Predikcija popularnosti Instagram objava

Olivera Blagojević
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad
olja_blagojevic@hotmail.com

Marko Dobrić
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad
dobric.marko23@gmail.com

Milica Damjanović
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad
milicaa031@gmail.com

***Apstrakt* — Zahvaljujući ekspanziji Interneta i društvenih mreža, danas veliki broj ljudi odlučuje da pokreće raznorazne poslove (prodaja, promocija, pisanje) uz pomoć najpopularnije društvene mreže – Instagram. Kako bi Instagram bio efikasan u marketinške svrhe i u tom smislu donio zaradu onima koji to žele, potrebno je imati popularne objave, tj. objave sa velikim brojem lajkova, što je znak da ih je veliki broj ljudi vidio, a to samim tim donosi i veliki broj pratilaca.**

U ovom radu su predstavljene tehnike predikcije popularnosti Instagram objava, kako bismo istražili to tržište, ali i pomogli zainteresovanim korisnicima da vide od čega sve zavisi broj lajkova (popularnost) na određenoj Instagram objavi. Korišćeni su Decision Tree, Random Forest i K-Nearest Neighbor (KNN) algoritmi i njihovi modeli za predikciju broja lajkova na Instagram objavi. Za kategorizaciju hashtag-ova na objavama i njihovu analizu korišćene su Natural Language Processing (NLP) tehnike. Takođe, korišćene su metode za određivanje važnosti i povezanosti atributa i za izračunavanje grešaka.

Na kraju su predstavljeni dobijeni rezultati, zaključci ali i predlozi i pravci budućeg rada, koji su se u toku procesa rada na ovom istraživanju sami nametnuli i pokazali bitnim za dalju predikciju.

Ključne riječi— Instagram; objava, post; hashtag; pratioci; lajk, lajkovi

I. UVOD

Instagram je besplatna i, trenutno, najpopularnija društvena mreža, koja svojim korisnicima omogućava obradu i dijeljenje fotografija i videa (dužine do 15 sekundi na Instagram priči i do jedan minut na objavi) putem Android i iOS platformi. Instagram je kreiran i pokrenut u oktobru 2010. godine, a u aprilu 2012. godine je kupljen od strane kompanije Facebook, Inc. Online servis je veoma brzo pridobio veliku popularnost, sa više od 100 miliona aktivnih korisnika u februaru 2013. godine, a aplikacija je četvrta najviše preuzimana aplikacija 2010-ih godina. [1]

Trendovi i popularnost u velikoj mjeri upravljaju društvenim medijima u modernom dobu. Danas Instagram ima preko milijardu korisnika, postao je tržište velikih razmjera sa potencijalom da se optimizuje za povećanje popularnosti, gledanosti i prihoda. U ovom istraživanju smo se bavili predikcijom popularnosti Instagram objava. Pod „popularnošću“ se u ovom slučaju smatra broj lajkova na objavama korisnika. Predikcija je vršena na osnovu analize profila Instagram influensera (eng. *influencer*), prehodnih uspješnih objava ovih korisnika (veliki broj lajkova), na i tagova koji su korišćeni. Cilj istraživanja je identifikovanje ključnih varijabli za povećanje odnosa broja lajkova na objavi do broja prosječnih lajkova koje korisnik dobije na objavama. U radu su predstavljene metodologije koje su nam pomogle da dođemo da zaključka kako bi objave mogle biti optimizovane da dobiju veću pažnju Instagram korisnika, kao i neočekivani šabloni u karakteristikama popularnih Instagram objava.

II. SLIČNA ISTRAŽIVANJA

Iako Instagram platforma postoji već deset godina, trenutno nema mnogo istraživanja koja se bave ovom temom, ali zato postoje mnoga istraživanja koja se bave predikcijom generalno, kao i predikcijom popularnosti korisničkih profila na društvenim mrežama. Istraživanja koja su koristila za ovaj rad su ona koja sadrže analizu društvenih mreža koje imaju opciju postavljanja fotografija (kao Instagram) i predikciju popularnosti tih fotografija i korisnika generalno.

U [2] je istraživano šta stoji iza popularosti određenih korisnika, prikupljani su podaci i na osnovu njih je rađena predikcija. Podaci koji su prikupljeni su ručno sastavljeni i sadrže 441 korisnika Instagrama i 60.785 parova fotografija sa opisima. Obilježja koja su korišćena su: korisničko ime, broj pratilaca, slika, opis slike i broj lajkova. Za istraživanje popularnosti objave jednog odeđenog korisnika korišćene su metode neuronskih mreža na neobrađenim podacima: ResNet-50, LSTM. Performanse različitih modela su procjenjivane pomoću četiri metrike: preciznost, povlačenje (eng. *recall*), F-mjera i tačnost. Validacija rezultata je vršena na osnovu broja lajkova, koji je korišćen kao indeks za mjerenje popularnosti. Da bi se uzela u obzir popularnost svakog korisnika pojedinačno, selektovano je 25% pozitivnih uzoraka (popularnih objava) i 25% negativnih uzoraka.

Nasumično je odabrano 20% tih objava kao test skup i nasumično je podijeljeno 10% trening skupa kao validacijski skup za odlučivanje hiper-parametara. Dobijen je skup od 21874 para fotografija sa opisima za trening set, 2430 parova fotografija sa opisima za validaciju i 6064 para fotografija sa opsima za test set. Odnos između pozitivnih i negativnih uzoraka je 1:1. U ovom istraživanju na kraju vidimo ostvarene rezultate korišćenih algoritama, ali ne i same predikcije. Autori su naveli da su na kraju analizirali fotografije i tekstualne informacije na osnovu statističkih rezultata i izvukli zaključke o povezanosti fotografije, opisa i popularnosti, ali tek u budućnosti namjeravaju razviti efikasniji model, koji bi obuhvatao korisničko okruženje.

Tema u [3] je prezentovanje pristupa za identifikaciju aspekata koji odlučuju popularnost objave na društvenoj mreži, a fokus istraživanja je na bjavama određenih brendova, kako bi se pomoglo kompanijama u ostvarivanju efikasnog marketinga. Analizom je utvrđeno da je najpopularnija kategorija na Instagramu brza hrana (eng. *fast food*). Obzirom da je autore rada zanimao brend, pronašli su skupove podataka šest najpopularnijih kompanija brze hrane, na osnovu godišnjeg prihoda iz 2014. godine, a to su bili: McDonald's, Burger King, Culver's, Wendy's, Sonic Drive In i Jack In The Box. Iz ovoga su sastavili skup podataka pretraživanjem 75.000 objava korisnika vezanih za ovih šest kategorija. Obilježja koja su korišćena u dobijenom skupu podataka su: brend (binarna indikacija koja se odnosi na logo brenda), lica (broj detektovanih lica na slici), proizvod (da li je proizvod korišćen na slici), osoba-proizvod (binarna indikacija slučaja u kome su na slici jedna osoba i jedan proizvod), osobe-proizod (vrijednost ovog parametra je 1 ako su na slici dva ili više lica i proizvod), sentiment (vizuelni ili tekstualni), estetika slike (42-dimenzionalni binarni vektor sa informacijama o 42 Instagram filtera) i broj pratilaca. Korišćen je algoritam Vector Regression (SVR) i k-fold cross-validation tehnika validacije. Za validaciju su napravljena tri eksperimenta u kojima se analizira uticaj multimodalnih *feature*-a na popularnost objave, uticaj parametara angažovanja (eng. *engagement parameters*) na popularnost objave i analiza parametara angažovanja na svaku kategoriju posebno. Nakon predikcije popularnosti svake objave u testnom vremenu, računali su Spirmenov rang korelacije (eng. Spearman's rank corellation) između predikcije i istinitosti, koja vraća vrijednost između [-1, 1]. Vrijednost blizu 1 odgovara idealnoj korelaciji. Rezultati eksperimenata pokazuju da predikcija popularnosti korišćenjem parametara angažovanja nadmašuje direktno modelovanje popularnosti korišćenjem vizuelnih i tekstualnih *feature*-a samo za faktor od dva. Parametri angažovanja, kao što su postojanje logoa brenda, sentimenta i fotografije, igraju važnu ulogu u predviđanju popularnosti objave. Takođe, rezultati su pokazali da korišćenjem korisničkih parametara angažovanja dobijamo mogućnost ne samo boljeg predviđanja popularnosti objave, nego i isticanja svojstava specifičnih za svaku kategoriju brenda.

U istraživanju [4] su analizirani specifični vizuelni sentimenti i tema je na razumijevanju njihovog pozitivnog ili negativnog uticaja na eventualnu popularnost fotografija na društvenim mrežama. Korišćena su dva skupa podataka:

- po korisniku – nasumično odabrano 520 hiljada fotografija iz VSO Flickr Dataset-a (skup podataka koji predstavlja scenario pretraživanja, gdje fotografije pripadaju različitim korisnicima);
- specifikacije korisnika: 25 korisnika iz VSO Flickr skupa podataka se bira nasumično kako bi sačinili 25 različitih ispitivanja, za svako se bira nasumično 10 hiljada fotografija (skup podataka koji predstavlja scenario za korisnika koji želi da selektuje svoje fotografije koje trebaju biti postavljene da privuku veću pažnju drugih korisnika).

Za ovo je korišćen SVR (Support Vector Regression) algoritam. Izvedeni su eksperimenti za:

- vizuelna obilježja;
- kontekstna obilježja;
- kontekstna obilježja i vizuelna obilježja;
- kombinacije korisnika.

Obzirom da je teško precizno definisati jedan rezultat kao mjeru popularnosti, predložili su nekoliko načina. Koristili su broj pregleda na Flickr-u kao osnovnu metriku. Osim toga, koristili su i broj pregleda i broj komentara na svakoj fotografiji zajedno, jer su istraživanjem pronašli njihovu korelaciju sa popularnošću videa. Metrika popularnosti im je bila broj pregleda na Flickr-u, pa su, da bi se lakše nosili sa velikim varijacijama pregleda, podijelili metriku popularnosti na razliku vremena između korisničkog postavljanja objave i njihovog pronalaženja, i na to su dodali log funkciju.

Eksperimenti sugerišu da su određeni sentimenti u korelaciji sa popularnošću manji nego specifikacije korisnika. Popularnost fotografija je visoko povezana sa popularnošću samog korisnika.

III. SKUP PODATAKA

Skup podataka koji je korišćen u istraživanju je ručno sačinjen i sadrži sve potrebne informacije o objavama popularnih Instagram korisničkih profila, čime je uspostavljena veza između popularnosti objave i popularnosti samog korisnika.

A. *Formiranje skupa podataka*

Za ručno formiranje skupa podataka su korišćeni podaci o popularnim Instagram korisničkim profilima, preuzeti sa Instagrama i sajtova koji sadrže liste najpopularnijih Instagram korisnika, i podaci o 17 posljednjih objava svakog od tih korisnika. Dobijeni skup podataka se sastojao od preko 16 hiljada objava, 972 Instagram korisnika.

B. *Filtriranje i trasformacija*

Kako bi bilo otklonjeno prisustvo „prljavih podataka“ (eng. *dirty data*), korišćena je desktop aplikacija Open Refine, pomoću koje su izbrisani prazni redovi, dupli podaci i specijalni karakteri i interpunkcijski znaci, koji su ometali dalji rad sa podacima. Skup podataka je potom transformisan u .xlsx datoteku, kako bi rad bio dalje pojednostavljen. Ovim je izgubljeno manje od stotinu podataka i dobijen je konačni skup od 16540 podataka. [5]

Nasumično je izabrano 80% podataka (13232) za obučavanje, kao trening podaci, a 20% (3308) za procjenu tačnosti, kao test podaci.

C. *Obilježja*

Atributi koje sadrži skup podataka nakon formiranja su sljedeći:

- numberPosts* – broj postova određenog korisnika;
- website – link ka web stranici korisnika;
- urlProfile – link ka profilu;
- username – korisničko ime;
- numberFollowing* – broj ljudi koje korisnik prati;
- description – opis profila;
- alias – ime osobe;
- numberFollowers* – broj pratilaca;
- urlImgProfile – link ka profilnoj fotografiji korisnika;
- filename - naziv postavljene fotografije;
- date - datum postavljanja fotografije;
- urlImage – link ka fotografiji na koju se odnosi podatak;
- mentions – označeni profili na fotografiji;
- multipleImage – da li objava sadrži više fotografija;
- isVideo – da li je objava video ili fotografija;
- localization – lokacija objave;
- url – Instagram link ka objavi na koju se odnosi podatak;
- numberLikes* – broj lajkova na objavi;
- description – opis fotografije;
- tags* – nabrojani *hashtag*-ovi za svaku objavu.

Atributi označeni sa zvjezdicom (*) su smatrani ključnim za predikciju.

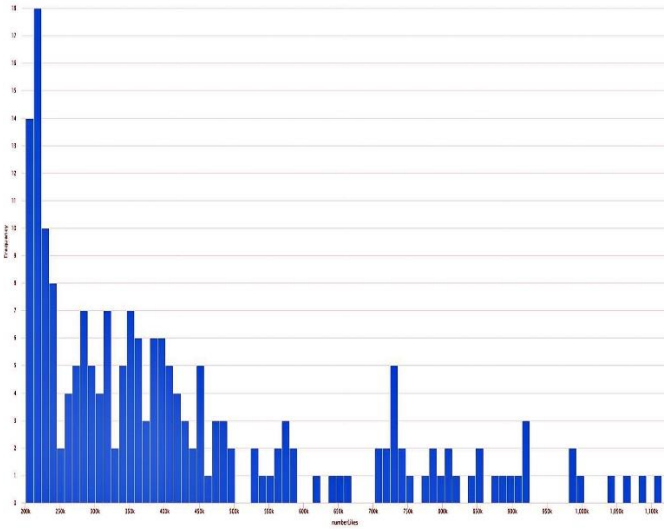
Skup podataka je obogaćen dodatnim obilježjima, a to su:

- country – država u kojoj korisnik živi;
- numberOfTags* – broj *hashtag*-ova.

Dobili smo ukupno 22 obilježja.

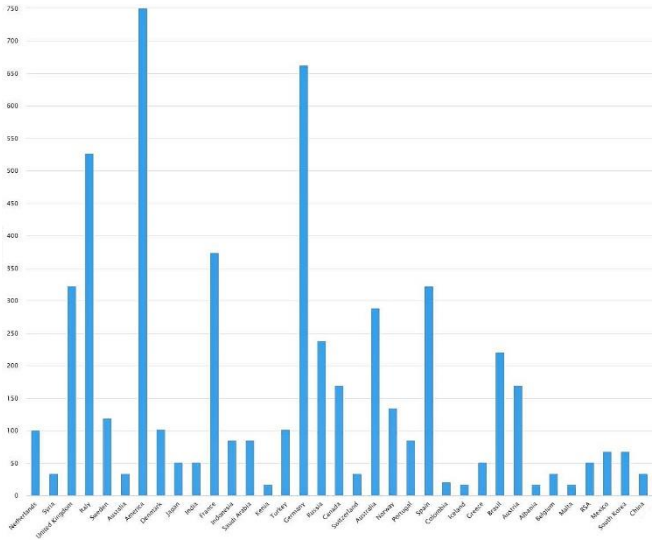
Geografske lokacije korisnika su pretraživane i unošene ručno, a broj *hashtag*-ova je izračunat. Prilikom računanja broja *hashtag*-ova, dodata je nova kolona u Excel datoteci u kojoj je korišćena formula za računanje broja riječi u svakom redu selektovane kolone; svaka pojedinačna riječ je na Instagram objavi bila jedan *hashtag*. Ovim je dobijen broj *hashtag*-ova na svakoj objavi, ali i ukupni broj *hashtag*-ova u skupu podataka, a to je 39254. Iz ovoga se odmah moglo zaključiti da je u prosjeku bilo 2.5 *hashtag*-a po objavi, odnosno da je skoro svaki popularni korisnik Instagrama koristio *hashtag*.

Nakon kreiranja svih obilježja, napravljena je statistika za neka od ključnih obilježja i predstavljena je na slikama 1, 2 i 3:



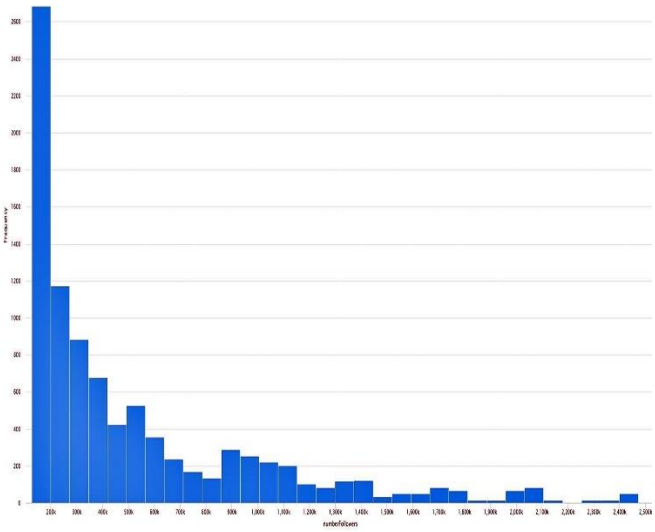
Slika 1: Broj objava sa preko 200.000 lajkova

Na Instagramu je mnogo više fotografija sa do dvije hiljade lajkova, dok je broj onih koje imaju preko 200.000 lajkova znatno manji.



Slika 2: Broj korisničkih profila po državama

Ubjedljivo najviše Instagram korisničkih profila je iz Sjedinjenih Američkih Država, a na drugom mjestu, sa znatno manje korisničkih profila, je Njemačka, a zatim Italija.



Slika 3: Broj korisničkih profila sa preko 250.000 pratilaca

Preko 2600 korisnika ima preko 250.000 pratilaca na Instagramu, dok je znatno manji broj onih koji imaju preko dva i po miliona pratilaca.

IV. METODOLOGIJA

U metodologiji ovog rada plan je bio ustanoviti uticaje parametara jednih na druge i uticaje određenih parametara na broj lajkova, koji predstavlja popularnost određene objave. Stoga je metodologija podijeljena na četiri dijela:

- Prvi dio - Linearna regresija, pomoću koje je namjera bila da se uspostave veze među parametrima;
- Drugi dio - Primijenjeni algoritmi, gdje smo ispitivali kako korišćeni algoritam može uticati na predikciju;
- Treći dio - *Natural Language Processing* (NLP) tehnike nad *hashtag*-ovima objavljenih fotografija, kako bismo kategorisali *hashtag*-ove i iz dobijenih kategorija napravili nove skupove podataka nad kojima je rađena predikcija;
- Četvrti dio - *Permutation Feature,* za utvrđivanje važnosti svakog pojedinačnog atributa.

A. *Linearna regresija*

Regresija je jedna od metoda u okviru metoda koje čine statističko učenje (eng. *statistical learning*) i čini je veliki skup metoda, tehnika, alata statistike za modelovanje i razumijevanje kompleksnih skupova podataka. Predstavlja jedan od načina da se sastavi model za predviđanje i ocjenjivanje jedne ili više zavisnih promjenljivih na osnovu jedne ili više nezavisnih promjenljivih. Za razliku od drugih statističkih metoda, u regresiji postoji izlaz, tj. postoji zavisna promjenljiva. Kod regresione analize se ispituju zavisnosti između dvije ili više promjenljivih, odnosno sagledava se uticaj promjene jedne ili više promjenljivih na promjenu drugih promjenljivih i potrebno je unapred odrediti koja pojava će imati ulogu zavisne promjenljive a koja nezavisne promjenljive. [6] Linearna regresija se odnosi na svaki pristup modelovanja relacija između jednog ili više responsa (zavisnih promjenljivih) i jedne ili više nezavisnih promjenljivih, na način da takav model linearno zavisi od nepoznatih parametara procijenjenih iz podataka. [7]

Cilj regresione analize u ovom radu je bio da se pomoću njenih statističkih metoda otkrije da li postoje veze između posmatranih parametara (obilježja u skupu podataka). Kod linearne regresije su korišćeni svi podaci. Međutim, u ovom slučaju, dobijena je greška root_mean_squared_error: 69031.069. *Root mean squared error* (RMSE) ili srednja kvadratna grešla je inače mjera nepouzdanosti rezultata mjerenja. To je standardna devijacija reziduala (grešaka predviđanja), koji predstavljaju mjeru koliko daleko su tačke podataka od regresione linije. RMSE je zapravo mjerilo koliko su prošireni ovi reziduali, tj. govori nam koliko su podaci koncentrisani oko linije najboljeg prilagođavanja (eng. *line of best fit*). Iz dobijene srednje kvadratne greške, u ovom slučaju, je zaključeno da tačan model linearne regresije nije moguće napraviti, jer linearna zavisnost posmatranih parametara ne postoji. [8]

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value
numberPosts	0.148	0.269	0.005	0.999	0.552	0.518
numberOfTags	-1594.741	156.149	-0.102	0.999	-10.213	
(Intercept)	29201.192	1023.956	-	-	28.518	

Tabela 1: Rezultat linerane regresije (greška)

U tabeli su prikazani atributi: broj objava (numberPosts) i broj tagova (numberOfTags). Koeficijenti regresije (Coefficient) su procjene nepoznatih parametara populacije i opisuju vezu između predikcijske promjenljive i odgovora. U linearnoj regresiji koeficijenti su vrijednosti koje umnožavaju vrijednosti predikcije. Standardna greška regresije (Std. Error) pruža apsolutnu mjeru tipične udaljenosti pada tačaka podataka sa linije regresije; S je u jedinicama zavisne promjenljive. Standardni regresioni koeficijent (Std. Coefficient), koji se takođe naziva i beta koeficijent, je procjena koja rezultuje iz standardizovane regresione analize, tako da varijacije zavisnih i nezavisnih promjenljivih predstavljaju 1. Tolerancija (Tolerance) mjeri uticaj jedne nezavisne promjenljive na sve druge nezavisne promjenljive i računa se prilikom početne analize linearne regresije. Statistika t-vrijednosti (t-Stat) je koeficijent podijeljen svojom standardnom greškom. P-vrijednost (p-Value) za svaki slučaj testira nultu hipotezu da je koeficijent jednak nuli (nema efekta). Obzirom da je p-Value veliko, na osnovu toga smo zaključili da treba odbaciti linearnu regresiju.

$$numberLikes = 0.148 * numberPosts - 1594.741 * numberOfTags + 29201.192$$

Koeficijenti prikazuju da je broj lajkova (numberLikes) u vezi, ali da je greška za tu formulu prevelika.

B. *Primijenjeni algoritmi*

Za ovo istraživanje su korišćeni algoritmi za koje smo smatrali da bi mogli biti korisni za predikciju popularnosti Instagram objave, kao i za utvrđivanje bitnih atributa koji utiču na popularnost Instagram objave.

Prvo je korišćeno stablo odlučivanja (eng. *decision tree*), kako bismo vidjeli strukturu podataka koje imamo i parametre od kojih bi popularnost mogla zavisiti. Decision tree model je predložio broj pratilaca (numberFollowers) kao bitan atribut, koristeći GINI index. Nakon broja pratilaca, kao bitni parametri su izdvojeni broj lajkova (numberLikes), za koji smo mi očekivali da će biti prvi parametar, i državu (Country) iz koje korisnik dolazi, što takođe nismo očekivali jer smo smatrali da će više uticaja imati numerička obilježja.

Stablo odlučivanja je zapravo dio Random Forest modela, tako da smo njega koristili sljedećeg. Kao što mu i samo ime kaže, Random forest se sastoji od velikog broja nasumičnih odluka koje djeluju zajedno. Svako individualno stablo predloži jednu klasu i klasa sa najviše glasova postaje predikcija našeg modela.[9] Zanimljivo u ovom slučaju za naš skup podataka je to što je Random Forest u više svojih modela predložio državu (Country) kao bitno obilježje. Na slici 4 je prikazan isječak iz RapidMiner softverskog okruženja, u kome se vidi predikcija broja lajkova u Random Forest modelu.

numberLikes	prediction(n...	numberPosts
9431	9553.474	2740
15138	15958.921	1580
13121	14392.326	1580

Slika 4: Predikcija broja lajkova u Random Forest-u

Treći korišćeni algoritam je K-Nearest Neighbor (KNN), pomoću koga je rađena regresija, kako bismo ispitali broj lajkova na objavama. Koristili smo ga nad trening skupom podataka. Model koji je dobijen sadrži 8929 primjera sa 21 dimenzijom. Napravljena je predikcija i izračunate su greške.

C. *NLP tehnike*

Prvi korak u izradi NLP tehika je bio „prečišćavanje“ skupa podataka. To u ovom slučaju znači da smo se fokusirali samo na riječi i morali smo ukloniti brojeve, interpunkcijske znake, nulte vrijednosti i oznaku za hashtag. Iako nam se skup podataka sastoji od podataka o popularnim korisnicima širom svijeta, nismo imali problem sa različitim jezicima, obzirom da su skoro svi hashtag-ovi bili na engleskom jeziku. Izbrisali smo takođe „stop riječi“ kao što su „the“, „a“ i „is“ kako bi što veći fokus bio na samom korijenu riječi. Za ovo nam je, između ostalog, pomogao Stanford CoreNLP [10] softver.

Nakon prečišćavanja, rađeno je prvo računanje učestalosti određenih riječi, zatim kategorizacija *hashtag*-ova po učestalosti i izdvajanje najbitnijih *hashtag*-ova. U tabeli 3 su prikazani najčešće korišćeni *hashtag*-ovi.

	Naziv	Broj pojavljivanja
1.	wolfmillionaire	238
2.	madwhips	237
3.	europe	236
4.	liketkit	154
5.	architecture	109
6.	travel	105
7.	nature	120
8.	sponsored	95
9.	coachella	93
10.	beautifuldestinations	87

Tabela 2: Top 10 najčešće korišćenih hashtag-ova

Kategorizacija je vršena tako što su dobijene liste sortirane u „kante“ (eng. *bucket*) kako bismo dobili brzi i visoki pregled onoga što se nalazi u podacima. Za obučavanje modela kategorizacije teksta, korišćen je unapred sortirani sadržaj. Na osnovu ovoga, a i daljom analizom i izdvajanjem *hashtag*-ova, dobili smo tri osnovne kategorije, od kojih smo ručno napravili tri nova test skupa podataka. Kategorije, tj. podskupovi, koje smo dobili, su sljedeće:

1. Automobile
2. Travel
3. LifeStyle

Dakle, ručno su izdvojene objave korisnika u kojima su korišćeni ovi tagovi i njihova pripadajuća obilježa, podijeljeni su u tri osnovne kategorije i vršena je predikcija nad tako dobijenim skupovima podataka, sa svim njihovim obilježjima i ciljnom labelom broj lajkova (numberLikes), korišćenjem već pomenutih modela za predikciju.

Tagovi koji su korišćeni uz prvu kategoriju Automobile su: auto, car, instacar, design, automotive, vintage, drive, classic, classiccar, classiccarchasers, carphotography, forsale, drivetasetfully, chevrolet, corvette, toyota, itd.

Tagovi koji su korišćeni uz drugu kategoriju Travel su: travel, travelling, europe, visitcopenhagen, beautifulhotels, tasteintravel, architecture, mytinyatlas, landscapedesign, livelife, wonderfulindonesia, openmyworld, travelwithme, itd. Ovdje je interesantno to što je u kategoriji travel najpopularniji hashtag europe, koji je inače na našoj listi najviše korišćenih *hashtag*-ova. Iz ovoga smo zaključili da je to zbog toga što je najveći broj korisnika iz Sjedinjenih Američkih Država i koriste ovaj *hashtag* kada borave u Evropi.

Tagovi koji su korišćeni uz treću kategoriju Lifestyle su: sunset, spring, friends, teamcoffee, coffeetime, weekend, fashion, mensfashion, lifestyle, lifestyleblogger, happyfriday, itd. Zanimljivo kod ove kategorije je to da su ove *hashtag*-ove više koristili korisnici iz Italije i tu smo se susreli sa rijetkim *hashtag*-ovima u ovom skupu podataka koji nisu bili na engleskom, a to su: primavera, amore, belezafeminina, lamashermosa, itd. Osim ovoga, u ovoj kategoriji je bilo dosta *hashtag*-ova vezanih za indijska vjenčanja i običaje.

NLP tehnike su primijenjene na *hashtag*-ovima umjesto na opisima, jer smo analizom utvrdili da skoro svaki opis sadrži *hashtag*-ove, pa smo ih smatrali bitnijim za predikciju.

D. Permutation Feature

Na kraju smo željeli odrediti i rangirati atribute po njihovoj važnosti. Metoda koju smo koristili je *Permutation feature*. Koncept ove metode je direktan: mjerimo važnost obilježja računanjem greške u predikciji modela nakon uklanjanja obilježja. Obilježje je važno ako prilikom njegovog uklanjanja greška modela raste, jer bi to značilo da se model oslanjao na ovo obilježje za predikciju. Obilježje je irelevantno ako prilikom uklanjanja njegovih vrijednosti greška modela ostane nepromijenjena, jer to znači da je model ignorisao ovo obilježje za predikciju. Mana ove metode je to što je njen fokus greška modela, ali je prednost to što nudi visoko kompresovani, globalni uvid u ponašanje modela. [11]

V. REZULTATI

Iz rezultata koje smo dobili može se zaključiti da na broj lajkova, odnosno popularnost Instagram objave, najviše utiče broj pratilaca, ali i država iz koje korisnik dolazi. Pretpostavljamo da je to zbog toga što je u određenim državama veći broj stanovnika, pa samim tim i veći broj korisnika Intagrama.

A. Rezultati primijenjenih algoritama

Koristili smo, i u prethodnom poglavlju opisali, KNN i Random Forest modele, za predikciju popularnosti Instagram objava. Prije ovoga, za sagledavanje i analizu skupa podataka nad kojim vršimo predikciju, koristili smo Decision Tree. U tabeli 2 je prikazano poređenje performansi ova tri modela. Ono što nismo očekivali je to da je Decision Tree napravio bolju predikciju od Random Forest-a.

Kada bismo rangirali uspješnost ovih algoritama u predikciji Instagram objava, to bi izgledalo ovako:

1. KNN
2. Random Forest
3. Decision Tree

	D-Tree	Random Forest	KNN
root_mean_squared_error	20873.627	22296.257	20510.056
normalized_absolute_error	0.267	0.279	0.250
root_relative_squared_error	0.357	0.382	0.351
correlation	0.934	0.925	0.937
squared_correlation	0.872	0.855	0.877

Tabela 3: Performanse različitih modela

B. Rezultati NLP tehnika

Pomoću NLP tehika izdvojene su tri osnovne kategorije (Automobile, Travel, LifeStyle), na osnovu kojih smo napravili tri podskupa našeg početnog skupa podataka. Za ova tri skupa podataka smo istraživali iste greške, kod istih alogritama, kao u sekciji Primijenjeni algoritmi, nad svim pripadajućim obilježjima. Rezultati koje smo dobili su prikazani u tabeli 4.

	Automobile	Travel	LifeStyle
root_mean_squared_error	7112.243	11863.337	5878.808
normalized_absolute_error	0.226	0.266	0.334
root_relative_squared_error	0.361	0.311	0.318
correlation	0.933	0.951	0.951
squared_correlation	0.870	0.904	0.905

Tabela 4: Performanse modela nad svim obilježjima kategorija hashtag-ova

Grupisanje po kategorijama nije napravilo velike promjene u samoj predikciji, obzirom da se najpopularniji tagovi ne razlikuju mnogo jedan od drugog, ali je činjenica da su ove tri kategorije, tj. ovi *hashtag*-ovi doprinijeli broju lajkova, jer je prosječan broj lajkova na objavama koje koriste neki od ovih *hashtag*-ova 10 do 20 hiljada.

C. Rezultat Permutation Feature-a

Nakon primijenjene metode Permutation Feature, rezultat koji smo dobili je da je obilježje sa *hashtag*-ovima (Tags) uticalo dosta na greške u modelima i da je model bez ovog obilježja tačniji. Što se tiče ostalih obilježja, ne može se reći da jedan atribut mnogo važniji od drugih, ali postoje razlike. Greške smo poredili sa greškama Random Forest-a u tabeli 2 iz prethodnog poglavlja; očekivali smo da će se obilježje korisničko ime (username) ispostaviti kao važno, ali je poprilično neutralno obilježje što se tiče uticaja na sam model. U tabeli 3 je predstavljeno poređenje atributa i greške u modelima bez ovih atributa; greške su prikazane istim redoslijedom kao u tabeli 2.

numberFollowers	numberPosts	username	tags	country
23701.834	22665.826	22296.257	21028.152	23166.434
0.343	0.284	0.250	0.269	0.227
0.406	0.388	0.351	0.360	0.397
0.915	0.922	0.937	0.933	0.918
0.837	0.855	0.877	0.871	0.843

Tabela 4: Rezultati Permutation feature metode

Kada bismo na osvnovu ovoga rangirali važnost atributa, dobili bismo sljedeće:

1. numberFollowers
2. country
3. numberPosts
4. username
5. tags

VI. ZAKLJUČAK

U ovom poglavlju su obrazloženi dobijeni rezultati, kroz diskusiju, sumarizaciju rada i predloge pravca budućeg rada.

A. Diskusija

Metoda linerane regresije se pokazala kao loš model za ovaj skup podataka. Ostale korišćene metode su se pokazale kao približno jednake jedna u odnosu na drugu, prilikom poređenja.

Kada smo pravili kategorije *hashtag*-ova susreli smo se sa problemom nedostatka *hashtag*-ova za određenu kategoriju, tj. nije bilo dovoljno podataka za formiranje skupa podataka, tako da smo proširivali kategorije sličnim podkategorijama. Prilikom ovoga, došlo je do automatskog spajanja određenih kategorija, na primjer: automobili i destinacije na koje su išli.

Plan obogaćivanja skupa podataka dodatnim obilježjima je bio dodavanje obilježja sa podatkom o filterima koji su bili korišćeni na objavi, kako bi bilo utvrđeno koliko oni mogu uticati na popularnost Instagram objave. Međutim, do ovog podatka nije bilo moguće doći, obzirom da većina korisnika i ne koristi Instagram filtere, a veoma je teško odrediti o kom drugom filteru se radi, obzirom na veliki broj aplikacija za

editovanje fotografija. Broj filtera korišćenih na određenoj Instagram objavi je takođe podatak do koga nismo mogli doći, iz sličnog razloga.

B. Sumarizacija rada

U ovom radu je prvo opisana motivacija i cilj istraživanja kojim smo se bavili, a to je analizirati i pokušati predvidjeti popularnost Instagram objava i time možda pomoći onima kojima je broj lajkova na Instagram objavi bitan u poslovanju. Nakon toga, predstavljen je naš ručno sastavljeni skup podataka, opisani su svi postojeći atributi i razlozi dodavanja novih atributa, tj. obogaćivanja skupa podataka. Metode koje su korišćene su:

- metode linearne regresije;
- algoritmi: Decision Tree, Random Forest i KNN;
- NLP tehnike za frekvenciju riječi kod *hashtag*-ova;
- kategorizacija;
- metode za određivanje važnosti atributa;

Na osnovu ovoga, zaključeno je ono što smo na početku istraživanja i pretpostavljali, a to je da je broj pratilaca (numberFollowers) izuzetno bitan za broj lajkova na objavi. Osim ovoga, država (country) iz koje korisnik dolazi se ispostavila kao bitno obilježje ovog skupa podataka, a njome je skup podataka naknadno obogaćen. Broj objava (numberPosts) je takođe bitan parametar; utvrđeno je da su korisnici koji na svom profilu imaju veći broj objava, imali i veći broj lajkova. Ono što nije očekivano je to da je korisnično ime (username) manje bitno u odnosu na prehodno pomenute atribute, iz čega smo došli do zaključka da su korisnici ipak obraćali veću pažnju na sadržaj same objave, nego na osobu koja je postavila tu objavu na Instagram.

Prilikom rangiranja atributa, broj pratilaca je takođe određen kao najvažniji parametar, ali ono što je bilo iznenađujuće je da hashtag-ovi i nisu bitni za popularnost Instagram objave. Primjenom NLP tehnika ovo je i dokazano – iako su objave sa popularnim *hashtag*-ovima imale veliki prosječni broj lajkova, taj broj nije znatno veći od broja lajkova na objavama bez *hashtag*-ova ili sa manje popularnim *hashtag*-ovima. Iz čega je opet potvrđeno da je broj pratilaca najbitniji parametar.

C. Predlog pravca budućeg rada

Za budući rad bismo obogatili skup podataka atributom regularLikes (broj redovnih lajkova), kako bismo provjerili koliko svaki profil ima redovnih lajkova na objavama i time vidjeli stvarnu popularnost određene objave, ali i samog korisničkog profila.

Takođe, obzirom da već postoji atribut mentions (označene osobe na fotografiji), nad tim obilježjem bi se mogle generisati dalje kombinacije obilježja za NLP tehnike, kako bi se vidjelo da li su određeni korisnički profili uticali na veći broj lajkova objave jednog korisnika. Uz ovo bi se mogao dodati novi atribut mentionsNumber (broj označenih osoba na fotografiji), na isti način kao što smo dodali atribut numberOfTags na osnovu atributa tags, u poglavlju 3 ovog rada, i vidjeli bismo broj osoba na fotografiji i uticaj broja osoba na lajkove.

Broj komentara na objavi bi bio takođe zanimljiv atribut za obogaćivanje skupa podataka, ali nismo sigurni koliko bi uticao na predikciju.

LITERATURA

[1] Wikipedia članak, Instagram, https://en.wikipedia.org/wiki/Instagram, posjećeno 10. 05. 2020. godine

[2] Zhongping Zhang, Tianlang Chen, Zheng Zhou, Jiaxin Li, Jiebo Luo; How to Become Instagram Famous: Post Popularity Prediction with Dual-Attention, University of Rochester: Department of Electrical Engineering and Department of Computer Science, https://www.groundai.com/project/how-to-become-instagram-famous-post-popularity-prediction-with-dual-attention/

[3] Masoud Mazloom, Robert Rietveld, Stevan Rudinac, Marcel Worring, Willemijn van Dolen; Multimodal Popularity Prediction of Brand-related Social Media Posts, University of Amsterdam: Informatics Institute and Amsterdam Business School,2016, https://staff.fnwi.uva.nl/m.mazloom/Papers/mazloom-ACM-MM-2016.pdf

[4] Francesco Gelli, Tiberio Uricchio, Marco Bertini, Alberto Del Bimbo MICC; Image Popularity Prediction in Social Media Using Sentiment and Context Features, Università degli Studi di Firenze and Shih-Fu Chang, Columbia University, https://www.academia.edu/29744719/Image_Popularity_Prediction_in_Social_Media_Using_Sentiment_and_Context_Features

[5] Skup podataka korišćen u radu: https://drive.google.com/open?id=1WEhnOAgm9nHG-ZAaYoLMsb8lBXLdKfGa

[6] Regresija, https://people.dmi.uns.ac.rs/~zlc/fajlovi/Regresija.pdf?fbclid=IwAR1EKVjY8azbfD47TtyMJ5bMQ2JU1Fqafu1UJxJxiIAUtZScn9FwdOBgsP8, posjećeno 01. 06. 2020.

[7] Linearna regresija, https://sr.wikipedia.org/wiki/Linearna_regresija, posjećeno 01. 06. 2020.
[8] RMSE: Root Mean Squared Error, Statistics How To, https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/?fbclid=IwAR0vgtI8ZEDeOUt6-38RpyLwr9C8j0W5MuaQL4igfhhT21OO_VvTeNkIv0c, posjećeno 01. 06. 2020.
[9] Towards Data Science, Understanding Random Forest, https://towardsdatascience.com/understanding-random-forest-58381e0602d2, posjećeno 13. 05. 2020.
[10] Stanford CoreNLP softver: https://stanfordnlp.github.io/CoreNLP/
[11] Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable",2019. https://christophm.github.io/interpretable-ml-book/.