Predikcija popularnosti objava na sajtu 9gag korišćenjem multimodalnih podataka

Nikola Ilić
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Novi Sad, Srbija
nikola_ilic96@hotmail.com

Nenad Gligorov
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Novi Sad, Srbija
ngligorov@uns.ac.rs

Svetislav Simić
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Novi Sad, Srbija
simicsvetislav@uns.ac.rs

***Apstrakt* — Predikcija popularnosti objava na sajtu *9gag* u vremenu kada su društvene mreže, forumi i portali veliki deo onoga što jedna prosečna osoba konzumira u toku svog dana, bi mogla da bude od velikog značaja za oblast marketinga. Analiza glavnih elemenata objava postavljenih na *9gag* sajt bi mogla da pokaže koji faktori najviše utiču da neka objava postane popularna. Metode kojim će se ovo izvesti su metode koje se tiču analize slike i tekstualnih podataka same objave. Konkretno, u radu se analiziraju sledeći faktori: prisustvo određenih tipova objekata na slici objave, informacija da li slika pripada grupi poznatih šablona (takozvanih *memes*), prisustvo teksta na slici, sentiment komentara korisnika na objavu, ključne reči iz naslova objave i analiza *hashtag*-ova. Ekstrahovana obeležja iskorišćena su za treniranje različitih regresionih modela za predikciju popularnosti objave. Po našem znanju, ovo je prvi rad koji se bavi problemom predikcije popularnosti *9gag* objave. Najbolji model je postigao vrednost od 0.4918 za R^2 meru. Izvršeni eksperimenti daju indikaciju da, pored broja komentara, obeležje koje ima najveći uticaj na predikciju jeste sentiment komentara.**

Ključne reči — popularnost objave; analiza teksta; analiza slike; multimodalni podaci; 9gag

I. Uvod

Svaka objava na *9gag*-u [1] ima određene karakteristike koje je prate. Neke su napravljene po određenom, prethodno ustanovljenom šablonu (takozvani *meme*), neke su potpuno originalne, kod nekih je glavni element slika objave, kod nekih je u pitanju tekst, dok je kod nekih kombinacija ta dva. Sve objave mogu biti ocenjene pozitivno i negativno, na svakoj objavi se može ostaviti komentar, što doprinosi određenoj popularnosti objave. Popularnost objave uzrokuje da se objava nađe u određenoj kategoriji na sajtu na osnovu toga koliko je popularna. Problem kojim se bavi ovaj rad je predikcija popularnosti neke objave na sajtu *9gag*. Odnosno, koji sve elementi koji karakterišu jednu objavu utiču na popularnost objave i u kolikoj meri. Ovo bi moglo da bude od značaja u oblasti marketinga, prvenstveno u onim delovima koji se bave reklamiranjem kao i predstavljanjem određenog zabavnog sadržaja kako na društvenim mrežama, tako i onlajn uopšte.

Cilj ovog rada je analiza faktora koji utiču na popularnost objave postavljene na *9gag* sajt. U trenutku pisanja ovog rada nije pronađen ni jedan rad na sličnu temu kada je u pitanju sajt *9gag*. Samim tim bi rešavanje problema kojim se bavi ovaj rad moglo da bude od značaja u nekim budućim istraživanjima.

Ideja prikazana u ovom radu je da se prvo izvrši ekstrakcija obeležja u vidu elemenata objave za koje je pretpostavka da su od značaja kada je u pitanju njena popularnost. Ekstrahovana obeležja se koriste za treniranje regresionog modela za predikciju popularnosti objava.

U cilju treniranja modela prikupljeni su podaci sa sajta *9gag* od kojih su glavni bili: multimedijalni sadržaj (koji može biti u ovim formatima: slika, gif, video), naslov objave, *hashtag*-ovi, komentari, kao i podaci vezani za komentare (broj komentara i sl.), pozitivne i negativne ocene objave itd. Metodi za ekstrakciju obeležja koji su korišćeni u ovom radu mogu da se razvrstaju na one koji se odnose na analizu slike objave i one koji se odnose na analizu teksta objave. Iz slike objave ekstrahovana su sledeća obeležja:

- Detektovano je prisustvo određenih tipova objekata na slici. Prepoznavanje objekata vršeno je primenom *SSD MobileNetV2* neuronske mreže [2] pretrenirane na *Open Images* skupu podataka [3],
- Detektovano je da li slika pripada nekom od klasičnih popularnih šablona. Ovaj problem se sveo na problem klasifikacije pomoću neuronske mreže sa VGG16 arhitekturom [4], pretreniranom nad *ImageNet* skupom podataka [5],
- Detektovano je prisustvo teksta na slici. Nažalost, usled ograničenja dostupnog skupa podataka, problem razumevanja teksta na slici se pokazao kao kompleksan i nije davao dobre rezultate. Stoga je kao obeležje objave iskorišćena samo dužina tekstova na slikama, koja je dobijena korišćenjem *Pytesseract* biblioteke [6] u *Python* programskom jeziku.

Iz tekstualnih podataka koji prate objavu, ekstrahovana su sledeća obeležja:

- Određen je prosečan sentiment komentara uz pomoć *Standford CoreNLP* API-ja [7] za analizu prirodnog jezika, a takođe je, sa istim ciljem, korišćen i pretrenirani BERT model
- Ekstrahovane su ključne reči objave, delimično iz njenog naslova, a delimično iz *hashtag*-ova.

Nakon ekstrakcije obeležja, ona su iskorišćena za treniranje regresionih modela. Obeležja su kombinovana na dva načina. Prvi način je da se sva obeležja konkateniraju i proslede modelu. Drugi način kombinovanja obeležja je stekovanje (eng. *stacking)* modela. Treći pristup rešavanju problema se bazira na treniranju autoenkodera na slikama objava, na kome se potom

dekoder zameni novom neuronskom mrežom kako bi se dobila procena popularnosti. Za ciljno obeležje (popularnost objave) korišćen je odnos pozitivnih i negativnih glasova neke objave.

Glavno zapažanje kada su u pitanju konačni rezultati predikcije je to da modeli nisu uspeli dovoljno dobro da „nauče" na osnovu ekstraktovanih obeležja i da za to postoji više razloga koji će biti pomenuti.

Nastavak ovog rada je organizovan na sledeći način. U drugom poglavlju prikazan je pregled istraživanja koja na određeni način mogu dati smernice za rešavanje problema koji je predmet ovog istraživanja. Način formiranja početnog skupa podataka i kako taj skup podataka izgleda opisano je u trećem poglavlju, U četvrtom poglavlju opisane su metode koje su korišćene za analizu podataka, izvlačenje određenih informacija iz početnog skupa podataka, njihovo kombinovanje i način dobijanja konačne predikcije na osnovu njih. Dobijeni rezultati su prikazani i komentarisani u petom poglavlju. Pokušaj da se kreira model koji bi generisao komentare je opisan u šestom poglavlju. Konačno, u sedmom poglavlju izvršena je sumarizacija i opisane su neke ideje i mogućnosti za dalji rad, koje bi mogle da doprinesu unapređenju postojećeg rešenja, kao i razvoju rešenja u potpuno novim pravcima.

II. Pregled literature

Po našem znanju, ovo je prvi rad koji se bavi problemom predikcije popularnosti objava na *9gag* sajtu. S toga će u ovoj sekciji biti predstavljena istraživanja koja se bave problemom predikcije popularnosti objava na drugim društvenim mrežama, istraživanja koja se bave određenim aspektima problema rešavanog u ovom radu (analiza slike i teksta), kao i problemima koji bi potencijalno mogli biti adresirani prilikom proširenja ovog istraživanja.

Cilj [8] je predviđanje popularnosti objava na društvenoj mreži *Flickr*[1]. Opisani pristup prilikom predikcije koristi vizuelna obeležja, koja se dobijaju analizom slike, zatim obeležja dobijena analizom teksta objave i društvena obeležja, poput prosečnog broja pregleda korisnika koji je objavu postavio. Korišćen je postojeći SMP-TI skup podataka, koji ukupno sadrži 432 hiljade objava sa društvene mreže *Flickr*. U ovom skupu podataka postoje kontekstne informacije, kao što su broj komentara na objavi, informacija o tome da li se na slici nalaze ljudi, dužina naslova, dužina opisa, broj tagova itd. Autori su dodatno proširili skup obeležja tekstualnim informacijama, u koje spadaju naslov, tagovi i opis fotografije. Za predikciju popularnosti iskorišćen je kombinovani pristup, u kome se neki podaci direktno šalju modelu koji vrši konačnu predikciju, dok se za druge prvo vrši izolovana predikcija popularnosti, pa se dobijeni rezultati šalju modelu koji vrši konačnu predikciju. Predikcija popularnosti na osnovu slike vršena je pomoću modifikovane verzije duboke konvolutivne neuronske mreže *InceptionResnetV2* arhitekture. Predikcija popularnosti na osnovu naslova i tagova se bazirala na upotrebi rečnika, a sentiment opisa je određen pomoću *Stanford CoreNLP* biblioteke. Ostala obeležja su direktno korišćena u konačnoj predikciji. Ovakvom analizom se došlo do 15 obeležja, koja se potom normalizuju, i vrši se predviđanje popularnosti objave korišćenjem konvolutivnog modela, koji se sastoji od dva konvolutivna i dva potpuno povezana sloja. Za evaluaciju rešenja korišćeni su Spirmanov koeficijent korelacije, srednja apsolutna greška i srednja kvadratna greška. Zaključak opisanog rada je da je korišćenje multimodalnog pristupa bilo opravdano, jer su rezultati bili bolji u odnosu na slučaj kada se ne koriste podaci iz svih izvora. Ovaj rad se može smatrati u značajnoj meri sličan našem istraživanju i zbog toga nam je dao značajne putokaze na koji način se mogu prikupljati i koristiti multimodalni podaci, tako da su analize naslova, tagova i sentimenta vršene na sličan način i u našem istraživanju.

Rad [9] predstavlja težnju autora da identifikuju koji aspekti u objavama koji se odnose na brendove brze hrane na društvenoj mreži *Instagram*[2] utiču na to da objave budu popularnije. Za to se koristi devet obeležja koje su autori nazvali *engagement parameters*, a to su:

- podatak o tome da li je logo brenda prisutan na slici,
- broj prepoznatih lica na slici,
- podataka o prisustvu barem jednog proizvoda na slici,
- da li se na slici zajedno pojavljuju jedna osoba i proizvod,
- da li se na slici zajedno pojavljuje više osoba i proizvod,
- sentiment objave, koji se dobija analizom *hashtag*-ova, natpisa uz objavu i komentara,
- upotrebljeni filteri na slici,
- prisustvo objekata, koji pripadaju nekoj od 15293 kategorija iz *ImageNet* skupa podataka,
- broj osoba koji „prati" osobu koja je postavila objavu.

Za analizu da li se na slici nalaze logo brenda, osobe i proizvod korišćen je *Google Vision* API. Određivanje sentimenta objave rađeno je na osnovu vizuelnog sentimenta, koji je određivan analizom slike korišćenjem *Sentibank* detektora, i sentimenta teksta koji je dobijen analiziranjem teksta iz *hashtag*-ova, naslova i komentara na objavi pomoću *SentiStrength* [10] metode. Skup podataka je dobijen skupljanjem informacija o 75 hiljada objava koje se odnose na šest poznatih lanaca brze hrane. Evaluacija je vršena korišćenjem Spirmanovog koeficijenta korelacije. Rezultati eksperimenata su pokazali da je značajno koristiti i vizuelna i tekstualna obeležja za predikciju popularnosti objave i da pojava jedne ili više osoba na slici zajedno sa proizvodom doprinose popularnosti objave. Ova saznanja su bila od značaja pri odabiru kategorija objekata koje treba detektovati na slikama objava, odnosno postojanje distinkcije toga da li je na slici jedna osoba ili više njih, kao i postojanje značaja u analizi naslova, komentara i *hashtag*-ova.

Kreiranje sistema za prepoznavanje humora u rečenicama prirodnog jezika je tema [11]. Korišćeni skup podataka se sastoji iz dva izvora koji su sadržali tekstove sa humorom i nekoliko izvora za koje se smatralo da ne sadrže smešne rečenice. Analizirane su semantičke i stilske osobine rečenica, kao i sličnosti među rečenicama. Iako je bilo planova, saznanja ovog

[1] https://www.flickr.com/

[2] https://www.instagram.com/

rada nisu korišćena, s obzirom na to da tekst nije ekstraktovan sa slika i toga da humor u naslovima objava zavisi od konteksta i često je prožet ironijom, a ponekad čak i nema preterane veze sa samom objavom.

Autori [12] su sproveli istraživanje o aspektima koji doprinose popularnosti *Facebook*[3] objava. Koristili su postojeći *Facebook News Dataset* [13]. Pored obeležja koja postoje u skupu podataka, dodali su nova obeležja za koja su smatrali da mogu doprineti predikciji, kao što su dan u nedelji i doba dana kada je objava postavljena i prosečan broj deljenja i reakcija, ukupno i svakog tipa pojedinačno, u sekundi. Dodatna obeležja su dobijena i nakon analize sentimenta objave korišćenjem VADER biblioteke [14] i analizom koliko je vremena proteklo pre nego što je objava dobila prvi i stoti komentar. Kao mera popularnosti objave uzeto je upravo vreme potrebno da objava dobije 100 komentara. Za predikciju popularnosti korišćene su linearna regresija, MARS (*Multivariate Adaptive Regression Spline*) i SVR (*Support Vector Regression*), a na osnovu R^2 mere, kao najbolji model se pokazao SVR. Analizom je utvrđeno da su popularnije objave sa negativnim sentimentom, a da je obeležje koje obično ima najveći uticaj na predikciju broj „tužnih" (*sad*) reakcija u jedinici vremena. Ovaj rad bi mogao da bude putokaz ka drugačijem rešavanju problema, pošto se popularnost ne posmatra statički, nego se kroz vreme prate određeni podaci i samo ciljno obeležje je zbog toga drugačije. Sa druge strane, još jednom se pokazuje da analiza sentimenta objave ima značaj za samu popularnost i da bi takva analiza trebalo svakako da bude obavljena.

Predviđanje popularnosti multimedijalnog sadržaja na osnovu multimodalnih podataka je problem koji se rešava u [15]. Korišćen je skup podataka YFCC100M [16], koji se sastoji od 100 miliona javnih slika i video klipova sa društvene mreže *Flickr*. Analizirani su podaci o tagovima na objavama i vizuelne osobine slika. Podaci o tagovima su reprezentovani rečnikom od 2000 najpopularnijih tagova. Metodama dubokog učenja izvlačene su vizuelne osobine sa svake slike. Broj pregleda je bio ciljno obeležje u ovom istraživanju i nad njime je primenjena logaritamska funkcija, zbog velike varijacije u različitim rasponima. Za predviđanje popularnosti korišćena je SVR metoda, a probani su multimodalni i unimodalni pristupi. Rezultati su pokazali da multimodalni pristup daje bolje rezultate od unimodalnog, koji uzima u obzir samo vizuelna obeležja, međutim, korišćenje podataka o tagovima daje najbolje rezultate, koji su bolji i od multimodalnog pristupa. Ovo istraživanje je pokazalo da vizuelna obeležja imaju manji značaj od tagova koji su povezani sa objavom i da bi, imajući to u vidu, trebalo na neki način uzeti u obzir podatke o tagovima prilikom određivanja popularnosti objava. Takođe nam je ukazalo na to da bismo mogli i mi da primenimo logaritamsku funkciju nad ciljnim obeležjem, zbog uočene disproporcije vrednosti u različitim intervalima.

Istraživanje [17] analizira diskusije sa veb sajta *Reddit*[4] u cilju uočavanja onih koje bi proizvele najviše kontraverze. Pod kontraverznim diskusijama se smatraju one koje su dobile dosta pozitivnih i dosta negativnih reakcija, odnosno postoji izrazito podeljeno mišljenje. Kako bi se napravilo predviđanje, pored teksta originalne objave, analiziraju se i grupe najranijih komentara, za koje su probani različiti vremenski periodi, i struktura komentara, to jest, analiziran je graf na kome su čvorovi pojedinačni komentari, a veze predstavljaju odnos da je jedan od komentara odgovor na drugi. Rezultati pokazuju da inicijalni komentari i njihova struktura imaju značajnu ulogu na predikciju da li će neka diskusija izazvati kontraverze. Iako ovaj rad nije sličan našem istraživanju, pogotovo zbog toga što kontraverznost i popularnost nisu iste stvari, on pokazuje značaj praćenja ranih reakcija na neku objavu, što bi moglo da bude uzeto u obzir tokom budućeg rada na našem projektu.

Sličan zaključak i značaj za naše istraživanje ima [18], u kojem se predviđa popularnost objava na *Reddit*-u na osnovu prvih deset komentara. Problem je posmatran kao binarna klasifikacija, zbog, kako autori navode, čudne distribucije ciljnog obeležja *score*, koje predstavlja razliku pozitivnih i negativnih reakcija. Kao obeležja za klasifikaciju korišćeno je više obeležja koji se odnose na sentiment komentara, kao i dužina komentara. Zaključak rada je da postoji određena veza između inicijalnih komentara i popularnosti objava. Pored značaja ranih reakcija na neku objavu za njenu popularnost, ovaj rad pokazuje da sentiment komentara treba uzeti u razmatranje prilikom predikcije popularnosti objava.

Rad [19] predstavlja istraživanje o tome šta jednu sliku na društvenoj mreži *Flickr* čini popularnom. Analiziraju se dva aspekta, a to su socijalni, kao, recimo, broj prijatelja osobe koja je sliku postavila, i, nama zanimljiviji, sam sadržaj slike. Analizirane su brojne stvari vezane za sadržaj slika, počevši od obeležja niskog nivoa do objekata na slikama. Od jednostavnih obeležja na slici, pokazalo se da dominantna boja u određenoj meri može imati veze sa popularnošću, konkretno, crvenkastije boje imaju veći značaj za popularnost od plavih i zelenih nijansi. Međutim, veći značaj imaju neka malo složenija obeležja, kao što su gradijent i teksture, a isto važi i za koncepte visokog nivoa, u vidu prepoznatih objekata. Zaključak rada je da, pored socijalnih obeležja, i obeležja srednjeg i visokog nivoa na slikama imaju značaj na njihovu popularnost. Glavni značaj za naše istraživanje proizilazi iz temeljne analize slika u opisanom radu. Ono navodi na zaključak da, pored objekata na slikama, možda treba ispitati i malo jednostavnija obeležja, koja mogu imati značaj za predviđanje njihove popularnosti.

III. Skup podataka

Kako nije pronađen relevantan postojeći skup podataka, za potrebe ovog istraživanja formiran je novi, prikupljanjem podataka (eng. *scrape*) direktno sa sajta *9gag*. Prikupljene su informacije o pojedinačnim objavama, kao i informacije o svim njihovim komentarima.

Isprobano je više metoda prikupljanja podataka. Prvobitna ideja je bila *scrape*-ovanje celih stranica sajta u HTML formatu korišćenjem *Jsoup* biblioteke [20], da bi se kasnije vršilo parsiranje iz HTML formata u objekte *Java* klasa. Ovaj pristup se ispostavio kao neadekvatan. Razlog tome je dinamičko učitavanje stranica sajta *9gag*, što znači da su u trenutku učitavanja sve HTML stranice prazne. Generisanje sadržaja vrši se nakon aktivacija *Javascript* skripti. Odabrana biblioteka ne podržava mogućnost automatske aktivacije ovih skripti, tako da

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

[4] https://www.reddit.com/

su sve prikupljene stranice sadržale samo osnovne *head* i *body* tagove.

Sledeća isprobana ideja se takođe zasnivala na vršenju *scrape*-a HTML stranica. Ovog puta korišćena je *Selenium* biblioteka [21]. Ova biblioteka namenjena je za testiranje *Web* sajtova, prevashodno se koristi za automatizaciju ranije osmišljenih manuelnih testova. Njena prednost u odnosu na *Jsoup* biblioteku je ta što podržava korisničku upotrebu sajta, koristeći *web driver*-e. *Selenium WebDriver* [22] upravlja pretraživačem manuelno, kao korisnik, time dajući priliku za generisanje podataka na stranici.

Ovakav pristup se pokazao kao uspešan, ali veoma spor. Analizom REST zahteva razmenjenih između *frontend*-a i *backend*-a sajta došlo se do saznanja da, na zahtev *Javascript* skripte, *backend* dostavlja informacije o objavama koje treba prikazati u JSON formatu. Svaka JSON datoteka sadrži u sebi listu objava sa svim potrebnim informacijama, kao i pokazivač, odnosno URL ka sledećoj datoteci. Na taj način je formirana jednostruko spregnuta lista, tako da je potrebno samo pratiti URL pokazivača (korišćenjem *Jsoup* biblioteke), da bi se došlo do naredne datoteke.

Svaka objava sadrži sledeće korisne informacije:

- broj komentara,
- *timestamp* kreiranja objave,
- broj negativnih ocena (eng. *downvote*),
- broj pozitivnih ocena (eng. *upvote*),
- *hashtag*-ovi,
- URL ka slici,
- sekcija u kojoj se nalazi – *hot*, *trending* ili *fresh*
- naslov,
- tip objave – da li je u pitanju slika ili animacija,
- URL ka originalnoj objavi na sajtu.

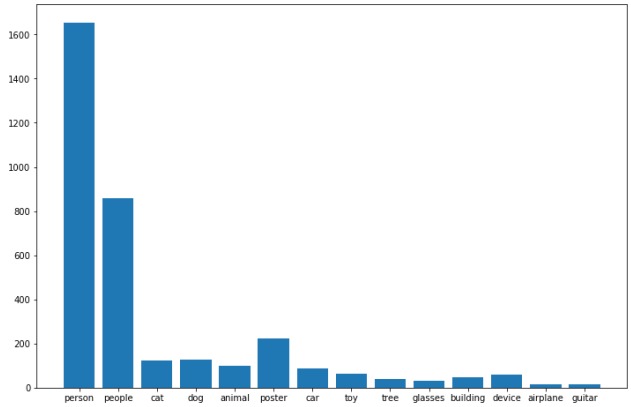
Prikupljanje komentara za pojedinačne objave analogno je prikupljanju informacija o objavama. Svaki komentar sadrži sledeće korisne informacije:

- broj potkomentara,
- broj pozitivnih ocena (eng. *like*),
- broj negativnih ocena (eng. *dislike*),
- tekst komentara – ako je u pitanju slika ili animacija (GIF) u ovom polju se nalazi URL,
- *permalink* – veza ka objavi kojoj komentar pripada (URL).

Prikupljeno je 6038 objava i 45036 komentara. Korišćena je MySQL baza za skladištenje podataka.

IV. Metod

Raznorodne metode su korišćene kako bi se iz podataka o objavama pribavile željene informacije i izvršila predikcija njihove popularnosti. Informacije su ekstraktovane iz dve grupe izvora, a to su same slike i tekstualni podaci iz objava. U ovom poglavlju opisana su i neka zapažanja nakon eksplorativne analize podataka, kao i metode koje su korišćene za dobijanje konačne predikcije popularnosti.



Slika 1 - Broj detektovanih objekata za odabrane kategorije

A. Analiza slika

Pod analizom slika podrazumeva se ekstrakcija informacija, za koje se pretpostavlja da su relevantne i da mogu doprineti uspešnosti predviđanja popularnosti, iz multimedijalnog sadržaja koji je povezan sa svakom objavom. Analiza animacija se svodila na analizu slika, tako što je analiziran samo prvi frejm animacije, odnosno frejm koji se dobije praćenjem URL-a koji je u bazi podataka sačuvan kao adresa multimedijalnog sadržaja. Od ukupno 6038 objava o kojima su sakupljeni podaci, za njih 31 nije bilo moguće dobaviti slike, tako da je kompletna analiza slika i konačna predikcija vršena za 6007 objava.

1) Prepoznavanje objekata

Analizom literature se pokazalo da je prepoznavanje objekata na slikama objava nezaobilazna stavka, kao i zbog pretpostavke da ljudi različito reaguju na vizuelni sadržaj u zavisnosti od toga šta je na njemu prikazano.

Detekcija objekata je izvršena korišćenjem *SSD MobileNetV2* neuronske mreže za detekciju objekata, prethodno trenirane na *Open Images* skupu podataka, i to njegove četvrte verzije. Ovaj skup podataka se sastoji od oko 9 miliona slika na kojima je, između ostalog, anotirano prisustvo objekata, koji mogu pripadati nekoj od 600 kategorija. Zbog hardverskih ograničenja nije korišćen neki od *state-of-the-art* modela, poput *Faster R-CNN Inception ResNet.*

Analiza svih 6007 slika je trajala oko 36 sati. Detektovano je ukupno 7459 objekata na 3581 slici, što je nešto manje od 60% svih slika. Analizom dobijenih rezultata su izdvojene kategorije koje se najčešće pojavljuju, a izvršena su i određena grupisanja postojećih i stvaranje jedne nove kategorije. Najveći broj detekcija je bio vezan za ljude na slici. Dve ubedljivo najprisutnije kategorije su bile ljudsko lice (*human face*) i čovek (*man*), sa 2404 i 1491 pojavljivanjem. Postojale su i druge kategorije koje ukazuju na prisustvo osoba na slikama, tako da su kategorije *person*, *man*, *woman*, *boy* i *girl* zajedno grupisane u kategoriju *person*, dok je *human face* ostala zasebna. Slična grupisanja su izvršena i za neke manje frekventnije, a logički

povezane, kategorije. U srodnim istraživanjima [9] je primećeno da se pravi razlika da li je na slici prisutna jedna osoba ili više njih, tako da je napravljena nova kategorija za ljude (*people*), koja je ukazivala na prisustvo više ljudi na slici. Ona je aktivirana kada su ispunjeni određeni uslovi, prvenstveno kada je prepoznato više od jedne osobe na slici, i isključiva je sa kategorijom *person*. Rezultat transformacija je 15 kategorija koje obuhvataju 6956 originalnih detekcija, što znači da su osačuvani podaci o oko 93% prvobitnih detekcija. Eksplorativnom analizom je naknadno utvrđena značajna korelacija između objekata *person* i *clothing*, tako da prilikom predikcije nije u obzir uziman podatak o tome da li je prepoznat objekat koji pripada kategoriji *clothing*.

Vrednost za svaku kategoriju predstavlja *one-hot encoded* vektor, a kada je na jednoj slici pristuno više objekata, vrši se *bitwise* ILI operacija nad odgovarajućim vektorima.

Distribucija broja pojava za 14 korišćenih kategorija, dobijenih nakon transformacija, je prikazana na Slika 1. Na grafiku se može videti da su prepoznati objekti pretežno osobe, s obzirom na to da su kategorije *person* i *people* ubedljivo najzastupljenije. Postoji i značajan broj prepoznatih životinja, koje su predstavljene kategorijama *dog*, *cat* i *animal*.

2) Prepoznavanje šablona

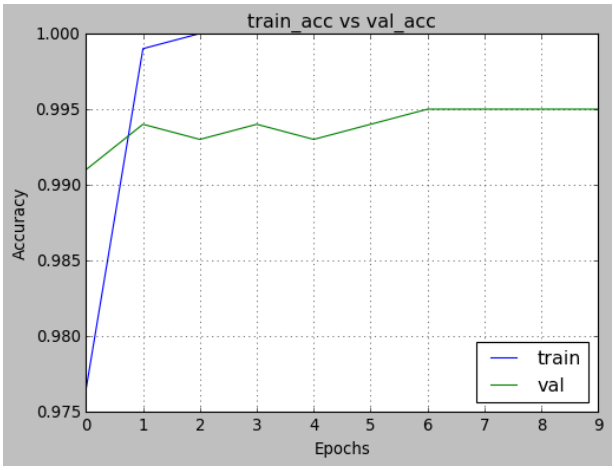
Ideja iza prepoznavanja da li je slika uz objavu nastala na osnovu nekog od često korišćenih šablona je pretpostavka da neki šabloni potencijalno imaju veću popularnost od drugih. Na osnovu iskustva prilikom posećivanja sajta, odabrano je 19 šablona za koje se smatra da su frekventni. Podaci za treniranje su *scrape*-ovani sa sajta *imgflip* [23], koji omogućava jednostavan pristup grupi objava koje su nastale po odabranom šablonu. Skup slika je proširen sa 1472 nasumične objave sa sajta *9gag*, pri čemu su prethodno ručno izbacivane slike koje su nastale na osnovu šablona, tako da se smatra da su ostale samo slike koje nisu nastale na osnovu šablona.

Detekcija šablona predstavlja klasifikacioni problem sa 20 klasa. Korišćena je konvolutivna neuronska mreža sa VGG16 arhitekturom, pretrenirana na *ImageNet* skupu podataka, ali je vršeno dodatno treniranje potpuno povezanih slojeva. Treniranje je obavljeno na ukupno 23585 slika u *mini-batch* režimu i veličine *batch*-a od 32. Skup podataka je podeljen na trening i test skup u odnosu 80:20, što je standardni način podele podataka kod ovakvih problema. Nakon 10 epoha dobijena je tačnost od 99.5% na test skupu. Broj epoha je određen empirijski, imajući u vidu brzu konvergenciju modela, što se može videti na grafiku na Slika 2.

Problem za ovaj deo rešenja je postojanje šablona za čije preoznavanje nije trenirana neuronska mreža, s obzirom na to da je skup korišćenih šablona znatno širi od onoga koji je korišćen pri treniranju, a treniranje na svim dostupnim šablonima je odbačeno zbog hardverskih ograničenja.

3) Prepoznavanje teksta

Početna ideja što se tiče prepoznavanja teksta na slikama je bila da se prepoznati tekst analizira na sličan način kao i drugi tekstualni sadržaji, međutim od toga se odustalo zbog relativne nepouzdanosti u potpuno preciznom prepoznavanju teksta, koje je osetljivo na stvari poput kvaliteta slike, fonta i boje slova. Međutim, iskorišćena je dužina prepoznatog teksta, kao obeležje

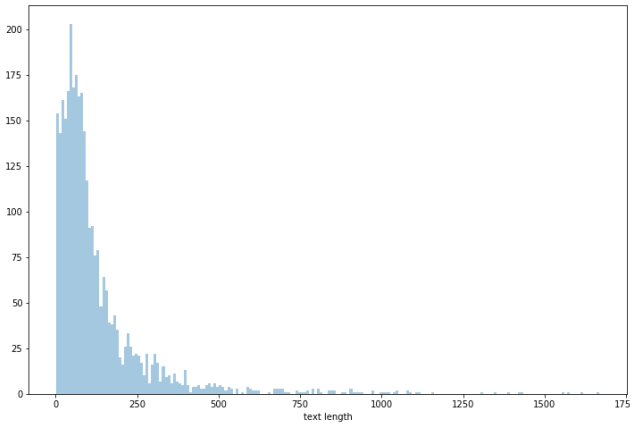


Slika 2 - Preciznost modela za prepoznavanje šablona kroz epohe

koje bi trebalo da ostane prilično dobro prepoznato, čak i ako se čitav tekst ne prepozna potpuno ispravno. Potencijalna veza dužine teksta sa popularnošću je u tome što ljudi možda ne vole da čitaju relativno dugačke objave [24] ili ne žele uopšte da čitaju tekst na slikama. Za prepoznavanje teksta na slikama korišćena je biblioteka *Pytesseract* za *Python* programski jezik. Distribucija dužina teksta na slikama prikazana je na Slika 3. Sa grafika su izbačeni podaci o objavama na kojima nema teksta, to jest, oni gde je dužina teksta nula.

B. *Analiza tekstualnih podataka*

Za analizu tekstualnih podataka, pretpostavka je bila da bi od značaja bili naslovi, *hashtag*-ovi i komentari objava. Pored ekstrakcije ključnih reči iz naslova, rađena je i sentiment analiza naslova pomoću *SentiStrength* API-ja [25]. *SentiStrength* je pogodan za analizu sentimenta kraćih tekstova. Rezultati koje daje su u formi koja određuje vrednost na pozivitnoj i negativnoj skali (od –1 do –5 za negativnu, od 1 do 5 za pozitivnu). Prvobitno testiranje nad naslovima je dalo rezultate koji su pokazali da ovaj metod neće biti od pomoći jer su naslovi specifični prvenstveno po tome da veliki broj sadrži sarkazam, što *SentiStrength* nije umeo da prepozna. Samim tim rezultati nisu bili verodostojni pa se nisu ni koristili u predikciji.



Slika 3 - Distribucija dužina teksta na slikama

1) Određivanje sentimenta komentara

Sentiment komentara je određivan na dva načina. Prvi koristi *Stanford CoreNLP* biblioteku, a drugi pretrenirani BERT model.

a) Stanford CoreNLP

Analiza sentimenta komentara je rađena pomoću *Stanford CoreNLP* API-ja za analizu i obradu prirodnog jezika. Rezultati koji se dobijaju pomoću *CoreNLP* API-ja su na skali od 0 do 4, gde je 0–veoma negativno, 1-negativno, 2-neutralno, 3-pozitivno i 4-veoma pozitivno. Svaka od tih pet ocena dobija procentualno vrednost, gde ukupan zbir čini 100%, a ocena koja ima najveću procentualnu vrednost se uzima kao krajnji rezultat sentimenta. Testiranjem nad tekstom koji nije bio deo komentara se pokazalo da API daje dobre rezultate. Ipak, kada je rađena analiza nad komentarima, primećeno je da rezultati nisu bili pouzdani za sve komentare iz više razloga, od kojih su neki: korišćenje smajlija/emotikona (eng. *emoji)* u komentarima, komentari koji su sadržali nestandardne karaktere, komentari koji nisu sadržali nikakav tekst, odnosno prazan string itd. Zbog toga su komentari bili filtrirani pre same analize, što je uticalo na to da broj komentara za analizu bude manji sa početnih 45036 komentara za 2928 objava, na 39305 komentara, za isti broj objava. Veliki broj komentara je ocenjen kao neutralan, čak preko 80% na celom skupu, dok se taj broj kod komentara koji pripadaju samo *trending* sekciji kretao do 90,5%. Jedan od glavnih razloga za to je način na koji API dolazi do kranje ocene sentimenta. Iz tog razloga kao obeležje nije uzimana srednja vrednost krajnje ocene sentimenta svih komentara za jednu objavu, što je bio pokušaj, ali oni nisu uticali da se rezultat predikcije značajno poboljša. Zbog toga je pokušan drugi način, koji je dao bolji rezultat, gde se pravio poseban model za predikciju, gde je za obeležja uzimana srednja vrednost procenata svakog komentara jedne objave, za svaku od pet ocena.

b) BERT

Drugi pristup analizi sentimenta komentara se zasniva na korišćenju BERT jezičkog modela (eng. *language model*), pretreniranog na skupu podataka koji se sastoji od komentara sa platforme *Google Play*[5] i koji ima tačnost od oko 88% na test skupu, prilikom određivanja sentimenta. Prilikom treniranja kao optimizator je korišćen *AdamW*, sa preporučenim parametrima [26], i *cross-entropy* kao *loss* funkcija.

Određivanje sentimenta za svaki komentar se svodi na klasifikacioni problem, gde se svaki tekst komentara klasifikuje kao pozitivan, neutralan ili negativan. Pre same klasifikacije tekst se pretvara u odgovarajući oblik, koji je kompatibilan sa BERT modelom, to jest, vrši se tokenizacija, reči se pretvaraju u brojeve, koji predstavljaju ID-eve reči i dužina ulazne sekvence se fiksira na određenu dužinu, konkretno na 160. Kompletan model se, osim od BERT-a, sastoji i od *dropout* sloja, kako bi se sprečio *overfitting*, i linearnog izlaznog sloja sa tri izlaza. Korišćenjem ovakvog modela se dobija sentiment svakog komentara, a kumulativan sentiment komentara na neku objavu se dobija uprosečavanjem predikcija za pojedinačne komentare, pri čemu su komentari prepoznati kao pozitivni označeni brojem 1, neutralni sa 0, a negativni sa -1. Najveći broj komentara je ponovo prepoznat kao neutralan, njih nešto manje od 21 hiljade, broj pozitivnih je nešto manji od jedanaest hiljada, a negativnih nešto preko trinaest hiljada.

Implementacija ovog dela projekta se oslanja na upotrebu biblioteka *PyTorch*[6] i *Transformers*[7], koje obezbeđuju BERT model, tokenizatore za pretprocesiranje teksta i rad sa samim modelom.

2) Ekstrakcija ključnih reči iz naslova

Ključne reči su izvlačene TFIDF metodom na korpusu podataka koji je obuhvatao naslove svih dobavljenih objava. Sve reči su lematizovane na svoje izvorne oblike, kako bi se smanjila nepotrebna raznovrsnost reči. Za svaku ključnu reč je povezana i vrednost, koja ukazuje na njen značaj, ali ona na kraju nije uzimana u obzir. Broj ključnih reči za naslove je varirao od nijedne, pa čak do 28. Implementacija se oslanja na upotrebu *NLTK* biblioteke[8] za *Python* programski jezik.

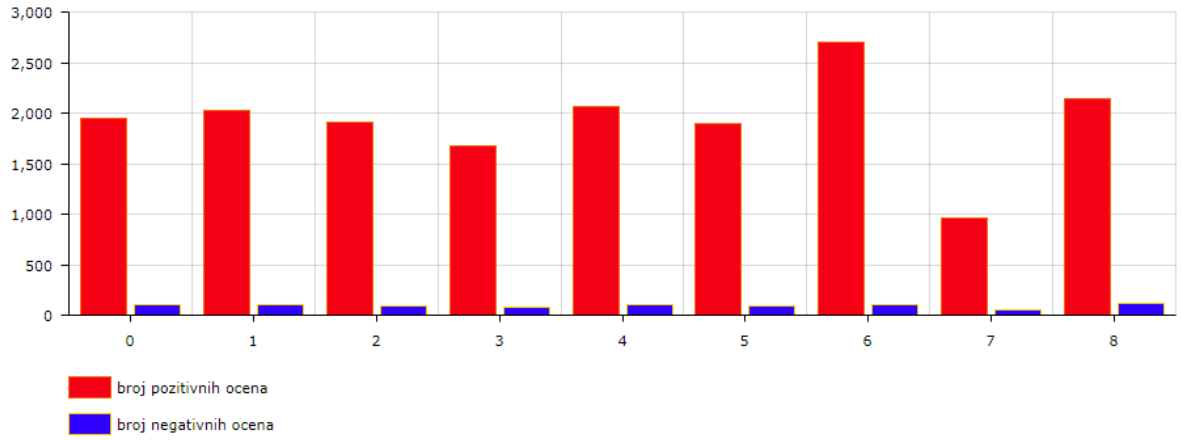
Pošto nije primećeno kako bi same ključne reči iz naslova mogle da utiču na predikciju popularnosti objave, jedina ideja bila je da se odabere određeni broj reči koje se najviše pojavljuju u naslovima, u ovom slučaju uzeto je 20 najfrekventnijih, i da se one koriste za predikciju. Kod njihovog odabira, primećeno je da se najfrekvetnije reči iz hiljadu najbolje ocenjenih objava javljaju i kao najfrekventnije reči u hiljadu najlošije ocenjenih objava, pa se tu već moglo zapaziti da one neće imati veliki značaj u konačnoj predikciji.

Druga metoda nije dala dobre rezultate već u samom koraku izvlačenja ključnih reči. U pitanju je RNN *sequence to sequence* neuronska mreža, sa *encoder-decoder* arhitekturom. Model je napravljen na osnovu koda [27] koji je prvobitno bio napisan za prevođenje teksta, a zatim je modifikovan sa ciljem da prepoznaje ključne reči u kraćim tekstovima kao što su naslovi. Treniran je nad korpusom od 25369 naslova sa njihovim ključnim rečima. Model na ulazu prima samo tekstove određenog formata (bez znakova interpunkcije, velikih slova itd.) što je jedna od mana. U retkim slučajevima daje dobre rezultate, dok za neke primere kao ključnu reč izdvaja reč koja nije deo naslova. Pretpostavka je da je ovakvo loše ponašanje modela uzrokovano kombinacijom toga da je model, prvobitno namenjen za prevođenje teksta, modifikovan u svrhu nalaženja ključnih reči, ali i nedovoljno velikim skupom podataka.

Tabela 1 - Distribucija objava po broju *hashtag*-ova

Broj *hashtag*-ova	0	1	2	3	4	5	6	7	8
Broj objava	3067	1129	583	848	231	133	34	8	5

[5] https://play.google.com/
[6] https://pytorch.org/
[7] https://huggingface.co/transformers/
[8] https://www.nltk.org/



Slika 4 - Odnos lajkova i dislajkova prema broju tagova

3) Analiza hashtag-ova

Posmatran je odnos između broja *hashtag*-ova po objavi i broja pozitivnih, odnosno negativnih ocena, kao i broja komentara. Broj *hashtag*-ova postavljenih na pojedinačnim objavama kreće se od 0 do 8. Od 6038 prikupljenih objava, njih 3067 nije imalo nijedan *hashtag*, dok svega 5 objava ima 8 hashtag-ova. Celokupna raspodela hashtag-ova na ovom skupu podataka prikazana je na Tabela 1.

Prosečan broj pozitivnih, odnosno negativnih ocena objava sa istim brojem *hashtag*-ova varira sa tolerancijom od ±10%, kao što je prikazano na Slika 4. Slična je situacija i sa prosečnim brojem komentara kod objava sa istim brojem *hashtag*-ova. Odstupanje od ove tolerancije nastaje kod objava sa 7 i 8 *hashtag*-ova. Međutim, ni ovde se ne da primetiti određena zavisnost. Zaključak je da je broj takvih objava previše mali, pa da zbog toga dolazi do odstupanja.

Poređene su i objave sa *hashtag*-ovima i one bez. Njihov odnos je prikazan na Slika 5.

Iz svega navedenog dolazi se do zaključka da broj *hashtag*-ova ni na koji način ne utiče na popularnost objave.

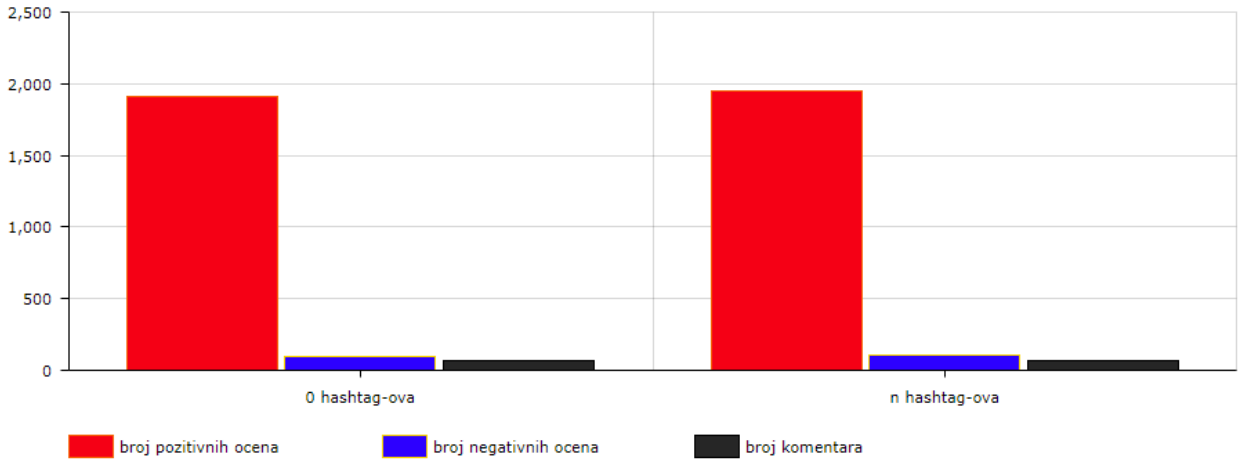
Pretraživanje najpopularnijeg *hashtag*-a i istraživanje nekih zavisnosti vezanih za to nisu pouzdani, budući da se radi o vremenski promenljivoj stavki. Kao takva, ona ne može uticati na ispravnu predikciju popularnosti objava u budućnosti.

C. *Predikcija popularnosti*

Konačni skup obeležja se sastoji od nekih od podataka vezanih za objave koji su *scrape*-ovani i onih koji su dobijeni različitim analizama. Korišćeni podaci dobijeni *scrape*-ovanjem su broj komentara na objavi i tip objave, to jest, da li je multimedijalni sadržaj vezan za objavu slika ili animacija.

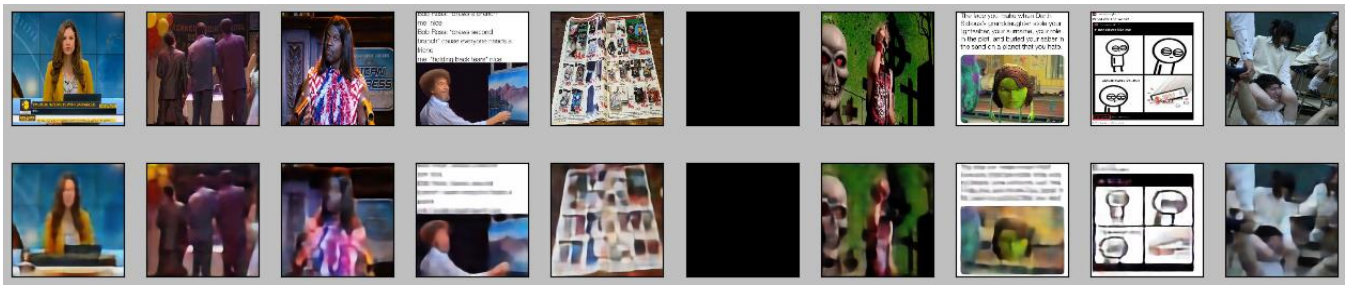
Podaci dobijeni analizama su:

- binarni vektor dužine 15, koji ukazuje na prisustvo odabranih kategorija objekata na slikama,
- binarni broj, koji pokazuje da li je na slici prepoznat šablon po kome je nastala,
- dužina teksta prepoznatog na slici,
- vektor od pet elemenata, koji sadrži podatke o sentimentu komentara na objavi, gde prvi broj



y

Slika 5 - Odnos lajkova, dislajkova i komentara prema tome da li postoje tagovi



Slika 6 - Primeri originalnih i rekonstruisanih slika

predstavlja prosečnu verovatnoću da komentar vrlo negativan, a peti element da je vrlo pozitivan,

- binarni vektor dužine 20, koji ukazuje na prisustvo najfrekventnijih ključnih reči u naslovima.

Ciljno obeležje, koje reprezentuje popularnost objave, predstavlja odnos pozitivnih (eng. *upvotes*) i negativnih (eng. *downvotes*) glasova za objavu, pri čemu se negativnim glasovima dodaje jedan fiktivan glas, kako bi se izbeglo deljenje sa nulom.

Treniranje modela radi dobijanja konačne predikcije vršeno je na dva različita pristupa. Jedan se zasniva na prosleđivanju svih obeležja modelima koji se treniraju, a drugi na stekovanju modela.

Za prvi pristup su isprobani razni modeli. Kao apsolutno osnovni model (eng. *baseline*) korišćen je model koji za svaku objavu predviđa srednju vrednost ciljnog obeležja, imajući u vidu ceo skup podataka. Kao osnovni modeli korišćeni su i multivarijabilna linearna regresija, kao i multivarijabilna linearna regresija sa kombinovanom L1 i L2 regularizacijom, to jest, model pod nazivom *Elastic net*. Modeli od kojih su se očekivali bolji rezultati su SVR, ansambl regresionih stabala (eng. *Random forest*) i regresor koji odluku donosi na osnovu uprosečavanja predikcija drugih modela (zvaćemo ga *voting* model). Što se tiče poslednjeg modela, uprosečavane su vrednosti modela koji su se najbolje pokazali pojedinačno, a to su SVR i ansambl regresionih stabala.

Prilikom treniranja svih modela 20% podataka je izdvojeno za test skup, a 80% za trening skup, što predstavlja uobičajeni odnos. Probavanje drugačijih odnosa nije dovelo do značajnije promene rezultata. Optimizacija hiperparametara je vršena unakrsnom validacijom sa deset podela (eng. *folds*), to jest, određeni delovi trening skupa su u različitim iteracijama predstavljali validacioni skup. Za *Elastic net* optimizovani su hiperparametri *l1* i α, pretragom koja u 100 iteracija, nasumičnim kombinovanjem parametara, traži najbolju kombinaciju (eng. *randomized search*). Na isti način su optimizovane vrednosti za broj estimatora i maksimalnu dubinu stabla kod ansambla regresionih stabala. Granice za broj estimatora su bile 10 i 320, a dubinu 2 i 50. Optimizacija za SVR je bila malo drugačija, pošto je korišćena „iscrpljujuća" pretraga (eng. *brute-force* ili *grid search*), gde se isprobavaju sve moguće kombinacije zadatih potencijalnih vrednosti. Optimizovani su parametri *C* i γ. Potencijalne vrednosti za parametar *C* su stepeni broja 10 u rasponu od 10^{-3} do 10^{2}, a za γ stepeni broja 10 u rasponu od 10^{-4} do 10^{2}.

Drugi pristup se sastoji od posebne predikcije popularnosti objava za podatke iz različitih izvora, a to su podaci dobijeni analizom slike, podaci o sentimentu komentara i podaci o ključnim rečima u naslovu. Tri posebna ansambla regresionih stabala predviđaju popularnost svake objave, a potom su dobijene vrednosti, zajedno sa brojem komentara i tipom objave, korišćene kao obeležja za novi model koji donosi konačnu procenu o popularnosti objave. Model koji donosi konačnu procenu je takođe bio ansambl regresionih stabala. Ovaj metod je izabran zato što je davao najbolje rezultate u prvom pristupu, kao pojedinačni regresor. Kao hiperparametri ponovo su optimizovani broj estimatora i maksimalna dubina, na isti način kao i u prvom pristupu.

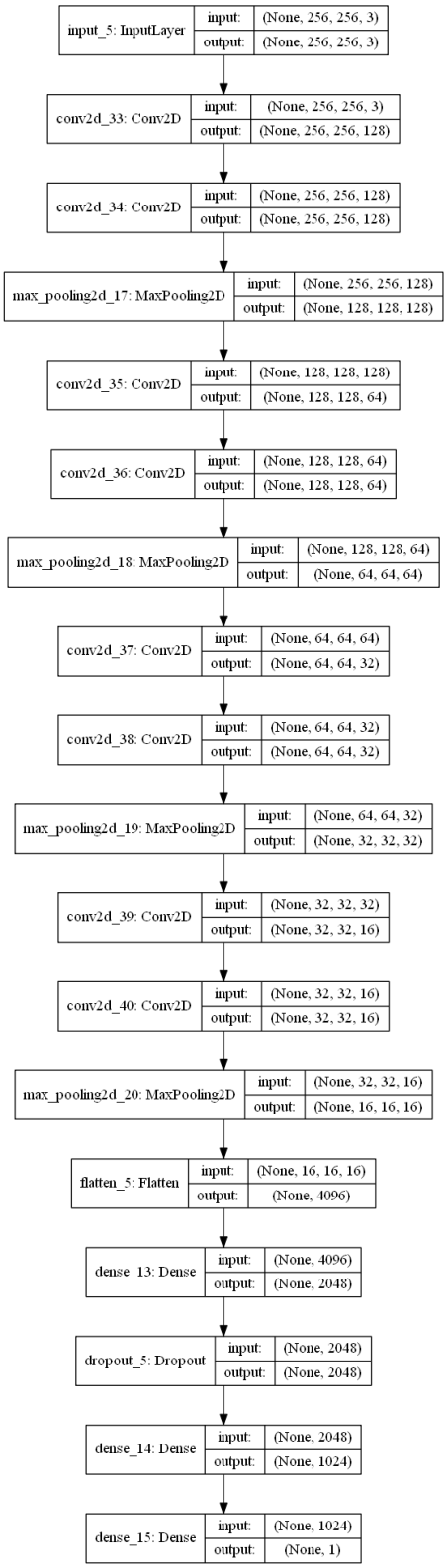
Implementacija ovog dela projekta je izvršena prvenstveno oslanjajući se na biblioteku *scikit-learn* [28] za programski jezik *Python*.

1) End-to-end pristup

Isproban je još jedan, značajno drugačiji, pristup predviđanju popularnosti objava, koji se bazira na donošenju odluke analizom isključivo slike objave upotrebom dubokog učenja. Ovaj, praktično *end-to-end* pristup, se sastoji od dve faze. Prva faza je upotreba konvolutivnog autoenkodera, kako bi se ulazna slika enkodirala, to jest, kako bi se stvorila reprezentacija slike koja je kompaktnija od originalne slike, ali dovoljno ekspresivna

Tabela 2 - Rezultati za prvi pristup na test skupu

Metod / Mera	*Baseline*	Linearna regresija	*Elastic net*	SVR	*Random forest*	*Voting*
MSE	0.9602	0.7874	0.7843	0.5122	0.4983	**0.4617**
R^2	-0.0307	0.1332	0.1367	0.4362	0.4515	**0.4918**
ρ	/	0.431	0.4606	0.6534	0.6586	**0.6935**



Slika 7 - Arhitektura regresionog modela kod *end-to-end* pristupa

da se na osnovu te reprezentacije može, u određenoj meri, rekonstruisati originalna slika. Druga faza se sastoji od iskorišćavanja istreniranog enkodera na koji se nadovežu dva potpuno povezana sloja, koji rade regresiju na osnovu enkodirane reprezentacije slike.

Arhitekure korišćenih modela su kreirane tako da enkoder i dekoder budu simetrični. Isprobane su različite arhitekture, koje uglavnom nisu bile previše kompleksne. Sa jedne strane, nisu korišćeni kompleksni modeli jer se model u kome je enkoder sastavljen od konvolutivnih slojeva VGG16 mreže nije dobro pokazao, a sa druge, zbog hardverskih ograničenja. Iz istog razloga veličina *batch*-a je tokom treniranja najčešće bila 4.

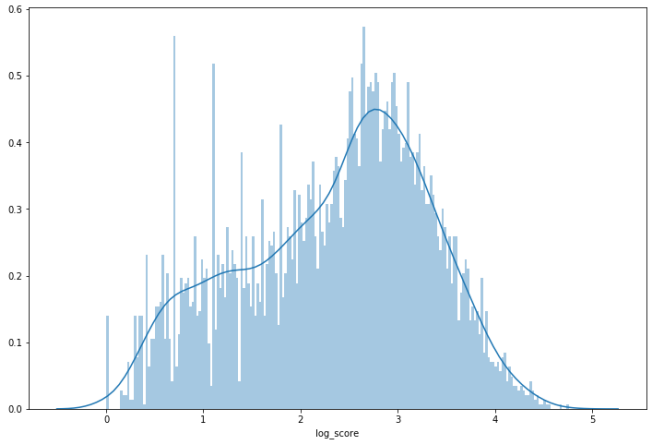
Uspešnost rekonstrukcije dosta zavisi od veličine sloja sa enkodiranom reprezentacijom (eng. *bottleneck layer*). Potpuno precizna rekonstrukcija nije primarni cilj prve faze, već je to reprezentacija koja je kompaktna, a ipak dovoljno ekspresivna. Na Slika 6 je prikazano kako su rekonstruisane neke od slika, korišćenjem autoenkodera čiji će enkoder kasnije biti korišćen za regresiju. Postojale su rekonstrukcije koje su bolje, iako je arhitektura autoenkodera jednostavnija, međutim, odabrana je arhitektura sa nešto većim brojem slojeva, zbog pretpostavke da će na taj način enkodirana reprezentacija ipak biti ekspresivnija i samim tim korisnija kada se bude radila regresija.

U drugoj fazi je između dva potpuno povezana sloja dodat i *dropout* sloj, kako bi se smanjile šanse za *overfitting*. Za broj neurona u potpuno povezanim slojevima su isbrobane različite vrednosti, ali nije uočeno da neke daju značajno bolje rezultate. Kompletan model koji je korišćen u drugoj fazi je prikazan na Slika 7. Jedini izlaz modela je vrednost koja reprezentuje prediktovanu popularnost.

Implementacija je izvršena korišćenjem biblioteka *keras*[9], *tensorflow*[10] i *keras-vis*[11].

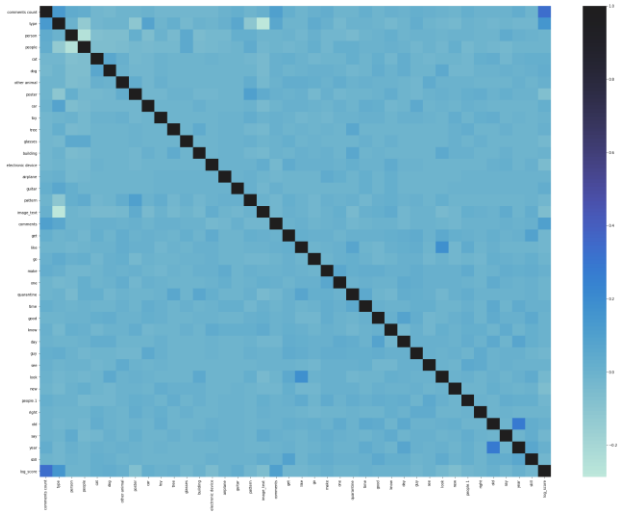
V. Rezultati i diskusija

Prvobitno formirano ciljno obeležje je imalo distribuciju „dugog repa”, pa su vrednosti transformisane upotrebom

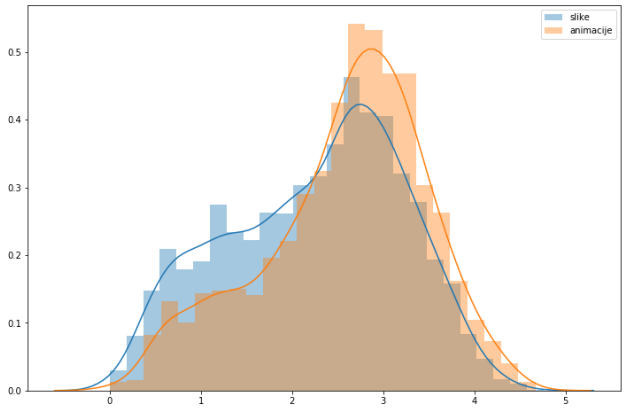


Slika 8 - Distribucija ciljnog obeležja

[9] https://keras.io/
[10] https://www.tensorflow.org/
[11] https://raghakot.github.io/keras-vis/



Slika 9 - Matrica korelacije za obeležja korišćena u predikciji



Slika 10 - Popularnost objava po tipu multimedijalnog sadržaja

logaritamske funkcije, kako bi distribucija postala približnija normalnoj raspodeli. Distribucija ciljnog obeležja nakon transformacije je prikazana na Slika 8.

Eksplorativnom analizom je utvrđeno da i obeležja koja predstavljaju broj komentara i dužinu teksta na slikama imaju raspodelu „dugog repa", pa je pokušana normalizacija njihovih vrednosti, korišćenjem modula *stats* iz *scipy* biblioteke, ali nije primećeno poboljšanje u konačnoj predikciji, pa su te transformacije odbačene.

Tokom testiranja praćene su različite mere evaluacije, ali zbog ne tako jednostavne interpretacije ciljnog obeležja kao glavna mera evaluacije odabran je koeficijent determinacije (R^2). Takođe je računat Spirmanov koeficijent korelacije (ρ), prvenstveno zbog toga što se kod nekih sličnih istraživanja saopštava baš Spirmanov koeficijent.

U Tabela 2 predstavljeni su rezultati za različite modele, kada je u pitanju prvi pristup, a u Tabela 3 rezultati za pristup koji se bazira na stekovanju modela.

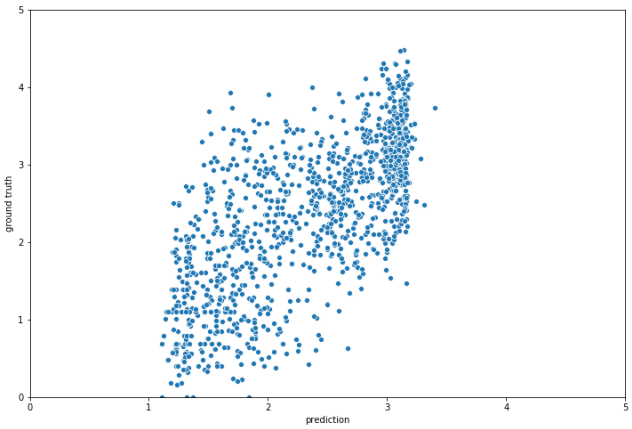
Najbolje rezultate kod prvog pristupa daje model koji konačnu predikciju donosi na osnovu glasova *Random forest* i SVR metoda. Kod pojedinačnih metoda najbolje rezultate ima *Random forest*, ali ni SVR ne zaostaje mnogo. Linearna regresija i *Elastic net* daju značajno lošije rezultate, ali i te metode imaju određenu prediktivnu moć. Očekivano, najlošije rezultate, sa negativnim koeficijentom determinacije, ima *baseline* predikcija. Vredi napomenuti da je praćena i p-vrednost (eng. *p-value*) i da je za sve metode kod prvog pristupa, osim za *baseline* metodu, važilo da je ona značajno manja od 0.001.

Kod rezultata za pristup zasnovan na stekovanju modela moguće je primetiti da, kada se posmatraju pojedinačni izvori informacija, najveći doprinos za tačnu predikciju ima sentiment komentara, pri čemu je analiza sentimenta upotrebom BERT modela nešto uspešnija. Informacije ekstraktovane sa slika imaju određeni značaj, ali znatno manji. Ključne reči u naslovu ne doprinose tačnosti predikcije pošto je koeficijent determinacije obično po vrednosti vrlo blizu nuli. Ovakav ishod za ključne reči je bio očekivan, s obzirom na to da je na osnovu domenskog znanja zapaženo da naslovi često nemaju uske veze sa samom objavom i da su zbog toga verovatno irelevantni prilikom predviđanja popularnosti. Konačna predikcija je nešto lošija nego kod najuspešnije metode prvog pristupa, ali ne značajno. Prilikom konačne predikcije korišćeni su rezultati za sentiment dobijeni upotrebom *Stanford CoreNLP* biblioteke. Sem za predikciju na osnovu ključnih reči, p-vrednosti su kod drugog pristupa bile značajno manje od 0.001.

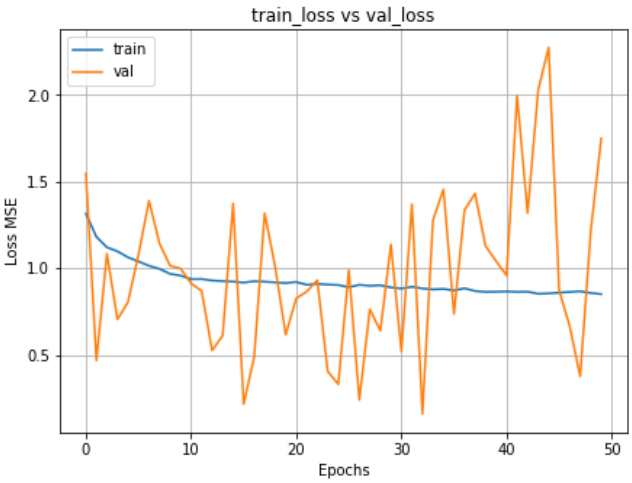
Očito je, međutim, da na osnovu ekstraktovanih multimodalnih podataka modeli nisu na najbolji način mogli da „nauče" da predviđaju popularnost objava. Modeli se pri predikciji u značajnoj meri oslanjaju na podatak o broju komentara na objavama, koji je bio dostupan i pre bilo kakve dodatne analize. To je moguće videti i u matrici korelacije za sva korišćena obeležja na Slika 9 (kvadrati u gornjem desnom i donjem levom uglu matrice).

Tabela 3 - Rezultati na test skupu za pristup baziran na stekovanju modela

Podaci / Mera	Informacije sa slika	Sentiment komentara – *Standford CoreNLP*	Sentiment komentara – BERT model	Ključne reči u naslovu	Konačna predikcija
MSE	0.8893	0.7309	0.1007	0.9702	0.4703
R^2	0.0211	0.1443	0.1498	0.0014	0.4843
ρ	0.1255	0.3877	0.4117	0.002	0.6782



Slika 12 - Prediktovane i stvarne popularnosti za *voting* metod



Slika 11 - Grafik loss funkcije prilikom treniranja regresionog modela

Na Slika 12 grafički su prikazana predviđanja za *voting* model, gde su na x-osi prediktovane, a na y-osi stvarne vrednosti ciljnog obeležja. Grafik ukazuje na to da pripremljeni podaci i dalje nisu potpuno pogodni za primenu regresionih metoda i da bi verovatno trebalo da bude izvršeno njihovo dodatno pretprocesiranje.

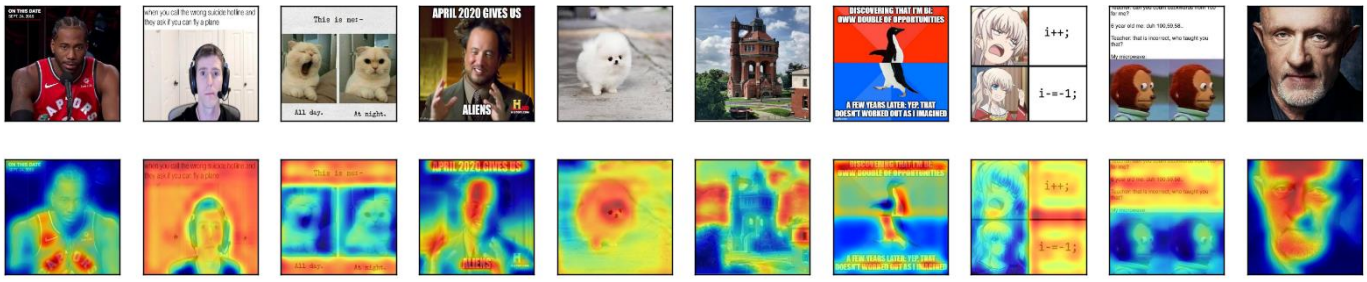
Eksplorativnom analizom je takođe uočeno da dve trećine objava kao multimedijlani sadržaj ima sliku, ali da animacije imaju bolji odnos glasova. Na Slika 10 uporedno su prikazani grafici popularnosti za objave koje imaju slike, odnosno animacije, kao multimedijalni sadržaj.

Rezultati *end-to-end* pristupa nisu stabilni na testnom skupu. Grafik kretanja *loss* funkcije, koja je bila MSE, tokom treniranja regresionog modela je prikazan na Slika 11. Iako tokom treniranja vrednost *loss* funkcije opada do vrednosti od otprilike 0.85, metrike koje su praćene pokazuju nestabilnost na testnom skupu. Može se videti da je MSE na testnom skupu u nekim trenucima značajno bolja, a u nekim značajno lošija u odnosu na trening skup, i to se menja bez jasnih pravilnosti. Uzrok za ovakvo ponašanje može biti težina problema koji se rešava i relativno mali skup podataka za primenu dubokog učenja, posebno imajući u vidu raznolikost slika objava. Pretpostavka je navedene stvari utiču na to da za određeni broj slika iz test skupa model ne može da donese pouzdanu predikciju i zbog toga za neke slike budu prediktovane relativno nasumične vrednosti, koje nekad mogu biti bolje, a nekad lošije.

Dodatno su iscrtane *heatmap*-e za istrenirani regresioni model, kako bi se ustanovilo koji delovi slike doprinose da se prediktovana popularnost poveća ili smanji. Na Slika 13 prikazane su neke od originalnih slika uz iscrtane *heatmap*-e. Može se zapaziti da izrazito svetli regioni na slikama često utiču na povećanje popularnosti. Na povećanje popularnosti takođe utiču regioni sa ljudskim licima bele puti i tekstovima, mada se obe ove grupe u određenoj meri svode na prosto prisustvo svetlih regiona, ali treba zapaziti da svetli regioni ne doprinose uvek povećanju popularnosti, kao recimo na osmoj sličici.

Postoji problem sa stavljanjem u kontekst dobijenih rezultata, s obzirom na to da nisu pronađena dovoljno slična istraživanja, sa kojima bi direktno mogle da se porede dobijene performanse. U istraživanju [8], koje je prepoznato kao najsličnije našem, dobijen je koeficijent korelacije od 0.83, što je nešto više od onoga su ostvarili naši modeli. Istraživanje [19] saopštava koeficijent korelacije od 0.81, međutim ono ima značajne razlike u odnosu na naše. Rad [15], koji analizira popularnost objave na osnovu slike i povezanih tagova, saopštava najbolji koeficijent korelacije od 0.464, a rad [9], koji koristi podatke iz više izvora, koeficijent korelacije od oko 0.5, u zavisnosti od postavke. Naše rešenje daje bolje performanse od oba pomenuta.

Razlozi za dobijene rezultate su verovatno posledica združenog delovanja različitih faktora. Pre svega, korišćeni su podaci o nešto više od 6000 objava. Iako ovaj broj na prvi pogled nije mali, povećanje skupa podataka bi verovatno moglo značajno da doprinese performansama modela. Trebalo bi razmisliti i o korekciji ciljnog obeležja, pošto sam odnos pozitivnih i negativnih glasova u nekim situacijama ne mora biti dovoljno dobar pokazatelj popularnosti neke objave. Posebno bi



Slika 13 - Originalne slike i njihove *heatmap*-e

bilo zanimljivo pratiti podatke o objavama tokom vremena, jer bi na taj način ciljno obeležje moglo da se definiše na drugačiji, verovatno bolji, način i mogli bi se steći novi uvidi u to šta doprinosi popularnosti.

Ironija, sarkazam, prenesena i skrivena značenja, kao i stereotipi su sveprisutni na objavama, počevši od naslova objave, preko multimedijalnog sadržaja, pa do komentara. Detekcija ovih aspekata predstavlja problem za sebe, koji nije posebno adresiran u ovom istraživanju.

Popularnost neke objave veoma zavisi i od trenutnog konteksta i predznanja pojedinca [29]. Obično postoje aktuelne teme, za koje u nekom vremenskom periodu postoji povećan broj objava koje postaju popularne, dok u nekoj drugoj situaciji tim temama se ne pridaje značaj. Treba imati u vidu i ogroman broj korisnika, koji mogu biti vrlo različiti po svojim predznanjima, shvatanjima i ubeđenjima. Zbog svega navedenog, prepoznavanje koji elementi generalno utiču na popularnost neke objave predstavlja pitanje na koje nije moguće dati precizan odgovor, a samim tim je značajno otežano kreiranje modela koji bi to trebao da uradi.

VI. Generisanje teksta komentara

U okviru ovog projekta napravljen je i model za generisanje teksta na osnovu komentara sa *9gag*-a. Kao osnova poslužio je kod [30] sa zvaničnog *Tensorflow* sajta. Zbog tehničkih nemogućnosti korišćen je *Google Colaboratory* [31]. Treniranje modela nad originalnim kodom je trajalo predugo, čak i sa samo 10 epoha. Zbog toga je kod u određenoj meri izmenjen kako bi trening bio brži. Pokušano je treniranje raznih oblika modela, od kojih je najbolje rezultate u trening fazi dao model sa tri GRU sloja od po 512 ćelija. Skup podataka nad kojim je trenirano sastoji se od korpusa od po 39695 komentara, sa čak 314 jedinstvenih karaktera. Kada se model istrenira na 300 epoha, dobije se *loss* 0.7604 i *accuracy* 0.8120. Ipak, kada se nad istreniranim modelom pokuša generisanje nekog teksta, dobije se nejasan niz karaktera bez ikakvog smisla. Prvobitno je pretpostavka bila da je u pitanju loš trening skup koji se sastoji od prevelikog broja karaktera nad kojim model ne može dobro da „nauči", ali isti trening skup nad drugim već poznatim modelom za generisanje teksta daje mnogo bolje rezultate, pa je zaključak bio da trening skup nije problem. Druga pretpostavka je da je problem kada se izmeni dimenzija na ulazu modela nakon treninga, da bi na ulaz mogao da se dovede tekst početnog niza karaktera za generisanje teksta. Za trening fazu je ulazna dimenzija 3, dok se kod generisanja teksta koristi dimenzija 2. Verovatno najveći uzročnik problema jeste sama konfiguracija modela iako u trening fazi daje zadovoljavajuće rezultate.

VII. Zaključak

Problem koji se rešavao u ovom radu tiče se predikcije popularnosti objava na sajtu *9gag*. Rešavanje ovog problema bi moglo da bude od značaja u daljim istraživanjima u oblasti marketinga, pre svega reklamiranja i predlaganja određenog zabavnog sadržaja putem društvenih mreža.

Problem je rešavan tako što je rađena predikcija popularnosti neke objave, gde su za modele koji su trenirani korišćena obeležja dobijena analizom elemenata od značaja za neku objavu. Predikcija je rađena na dva načina, a to su: treniranje modela tako što su prosleđena sva obeležja i stekovanje modela Pretpostavka je bila da su to slika i tekstualni podaci. Analize koje se odnose na sliku su bile prepoznavanje objekata na slici, prepoznavanje šablona, odnosno klasifikacija, kao i prepoznavanje teksta na slici, što se zbog kompleksnosti svelo na određivanje dužine teksta. Što se tiče analize tekstualnih podataka, rađena je sentiment analiza komentara, ekstrakcija ključnih reči iz naslova kao i analiza *hashtag*-ova na osnovu njihovog odnosa i veze sa brojem komentara i odnosom pozitivnih i negativnih glasova za neku objavu. Primenjen je i poseban pristup koji analizira isključivo sliku objave i zasnovan je na upotrebi autoenkodera.

Za rezultate se može primetiti da nisu loši, ali ne i u kojoj meri su dobri, jer ne postoje referentne vrednosti, odnosno rezultati iz nekih drugih dovoljno sličnih istraživanja sa kojima bi mogli da se direktno uporede. Ipak, primetno je bilo da pored broja komentara, obeležje koje ima najveći uticaj na popularnost neke objave jeste ono koje se tiče sentimenta komentara, što bi moglo da bude od značaja u nekim drugim istraživačkim radovima koji se tiču analize teksta i određivanja sentimenta teksta.

Dalji rad na ovom sistemu bi mogao da se sastoji od usavršavanja metoda koje su korišćene i ponovnog pokušaja da se analiziraju neki aspekti objava, koji do sada nisu preterano uspešno analizirani, uz eventualno uvođenje novih vidova analize i preispitivanje načina na koji se formira ciljno obeležje.

Potpuno nove mogućnosti, kao i problemi koje bi trebalo rešiti, se otvaraju ukoliko se u obzir uzme i vremenska komponenta. To bi podrazumevalo da se podaci o objavama prate na određenim vremenskim intervalima, kao i uzimanje vremenske komponente u obzir, na neki način, prilikom računanja ciljnog obeležja. Ovim pristupom bi moglo više da se sazna o promeni podataka vezanih za objave kroz vreme i tokovima događaja koji utiču na popularnost objave.

Literatura

[1] 9gag - https://9gag.com/ (posećeno 15. maja 2020.)

[2] SANDLER, Mark, et al. Mobilenetv2: Inverted residuals and linear bottlenecks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018. p. 4510-4520.

[3] Open Images Dataset – skup podataka https://storage.googleapis.com/openimages/web/index.html (posećeno 15. maja 2020.)

[4] SIMONYAN, Karen; ZISSERMAN, Andrew. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[5] ImageNet – skup podataka – http://www.image-net.org/ (posećeno 12. maja 2020.)

[6] Pytesseract – biblioteka za detekciju karaktera na slikama za *Python* programski jezil – https://github.com/madmaze/pytesseract (posećeno 11. maja 2020.)

[7] Stanford CoreNLP – softver za analizu prirodnog jezika https://stanfordnlp.github.io/CoreNLP/ (posećeno 15. maja 2020.)

[8] MEGHAWAT, Mayank, et al. A multimodal approach to predict social media popularity. In: *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2018. p. 190-195.

[9] MAZLOOM, Masoud, et al. Multimodal popularity prediction of brand-related social media posts. In: *Proceedings of the 24th ACM international conference on Multimedia*. 2016. p. 197-201.

[10] THELWALL, Mike, et al. Sentiment strength detection in short informal text. *Journal of the American society for information science and technology*, 2010, 61.12: 2544-2558.

[11] YANG, Diyi, et al. Humor recognition and humor anchor extraction. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2015. p. 2367-2376.

[12] REDDY, Thanika. An Analysis of the Popularity of Facebook News Posts. *Projects in Applied Data Science: Fall 2019*, 2019, p. 88-101.

[13] *Facebook News* skup podatak – https://github.com/jbencina/facebook-news (posećeno 31. maja 2020.)

[14] VADER – biblioteka za analizu sentimenta – https://github.com/cjhutto/vaderSentiment (posećeno 10. maja 2020.)

[15] HU, Jiani; YAMASAKI, Toshihiko; AIZAWA, Kiyoharu. Multimodal learning for image popularity prediction on social media. In: *2016 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*. IEEE, 2016. p. 1-2.

[16] YFCC100M – skup podataka – http://www.yfcc100m.org/ (posećeno 31. maja 2020.)

[17] HESSEL, Jack; LEE, Lillian. Something's Brewing! Early Prediction of Controversy-causing Posts from Discussion Features. *arXiv preprint arXiv:1904.07372*, 2019.

[18] TERENTIEV, Andrei; TEMPEST, Alanna. Predicting Reddit Post Popularity Via Initial Commentary. *nd): n. pag*, 2014.

[19] KHOSLA, Aditya; DAS SARMA, Atish; HAMID, Raffay. What makes an image popular?. In: *Proceedings of the 23rd international conference on World wide web*. 2014. p. 867-876.

[20] Jsoup – biblioteka za pribavljanje *HTML* stranica sa interneta za *Java* programski jezik - https://jsoup.org/ (posećeno 13. maja 2020.)

[21] Selenium – biblioteka za testiranje *Web* sajtova - https://www.selenium.dev/ (posećeno 13. maja 2020.)

[22] Selenium WebDriver – https://www.selenium.dev/projects/ (posećeno 13. maja 2020.)

[23] Imgflip – veb sajt – https://imgflip.com/ (posećeno 12. maja 2020.)

[24] LEE, Kevan. The proven ideal length of every tweet, Facebook post, and headline online. *Fast Company*, *Apr*, 2014.

[25] Sentistrength - http://sentistrength.wlv.ac.uk/ (posećeno 15. maja 2020.)

[26] DEVLIN, Jacob, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[27] https://towardsdatascience.com/extracting-keywords-from-short-text-fce39157166b (posećeno 14. juna 2020.)

[28] PEDREGOSA, Fabian, et al. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 2011, 12: 2825-2830.

[29] VEATCH, Thomas C. A theory of humor. 1998.

[30] https://www.tensorflow.org/tutorials/text/text_generation (posećeno 14. juna 2020.)

[31] https://colab.research.google.com/notebooks/intro.ipynb (posećeno 14. juna 2020.)