Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.85
DOI: https://doi.org/10.24867/16BE38Ilic

PREDIKCIJA POPULARNOSTI OBJAVA NA SAJTU 9GAG KORIŠĆENJEM MULTIMODALNIH PODATAKA SA FOKUSOM NA TEKSTUALNE PODATKE

POPULARITY PREDICTION OF 9GAG POSTS WITH USAGE OF MULTIMODAL DATA AND FOCUS ON TEXTUAL DATA

Nikola Ilić, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIKA I RAČUNARSTVO

Kratak sadržaj – *Predikcija popularnosti objava na sajtu 9gag bi mogla da bude od značaja za prepoznavanje i analizu faktora koji su bitni kod popularnosti sadržaja na društvenim mrežama. Ova istraživanja mogu biti korisna pre svega za oblast marketinga. U ovom radu predstavljena je arhitektura sistema za predikciju popularnosti objava. Fokus je bio na analizi onih elemenata objava koje se tiču teksta objave (komentara, naslova,...). Isprobana su dva načina, od prvi predstavlja testiranje više različitih klasifikatora, a drugi stekovanje klasifikatora. Na kraju su predstavljeni i analizirani rezultati, a predložena su i dalja unapređenja sistema.*

Ključne reči: *Analiza sentimenta, obrada teksta, predikcija popularnosti*

Abstract – *Popularity prediction of 9gag posts could help recognize and analyze key elements of popular content on social media. This type of research could be necessary for areas such as marketing. This paper presents the system's architecture for 9gag post popularity prediction, where the focus is on analyzing the textual elements of the posts. Two approaches were used - training multiple different classifiers and classifier staking. Results are presented and analyzed at the end, and future work of the system is discussed.*

Keywords: *Sentiment analysis, text mining, popularity prediction*

1. 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 tekst, dok je kod nekih bitna kombinacija ta dva. Sve objave mogu biti ocenjene pozitivno i negativno i 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 jeste predikcija popularnosti neke objave na sajtu *9gag*. Odnosno, analizira se 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.

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: naslov objave, *hashtag*-ovi, komentari, kao i podaci u vezi sa komentarima (broj komentara i sl.). Ovde predstavljeni klasifikator teksta je deo većeg sistema i korišćen je u kombinaciji sa klasifikatorom koji radi na osnovu slike za koji su takođe bila ekstrahovana obeležja od značaja.

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. 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 se obuče na ekstrahovanim obeležjima. Za to postoji više razloga. Neki od njih su činjenica da se model u velikoj meri oslanja na broj komentara kao obeležje i da je utvrđeno da vrlo često sam naslov nema skoro nikakve veze sa samom objavom.

U poglavlju 2 će biti izložena postojeća literatura koja se bavi sličnim problemom. U poglavlju 3 biće govora o skupu podataka koji je korišćen, a u poglavlju 4 će biti detaljnije pojašnjenje arhitekture sistema i njegova implementacija. U poglavlju broj 5 će biti opisan postupak verifikacije i prikaz rezultata. Konačno, u poglavlju 6 iznesen je zaključak rada i opisane su neke ideje i mogućnosti za dalji rad koje bi mogle da doprinesu unapređenju postojećeg rešenja.

2. SRODNA ISTRAŽIVANJA

U toku pregleda literature relevantne za ovaj rad nije pronađen ni jedan rad koji se bavi problemom predikcije popularnosti objava na sajtu 9gag. Zato će u ovom poglavlju biti izložen prikaz radova koji se bave

---

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bila dr Jelena Slivka, vanr. prof.

problemom predikcije popularnosti na drugim društvenim mrežama, kao i radovi u kojima je fokus na tekst kao glavni aspekt rešavanja problema kojim se bave ti radovi.

Cilj rada [2] jeste predviđanje popularnosti objava na društvenoj mreži *Flickr*[1]. Opisani pristup prilikom predikcije koristi vizuelna obeležja, koja se dobijaju analizom slike, obeležja dobijena analizom teksta objave i društvena obeležja, poput prosečnog broja pregleda korisnika koji je objavu postavio. Korišćen je postojeći SMP-TI skup podataka [3], koji ukupno sadrži 432 hiljade objava sa društvene mreže *Flickr*. Za analizu kontekstnih i društvenih informacija korišćeni su *Random Forest* algoritam i namenski kreirana konvolutivna neuronska mreža sa šest slojeva. Analiza naslova i tagova se bazirala na upotrebi rečnika, a sentiment opisa je određen pomoću *Stanford CoreNLP*[2] biblioteke. Sentiment se ocenjuje sa pet opisnih ocena na osnovu čega se izvodi ocena opisa. Ovakvom analizom se došlo do 15 obeležja, koja se potom normalizuju i vrši se predviđanje popularnosti objave pomoću konvolutivnog modela. 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. Rad [2] se može smatrati najsličniji ovom istraživanju i zbog toga je dao značajne putokaze na koji način se mogu prikupljati i koristiti multimodalni podaci. Analiza objekata na slikama, naslova, tagova i sentimenta vršene su na sličan način i u ovom istraživanju.

Cilj rada [4] jeste da se utvrdi šta sve u okviru objave koja se odnosi na brend brze hrane na društvenoj mreži *Instagram*[3] utiče na to da objava bude popularnija. U radu su korišćeni parametri koje su autori nazvali *engagement parameters*, od kojih se neki odnose na sliku, a neki na tekst. S obzirom na to da je za ovaj rad relevantno samo to kako su iskorišćena obeležja u vezi sa tekstom, neće biti reči o obeležjima i metodama u vezi sa slikom. Prikupljeno je 75 hiljada objava koji se odnose na šest poznatih lanaca brze hrane i od njih je sačinjen skup podataka. Kao obeležja u vezi sa tekstom korišćeni su komentari, naslovi i *hashtag*-ovi nad kojima je izvršena analiza sentimenta. Obrada teksta je rađena slično kao i u ovom radu, gde je analiza sentimenta bila primenjena samo na komentare.

Sličan zaključak i značaj za ovo istraživanje ima rad [5], 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. Prikupljeno je oko 2220 primera koji čine skup podataka. Obeležija koja su izvučena su: *maximum*, *minimum*, *lower-half average*, *upper-half average* i *average* ocena dobijena analizom sentimenta komentara. Analiza sentimenta je vršena pomoću *Stanford NLTK* biblioteke. Slično kao u [2] i ova biblioteka daje rejting za svaku rečenicu, a rejtinzi su: veoma negativno, negativno, neutralno, pozitivno i veoma pozitivno. Svakom od rejtinga je data brojna vrednost i to redom –2, -1, 0, 1, 2. Svaki komentar je dobio svoju sentiment ocenu na osnovu srednje vrednosti sentiment ocena svih rečenica u tom komentaru. Razlika pristupa predstavljenog u radu [5] i pristupa prikazanog u ovom radu je ta da je u ovom radu korišćena procentualna vrednost za meru ocene, gde se kao krajnja ocena sentimenta uzima ona koja ima najveću procentualnu vrednost. Rezultati rada [5] su dobijeni primenom nekoliko metoda, a to su: *K-Neighbors*, *Perceptron*, *Support Vector Machine* i *Logistic Regression.* Pored značaja ranih reakcija na neku objavu za njenu popularnost, rad [5] pokazuje da sentiment komentara treba uzeti u razmatranje prilikom predikcije popularnosti objava.

Cilj rada [6] jeste predikcija popularnosti onlajn peticija putem multimodalnog dubokog regresionog modela. Skup podataka korišćen u radu [6] sastoji se od oko 75 hiljada peticija prikupljenih sa sajta „Avaaz.org“. Slično kao i u ovom radu, autori evaluiraju model na osnovu tri klase modela: tekstualni model, model slike i kombinovani model slike i teksta. Modeli korišćeni za tekst su CNN regresioni model i BERT regresioni model. Eksperimenti su vršeni na skupu podataka koji je bio nasumično podeljen na trening, validacioni i test skup. Takođe, valja napomenuti i da su eksperimenti posebno vršeni na skupu podataka koji je sačinjen samo od podataka na engleskom jeziku, a posebno vršeni na skupu podataka od nekoliko različitih jezika. Srednja apsolutna procentualna greška (eng. *mean absolute percentage error* - MAPE) i Spirmanov koeficijent korelacije su korišćeni za računanje evaluacije. Autori rada [6] na osnovu dobijenih rezultata zaključuju da tekstualni model ima nešto bolje performanse u odnosu na model slike, dok kombinovani model ima nešto bolje rezultate od model teksta. Na osnovu Spirmanovog koeficijenta korelacije vrednosti za tekstualni BERT model su $\rho = 0.385$, tekstualni CNN model $\rho = 0.363$, a dok su vrednosti za model slike $\rho = 0.235$. Spirmanov koeficijent korelacije za kombinovani model je $\rho = 0.405$. Ovde naznačeni rezultati su dobijeni na osnovu skupa podataka koji se sastoji samo od podataka na engleskom jeziku. S obzirom na nešto bolje performanse BERT modela u odnosu na CNN model, za dodatnu analizu sentimenta komentara u ovom radu odabran je BERT model.

3. 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.

Svaka objava sadrži sledeće korisne informacije:

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

[1] https://www.flickr.com/
[2] https://stanfordnlp.github.io/CoreNLP/
[3] https://www.instagram.com/

Tabela 1. *Rezultati za prvi pristup na test skupu*

Metod / Mera	*Baseline*	Linearna regresija	*Elastic net*	SVR	*Random forest*	*Voting*
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**

Tabela 2. *Rezultati za drugi pristup na test skupu*

Podaci / Mera	Informacije sa slika	Sentiment komentara – Core NLP	Sentiment komentara - BERT	Ključne reči u naslovu	Konačna predikcija
R^2	0.0211	0.1443	0.1498	0.0014	0.4843
P	0.1255	0.3877	**0.4117**	0.002	**0.6782**

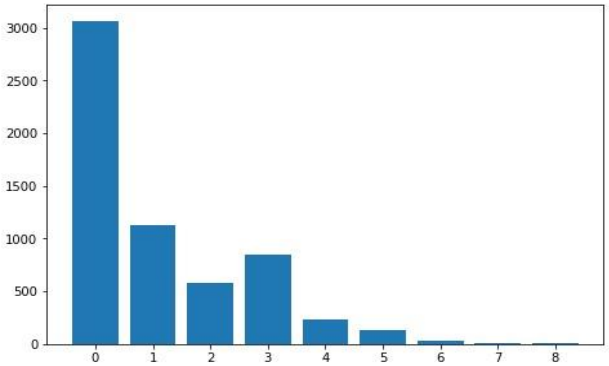
- URL ka originalnoj objavi na sajtu
- Informacije o tagovima.

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)

4. SPECIFIKACIJA I IMPLEMENTACIJA SISTEMA

Za analizu tekstualnih podataka, pretpostavka, utemeljena u analizu tekstualnih podataka u srodnim istraživanjima iz poglavlja 2, je da bi od značaja za popularnost objave bili naslovi, *hashtag*-ovi i komentari objava. Isprobano je nekoliko metoda kako bi se iz podataka pribavile informacije potrebne za izvođenje predikcije popularnosti objava.



Slika 1. *Broj objava koje imaju od 0 do 8 hashtag-ova*

4.1. 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 slici 1. Prosečan broj pozitivnih, odnosno, negativnih ocena objava sa istim brojem *hashtag*-ova varira sa tolerancijom od ±10%. 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.

Poređena je i popularnost objava sa *hashtag*-ovima i bez njih i nema primetne razlike u njihovom odnosu. Iz svega navedenog dolazi se do zaključka da broj *hashtag*-ova ni na koji način ne utiče na popularnost objave.

4.2. Ekstrakcija ključnih reči iz naslova

Ključne reči su izvlačene TFIDF metodom na korpusu podataka koji je obuhvatao naslove svih dobavljenih objava.

Kod njihovog odabira, primećeno je da se najfrekventnije 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.

4.3. Određivanje sentimenta komentara

Analiza sentimenta komentara je vršena pomoću *Stanford CoreNLP* API-ja [5] za analizu i obradu prirodnog jezika, a dodatno je korišćen i pretrenirani BERT model. 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 krajnja vrednost sentimenta.

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%.

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*[4] i koji ima tačnost od oko 88% na test skupu, prilikom određivanja sentimenta. Prilikom treniranja kao optimizator je korišćen *AdamW*, sa preporučenim parametrima iz rada [8], i *cross-entropy* kao *loss* funkcija. 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

[4] https://play.google.com/

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.

4.4. Opis celokupnog sistema

Celokupan sistem predstavlja kombinaciju različitih klasifikatora slike i teksta.
Isprobana su dva različita načina dobijanja konačne predikcije modela, a to su:

1. prosleđivanje svih obeležja modelima koji se treniraju
2. stekovanje modela zasnovanih na različitim skupovima obeležja (engl. *stacking*)

Što se tiče prvog načina, isprobano je nekoliko modela. Modeli koji su se pokazali najbolje bili su: SVR, ansambl regresionih stabala i *voting* model.

Drugi pristup se sastoji od posebne predikcije popularnosti objava za podatke iz različitih izvora. 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 ansambl regresionih stabala. Rezultati se mogu videti u tabeli 2.

5. VERIFIKACIJA I REZULTATI

Podela skupa podataka na trening skup i test skup je vršena tako što je 80 % skupa izdvojeno za trening, a 20 % za test. Ova podela je važila za sve modele.
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. Rezultati se mogu videti u tabeli 1.
Konačna predikcija zasnovana na stekovanju modela je nešto lošija nego kod najuspešnije metode prvog pristupa, ali ne značajno. Rezultati drugog pristupa se mogu videti u tabeli 2.
Dalji rad na usavršavanju performansi ovih modela bi mogao da se sastoji od poboljšanja analiza vršenih na tekstualnim podacima. Pre svega na tome da se reši problem sarkazma kada je u pitanju sentiment komentara.

6. ZAKLJUČAK

Problem koji se rešavao u ovom radu tiče se predikcije popularnosti objava na sajtu 9gag. Rešavanje ovog problema moglo bi 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 vršena predikcija popularnosti neke objave, gde su za modele koji su trenirani korišćena obeležja dobijena analizom elemenata od značaja za neku objavu.

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 uporede.

Nedostaci ovog rešenja kada je reč o tekstualnim podacima bi bili nemogućnost metoda da prepoznaju sarkazam, kako u naslovima, tako i u komentarima. Buduća proširenja i poboljšanja sistema mogla bi da se vrše na onim delovima tekstualne analize koji su dali najbolje rezultate, pokušajem analize teksta metodama koje nisu isprobane, a potencijalno bi mogle da daju bolje rezultate.

7. LITERATURA

[1] 9gag - https://9gag.com/
[2] 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.
[3] SMP-T1 in ACM Multimedia Grand Challenge. https://social-media-prediction.github.io/MM17PredictionChallenge/leaderboard.html, 2017
[4] 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.
[5] Stanford Core NLP - https://stanfordnlp.github.io/CoreNLP/
[6] TERENTIEV, Andrei; TEMPEST, Alanna. Predicting Reddit Post Popularity Via Initial Commentary. nd): n. pag, 2014.
[7] Kitayama, Kotaro et al. "Popularity Prediction of Online Petitions using a Multimodal DeepRegression Model." ***ALTA*** (2020).
[8] DEVLIN, Jacob, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

Kratka biografija:

Nikola Ilić rođen je 9.1.1996. godine u Novom Sadu. Osnovnu školu „Miroslav Antić" u Futogu, završio je 2011. godine. Iste godine upisao je gimnaziju „Isidora Sekulić" u Novom Sadu, na prirodno-matematičkom smeru, koju je završio sa odličnim uspehom. Osnovne akademske studije na Fakultetu tehničkih nauka u Novom Sadu, smer računarstvo i automatika, upisuje 2015. godine i završava 2019. godine sa prosečnom ocenom 8.57. Odbranio je diplomski rad na temu „Veb servisi i njihova implementacija upotrebom .NET tehnologije".