 66%. Godine 2016. više od 81% stanovništva Sjedinjenih Američkih Država imalo je profil na društvenim mrežama. Ukoliko sagledamo koliko su vremenski utrošili stanovnici SAD-a na društvenim mrežama, brojka se od drugog tromjesečja 2016. godine penje na 215 minuta tjedno putem pametnih uređenja, 61minuta tjedno putem računala te 47 minuta tjedno putem tabletnih uređaja.

Također, Statista je na temelju prošle i trenutne upotrebe društvenih mreža predstavila broj korisnika društvenih medija u svijetu od 2010. do 2021. (u milijardama), što je prikazano u grafikonu ispod.



Grafikon 1. Kretanje korisnika društvenih medija od 2010. Do 2021.
Izvor: Statista; https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/, 2018.

Povećana upotreba pametnih telefona i uređaja diljem svijeta otvorila je nove mogućnosti povezivanja sa društvenim mrežama s povećanim značajnostima usluga lokacije poput Foursquare ili Google Now. S više od 1,86 milijardi mjesečnih korisnika, Facebook je vodeća društvena mreža za komunikaciju i umrežavanje. Društveni mediji predstavljaju mjesto u kojem se korisnici otkrivaju digitalnom svijetu, otkrivajući svoje osobne detalje i uvide u svoje živote.

OSN (engl. *Online Social Network sites*) su postali vrlo brzo integrirani u središte suvremenih društvenih interakcija i široko se koriste kao primarni medij za komunikaciju i umrežavanje [21]. Unatoč sve većoj integraciji aktivnosti OSN-a u svakodnevni život, postavlja se pitanje - prenose li korisnikova aktivnost na društvenim medijima točne impresije vlasnika profila?

Velika pretpostavka podržana analizom sadržaja sugerira da se OSN profili rabe za stvaranje i komuniciranje idealiziranih ličnosti sebe [22]. Prema ovom idealiziranom virtualnom identitetu, vlasnici hipoteze pokazuju idealizirane karakteristike koje ne odražavaju njihove stvarne osobnosti. Stoga, pojavljivanja osobnosti koja se temelje na OSN profilima trebala bi odražavati vlasničke preglede idealnih vlasničkih prikaza, a ne ono što vlasnici zapravo vole.

Nadalje, OSN-ovi mogu predstavljati proširenu društveni kontekst u kojem će se izraziti stvarne karakteristike svojih osobnosti, potičući točnu interpersonalnu percepciju. OSN integriraju različite izvore osobnih podataka koji odražavaju one koji se nalaze u osobnim okruženjima, privatnim mislima, slikama lica i društvenom ponašanju, od kojih svi znaju sadržavati valjane informacije o osobnosti [23]. Štoviše, stvaranje idealiziranih identiteta trebalo bi biti teško ostvariti iz više razloga. Razlozi tome jesu da a) OSN profili uključuju informacije o reputaciji koje je teško kontrolirati kao npr. zidni postovi, b) prijatelji pružaju odgovornost i suptilnu povratnu informaciju o profilima jedni drugima. Prema tome, proširena hipoteza o stvarnom životu predviđa da ljudi koriste OSN-ove da komuniciraju svoju stvarnu osobnost. No, kako bi se provelo istraživanje može li se na temelju interakcije korisnika na društvenim mrežama ostvariti predikcija o osobnosti. Osobnost se odnosi na individualne razlike u karakterističnim obrascima razmišljanja, osjećaja i ponašanja. Proučavanje osobnosti usredotočeno je na dva široka područja: jedan je razumijevanje individualnih razlika u osobitim karakteristikama osobnosti, kao što su društvenost ili razdražljivost. Drugi se odnosi na razumijevanje kako se različiti dijelovi osobe dolaze zajedno kao cjelina [24]. U prošlosti, aspekti osobnosti proučavani su iz mnogih kutova, bilo analizom interpersonalnih odnosa ili analizom društvenih mreža ili kroz istraživanja u neuroznanosti koji otkrivaju biološku podlogu osobina ličnosti. Naime, pojavnost društvenim medija utjecala je na razvoj novih pristupa s ciljem otkrivanja

obrazaca, poboljšanja u predikciji ljudskog ponašanja i razvojem sustava personalizacije. Mnoga istraživanja imale su za cilj predvidjeti osobnost analizom obrasca u ponašanju, slika, pa čak i rukopisima. Nadalje, pojavljivanjem društvenih medija i sve više povezana *online* zajednica omogućuju analizu personaliziranih tekstualnih podataka sve više dostupnim. Osobnost kao jedna od najutjecajnijih istraživanja u psihologiji predviđa mnoge posljedične ishode poput tjelesnog i mentalnog zdravlja, kvalitete međuljudskih odnosa, prilagodbe i zadovoljstva u karijeri, općeg zadovoljstva i performanse na radnom mjestu [25]. Postoji mnogo načina mjerenja ličnosti, no najveći broj autora (kao skup dimenzija najvišeg reda kojima je moguće opisati ličnost svakog pojedinca) prihvaća petofaktorski model.

7. Razumijevanje ličnosti; Model pet dimenzija ličnosti

Opće prihvaćene osobine ličnosti, prema FFM modelu (engl. *Five – factor model*) kojeg su razvili mnogi istraživači, uključujući psihologe Robert McCrae i Paul Costa. U skladu s teorijama ličnosti koja analizira razlike u smislu individualnih stavova i obrasca ponašanja, model pet dimenzija usredotočuje se na pet mjera ličnosti.

Svaka je od pet širih dimenzija pet faktorskog modela sastavljena od više komponenti nižeg reda koje predstavljaju osnovna obilježja pojedinaca [26].

Model pet dimenzija ličnosti uključuje:

- Otvorenost prema iskustvu – spremnost da isprobate nove aktivnosti, iskusite različite načine života i prihvatite nekonvencionalne ideje. Mješavina različitih komponenti ličnosti među kojima se nalazi intelekt u užem smislu (npr. inteligencija, pronicljivost, kreativnost), otvorenost za iskustva (npr. radoznalost, maštovitost, liberalnost), te neki aspekti kulture, osobnih stavova, sklonosti i orijentacija (npr. umjetnički interesi, nekonformizam, progresivne i nekonvencionalne vrijednosti, potreba za raznolikošću iskustava i sl.).
- Savjesnost – tendencija da budemo svjesni vlastitog ponašanja i posljedica djelovanja. Ljudi s izraženom savjesnošću vole imaju tendenciju organizirati se, biti točni i usmjereni na ciljeve u svom djelovanju i ponašanju
- Ekstraverzija – ekstroverti su živahni, komunikativni i društveni. Otvoreni su prema upoznavanju novih ljudi. Karakterizira ju pozitivna emocionalnost, kao komponente sadrži socijalnost, poduzetnost, ambicioznost i asertivnost
- Ugodnost – konceptualizira se kao varijabla koja sumira tendencije ponašanja kao što su ljubaznost, kooperativnost i sklonost pomaganju
- Neuroticizam – kao negativna emocionalnost sastavljena od emocionalne reaktivnosti, iritabilnosti i nesigurnosti.

Predstavljen model uvod je u razumijevanje ličnosti u smislu specifičnih stavova i vrsta ponašanja. Razmotrili smo i predstavili teoriju i model koji je oblikovalo naše razumijevanje ljudske osobnosti. Poznato je da su komponente predstavljenog modela poput ekstraverzije, savjesti i neurocitizma relativno konzistentne tijekom čitavog života. Međutim, načini na koje

se naša ponašanja izražavaju riječima i djelovanjem, nisu uvijek određena temeljem ličnosti i impulsima samim. Ljudi učinkovito nauče modelirati svoje ponašanje kako bi se uskladili s vanjskim okolinom i okolnostima [27].

Sveukupno gledajući, osobine ličnosti vrlo su utjecajne na naše ponašanje, ali čitanje ponašanja drugih osoba nije dovoljno za točna predviđanja njihovih osobnosti. Budući da se svijet svakodnevno sve više oslanja na pisanu komunikaciju (komunikaciju temeljenu na tekstu) javlja se potreba za razvojem modela koji će automatski i precizno pročitati napisani tekst i profilirati osobu. Također, studije neuroznanosti otkrile su kako su za osobine ličnosti ekstraverzija i neuroticizam povezane s pisanom komunikacijom [28]. S obzirom na međusobnu povezanost neurona možemo vjerovati kako bi se temeljni obrasci mogli izdvojiti iz pisanog teksta i doprinijeti novim istraživanjima.

7.1 Analiza ličnosti kroz MBTI model

Sljedeći korak u analizi ličnosti jest predstaviti jedan od najpoznatijih modela za mjerenje osobnosti, MBTI (engl. *Myers-Briggs Type Indicator*), opisati funkcionalnosti modela, pretpostaviti da je stil pisanja pojedinaca u velikoj mjeri povezan s njihovim osobinama ličnosti te predstavljaju model učenja za predviđanje Myers Briggs Personality Type kroz tekstualne objave na društvenim medijima. Prethodni modeli predviđanja ličnosti fokusirali su se na primjenu tehnika strojnog učenja i neuronskih mreža kako bi predvidjele značajke ličnosti *Big Five* modela (otvorenost prema iskustvu, ekstraverzija, neuroticizam, savjesnost i ugodnost) iz objava na društvenim mrežama [29]. Nadalje, neke od studija također su na temelju tehnika računalnog vida pokušale predvidjeti osobine ličnosti iz profila rukopisa i ostalih podataka baziranih na slikama [30]. Koristeći tekstualne i fotografske podatke iz velike zbirke društvenih mreža bilo bi korisno u predviđanju prijateljstva (društvenih krugova, društvenih grupa i sličnim). Uzmimo za primjer svjetsku organizaciju Google, konkretnije, vaš korisnički račun, Google+ koji ima mogućnost grupiranja ljudi u krugove, dok na drugoj globalnoj kompaniji Facebook i Twitteru moguće je raditi liste, tj. grupe. Svaki krug se sastoji od podskupa određenog korisnikovim prijateljima. Takvi krugovi mogu biti razdvojeni, preklapati se ili se hijerarhijski postavljati [31].

Upravo na temelju raznolikosti dostupnih podataka brojnih studija, tekstualna analiza objava npr. Tweetova ili facebook objava ima prednost u velikoj varijabilnosti *inputa* i proznom stilu pisanja te omogućuje prediktivnu analizu u pomnom definiranju ekstroverata i introverata

prema Myers Briggs Personality Type (MBTI) podataka o posljednjim objavama nakon rješavanja najpoznatijeg svjetski priznatog testa osobnosti. Zahtjevniji zadatak predstavlja podizanje nijansi u individualnosti u formalnim i standardiziranim stilovima pisanja poput elektroničke pošte, eseja ili obrasca za posao. Studije koje su se usredotočile na osobine Big-Five modela imaju tendenciju predstaviti osobnost individualaca na temelju slike, dok je MBTI povezan s prototipovima koji su lakše usporedivi i imaju funkcionalne primjene u predviđanju kompatibilnosti zanimanja i odnosima, kao i u ponašanjima pojedinaca.

7.2 Osobine i Myers- Briggs

Pokazatelji tipova Myers-Briggs Type Indicator (MBTI) temelji se na psihološkoj teoriji tipova koje je predstavio Carl Jungs polazeći od toga da slučajne varijance u ponašanju ovise o načinu na koji ljudi donose procjenu i percipiraju. Postoji 16 vrsta osobnosti u četiri dimenzije: Ekstraverzija (E) - Introverzija (I), mjera u kojoj pojedinac preferira svoj vanjski i unutarnji svijet. Senzacija (S) – Intuicija (N), mjera u kojoj se ukazuju razlike onih koji obrađuju informacije kroz pet osjetila u odnosu na pojavljivanja kroz obrasce. Mišljenje (T) – Osjećanje (F), kao mjera kognitivnih procesa nasuprot mjeri subjektivnog prosuđivanja i procjenjivanja kojom određujemo koliko je nešto važno za nas. Konačno, Prosuđivanje (J) – Percipiranje (P), mjera koja ukazuje na razlike onih koji preferiraju planirati život i držati se životnog reda u odnosu na ljude koji su fleksibilni i spontani. Osobnost je samo jedna od mnogih čimbenika koji vode naše ponašanje, međutim, na naše djelovanje utječe okruženje, iskustva i individualni ciljevi. Na stranici testa osobnosti MBTI, test omogućava ljudima detaljan uvid i opis tipa kojem pripadaju te kako se prema dobivenom tipu ponašaju. Značajne razlike mogu postojati čak i kod ljudi koji dijele tip osobnosti. Cilj MBTI testa je potaknuti osobni rast pojedinca na način da bolje razumije sebe i odnose s okolinom [32].

Zahvaljujući jednostavnosti u korištenju modela, imenovanja četiri različita slova (temeljna opisna slova vaših osobnosti) prihvatile su mnoge različite teorije i pristupi tijekom posljednjih nekoliko desetljeća, uključujući Socionics (modifikacija modela ličnosti Carl Junga), Keirsey Temperament Sorter (KTS - upitnik vrlo usko povezan s pokazateljima MBTI s značajnim teorijskim i praktičnim razlikama) te Stilovi interakcije Linde Berens (skupina 16 vrsta MBTI instrumenta psihometrije i Junlove psihologije) i mnogi drugi. Iako svaka teorija ima identične ili vrlo slične akronime, njihova se značajnost i definicija ne preklapaju uvijek. Svaka teorija ih definira na svoj način i u potpunosti je moguće da, ukoliko se upoznate s pet ljudi koji tvrde da

imaju isti tip, uzmimo za primjer tip osobnosti INFJ, njihove definicije će se razlikovati s obzirom koji test su rješavali.

7.3 Tipovi ili osobine

Bez obzira na strukturu, svaka teorija temelji se na opisivanju ili karakterizaciji tipova čiji rezultati leže blizu linije razdvajanja. Prema MBTI-u drugačiji način gledišta osobnosti jest definiran putem prizme *karakteristika*, radije nego kroz model na *bazi tipa.* Umjesto stvaranja proizvoljnog broja kategorija i pokušaja uklapanja ljudi unutar njih, model temeljen na osobinama procjenjuje stupanj do kojeg ljudi pokazuju određene osobine.

U primjeru pojma ambivalentnosti objasnit ćemo kako se razlikuje MBTI teorija u odnosu na ostale. Ambivalencija je pojam u psihologiji koji označava istovremeno postojanje sasvim suprotnih osjećaja ili stavova u odnosu na neku osobu, ideju, predmet ili situaciju kod jedne osobe [33]. Ambivalentnost predstavlja vrstu unutarnjeg sukoba, u kojem se osoba nalazi u sredini na skali između Introverzije - Ekstraverzije. Teorije temeljene na osobnosti jednostavno bi rekle za ambivalentnu osobu da je umjereno introvertna ili umjereno ekstrovertna te ih ostavile definirano tako, bez dodjele tipa osobnosti.

8. MBTI

Pokazatelj tipa Myers-Briggs je introspektivni upitnik za samoprocjenu s ciljem ukazivanja na različite psihološke preferencije u načinu na koji ljudi vide svijet oko sebe i donose odluke. Također, polazište teorije Myers – Briggs pokazatelja jest da su neke od naizgled nasumičnih razlika u ponašanju pojedinaca zapravo logične i strukturirane te se ujedno mogu sistematizirati s obzirom na nekoliko temeljnih kategorija načina na koji ljudi percipiraju i prosuđuju svijet. Myers-Briggs tipologija preispituje preferencije pojedinaca unutar četiri kategorije kognitivnog funkcioniranja.

Identifikacija i opis 16 različitih tipova osobnosti proizlaze iz interakcija među preferencijama. 16 tipova obično se spominju skraćenicom od četiri slova - početnim slovima svake od četiri vrste preferencija (osim u slučaju intuicije, koja koristi kraticu "N" kako bi se razlikovala od introverzije). Na primjer:

ESTJ: Ekstraverzija (E), Senzacija (S), Mišljenje (T), Prosuđivanje (J)

INFP: Introverzija (I), Intuicija (N), Osjećanje (F), Percipiranje (P)

Ove kratice primjenjuju se na svih 16 vrsta, a u nastavku su predstavljene i opisane prema Myers Briggs-u u 4 kategorije po kojem se nakon rješavanja testa definira tip ličnosti [34].

1. Ekstraverzija (E) – Introverzija (I)

U kognitivnom procesu ekstraverzija je usmjerena prema vanjskom svijetu. S jedne strane, ekstraverzija je određena osobinama kao što su druželjubivost, komunikativnost i energičnost, a s druge strane osobinama poput zatvorenošću, povučenošću i rezerviranošću (koje definiraju njen negativni pol – introverziju). Ekstroverte stimuliraju vanjske informacije i potrebna im je kontinuirana interakcija s vanjskim svijetom kao izvorom informacija.

Introvertnu osobu moguće je prepoznati po tome što ju opisuje stimulacija prvenstveno vlastitog, unutarnjeg svijeta, bilo da su to misli ili osjećaji. Introverzija također znači da je to prvi korak u kognitivnom procesu usmjeravanja na vlastiti um i traženja informacije unutar njega, iako kasniji kognitivni koraci mogu biti ekstrovertirani.

2. Senzacija (S) - Intuicija (N)

Senzacija opisuje osobu koja obraća pozornost na stvarnost, odnosno na osjetila (miris, okus, dodir, sluh, vid). Odnosi se na sadašnji, aktualni i stvaran svijet. Primjećuje činjenice i pojedinosti koje su važne. Osobe ovog tipa imaju praktični pogled na svijet, te prilikom učenja implementiraju naučeno na praktičnom radu. Osoba ovog tipa može se detaljno prisjetiti nekog događaja, rješava probleme polazeći od činjenica, pragmatične naravi, stvara širu sliku temeljenu na činjenicama, vjeruje iskustvu te ponekad previše pažnje posvećuju činjenicama, bilo trenutnim ili prošlim ponekad propusti nove mogućnosti.

Intituitivni tip najviše posvećuje pozornost na značenja i obrasce informacija koje dobiva od vanjskog svijeta. Savladava proces učenja kroz razmišljanje o problemu, a ne na praktičnim iskustvima. Zanimaju ga nove stvari i ono što je moguće, stoga ima pogled u budućnost. Događaje pamti po osjećaju koji je imao u vrijeme samog događaja više nego kao stvarne činjenice ili pojedinosti o tome što se dogodilo.

3. Mišljenje (T) - Osjećaj (F)

Mišljenje opisuje osobu koja u trenutku donošenja odluke pronalazi osnovnu istinu ili načelo koje će primjenjivati, bez obzira na konkretnu situaciju. To je osoba koja analizira prednosti i nedostatke, logična je i dosljedna u odlučivanju. Ne dopušta da želje drugih ili vlastite želje utječu na nju.

Osobe koje pripadaju tipu Osjećaja mogu donositi najbolje odluke primjećujući što ljudi oko mene osjećaju i uzimajući u obzir stajališta svih uključenih u određeni događaj. Osobe ovog tipa zabrinute su za vrijednosti i što je najbolje za ljude koji su uključeni. To je osoba koja uspostavlja ili održava sklad. U odnosima s drugima su brižni, topli i susretljivi.

4. Prosuđivanje (J) - Percipiranje (P)

Posljednji par opisuje kako osoba želi živjeti svoj javni život, koja su to ponašanja koje drugi primjećuju. Preferirate li više strukturiran i odlučan način života (J) ili fleksibilniji i prilagodljiviji način života (P). Na taj način sagledava se koliko ste orijentirani prema vanjskom svijetu.

Svatko je u nekim trenucima ekstrovert, stoga ovaj par opisuje sklonost djelovanju u vanjskom svijetu, prilikom donošenja odluka ili prilikom uzimanja informacija. Stupaju u interakciju s vanjskim svijetom kada dobivaju informacije od njih. Bez obzira koriste li preferencije tipa Osjetljivost (S) ili Intuicije (N) oni još uvijek međusobno djeluju u vanjskom svijetu. Drugi

ljudi međusobno djeluju kada donose odluke, neovisno upotrebljavaju li preferencije za razmišljanje (T) ili preferencije osjećaja (F) još uvijek međusobno djeluju u vanjskom svijetu. Svatko u određenom trenutku prima informacije i donosi odluke u nekom n vremenskom razdoblju. Međutim, kada je u pitanju suočavanje s vanjskim svijetom, ljudi koji imaju tendenciju da se usredotoče na donošenje odluka imaju sklonost suditi s obzirom da polaze s odlukom kako stvari moraju biti odlučene. Ljudi koji se fokusiraju na preuzimanje informacija preferiraju Percipiranje (P) iz razloga što svoju odluku temelje na informacijama prije stvaranja konačnih odluka. Ponekad ljudi osjećaju da imaju obje preferencije, što je istinito. Preferencija J ili P samo govori koliko je osoba ekstrovertna s obzirom na preferenciju J ili P. Jedna osoba može se osjećati vrlo staloženo/ strukturirano (J) iznutra, ali njih vanjski život izgleda spontan i prilagodljiv (P). Druga osoba može se osjećati vrlo znatiželjnom i otvorenom (P) u svom unutarnjem svijetu, dok njihov vanjski život izgleda strukturiran ili više odlučujući (J).

Neke osobe koje koriste preferencije tipa prosuđivanja (J) u odlučivanju (neovisno jesu li preferencije tipa T ili F) u vanjskom svijetu. Za druge, čini se da preferiraju planirani ili uređeni način života, organiziran, osjećaju se ugodnije kada se donose odluke i vole držati svoj život pod kontrolom u mjeri u kojoj mogu. Dok s jedne strane ove preferencije opisuju odnos prema vanjskom svijetu, unutar sebe mogu osjetiti fleksibilnost i otvoriti se novim informacijama. Važno je da se preferencija tipa J ne miješa sa osudom, s negativnom konotacijom o ljudima i događajima. Oni nisu povezani na taj način.

Za razliku od njih tipovi s preferencijom Percipiranja (P) koriste ju (neovisno o S ili N tipu u sebi) u vanjskom svijetu. Za druge, čini se da preferiraju fleksibilniji i spontaniji način života, žele razumjeti i prilagoditi se svijetu više nego da ga organiziraju. Drugi ljudi ih doživljavaju otvorenim za nova iskustva i informacije. S obzirom da ovaj par opisuje samo što preferiraju u vanjskom svijetu, unutar osobnog svijeta osjećaju se da planiraju i odlučuju o njemu dovoljno. Potrebno je naglasiti da u jeziku tipova kojima pripadaju, percipiranje (P) na "preferenciju prikupljanja informacija". "Perceptivno" ne označava smisao brze i točne percepcije o drugima i događajima.

Prema modelu MBTI-a kombinirana su oba "svijeta". MBTI koristi akronime koje je predstavio Myers-Briggs zbog njegove praktičnosti i jednostavnosti s ekstra slovom za smještaj petog, umjesto četvrtog tipa osobnosti na ljestvici. Nadalje, za razliku od Myers-Briggsa ili drugih teorija temeljenih na Jungovom modelu, nisu uključili koncepte kao što su kognitivne funkcije. Jungove koncepte vrlo je teško mjeriti i testirati, stoga je predstavljen model u kojem su

preoblikovane i ponovno balansirane osobnosti zvane *Big Five* osobine ličnosti, model koji i dalje dominira modernim psihološkim i društvenim istraživanjima.

8.1 Pouzdanost i valjanost MBTI-a

Društvene znanosti koje su uključene u istraživanje osobnosti nailaze na problem u pogledu na pojedina ljudska bića s obzirom na poteškoće u konzistentnosti prilikom definiranja i predstavljanja. Pouzdanost i valjanost kao dosljedni rezultati te mjerenje onoga što mislimo da mjerimo, predstavljaju najveća dva izazova s kojima se svaka organizacija u ovom području suočava.

Pouzdanost se odnosi na sposobnost testa/ upitnika da omogući dosljedne rezultate, posebice prilikom istraživanja je li dosljedan tijekom vremena (pouzdanost ponovnog testiranja) te jesu li pitanja koja mjere skalu pripadnosti tipu međusobno konzistentna (pouzdanost interne konzistencije). Korelacija 0,7 je minimalna prihvatljiva vrijednost za upitnik ličnosti. Nezavisna, recenzirana studija [35] potvrđuje alat MBTI pouzdanim i valjanim. Na više 32,000 ispitanika Center for Applications of Psychological Type (CAPT) prikazao je koeficijente pouzdanosti s prosjekom EI=0.79, SN=0.84, TF=0.74 i JP=0.82 [36]. Također, Harvey je [37] proveo meta-analizu na studijama sažetim u MBTI priručniku [38] za koje su podaci dobiveni po spolu na uzorku od 102.174 ispitanika. Ova meta-analiza daje korigirane procjene razdvajanja na muškarcima i ženama: EI, 0.82 i .0.83; SN, 0.83 i 0.85; TF, 0.82 i .0.80; JP, 0.87 i 0.86. Test-ponovne pouzdanosti za MBTI rezultate sugeriraju postizanje dosljednosti tijekom vremena. Test-ponovnih koeficijenta u razdoblju od 1 tjedna do 2,5 godine variraju od 0.93 do 0.69 na SN skali, 0.93 do 0.75 na EI skali, 0.89 do 0.64 na JP skali i od 0.89 do 0.48 na TF skali [38]. Kada ispitanici pokazuju promjenu tipa, obično je to samo u jednoj preferenciji, a zatim u mjerilima gdje nisu bili izvorno jako razdijeljeni [38]. Sveukupno, najniža pouzdanost pronađena je u TF skali.

Valjanost

Valjanost provjerava mjeri li test/upitnik ono što bi trebalo mjeriti. Postoji dosta dokaza da alat MBTI točno opisuje stilove osobnosti, od kojih je jedna od njih navedena u nastavku. Postoji nekoliko načina da se pokaže valjanost, uključujući:

- Odnosi s drugim upitnicima
- Unutarnja struktura

- Odnosi s ponašanjem
- Tip opisa
- Praktična valjanost

Ukoliko MBTI instrument mjeri ono što bi trebao, prilikom rješavanja drugih upitnika uz druge alate koji mjere iste ili slične pojmove, trebao bi postojati visok stupanj korelacije između dva rezultata.

U istraživanju *The relationship between the revised NEO-Personality Inventory and the Myers-Briggs Type Indicator* [40] ukupno 900 sudionika završilo je upitnik NEO PI-R i MBTI. Korelacijska analiza mjera osobnosti pokazala je da je NEO-PI-R ekstraverzija povezana s MBTI Ekstraverzija – Introverzija. Otvorenost je bila u korelaciji s *Senzacija - Intuicija* (S – N), ugodnost s *Mišljenje - Osjećaj* (T-F) i savjesnost s *Prosuđivanje – Percipiranje*, replicirajući nalaze Costa i McCrae (1989).

9. Tekstualna analiza

Analiza teksta ili obrada prirodnog jezika (engl. *Natural Language Processing*) predstavlja način da se korisno i pametno analiziraju, razumiju i izvode značenje iz ljudskog jezika. Pomoću tehnika NLP-a stručnjaci mogu organizirati i strukturirati znanja za obavljanje zadataka kao što su automatsko sažimanje, prevođenje, prepoznavanje entiteta, ekstrakcija odnosa, sentiment analiza, prepoznavanje govora i segmentacija tema [41].

Prema Gartneru, tekstualna analiza predstavlja proces dobivanja podataka iz izvora teksta. Također, cesce se na NLP/Information Extraction gleda kao na proces dobivanja informacija iz podataka, a logika/ simbolicka AI bi trebala iz informacija dobiti "actionable insights" ili "knowledge". Ona se može primijeniti na bilo koji tekstualni skup podataka, uključujući društvene medije, forume, transkripte poziva i mnoštvo drugih izvora. Nova tehnološka postignuća u rudarenju podacima utjecala su na razvoj tehnika strojnog učenja (engl. *Machine Learning*) koje su usko vezane uz znanost o podacima (engl. *Data Science*). To se odnosi na široku klasu metoda koje se vrte oko modeliranja podataka na algoritamskim predviđanjima i na algoritamskom dešifriranju obrasca u podacima.

Primjer strojnog učenje za izradu predikcija – osnovni koncept temelji se na upotrebi označenih podataka za treniranje modela. Trenirani modeli predstavljaju automatsko označavanje podataka na način da se predvide oznake za nepoznate točke podataka. Primjer je model otkrivanja prevare s kreditnim karticama koji se može trenirati na temelju povijesnih zapisa o prevarama u kupnji. Rezultat modela jest procjena vjerojatnosti da je svaka nova kupnja lažna. Uobičajene metode za treniranje modela variraju od osnovnih regresijskih do složenih neuronskih mreža.

Glavno pitanje koje se postavlja jest što možemo učiniti s tekstualnom analizom općenito te kako je možemo primijeniti na analizu društvenih medija. Analiza teksta može se primijeniti na podatke s društvenih medija kako bi odgovorili na široku paletu pitanja o potrošačima, robnim markama, proizvodima te na temelju istog pokušali izraditi prediktivni model sa svrhom. Također, uz pomoć tekstualne analize možemo razumjeti opće mišljenje i specifične emocije, mjeriti i razumjeti sentimente robnim markama, proizvodima ili drugim temama.

Tipično, tekstualno miniranje obuhvaća tri glavne aktivnosti: [42]

1) prikupljanje informacija za prikupljanje relevantnog nestrukturiranog teksta među heterogenim bazama podataka, dokumentima i web stranicama,

2) ekstrakcija informacija (IE) za identificiranje i izdvajanje entiteta, činjenica i odnosa među tim entitetima, i

3) podatkovno rudarstvo kako bi pronašli udruge među informacijama izvađenim u različitim tekstovima koji se nalaze. Cilj tekstualnog pretraživanja jest izdvajanje i otkrivanje znanja skrivene u tekstu identificiranjem koncepata, pronalaženje činjenica/ odnosa u tekstovima, otkrivanjem implicitnih veza i stvaranjem hipoteza.

Prema tome, izdvojene su prednosti i nedostaci rudarenja tekstom:

Prednosti rudarenja tekstom:

- Rudarenje tekstom pomaže u sustavnom pregledavanju velikog broja literature
- Može pomoći u istraživanju, smanjujući rizik od propusta onoga što je važno
- Pomaže u otkrivanju obrazaca i trendova u podacima, povezanosti između entiteta, prediktivnih pravila, itd.
- Tekstualno miniranje ima mogućnost obogaćivanja nestrukturiranih teksta semantičkim oznakama i bilješkama (FOAF- eng. *Friend of a friend*, strojno čitljiva ontologija koja opisuje ljude, veze između njih i mnogih različitih stvari koje ih zanimaju)
- Rudarenje tekstom pomaže autorima s alatima za razvoj semantičkih anotacija
- Pomaže u upravljanju dokumentima i informacijama

Nedostaci rudarenja tekstom:

- Zahtijeva upravljanje velike količine "slobodnog teksta" (engl. *Free text*)
- Podaci najčešće nisu dobro organizirani i opisani
- Tekstovi sadrže nejasnoće i zahtijevaju ljudsku intervenciju, problematika leksičke, sintaktičke, semantičke i pragmatičke naravi
- Tehnike učenja za obradu teksta zahtijevaju komentare prije treniranja
- Razvoj resursa (otologija, korpusi) za poboljšanje rudarenjem teksta nije jednostavan zadatak.

Tekstualno rudarenje, poznato i kao tekstualno miniranje podataka ili tekstualna analiza olakšavaju organizacijama dobivanje vrijednog poslovnog uvida iz bogatstva tekstualnih informacija koje posjeduju. U nastavku je predstavljen set podataka na kojem ćemo temeljiti tekst analizu s ciljem : a) potvrđivanja ili pobijanja hipoteze, b) opisivanja i predstavljanja korištenih metode u rudarenju podacima, c) definiranja problema koji su se pojavili prilikom generiranja prediktora, d) provođenja semantičke analizu te d) modeliranja podataka putem tehnika Stabla odlučivanja i Random forest.

10. Metodologija istraživanja

U ovom dijelu obrazložena je metodologija istraživanja koja uključuje ciljeve istraživanja, postavljenu hipotezu, grafičke rezultate istraživanja te interpretaciju provedenog istraživanja.

10.1 Ciljevi istraživanja

- Informirati se o povezanosti podataka sa 4 tipa MBTI-a
- Izdvojiti značajke iz teksta (metodom Bag of Words)
- Provesti pre-procesuiranje podataka
- Izračunavanje frekvencija riječi u postovima
- Modeliranje podataka
- Usporedba prediktivnih klasifikacijskih modela

10.2 Hipoteze istraživanja

Na osnovi dosadašnjih istraživanja postavljena je sljedeća hipoteza:

H1… Može li se analizom postova na forumu predvidjeti MBTI ličnost.

10.3 Metoda istraživanja

Kako bi se istražili postavljeni ciljevi i testirala hipoteza, u radu su prikupljeni podaci iz sekundarnih izvora. Sekundarni podaci uključivali su znanstvena istraživanja na temu web profiliranja putem rudarenja podacima (engl. *Data Minning*), najčešće korištenih sustava profiliranja korisnika, koji kao i svi navedeni, primjenu imaju u personalizaciji usluga/proizvoda krajnjem korisniku. U ovom djelu prikazan je tekstualni okvir za tekstualnu analizu za koju je preliminarno potreban R [43] (besplatni programski jezik i softversko okruženje za statističku računalnu obradu i grafiku koju podržava Zaklada za statističku obradu) i RStudio [44] (nenaplatna i open-source integrirana razvojna okolina (IDE) za R, programski jezik za statističko računanje i grafiku).

Tekstualni podaci izvorno su preuzeti s platforme Kaggle [45]. Set podataka na Kaggle-u prikupljen je putem foruma *PersonalityCafe* (sadrži veliki broj korisnika na forumu). Također,

ovaj skup podataka sadrži 8675 redaka podataka, 2 kolone (*type* i *post*), a svaki red je predstavlja osobu s MBTI tipom te zadnjih 50 stvari koji su objavili na jednoj od društvenih mreža. Nadalje svaki post je odijeljen s "|||" te ima funkciju razdjelnika postova. Tehnike pre-procesuiranja, usredotočene su na identifikaciju i ekstrakciju značajki iz tekstualnih podataka te se potom koriste za transformaciju nestrukturiranih podataka u strukturirani set.

11. Metode pret- procesiranja

11.1 Metoda Bag of Words

Metoda “vreća riječi” (engl. *Bag of Words* (BoW)) predstavlja način izdvajanja značajki iz teksta za upotrebu u modeliranju, poput algoritama strojnog učenja. Pristup modelu je vrlo jednostavan i tehnički nije zahtjevan, a za cilj ima izdvojiti značajke iz dokumenta koji sadrži set podataka. Predstavljaju tekst koji opisuje pojavu riječi unutar dokumenta. To uključuje dvije stvari: a) Vokabular poznatih riječi i b) Mjera prisutnosti poznatih riječi.

To se naziva "vreća riječi" jer se brišu sve informacije o redoslijedu ili strukturi riječi u dokumentu. Model se bavi samo poznatim riječima u dokumentu. U nastavku kroz primjer temeljen na podacima MBTI-a, konkretnije u *postu* (kolumna podataka) predstavljena je metoda BoW. Nadalje, opisan je postupak upravljanja vokabularom te koje su ograničenja predstavljene metode.

19. redak u podacima *Posts:*

"I think we do agree. I personally don't consider myself Alpha, Beta, or Foxtrot (lol at my own joke). People are people."

```
Number of docs: 1
0 stopwords:  ...
ngram_min = 1; ngram_max = 1
Vocabulary:
          term term_count doc_count
 1:         at          1         1
 2:   consider          1         1
 3:         or          1         1
 4:      agree          1         1
 5:       joke          1         1
 6:    foxtrot          1         1
 7: personally          1         1
 8:     myself          1         1
 9:         we          1         1
10:          t          1         1
11:         my          1         1
12:       beta          1         1
13:        own          1         1
14:      think          1         1
15:         do          1         1
16:        are          1         1
17:        don          1         1
18:      alpha          1         1
19:        lol          1         1
20:     people          2         1
21:          i          2         1
          term term_count doc_count
```

Slika 1. Metoda BoW na primjeru objave
Izvor: Vlastiti rad

Kroz primjer predstavljene su jedinstvene riječi (ignorirane su interpunkcije), te se naš vokabular sastoji od 21 riječ u korpusu od 23 riječi. Iako je većina teksta kreirana i pohranjena tako da ga ljudi mogu razumjeti, za računala nije uvijek lako izvršiti tu obradu teksta. S povećanjem robusnosti teksta generiranih na društvenim medijima, očistiti takav tekst postalo je preko potrebno, također razlog tome jest činjenica kako NLP tehnike uopće ne funkcioniraju zbog nekoliko razloga kao što su oskudnost u podacima, riječi koje se pojavljuju izvan rječnika riječi i nepravilne sintaktičke strukture u takvim tekstovima. Problematika u našem setu podataka pojavila se kod riječi "www.youtube.com" koju je teško tretirati a pojavnost joj je bila visoko izražena. Moguće je razdvojiti je na 3 zasebne riječi no ona nam neće svejedno dati uvid u kontekst s obzirom da je ona najčešće spominjana kao link u postu ljudi. Ujedno je i uklonjena riječ *watch* koja je imala veliku pojavnost u korpusu podataka jer je vezana uz link youtuba (primjer: "http://www.youtube.com/watch?v=qsXHcwe3krw") te nam neće dati kvalitetan uvid u daljnju tekstualnu analizu.

Također, metoda "vreća riječi" ima određene nedostatke poput [46]. :

a) Rječnik: Vokabular zahtijeva pažljivi dizajn, posebno kako bi se upravljalo veličinom koja utječe na oskudnost prikazivanja dokumenata.
b) Oskudnost: Oskudne podatke teže je modelirati i iz računalnih razloga (prostor i vrijeme složenosti), ali i zbog informacijskih razloga, gdje je izazov za modele da koriste tako malo informacija u takvom velikom reprezentativnom prostoru.
c) Značenje teksta: Odbacivanje redoslijeda riječi ignorira kontekst, kao i značenje riječi u dokumentu (semantika). Kontekst i značenje mogu ponuditi mnogo modelu, ukoliko bi model mogao reći razliku između iste riječi različito raspoređene ("ovo je zanimljivo" vs "to je zanimljivo"), sinonimi ("stari bicikl" vs "rabljeni bicikl") , i mnogo više.

11.2 Tokenizacija

Na utreniranim podacima koje ćemo koristiti za kreiranje vokabulara također je provedena *Tokenizacija* [47] (engl. *Tokenization*). Ona predstavlja proces razbijanja komada teksta u manje komadiće kao što su riječi, fraze, simboli i drugi elementi koji se zovu tokeni. Čak se cijela rečenica može smatrati znakom. Tijekom procesa tokenizacije mogu se ukloniti neki

znakovi kao što su znakovi interpunkcije. Znakovi zatim postaju *input* za ostale procese u tekstualnom rudarstvu kao što je raščlanjivanje. Oni se razdvajaju u nekoliko koraka:

Tokeni ili riječi odvojeni su razmakom, interpunkcijskim znakovima ili prekidima retka. Interpunkcijski znakovi mogu ili ne moraju biti uključeni ovisno o potrebi. Svi znakovi unutar susjednog niza su dio token-a. Tokeni mogu biti sastavljeni od svih alfa znakova, alfanumeričkih znakova ili samo numeričkih znakova. Ono mogu također biti i separatori. Na primjer, na većini programskih jezika, identifikatori mogu biti postavljeni zajedno s aritmetičkim operatorima bez razmaka. Iako se čini da se to pojavljuje kao jedna riječ ili token, gramatika jezika zapravo smatra matematički operator (token) kao razdjelnik, pa čak i kada se više tokena skupljaju zajedno, one se ipak mogu razdvojiti matematičkim operaterom.

Tokenizacija je proces koji se smatra među prvim koracima u obradi teksta. Vrlo je teško izdvojiti korisne informacije o visokoj razini iz teksta bez da se identificiraju tokeni. Svaki tok je primjer vrste, pa je broj tokena puno veći od broja tipova. Kao primjer, u prethodnoj rečenici nalaze se dva znaka "." Ovo su oba slučaja tipa "," što se dvaput pojavljuje u rečenici. Pravilno govoreći, uvijek se treba odnositi na učestalost pojavljivanja nekog tipa, ali slobodna upotreba također govori o učestalosti tokena. Prostor znakova, kartica i novi redak koji pretpostavljamo uvijek su graničari i ne broje se kao tokeni. Često se često nazivaju bijelim prostorom. Likovi () <>! ? "uvijek mogu biti razgraničenja i mogu biti znakovi, a znakovi mogu biti, ili ne moraju biti, graničnici, ovisno o njihovoj okolini. Bilo koji drugi zarez je graničnik i može biti token. Tokenizacija je provedena nad našim utreniranim podacima i kreiran je vokabular za daljnju tekstualnu obradu.

Nadalje, nakon kreiranog vokabulara sljedeći potez u pret- procesuiranju podataka je uklanjanje uobičajenih riječi iz teksta po metodi filtriranja Zaustavnih riječi (engl. *Stop Words*). To uključuje riječi kao što su članovi (a, an, the), veznici (i) uobičajeni glagoli (is), prepozicije (of) i drugi. Razlog uklanjanja tih riječi jest činjenica kako neke od riječi u korpusu dokumenata nisu informativne, odnosno u sljedećim koracima one nam neće dati nikakav uvid u značajnosti teksta u daljnjoj analizi. Ne postoji univerzalni popis zaustavnih riječi koje svi alati koriste jer neki od alata posebno izbjegavaju uklanjanje riječi kako bi podržali pretraživanje fraza.

Problematika skupa podataka na kojima se radi analiza MBTI-a je pristranost te je upravo zbog pristranosti teško generalizirati zaključke s obzirom da su oni prikupljeni s foruma. Iako su prikupljeni s Kaggle-a, kao platforma za koju je globalna tvrtka Google najavila preuzimanje

[48], govori o vjerodostojnosti seta objavljenih podataka Kaggle zajednice. Pret- procesiranje podataka ima važnu ulogu u pretvaranju nestrukturiranih tekstualnih podataka iz zbirke dokumenata u više rukovodeći sustavni prikaz temeljem raznih besplatnih paketa korisnih alata za tekst analizu. U pret-procesiranju, prilikom kreiranja utreniranog seta podataka upotrjebljena je funkcija koja filtrira ulazni vokabular i izbacuje vrlo česte ili vrlo rijetke pojmove. Na taj način možemo smanjiti vokabular i značajno poboljšati točnost rječnika. Nadalje, postavili smo uvjete u setu podataka, poput:

a) minimalni broj pojavljivanja na svim dokumentima (engl. Term_count_min) – 100
b) minimalni udio dokumenata koji bi trebali sadržavati termine (engl. Doc_proportion_min) – 0.1
c) maksimalni udio dokumenata koji bi trebali sadržavati pojam (eng. doc_proportion_max) – 0.5

Sljedeća funkcija u daljnjem pret- procesuiranju podataka *Vectorizer* stvara objekt koji definira kako transformirati popis tokena u vektorski prostor. Sami tekstovi mogu zauzeti puno memorije, no vektorizirani tekstovi obično ne, jer se tako pohranjuju kao rijetke matrice. Zbog R *copy-on-modify* semantike nije lako iterativno razviti DTM (engl. *Document-term matrix*). Stoga, izgradnja DTM-a, čak i za male zbirke dokumenata može biti usko grlo za analitičare. To uključuje čitanje cjelokupne zbirke tekstualnih dokumenata u RAM i njegovu obradu kao jednog vektora, što može povećati potrošnju memorije faktorom od 2 do 4. Paket "Text2vec" rješava taj problem, pružajući bolji način izgradnje matrice dokumenta [49]. Prije same analize korisno može biti pretvoriti DTM, odnosno primijeniti normalizaciju. Na primjer, duljine dokumenata u zbirci mogu se znatno razlikovati.

11.3 Frekvencija riječi u postu

Za izračunavanje frekvencija riječi moramo imati naše podatke u urednom obliku. To je razlog zašto smo vokabular iterativno pretvorili u DTM format. U nastavku predstavljene su frekvencije riječi u postovima MBTI –a.

Slika 2. Frekvencija riječi u postu
Izvor: autorov rad

Na slici iznad predstavljene su 100 najzastupljenijih riječi u setu podataka, konkretnije u objavama. Top 3 riječi s najvećim brojem pojavljivanja su:

- Time (27582)
- Feel (23344)
- Watch (19258)

11.4 Sentiment analiza

Analiza sentimenta ili mišljenja javnog mnijenja su zadaci za obradu prirodnih jezika ili ekstrakcija informacija koji pomažu u izdvajanju mišljenja ili sentimenta koje su izražavali ljudi u zbirci dokumenata. Cilj analize sentimenta na primjeru MBTI postova je uvid u analizu tipova MBTI-a koji je povezan s postovima ljudi, odnosno uvid u ukupnoj kontekstualnoj polarnosti sadržanoj u dokumentu.

Sentiment u analizu sentimenta može se dobiti na razini dokumenta klasifikacijom polariteta izraženog mišljenja u tekstu dokumenta, na rečenici ili entitetskoj značajci kako bi se utvrdilo je li izraženo mišljenje pozitivno, negativno ili neutralno. Tekst može sadržavati mnoge entitete, ali je nužno pronaći entitet prema kojem su usmjereni osjećaji. Identificira polarnost i stupanj

sentimenata. Sentimenti se klasificiraju kao objektivni (činjenični), pozitivni (označavaju stanje sreće, blaženstvo ili zadovoljstvo dijela pisca) ili negativni (označava stanje tugovanja, odbijanja ili razočaranja dijela pisca).

Analiza sentimenta računalna je zadaća automatskog utvrđivanja osjećaja koje ljudi pišu u tekstu. Osjećaj je često uokviren kao binarna razlika (pozitivno ili negativno), ali može biti i preciznija, poput prepoznavanja specifičnih emocija koje autor izražava (poput straha, radosti ili ljutnje).

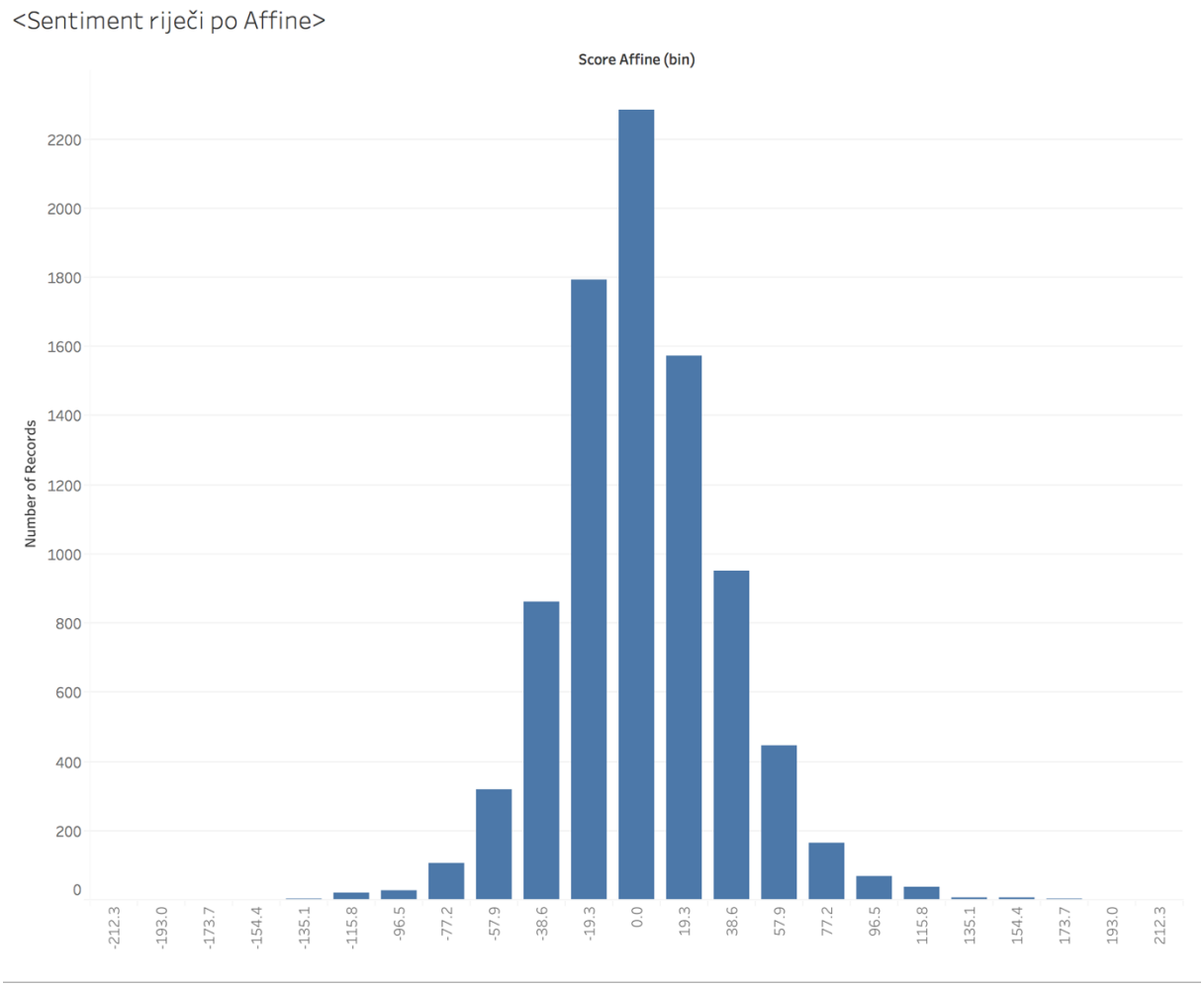
Drugi način saznavanja sentimenta je korištenjem metode ocjenjivanja gdje se sentimenti daju na temelju njihovog stupnja pozitivnosti, negativnosti ili objektivnosti. U ovoj metodi analizira se tekst, a naknadna analiza pojmova sadržanih u tekstu se provodi kako bi se razumjele sentimentalne riječi i kako se te riječi odnose na koncepte. Svaki je koncept tada dobiven rezultat baziran na odnosu između sentimentalnih riječi i povezanih pojmova.

Postoje različite metode i rječnici za procjenu mišljenja ili sentimenta u tekstu. Tidytext paket sadrži nekoliko leksikona, od kojih su navedeni korišteni prilikom određivanja sentimenta u skupu podataka.

Tri općenita leksikona su:

- Affin (prema Finn Arup Nielsen) – rangira riječi od: -5 (vrlo negativno) do +5 (vrlo pozitivno)
- Bing (prema Bing Liu and collaborators) - daje oznaku "negativnih" ili "pozitivnih"
- Nrc (prema Saif Mohammad i Peter Turney) – daje etiketu entitetima (ljutnju, predviđanje, nelagodu, strah, radost, negativnost, pozitivan osjećaj, tugu, iznenađenje ili povjerenje)

Primjenom paketa određivanja sentimenta na podatke rezultati ukazuju sljedeće:



Grafikon 2. Analiza sentimenta riječi po leksikonu Affin

Ukoliko sagledamo sentiment po leksikonu Affin možemo uočiti kako najviše zapisa, točnije njih 2,283 ima pozitivni sentiment od 0 do 19.3 (ne uključeno). Nadalje, 1,794 zapisa ima sentiment od -19.3 do -38.6, dok za razliku od njih, manje pozitivne vrijednosti vidimo za 1,572 zapisa sa vrijednostima 19,3 – 38.6 (ne uključeno). Kako broj zapisa pada istodobno se smanjuje i vrijednost sentimenta, kako pozitivnih tako i negativnih.

Problematika u određivanju sentimenta pojavljuje se u slučaju kada je skup podataka opsežan jer uklanjanjem neinformativnih riječi zanemaruje se kontekst. Nadalje, ukoliko u podacima nema sentimenta definirali smo da je on tada neutralan s ciljem dobivanja kvalitetnijih rezultata.

12. Modeliranje podataka

12.1 Klasifikacija algoritma

Rudarenje podataka (engl. *Data Mining*) definirano je kao proces izdvajanja informacija iz baze podataka koja se smatra ključnom i ima potencijalnu vrijednost. Nova tehnologija za analizu podataka i ima široku primjenu u financijama, osiguranjima, državnim upravama, nacionalnom segmentu te prometu [50]. Rudarenje podacima korisno je za dobivanje velikih količina nestrukturiranih podataka u strukturirane i korisne, čime se generira ne samo vrijednost, već ukazuje na ROI-a (engl. *Return of investment*) od nestrukturiranog upravljanja podacima. Kroz tehnike kategorizacije, ekstrakcije entiteta, sentiment analize, tekstualno rudarenje izvlači korisne informacije i znanja skrivenih u tekstualnom sadržaju. U poslovnom svijetu stječe se uvid u otkrivanju obrasca i trendova u većim količinama nestrukturiranih podataka. Sposobnost rudarenja podataka ogleda se u funkcionalnosti, sposobnosti da odbaci sav ne relevantni materijal i ponudi odgovore koji vode k brzom usvajanju, posebice u velikim organizacijama.

Kroz tehnike poput kategorizacije, entiteta ekstrakcije, analize osjećaja i ostalih, tekstualno rudarstvo izvlači korisne informacije i znanja skrivene u tekstualnom sadržaju. U poslovnom svijetu ovo se odnosi na sposobnost otkrivanja uvida, obrazaca i trendova u većim količinama nestrukturiranih podataka. Zapravo, to je ta sposobnost da odbaci sav ne-relevantni materijal i daju odgovore koji vode do brzog usvajanja, osobito u velikim organizacijama. S obzirom da smo u prvom djelu definirali primjere personalizacije i u koje se sve svrhe mogu koristiti, predstavljeni su neki od primjera u kojima ova tehnologija danas pomaže:

1. Detekcija prijevara kroz istragu zahtjeva

Tekstualna analitika iznimno je učinkovita tehnologija u bilo kojoj domeni gdje se većina informacija prikuplja kao tekst. Osiguravajuća društva iskorištavaju ove tehnologije rudarenja kombiniranjem rezultata analize teksta s strukturiranim podacima kako bi se spriječile prijevare i brzo obradili zahtjevi.

2. Digitalno oglašavanje - kontekstualno

Novo i sve veće područje primjene za tekstualnu analizu predstavlja digitalno oglašavanje (kontekstualno). Posljednjih godina Informacijska ekstrakcija [51] koja predstavlja aktivno

područje istraživanja analize teksta bilježi stalni rast. Rješenje koje je razvila tvrtka AdmantX [52] omogućila je bolje pozicioniranje oglasa na web stranicama koje su temeljene na semantici glavnog sadržaja stranice, entiteta navedenih u tekstu, izraženih mišljenja, emocija i poruka koju prenosi tekst. Uspoređujući s tradicionalnim pristupima temeljenih na kolačićima, kontekstualno oglašavanje ima veću točnost te u potpunosti čuva privatnost korisnika.

3. Analiza podataka društvenih medija

S gledišta organizacija, društveni mediji jedan su od najplodnijih izvora nestrukturiranih podataka. Društvene mreže sve više postaju prepoznatije kao vrijedan izvor te ih koriste za analizu ili predviđanje potreba kupaca. Tekstualna analiza može obrađivati veliku količinu nestrukturiranih podataka, analizirati mišljenja, izlučivanje emocija i osjećaja te njihov odnos s robnim markama, njihovim proizvodima i uslugama.

12.2 Stablo odlučivanja

Prema Apte i Weiss, Stablo odluke je veoma poznata metoda klasifikacije [53]. Struktura čiji su glavni dijelovi: izvorni čvor (engl. *root node*), grane (ogranci) te granični čvor (engl. *leaf nodes*). Svaki čvor označava ishod te svaki ogranak ima oznaku klase. Algoritam stabla odluke je model rudarenja podacima kroz koji se stvara algoritam indukcijskog učenja baziranih na primjerima. Primijenjeno je stablo odlučivanja na setu utreniranih podataka kako bi se vidjela potencijalna vrijednost seta. Cilj metode Stabla odlučivanja je potvrditi ili pobiti definiranu hipotezu *H1 Može li se analizom postova na forumu predvidjeti MBTI ličnost*. Stablo odluke iskorišteno je kao klasifikacijski model i ima svrhu prepoznavanje tipova u MBTI postovima.

Na sljedećem primjeru predstavljen je utrenirani model na stablu odluke kao i rezultati dobiveni modelom.

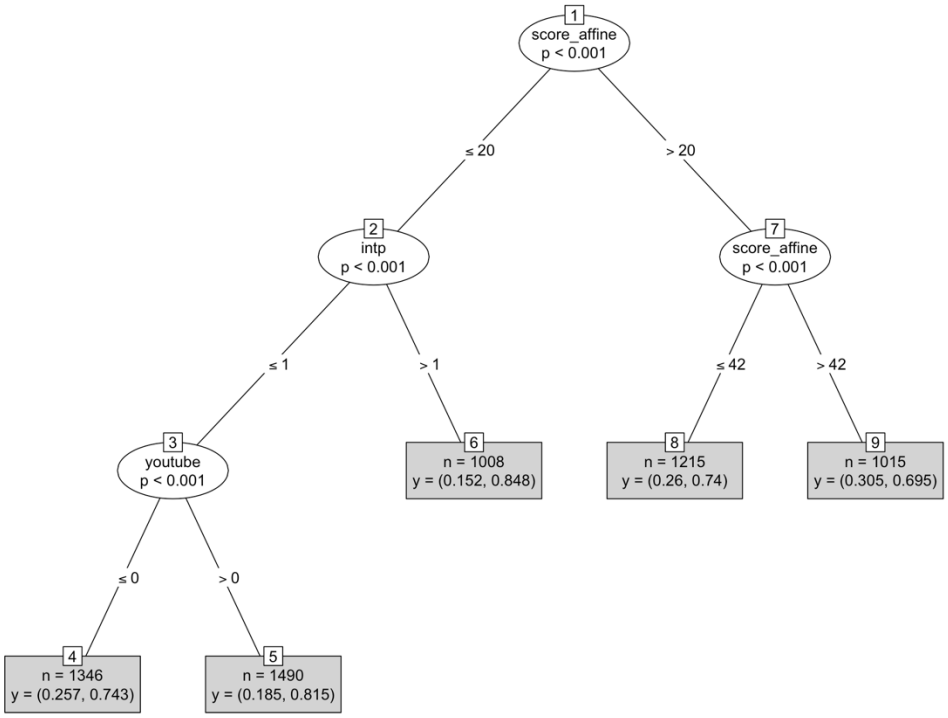


Figure 1. Model Stabla odlučivanja nad odabranim entitetima
Izvor: Autorov rad

Kreirani model stabla odluke uključio je nasumično odabrane riječi poput: youtube, intp, score_affine i htpp. Analiza stabla odlučivanja temelji se na metodi povratne indukcije, započinje u krajnjim granama i nastavlja se u pravcu početnog čvora odlučivanja. Na temelju rezultata analize, u kojoj je uključen minimalni broj opservacija u graničnom čvoru 1000, dana je prosudba o vrijednosti svake od navedenih riječi.

Upute za čitanje stabla odluke:

N – predstavlja broj ljudi

Y – postotak korisnika I (Introvert) ili E (Ekstrovert)

Prema tome, skup od 1015 ljudi riječi koje je paket score_affin definirao kao pozitivne ili negativne (-5 do 5) koristilo je više od 20 puta te više od 42 puta, stoga među takvim korisnicima 30,5 % ljudi model je stavio u E, dok je njih čak 69,5 % model stavio u grupu I. Nadalje, korisnici koji su više od 20 puta te manje ili jednako 42 puta iskoristili riječ iz skupa riječi score_affin, na uzorku od 1215 takvih ljudi, model svrstao 26% u E, dok njih 74% pripada I grupi. 10008 korisnika koji su manje ili jednako koristili riječi score_affina od 20 puta, a

koristili riječ intp u svojim postovima više od jednog puta model ih je stavio, njih 84,4 % u I, a njih 15,2 % u E. Također, korisnici, njih 1490 koji su iskoristili riječ intp manje ili jednako jednom putu, te koristili riječ youtube više od 0, model je svrstao njih 81,5 %u E dok je njih 18,5 % svrstao u I tip. Nadalje, korisnici koji su manje od jednom koristili riječ intp, a riječ youtube nisu koristili u svom opusu riječi u postovima, njih 1346 model je svrstao u I (74,3%) i E (25,7 %). Prema tome, rezultati dobiveni modelom stabla odlučivanja dobri su na skupu za treniranje, dok su se pokazali lošiji na testnom skupu.

Poteškoće u primjeni stabla odlučivanja na modelu jest što dobro stablo odlučivanja ne može "podnijeti" preveliki broj inačica, stoga je model predstavljen samo na primjeru. U nastavku, biti će definirana tehniku modeliranja na *training* skupu podataka za generiranje (učenje) modela, koja je namijenjena za veći broj inputa – Random Forest. Razlog odabira modela Random Forest leži u činjenici kako neke tehnike modeliranja imaju određeni broj parametra koji utječe na stvaranje modela, a na taj način utječe i na oblik i kvalitetu generiranog modela.

Proces generiranja modela iterativne je prirode, mijenjanjem parametra traži se optimalna kombinacija koja će dati najbolji rezultat na testnom uzorku podataka (training set).

12.3 Random Forest modeliranje

Random Forest je klasifikacijski algoritam. Djeluje tako da gradi mnoštvo stabla odlučivanja u vrijeme treniranja te proizvodi klasifikaciju ili srednja predviđanja (regresija) pojedinih stabala. S bilo kojim modelom stabla odluke koji koristi malu veličinu uzorka, uvijek postoje pitanja o "pretreniranosti" stabla (engl. *overfitting*) kada model ne odražava stvarne zavisnosti među ulaznim i izlaznim varijablama. Modeli poput Random Foresta općenito će bolje generalizirati, a više stabla koje koristimo prilikom korištenja Random Forest modela predstavljaju veću šansu da prediktivni model bude pretreniran sada se smanjuje jer sada stavljamo izvedbu na nekoliko različitih modela. Postoji izravan odnos između broja stabala u šumi (engl. *Random Forest*) i dobivenih rezultata: što je veći broj stabala, točniji je rezultat. No, potrebno je imati na umu da stvaranje šume nije isto što i konstruiranje odluke s informacijskim dobitkom ili dobivanjem indeksnog pristupa. Razlika između algoritma Random Forest i algoritma Stabla odluke leži u činjenici da se u Random Forest algoritmu postupci pronalaženja izvornog čvora (engl. *root node*) te podjela čvorova značajki izvode nasumično.

U usporedbi s drugim klasifikacijskim tehnikama prednosti [54] Random Forest tehnike su:

1. Za aplikacije u klasifikacijskim problemima, algoritam Random Forest izbjeći će preveliki problem pretreniranosti.
2. Za zadatak klasifikacije i regresije, može se koristiti isti slučajni Random Forest algoritam
3. Algoritam Random Forest može se koristiti za identifikaciju najvažnijih značajki iz skupa podataka za obuku, drugim riječima, značajka inženjeringa.

U nastavku prije cjelokupnog modeliranja podatka predstavljena je ilustracija algoritma Random Forest u dva stabla s ciljem uvida i shvaćanja funkcionalnosti algoritma.

12.4 Interpretacija utreniranog seta podataka

Na *training* set su primijenjene nasumično izabrane riječi poput: youtube, intp score_affine http. *Training set* koji je uključivao minimalni broj opservacija u graničnom čvoru 500 uključio je predikcijski model, svrstavanje Introverata i Ekstroverata prema broju korištenja riječi *p <0.001 i >0.005*. S naglaskom na cilj primjera – ilustracija metode, predstavljeni model temelji se interpretaciji tekstualne analize primjenom modela, Random Forest algoritma koji na primjeru predviđa postotak zastupljenosti u skupu s obzirom na definirane riječi u utreniranom setu. Primjenom algoritma u R format, dobiveno je stablo odlučivanja s sljedećim rezultatima:

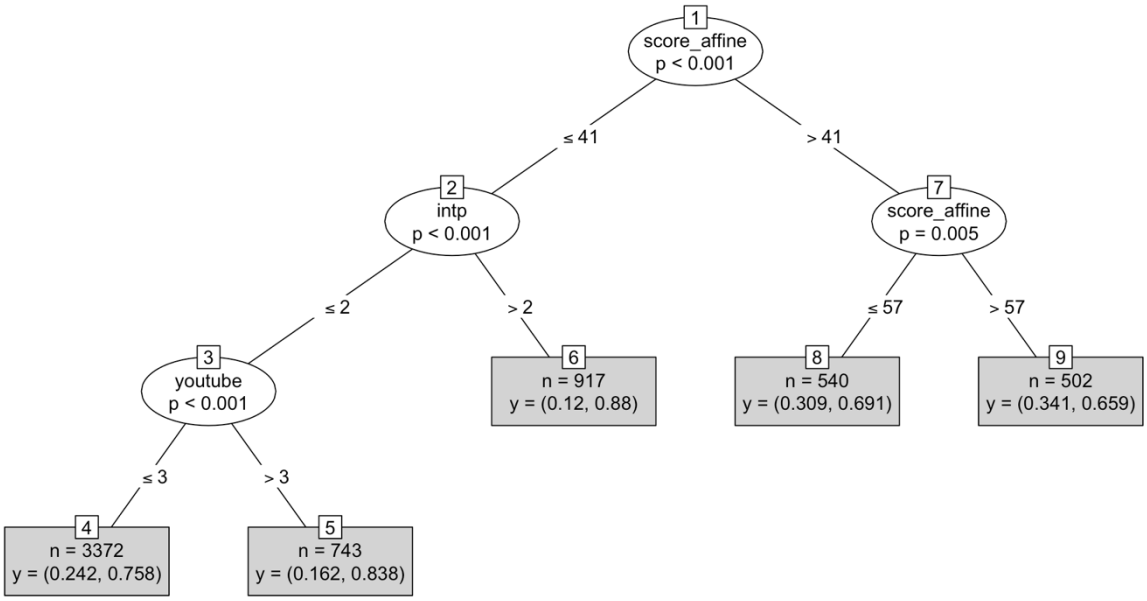


Figure 2. Model Random Forest nad nasumičnim setom podataka
Izvor: Autorov rad

Upute za čitanje stabla odluke:

N – predstavlja broj ljudi

Y – postotak korisnika I (Introverata) ili E (Ekstroverata) prema *p <0.001 i >0.005*

Broj ljudi, njih 502 koji su u svom opusu riječi koristili set riječi affine_score > 41 puta te >57 puta model ih svrstava njih 66% u I (Introvert) dok njih 34,1 % svrstava u E (Ekstravert). Nadalje, 540 ljudi koji su koristili isti set riječi affine_score > 41 puta te ≤ 57 puta sustav ih je svrstao u E njih 31 % a u I 70 % ljudi. Također, 917 ljudi koji su ≤ 41 puta iskoristili riječi iz score_affina te koji su >2 iskoristili uz to i riječ intp model ih svrstava u I njih 88% dok njih 12% svrstava u E. Od njih 743 koji su riječ youtube iskoristili > 3 puta, a riječ intp ≤ 2, te ≤ 41 riječi score_affin sustav je filtrirao 83,8 % ljudi u I dok je 16,2 % njih svstao u E. Na uzorku od 3372, najveći uzorak, njih 75,8% sustav je svrstao u I na temelju korištenja riječi score_affin ≤ 41, te intp ≤ 2 i youtube ≤ 3.

U sljedećem primjeru utrenirani set podataka nasumično je odredio riječi na koje ćemo se fokusirati, konkretnije to su riječi: nfj, score_affine, fear te joy.

Training set koji je uključivao minimalni broj opservacija u graničnom čvoru 500 uključio je predikcijski model, svrstavanje Introverata (I) i Ekstroverata (E) prema broju korištenja riječi *p <0.001 i =0.008*. S naglaskom na cilj primjera – ilustracija metode, predstavljeni model temelji se interpretaciji tekstualne analize primjenom Random Forest algoritma koji na primjeru predviđa postotak zastupljenosti u skupu s obzirom na definirane riječi u utreniranom setu. Primjenom algoritma u R format, dobiveno je stablo odlučivanja s sljedećim rezultatima:

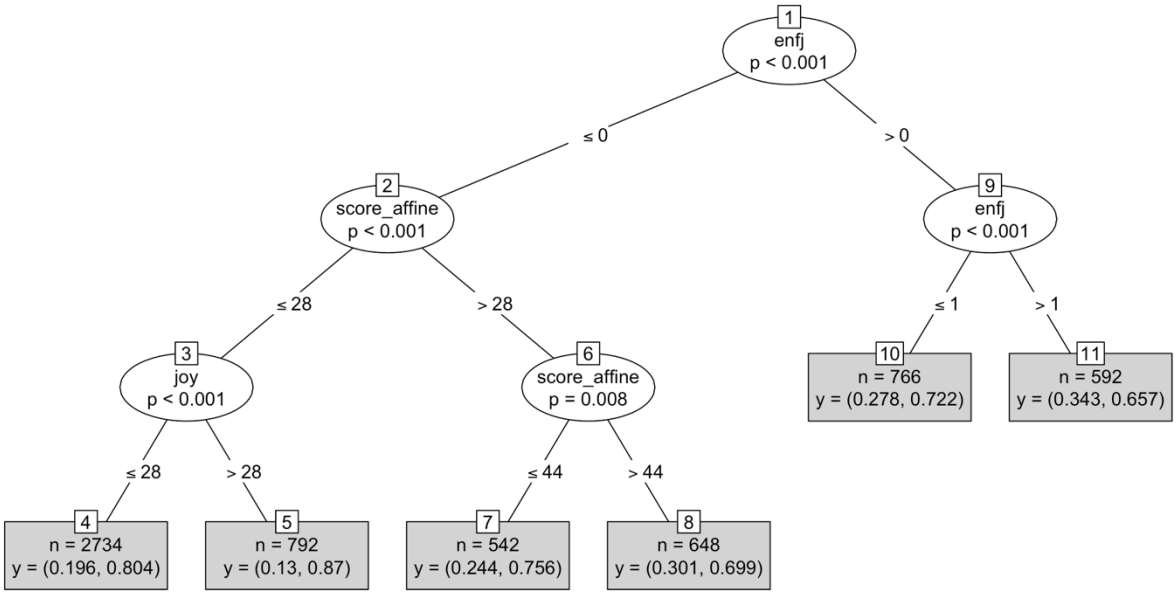


Figure 3. Model Random Forest na nasumičnom setu podataka II.

Na uzorku 592 ljudi, oni koji su spomenuli riječ enfj > 1 algoritam je svrstao njih 65,7 % u I, te njih 34.3 % u E. S obzirom na 766 ljudi koji su > 0 te ≤ 1 iskoristili riječ enfj sustav je filtrirao 72 % ljudi u I, a njih 27,86 u E. Nadalje, ljudi koji su ≤ 0 iskoristili riječ enfj, ali su iskoristili riječi score_affin ≤ 28 puta te riječ joy > 28 puta, njih 792 sustav je svrstao 87% u I te 13% u E. Također, najviše 2374 ljudi koji su iskoristili ≤ 0 riječ enfj, ali su riječi score_affin i joy iskoristili također ≤ 28 puta, sustav ih je svrstao, njih 80% u I dok je njih 19,6% svrstao u E.

13. Predviđanje točnosti modela

Random Forest jedan je od najpoznatijih algoritama klasifikacije - sposoban je velike količine podataka klasificirati s točnošću. Polazeći od te činjenice, izračunati će se prototipovi koji daju informacije o odnosu između varijabli i klasifikacije. Predstavljena metode klasifikacije ima za cilj izgraditi model koji će potvrditi ili pobiti predstavljenu hipotezu *H1 Može li se analizom postova na forumu predvidjeti MBTI ličnost.*

Na svakom tipu: Introverzija (I) – Ekstraverzija (E), Intuicija (N) – Senzacija (S), Mišljenje (T) – Osjećaj (F), Prosuđivanje (J) – Percipiranje (P) dobiveni su prototipovi s informacijama o njihovim odnosima u skladu s našom hipotezom.

Najjednostavnije raščlanjivanje seta podataka jest slučajna podjela koja generira skup učenja (ili treninga, *eng. Training set*) i test set (*engl. Test set*). Razmišljanje na temelju takve podjele jest sljedeće:

Podaci dostupni za analitiku prilično predstavljaju procese stvarnog svijeta koji želimo modelirati. Podjela podataka u učenje i ispitivanje mora se izvršiti pažljivo kako bi se izbjeglo uvođenje bilo kakvih sustavnih razlika između učenja i testiranja. Podjela podataka (engl. *Data Partition*) cjelokupnog seta je proces logičkog i/ ili fizičkog podjeljivanja podataka u segmente koji su lakše održavani ili dostupni. Upravo je time kreirana serija raščlanjivanja podataka na *test* i *training* skup u kojima p – predstavlja postotak podataka za treniranje. P iznosi 0,7, tj 70% podataka odlazi na testiranje, dok preostalih 0,3, tj 30 % odlazi na treniranje modela. Prilikom pozivanja Random Forest algoritma na modele IE, NS, TF i JP broj stabla (engl. *ntree*) iznosi 301, dok je broj slučajno uzrokovanih varijabli u svakom odjeljku 25.

13.1 Rezultati modela IE

Konfuzijska matrica (engl. *Confusion matrix*) je tablica koja se često koristi za opisivanje izvedbe klasifikacijskog modela (ili "klasifikatora") na skupu testnih i treniranih podataka za koje su poznate prave vrijednosti. U nastavku je prikazana i interpretirana konfuzijska matrica modela IE, NS, PJ i TF – zastupljenost tipova

```
                Type of random forest: classification
                      Number of trees: 301
No. of variables tried at each split: 25

        OOB estimate of  error rate: 20.6%
Confusion matrix:
    E    I class.error
E 177 1226 0.873841768
I  26 4650 0.005560308
                Test set error rate: 20.53%
Confusion matrix:
   E    I class.error
E 68  528    0.885906
I  5 1995    0.002500
```

Figure 4. Prediktivni model IE

Prediktivi model IE na utreniranom skupu (engl. Training set)

- Ukupno broj E (Ekstroverata) u skupu je 203
- Ukupni broj I (Introverata) u skupu je 5876
- Ispravne vrijednosti organizirane su u dijagonalnoj liniji od vrha lijevo do dna desno od matrice (4827).
- Klasifikacijska pogreška modela: Model pogrešno predviđa "Ne" u 0,87% slučajeva
- Klasifikacijska pogreška modela: Model pogrešno predviđa "Da" u 0,005% slučajeva

Prema predstavljenoj konfuzijskoj matrici i njezinoj interpretaciji, Random Forest algoritam je za model IE u skupu za treniranje predstavio:

- Za 177 ljudi model je potvrdio da su Ekstroverti
- Za 1226 ljudi koji su bili Ekstroverti model je potvrdio da su Introverti
- Za 26 ljudi koji su bili Introverti model je potvrdio da su Ekstroverti
- Za 4650 ljudi model je predvidio da su Introverti

S obzirom na rezultate klasifikacijskog modela i njegove pogreške, predikcija modela IE pokazala se kao loša predikcija s obzirom na pogrešno predviđanje Introverata u 87% slučajeva.

Prikazana je točnost klasifikacije za ovaj skup predviđanja, koja ukazuje na gotovo 80 % točnosti prediktivnog modela u *training* setu.

Accuracy = total correct predictions/ total predictions made * 100

Accuracy = $\frac{4827}{6079}$ = 79,4 %,

Prediktivi model IE na testiranom skupu (engl. Test set)

- Ukupno broj E (Ekstroverata) u skupu je 73
- Ukupni broj I (Introverata) u skupu je 2523
- Ispravne vrijednosti organizirane su u dijagonalnoj liniji od vrha lijevo do dna desno od matrice (2063).
- Klasifikacijska pogreška modela: Model pogrešno predviđa "Ne" u 0,88% slučajeva
- Klasifikacijska pogreška modela: Model pogrešno predviđa "Da" u 0,002% slučajeva

Prema predstavljenoj konfuzijskoj matrici i njezinoj interpretaciji, Random Forest algoritam je za model IE u skupu za testiranje predstavio:

- Za 68 ljudi model je potvrdio da su Ekstroverti
- Za 528 ljudi koji su bili Ekstroverti model je potvrdio da su Introverti
- Za 5 ljudi koji su bili Introverti model je potvrdio da su Ekstroverti
- Za 1995 ljudi model je predvidio da su Introverti

Prikazana je točnost klasifikacije za ovaj skup predviđanja, koja ukazuje na gotovo 80 % točnosti prediktivnog modela u *test* setu.

Accuracy = total correct predictions/ total predictions made * 100

$$\text{Accuracy} = \frac{2063}{2596} = 79{,}4\ \%$$

13.2 Rezultati modela NS

```
               Type of random forest: classification
                     Number of trees: 301
No. of variables tried at each split: 25

        OOB estimate of  error rate: 13.74%
Confusion matrix:
     N S class.error
N 5236 1 0.000190949
S  834 8 0.990498812
                Test set error rate: 13.64%
Confusion matrix:
     N S class.error
N 2241 0   0.0000000
S  354 1   0.9971831
```

Figure 5. Prediktivni model NS

Prediktivi model NS na utreniranom skupu (engl. Training set)

- Ukupni broj N (Intuition) u skupu je 5590
- Ukupni broj S (Sensing) u skupu je 9

- Ispravne vrijednosti organizirane su u dijagonalnoj liniji od vrha lijevo do dna desno od matrice (5244).
- Klasifikacijska pogreška modela: Model pogrešno predviđa "Ne" u 0,00019% slučajeva
- Klasifikacijska pogreška modela: Model pogrešno predviđa "Da" u 0,99% slučajeva

Prema predstavljenoj konfuzijskoj matrici i njezinoj interpretaciji, Random Forest algoritam je za model NS u skupu za treniranje predstavio:

- Za 5236 ljudi model je potvrdio da su Intuitivni
- Za 1 osobu koja je bila Intuitivna model je potvrdio da je Osjećajna
- Za 834 ljudi koji su bili Osjećajni model je potvrdio da su Intuitivni
- Za 8 ljudi model je predvidio da su Osjećajni

S obzirom na rezultate klasifikacijskog modela i njegove pogreške, predikcija modela NS pokazala se kao loša predikcija s obzirom na pogrešno predviđanje Osjećajnosti u 99% slučajeva.

Prikazana je točnost klasifikacije za ovaj skup predviđanja, koja ukazuje na 86 % točnosti prediktivnog modela u *training* setu.

Accuracy = total correct predictions/ total predictions made * 100

Accuracy = $\frac{4827}{6079}$ = 86,26 %

Rezultat modela NS

Prediktivi model NS na testiranom skupu (engl. Test set)

- Ukupno broj N (Intuicija) u skupu je 2595
- Ukupni broj S (Senzacija) u skupu je 1
- Ispravne vrijednosti organizirane su u dijagonalnoj liniji od vrha lijevo do dna desno od matrice (2242).
- Klasifikacijska pogreška modela: Model pogrešno predviđa "Ne" u 0 % slučajeva
- Klasifikacijska pogreška modela: Model pogrešno predviđa "Da" u 99,71% slučajeva

Prema predstavljenoj konfuzijskoj matrici i njezinoj interpretaciji, Random Forest algoritam je za model NS u skupu za testiranje predstavio:

- Za 2241 ljudi model je potvrdio da su Intuitivni
- Za 1 osobu koja je bila Intuitivna model je potvrdio da je Osjećajnog tipa
- Za 834 ljudi koji su bili Osjećajni model je potvrdio da su Intuitivni
- Za 8 ljudi model je predvidio da su Osjećajni

Prikazana je točnost klasifikacije za ovaj skup predviđanja, koja ukazuje na gotovo 80 % točnosti prediktivnog modela u *test* setu.

Accuracy = total correct predictions/ total predictions made * 100

$$\text{Accuracy} = \frac{2063}{2596} = 79{,}4\ \%$$

13.3 Rezultati modela PJ

```
                    Number of trees: 301
No. of variables tried at each split: 25

        OOB estimate of  error rate: 24.92%
Confusion matrix:
     J    P class.error
J 1081 1326  0.55089323
P  189 3483  0.05147059
               Test set error rate: 25.39%
Confusion matrix:
    J    P class.error
J 440  587   0.5715677
P  72 1497   0.0458891
```

Figure 6. Prediktivni model JP

Prediktivi model JP na utreniranom skupu (engl. Training set)

- Ukupno broj P (Percipiranje) u skupu je 4809
- Ukupni broj J (Prosuđivanje) u skupu je 1270
- Ispravne vrijednosti organizirane su u dijagonalnoj liniji od vrha lijevo do dna desno od matrice (4564).
- Klasifikacijska pogreška modela: Model pogrešno predviđa "Ne" u 55,09 % slučajeva
- Klasifikacijska pogreška modela: Model pogrešno predviđa "Da" u 5,15 % slučajeva

Prema predstavljenoj konfuzijskoj matrici i njezinoj interpretaciji, Random Forest algoritam je za model JP u skupu za testiranje predstavio:

- Za 1081 ljudi model je potvrdio da su Prosuđujući
- Za 1326 osoba koje su bile Prosuđujući model je potvrdio da je Perceptivni

- Za 189 ljudi koji su bili Perceptivni model je potvrdio da su Prosuđujući
- Za 3483 ljudi model je predvidio da su Prosuđujući

Prikazana je točnost klasifikacije za ovaj skup predviđanja, koja ukazuje na 75 % točnosti prediktivnog modela u *training* setu.

Accuracy = total correct predictions/ total predictions made * 100

Accuracy = $\frac{4564}{6079}$ = 75,09 %

Prediktivi model PJ na testiranom skupu (engl. Test set)

- Ukupno broj P (Percipiranje) u skupu je 2084
- Ukupni broj J (Prosuđivanje) u skupu je 512
- Ispravne vrijednosti organizirane su u dijagonalnoj liniji od vrha lijevo do dna desno od matrice (1937).
- Klasifikacijska pogreška modela: Model pogrešno predviđa "Ne" u 57,16 % slučajeva
- Klasifikacijska pogreška modela: Model pogrešno predviđa "Da" u 4,69 % slučajeva

Prema predstavljenoj konfuzijskoj matrici i njezinoj interpretaciji, Random Forest algoritam je za model PJ u skupu za testiranje predstavio:

- Za 440 ljudi model je potvrdio da su Prosuđujući
- Za 587 osoba koje su bile Prosuđujući model je potvrdio da je Perceptivni
- Za 72 ljudi koji su bili Perceptivni model je potvrdio da su Prosuđujući
- Za 1497 ljudi model je predvidio da su Prosuđujući

Prikazana je točnost klasifikacije za ovaj skup predviđanja, koja ukazuje na gotovo 74 % točnosti prediktivnog modela u *test* setu.

Accuracy = total correct predictions/ total predictions made * 100

Accuracy = $\frac{1937}{2596}$ = 74,61 %

13.4 Rezultat modela TF

```
               Type of random forest: classification
                     Number of trees: 301
No. of variables tried at each split: 25

        OOB estimate of  error rate: 19.36%
Confusion matrix:
     F    T class.error
F 2916  373   0.1134083
T  804 1986   0.2881720
                Test set error rate: 18.68%
Confusion matrix:
     F   T class.error
F 1257 148   0.1053381
T  337 854   0.2829555
```

Figure 7. Prediktivni model TF

Prediktivi model TF na utreniranom skupu (engl. Training set)

- Ukupno broj T (Mišljenje) u skupu je 2359
- Ukupni broj F (Osjećaj) u skupu je 3720
- Ispravne vrijednosti organizirane su u dijagonalnoj liniji od vrha lijevo do dna desno od matrice (4902).
- Klasifikacijska pogreška modela: Model pogrešno predviđa "Ne" u 11,34 % slučajeva
- Klasifikacijska pogreška modela: Model pogrešno predviđa "Da" u 28,82 % slučajeva

Prema predstavljenoj konfuzijskoj matrici i njezinoj interpretaciji, Random Forest algoritam je za model TF u utreniranom skupu predstavio:

- Za 2916 ljudi model je potvrdio da su Osjećajni
- Za 373 osoba koje su bile Osjećajni model je potvrdio da pripadaju tipu Mišljenja
- Za 804 ljudi koji su pripadali tipu Mišljenja model je potvrdio da su Osjećajni
- Za 1986 ljudi model je predvidio da su tip Mišljenja

Prikazana je točnost klasifikacije za ovaj skup predviđanja, koja ukazuje na 80 % točnosti prediktivnog modela u *training* setu.

Accuracy = total correct predictions/ total predictions made * 100

Accuracy = $\frac{4902}{6079}$ = 80,6 %

Prediktivi model JP na testiranom skupu (engl. Test set):

- Ukupno broj T (Mišljenje) u skupu je 10021

- Ukupni broj F (Osjećaj) u skupu je 1594
- Ispravne vrijednosti organizirane su u dijagonalnoj liniji od vrha lijevo do dna desno od matrice (2111).
- Klasifikacijska pogreška modela: Model pogrešno predviđa "Ne" u 10,53 % slučajeva
- Klasifikacijska pogreška modela: Model pogrešno predviđa "Da" u 28,29 % slučajeva

Prema predstavljenoj konfuzijskoj matrici i njezinoj interpretaciji, Random Forest algoritam je za model TF u skupu za testiranje predstavio:

- Za 1257 ljudi model je potvrdio da su Osjećajni
- Za 148 osoba koje su bile Osjećajni model je potvrdio da pripadaju tipu Mišljenja
- Za 337 ljudi koji su bili tip Mišljenja model je potvrdio da su Osjećajni
- Za 854 ljudi model je predvidio da pripadaju tipu Mišljenja

Prikazana je točnost klasifikacije za ovaj skup predviđanja, koja ukazuje na gotovo 80 % točnosti prediktivnog modela u *test* setu.

Accuracy = total correct predictions/ total predictions made * 100

Accuracy = $\frac{2111}{2596}$ = 81,32 %

14. Evaluacija klasifikacijskog algoritma Random Forest na modelima

Nakon predstavljenog prediktivnog tipa u modelima IE, NS,TF i JP, algoritam Random Forest specificirao je utrenirane modele. U nastavku, prema slici precizirane su vrijednosti promjena u tipu nakon modeliranja, odnosno postotak ljudi u novom tipu ličnosti

ENFJ	**ENFP**	**ENTJ**	**ENTP**	**ESFJ**	**ESFP**	**ESTJ**	**ESTP**
2.2%	7,78%	2,66%	7,9%	0,46%	0,54%	4,23%	1 %
INFJ	**INFP**	**INTJ**	**INTP**	**ISFJ**	**ISFP**	**ISTJ**	**ISTP**
17%	21,14%	12,6%	15,06%	1,88%	3,12%	2,35%	3,89%

Figure 8. Vrijednosti promjena u tipu ličnosti izražena u postotku

Također, u sljedećoj tablici prikazan je rezultat broja ljudi prema tipovima s obzirom na prediktivni modele algoritma Random Forest. Prikazani su tipovi s najvećim vrijednosnim promjenama.

	ENFJ	ENFP	ENTP	INFJ	INFP	INTJ	INTP	ISTP
ENFJ	0	2	1	5	23	4	22	0
ENFP	0	2	1	24	101	16	58	0
ENTJ	0	1	1	7	35	1	24	0
ENTP	0	2	2	23	105	13	60	0
ESFJ	0	0	0	3	5	2	2	0
ESFP	0	0	0	4	4	4	2	0
ESTJ	0	0	0	1	4	2	4	0
ESTP	0	0	0	2	17	1	6	0
INFJ	0	6	5	55	213	31	131	0
INFP	0	7	4	73	261	45	159	0
INTJ	1	4	0	35	159	27	101	0
INTP	0	4	7	43	182	27	128	0
ISFJ	0	0	1	5	20	9	13	1
ISFP	0	1	1	9	37	3	30	0
ISTJ	0	0	0	8	33	1	19	0
ISTP	0	4	2	18	46	10	21	0

Figure 9. Vrijednosne promjene tipova ličnosti

Prema analizi teksta, rezultati utreniranih modela (IE, NS, JP,TF) algoritma Random Forest ukazuju kako je za 261 osobu koja je prije tekstualne analize bila tip INFP sada također pripada tipu INFP. Nadalje, 213 osoba tipa INFJ pripada tipu INFP, što ukazuje kako je prema tekstualnoj analizi i sentimentima (Affin, Bing i Nrc) model predvidio veću pripadnost u tip P (Percipiranje) nego u J (Prosuđivanje). Za tipove INTP, 182 ljudi model je svrstao u INFP, što ukazuje na veću pripadnost tipu F (Osjećaj) nego tipu T (Mišljenje). Zanimljivost u podacima ogleda se u tipovima ENTP i ENFP koji nakon tekstualne analize pripadaju tipu I (Introvert). Preko 100 ljudi po modelu nisu Ekstroverti (E) već pripadaju tipu Introverti (I).

U nastavku, Isabel Briggs Myers predstavila je distribuciju tipa osobnosti u općoj populaciji. Prema podacima, u općoj populaciji najviše ljudi pripadaju tipu ISFJ za 13,8 % te ESFJ sa 12,3 % te ISTJ sa 11,6 %.

Type	Frequency in Population	
ISFJ	■■■■■■■■■■■■■	13.8%
ESFJ	■■■■■■■■■■■■	12.3%
ISTJ	■■■■■■■■■■■■	11.6%
ISFP	■■■■■■■■■	8.8%
ESTJ	■■■■■■■■■	8.7%
ESFP	■■■■■■■■■	8.5%
ENFP	■■■■■■■■	8.1%
ISTP	■■■■■	5.4%
INFP	■■■■	4.4%
ESTP	■■■■	4.3%
INTP	■■■	3.3%
ENTP	■■■	3.2%
ENFJ	■■■	2.5%
INTJ	■■	2.1%
ENTJ	■■	1.8%
INFJ	■■	1.5%

Figure 10. Distribucija tipa osobnosti u općoj populaciji
*Izvor: "*MBTI Manual*" objavljeno od strane CPP*

Cilj tekstualne analize jest izdvajanje i otkrivanje znanja u podacima identificiranjem koncepata, pronalaženje odnosa u tekstovima, otkrivanjem implicitnih veza i stvaranjem hipoteza koje potvrđujemo ili opovrgavamo, ovisno o rezultatima. Iako se Random Forest pokazao kao dobar prediktivni model, prvi razlog pobijanja H1 hipoteze lezi u činjenici da je nemoguće generalizirati zaključke s obzirom na količinu podataka o tipu MBTI i njihovim postovima, stoga ujedno vežemo pobijanje hipoteze s problemom pristranosti podataka. Nadalje, prediktivni model za svaki tip osobnosti ukazuje na visoku točnost modela, što potvrđuje činjenicu kako bi nam upravo ova analiza metodom Random Forest, da je napravljena na većem skupu podataka mogla dati jasnije i preciznije informacije o predviđanju tipa ličnosti putem objava po tipu.

15. Zaključak

Razvoj podatkovne znanosti, strojnog učenja, ekstrakcije znanja iz informacija doveli su do napretka tehnika s kojima je moguće otkrivati odnose i obrasce u svijetu personalizacije potrošača**.** Tehnologija nastavlja napredovati te nam omogućava razvoj svijeta personalizacije. Organizacije ulažu veliki napor prilikom pronalaženja novih i zanimljivih pružanja jedinstvenih iskustava za svoje klijente pritom koristeći nova znanja strojnog učenja, implementirajući ih u svoje svakodnevno poslovanje. Strojno učenje i prediktivna analiza koristi se kako bi organizacije mogle identificirati obrasce, prilike ili probleme na temelju podataka koje posjeduju. Nova tehnološka postignuća utjecala su na primjenu i razvoj tehnika strojnog učenja koja imaju primjenu u svim industrijama, a neke od izdvojenih su: sigurnost podataka, osobna sigurnost, financijsko trgovanje, zdravstvo, personalizacija u marketingu, otkrivanje prijevara kreditnih kartica, sustavi preporuke, automobilska industrija- *Smart cars* i ostali. U svim navedenim primjerima, tehnikama strojnog učenja izrađuju se predikcije, a odnosi se na široku klasu metoda koje se usmjerene na modeliranje podataka na algoritamskim predviđanjima i na algoritamskom dešifriranju obrasca u podacima. S obzirom na rast korisnika interneta u posljednjem desetljeću te njihovo umrežavanje na raznim društvenim medijima, potreba za razumijevanje tekstualnih podataka s društvenih medija odgovorila bi na široku paletu pitanja o potrošačima, robnim markama, proizvodima, ponašanju korisnika itd. Zbog velike količine podataka nužno je procesu personalizacije na temelju rudarenja podataka pristupiti na valjani način. Analiza teksta ili obrada prirodnog jezika predstavlja jedan od načina putem kojeg se podaci mogu korisno i pametno analizirati, razumjeti i izvesti značenje iz ljudskog jezika.

Glavno pitanje koje se postavlja što možemo učiniti s tekstualnom analizom te kako je primijeniti na analizu društvenih medija. Uvažavajući pravna i etička ograničenja, podaci se mogu koristiti za ekstrakciju informacija prilikom profiliranja korisnika u marketinške ali i druge svrhe. Stoga je cilj ovog rada bio povezati istraživanja u psihologiji na temu osobnosti i objava na društvenim medijima, kvantificirati i analizirati podatke koji će nas usmjeriti na pravovaljani prediktivni model putem kojeg će se odgovoriti na glavnu tezu, može li se analizom postova na forumu predvidjeti osobnost sukladno glavnim psihometrijskim modelima klasifikacije osobnosti. Prikupljeni su podaci s platforme Kaggle o tipu osobnosti po Myers-Briggs te njihove posljednje objave na društvenim medijima. Nadalje, u radu su opisani sustavi

personalizacije, predstavljene njihove prednosti i nedostaci kao i izazovi u pristupu te primjenu u praksi.

Kako bi se dobio skup značajki potrebni za preciznu analizu nužno je bilo kvantificirati i analizirati podatke kroz alat koji vrši analizu teksta. Također, nad podacima korištene su tehnike pret-procesiranja koje su usredotočene na identifikaciju i ekstrakciju tekstualnih podataka koje su potom transformirane u strukturirani set. Neke od korištenih metoda uključivale su metodu "Vreća riječi" (engl. *Bag of Words*), Tokenizacija, metoda filtriranja Zaustavnih riječi (engl. *Stop Words*). Nakon provedenih tehnika pret-procesiranja, tekstualnom ekstrakcijom predstavljena je frekvencija 100 najzastupljenijih riječi u setu podataka, objavama korisnika. Nadalje, provedena je sentiment analiza, ekstrakcija informacija koje pomažu u izdvajanju mišljenja ili sentimenta s ciljem utvrđivanja pozitivnih, negativnih ili neutralnih mišljenja. Veoma poznata metoda klasifikacije, stablo odlučivanja primijenjena je na setu utreniranih podataka s ciljem analiziranja potencijalne vrijednosti seta podataka, odnosno ima svrhu prepoznavanja tipova u MBTI postovima. Ograničenja koje Stablo odlučivanja kao klasifikacijski algoritam predstavlja ogleda se u činjenici kako ono ne može "podnijeti" preveliki broj inačica te je upravo iz tog razloga odabrana tehnika modeliranja Random Forest koja je namijenjena radu s većim brojem inputa. Sljedeći korak u analizi jest računanje prototipova koji će dati uvid u odnos između varijabli i klasifikacije. Izgrađeni su prediktivni modeli za sve tipove MBTI –a, model IE, NS, PJ i TF koji su nam dali uvid u zastupljenost tipova. Rezultati dobivenih modela pobijaju postavljenu hipotezu o predviđanju MBTI ličnosti na temelju postova na forumu. Iako se Random Forest pokazao kao dobar prediktivni model, nemoguće je, s obzirom na količinu podataka o MBTI i njihovim postovima generalizirati zaključke. Unatoč činjenici da svaki prediktivni tip osobnosti ukazuje na visoku točnost modela, trenutno možemo samo pretpostaviti“ kako bi na većem skupu podataka naš model ostvario mogućnost predviđanja tipa osobnosti.

„Pod punom odgovornošću pismeno potvrđujem da je ovo moj autorski rad čiji niti jedan dio nije nastao kopiranjem ili plagiranjem tuđeg sadržaja. Prilikom izrade rada koristio sam tuđe materijale navedene u popisu literature ali nisam kopirao niti jedan njihov dio, osim citata za koje sam naveo autora i izvor te ih jasno označio znakovima navodnika. U slučaju da se u bilo kojem trenutku dokaže suprotno, spreman sam snositi sve posljedice uključivo i poništenje javne isprave stečene dijelom i na temelju ovoga rada“.

U Zagrebu, 26.02.2018.

Popis slika

Slika 1. Metoda BoW na primjeru objave 31

Slika 2. Frekvencija riječi u postu 35

Popis grafikona

Grafikon 1. Kretanje korisnika društvenih medija od 2010. do 2021. 14

Grafikon 2. Analiza sentimenta riječi po leksikonu Affin 37

Popis tablica

Figure 1. Model Stabla odlučivanja nad odabranim entitetima 40

Figure 2. Model Random Forest nad nasumičnim setom podataka 42

Figure 3. Model Random Forest na nasumičnom setu podataka II. 44

Figure 4. Prediktivni model IE 46

Figure 5. Prediktivni model NS 47

Figure 6. Prediktivni model JP 49

Figure 7. Prediktivni model TF 51

Figure 8. Vrijednosti promjena u tipu ličnosti izražena u postotku 53

Figure 9. Vrijednosne promjene tipova ličnosti 53

Figure 10. Distribucija tipa osobnosti u općoj populaciji 54

Literatura

[1] Mulvenna, M., Anand, S.S., Buchner, (2000.), A.G.: Personalization on the net using web mining. Communication of ACM 43(8)

[2] (2018-01-17) https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/, 2018.

[3] L. Saulsman and A. Page, (2004.), The five-factor model and personality disorder empirical literature: A meta-analytic review* 1. Clinical Psychology Review, 23(8):1055–1085,

[4] T. Judge, C. Higgins, C. Thoresen, and M. Barrick, (1999.), The big five personality traits, general mental ability, and career success across the life span. Personnel psychology, 52(3):621–652,

[5] P. Shaver and K. Brennan., (1992.), Attachment styles and the "Big Five" personality traits: Their connections with each other and with romantic relationship outcomes. Personality and Social Psychology Bulletin, 18(5):536

[6] A. Chin, B. Xu and H. Wang, (2013.), "Who Should I Add as a "Friend"? A Study of Friend Recommendations using Proximity and Homophily", Proc. the 4th International Workshop on Modeling Social Media, Article 7

[7] A.T. Fiore, L.S. Taylor, G.A. Mendelsohn and M. Hearst, (2008.), "Assessing Attractiveness in Online Dating Profiles", Proc. the SIGCHI Conference on Human Factors in Computing System, pp. 797-806,

[8] S. Picazo-Vela, M. Fernandez-Haddad and L. F. LunaReyes, "IT's alive!! Social Media to Promote Public Health", Proc. the 14th Annual International Conference on Digital Government Research, pp. 111-119, 2013.

[9] P. Brusilovsky, A. Kobsa, W.Nejdl, (2007.), "Methods and Strategies of Web Personalization", Information Systems and Applications, incl. Internet/Web, and HCI, Springer, p. 697-719.

[10] (2018-01-17) https://hackernoon.com/spotifys-discover-weekly-how-machine-learning-finds-your-new-music-19a41ab76efe

[11] Robert Hoekman Jr., (2010.), Designing the Obivious: A Common Sense Approach to Web & Mobile Application Design (2 nd edition)

[12] Bamshad Mobasher, Robert Cooley, Jaideep Srivastava, (2000.), "Automatic Personalization based on web usage mining", Communications of the ACM, Vol. 43 No. 8, Pages 142-151

[13] Kohavi, R., Provost, F.: Applications of data mining to electronic commerce. Data Mining and Knowledge Discovery 5(1–2) (2001) 5–10

[14] (2018-01-17) https://www.packtpub.com/mapt/book/big_data_and_business_intelligence/9781787121423/7/ch07lvl1sec71/challenges-for-the-rule-based-system

[15] Salton, G., McGill, (1983.), M: Introduction to Modern Information Retrieval. McGraw-Hill, New York, NY

[16] (2018-01-17) https://www.tidytextmining.com/index.html

[17] (2018-01-17) https://export-x.com/2013/12/15/many-products-amazon-sell/

[18] https://www.upwork.com/hiring/data/what-is-content-based-filtering/

[19] Sinha, R., Swearingen, K., (2001.), Comparing recommendations made by online systems and friends. In: Proceedings of Delos-NSF Workshop on Personalization and Recommender Systems in Digital Libraries

[20] (2018-01-17) https://www.upwork.com/hiring/data/how-collaborative-filtering-works/

[21] Boyd, d.m., & Ellison, N.B. (2007.), Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication, 13, 210–230.

[22] Manago, A.M., Graham, M.B., Greenfield, P.M., & Salimkhan, G. (2008.), Self-presentation and gender on MySpace. Journal of Applied Developmental Psychology, 29, 446–458

[23] Ambady, N., & Skowronski, J. (Eds.). (2008). First impressions. New York: Guilford., Funder, D.C. (1999). Personality judgment: A realistic approach to person perception. San Diego, CA: Academic Press,

Hall, J.A., & Bernieri, F.J. (Eds.). (2001.), Interpersonal sensitivity: Theory and measurement. New York: Erlbaum

Vazire, S., & Gosling, S.D. (2004.), E-perceptions: Personality impressions based on personal websites. Journal of Personality and Social Psychology, 87, 123–132.

[24] Alan E. Kazdin, PhD, Encyclopedia of Psychology: 8 Volume Set, 2000.

[25] Li L, Li A, Hao B, Guan Z, Zhu T (2014.), Predicting Active Users' Personality Based on Micro-Blogging Behaviors.

[26] John, O.P. i Srivastava, S. (1999.), The Big five trait taxonomy: History, measurement, and theoretical perspectives. U: L.A. Pervin i O.P. John (Ur.), Handbook of personality (str. 102-138). New York: The Guilford Press.

[27] Read, Stephen J., Brian M. Monroe, Aaron L. Brownstein, Yu Yang, Gurveen Chopra, and Lynn C. Miller, (2010.), "A Neural Network Model of the Structure and Dynamics of Human Personality." Psychological Review 117.1: 61-92.

[28] Adelstein, Jonathan S. et al., (2011.), Personality Is Reflected in the Brains Intrinsic Functional Architecture. Ed. Mitchell Valdes-Sosa. PLoS ONE 6.11: e27633. PMC

[29] Schwartz HA, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones SM, Agrawal M, et al. (2013.), Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach.

[30] Liu L, Pietro D, Samani Z, Moghaddam M, Ungar L., (2016.), Analyzing Personality through Social Media Profile Picture Choice. AAAI Digital Library

[31] McAuley J., Leskovec J., (2013.) Learning to Discover Social Circles in Ego Networks Stanford, USA

[32] (2018-01-17) https://www.16personalities.com/articles/our-theory

[33] (2018-01-17) https://hr.wikipedia.org/wiki/Ambivalentnost

[34] (2018-01-17) http://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/

[35] Robert M. Capraro and Mary Margaret Capraro, Myers-Briggs Type Indicator Score Reliability Across: Studies a Meta-Analytic Reliability Generalization Study, Educational and Psychological Measurement 2002.;
http://people.wku.edu/richard.miller/MBTI%20reliability%20validity.pdf

[36] Myers, I. B., & McCaulley, M. H. (1985.), Manual: A guide to the development and use of the Myers-Briggs Type Indicator. Palo Alto, CA: Consulting Psychologists Press.

[37] Harvey, R. J. (1996.), Reliability and validity. In A. L. Hammer (Ed.), MBTI applications: A decade of research on the Myers-Briggs Type Indicator (pp. 5-29). Palo Alto, CA: Consulting Psychologists Press.

[38] Myers, I. B., & McCaulley, M. H. (1985.), Manual: A guide to the development and use of the Myers-Briggs Type Indicator. Palo Alto, CA: Consulting Psychologists Press.

[39] Myers, I. B., & McCaulley, M. H. (1989.), Manual: A guide to the development and use of the Myers-Briggs Type Indicator. Palo Alto, CA: Consulting Psychologists Press

[40] Furnham, Adrian, Moutafi, Joanna, Crump, John, (2003.), The Relationship between the Revised NEO-Personality Inventory and the Myers-Briggs Type Indicator, Social Behavior and Personality: an international journal

[41] (2018-01-17) https://en.wikipedia.org/wiki/Text_mining

[42] (2018-01-17) http://hlwiki.slais.ubc.ca/index.php/Text-mining

[43] R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org

[44] (2018-01-17) https://www.rstudio.com/about/

[45] (2018-01-17) https://www.kaggle.com/datasnaek/mbti-type

[46] J. Brownies, A Gentle Introduction to the Bag-of-Words Model, 2017. https://machinelearningmastery.com/gentle-introduction-bag-words-model/

[47] (2018-01-17) http://www.mjdenny.com/Text_Processing_In_R.html

[48] (2018-01-17) https://techcrunch.com/2017/03/07/google-is-acquiring-data-science-community-kaggle/

[49] (2018-01-17) https://cran.r-project.org/web/packages/text2vec/vignettes/text-vectorization.html

[50] (2018-01-17) http://www.expertsystem.com/10-text-mining-examples/

[51] S. Feldman, (2012.) The Answer Machine: Synthesis Lectures on Information Concepts, Retrieval, and Services, Morgan & Claypool Publishers

[52] (2018-01-17) http://www.admantx.com/

[53] Apte and Weiss, (1997) C. Apte and S. Weiss. Data Mining with Decision Trees and Decision Rules. Future Generation Computer Systems, 13:197- 210

[54] (2018-01-17) https://dataaspirant.com/2017/05/22/random-forest-algorithm-machine-learing/