Predlog projekta iz predmeta
“Sistemi za istraživanje i analizu podataka”

Jelena Budiša	Stevan Rašković	Stefan Jokić
E2 47/2021	E2 54/2021	E2 57/2021

1. Definicija problema

Tema ovog projekta biće predviđanje broja lajkova koji post ima u određenom vremenskom trenutku na osnovu parametara kao što su broj pratilaca osobe koja ga objavljuje, proteklo vreme od objavljivanja, broj tagovanih osoba, opis posta, opis profila, prosečan broj komentara... Projekat će se baviti analizom uticaja pomenutih parametara na popularnost posta.

2. Motivacija

Popularnost nekog posta je bitan faktor prilikom angažovanja Instagram influensera za marketinške kampanje. Poznavanje faktora koji utiču na to koliko će post biti popularan može imati značajnu ulogu prilikom odabira Instagram profila sa kog će neki proizvod biti reklamiran, kao i veliki uticaj na sam izgled i sadržaj posta. Takođe, influenserima bi poznavanje utica
ja određenih faktora na popularnost posta omogućilo bolje plasiranje svojih postova, veću popularnost i potencijalno veći broj novih pratilaca.

3. Relevantna literatura

- *[1] “Multimodal Deep Learning Framework for Image Popularity Prediction on Social Media” Fatma S. Abousaleh, Wen-Huang Cheng, Neng-Hao Yu, and Yu Tsao, Senior Member, IEEE, 2021.* *https://arxiv.org/pdf/2105.08809.pdf*

  Osnovni zadatak ovog rada jeste da se predvidi popularnost objave slike na društvenoj mreži Flickr. Predviđanje popularnosti slike se vrši na osnovu vizuelnog sadržaja i društvenog konteksta slike. Društveni kontekst slike čine obeležja o korisniku, obeležja o metapodacima objave, kao i obeležja o vremenu objave slike. Društveni kontekst se dobija analizom korisnika i teksta objave slike, dok se vizuelni sadržaj izvlači iz same slike, tj. gledaju se podaci poput kvaliteta slike, zasićenosti boja, jednostavnosti pozadine i dobijaju se

low-level, high-level i deep learning obeležja. Za dataset je korišćen Flickr-ov dataset, koji sadrži oko 432 hiljade slika. U ovom radu je korišćen VSCNN, odnosno visual-social konvoluciona neuronska mreža, koja koristi vizuelna i društvena obeležja kako bi prediktovala popularnost objavljene slike. To je zapravo mreža, koja je nastala spajanjem dve konvulucione neuronske mreže, gde jedna mreža ekstraktuje vizuelna obeležja, a druga ekstraktuje društvena obeležja. Obe mreže se sastoje od 3 konvuluciona sloja. Nakon ekstrakcije obeležja dobija se vektor obeležja čija se dimenzija smanjuje postupkom analize glavnih komponenti. Ovaj postupak se koristi i zbog selektovanja samo preovlađujućih obeležja. Obeležja se na kraju normaliziju i kao takva se koriste za treniranje VSCNN modela.
Što se tiče metoda evaluacije ovog rada, korišćene su tri metode, a to su Spirmanov koeficijent korelacije, srednja kvadratna greška i srednja apsolutna greška. Rezultati rada su pokazali da je VSCNN model poprilično outperformovao SOTA modele u pogledu predvidjana popularnosti slike. Ovaj rad je upotrebljen zato što se bavi istom tematikom kao i naš rad.

- *[2] Instagram Post Popularity Trend Analysis and Prediction using Hashtag, Image Assessment, and User History Features*
  https://iajit.org/portal/PDF/Vol%2018,%20No.%201/19395.pdf

  Rad se bavi procenom popularnosti Instagram posta odnosno prosečnim brojem interakcija koje korisnik ostvari sa postom (Engagement Rate). Osnovna motivacija za rešavanje ovog problema u radu je predstavljena kao potreba za određivanje korisnika koji su pogodni za marketing putem Instagrama. Kao ulazne podatke rad koristi hashtagove, kvalitet i sadržaj fotografije, istoriju korisnika… Rad je koristio podatke o 19324 posta objavljenih od strane 16804 korisnika. Neka od najvažnijih obeležja koja su bila u upotrebi su broj pratilaca, broj postova, broj hastagova u postu, dužina opisa posta, vreme objavljivanja. Pored toga, uzeti su u obzir parametri vezani za samu sliku kao što su kvalitet slike, poza u kojoj se osoba na slici nalazi, mesto gde je slikana… Za parametre vezane za istoriju korisnika najvažniji su bili procenat pratilaca koji su lajkovali bar jednu sliku korisnika, prosečan broj lajkova kao i prosečan broj komenatra koji ostavi jedan pratilac.
  Za rešavanje problema u radu su upotrebljeni Random Forest, Linear Regression i Support Vector Machine algoritmi.
  Kao metode evaulacije u radu su korišteni R2, Mean Squere Error, Mean Absoulte Error i druge. Studija je pokazala da kvalitet slike, vreme objave, kao i tip slike imaju najveći uticaj na broj interakcija sa postom, dok se istorija korisnika pokazala kao manje relevantan faktor. Preciznost predicije je dostigla čak 73,1% koristeći SVR, što je veći rezultat nego prethodne studije nad korišćenim datasetom. Ovaj rad nam je značajan jer se bavi jako sličnim

problemom koji i mi pokušavamo rešiti. Takođe neke od navedenih metoda biće korištene i u našem pristupu.

- *[3] Zhongping Zhang, "How to Become Instagram Famous: Post Popularity Prediction with Dual-Attention", 2018 IEEE International Conference on Big Data, 2019.*
  https://arxiv.org/pdf/1809.09314.pdf

  Cilj ovog rada je da se prediktuje popularnost objave na Instagramu za određenog korisnika, na osnovu objavljene slike, teskta slike, kao i podataka o samom korisniku. Predikcija se obavlja pomoću dual-attention modela, koji se sastoji od dva dela, tj eksplicitnog i implicitnog modela. Oni primaju različite informacije kao input, a nakon toga se konkateniraju. Eksplicitni model se bavi ekstraktovanjem obeležja iz teksta i slike objave, gde se za slike koristi konvoluciona neuronska mreža, a za tekst LSTM. Implicitni model se bavi ekstraktovanjem obeležja iz podataka o korisniku pomoću LDA. Pošto ne postoji javni dataset za Instragram, u ovom radu je rađen data scraping Instragrama, gde je prikupljeno 441 korisnika i njihovih 60785 parova slika-teskt. Za evaluaciju rešenja se koristi klasifikaciona evaluacija i to precision, recall, F-measure i accuracy. Rezultati klasifikacije su pokazali da model daje dobre predikcije popularnosti posta, kao i kakve vrste slika i teksta treba koristiti u objavi, kako bi se pomoglo korisniku da ostvari veći broj lajkova. Ovaj rad je upotrebljen jer se bavi istom tematikom poput našeg rada.

4. Skup podataka

Skup podataka biće prikupljen scrapovanjem Instagram sajta. Tom prilikom biće preuzeti svi relevantni podaci o korisnicima i postovima (broj pratilaca, broj praćenih korisnika, broj postova, broj lajkova po postu, broj komentara po postu, vreme objave posta, opis posta, tagovane osobe na postu, slika posta). Ograničićemo se samo na postove koji su fotografije, i to samo jedna fotografija u postu. Pored scrapovanja, podaci će se prikupljati u upotrebom Instagram API-a a po potrebi i biti ručno dopunjeni.

5. Metodologija

Prvi korak biće obrada prikupljenih podataka. Svi sirovi podaci biće analizirani, iz njih će biti izbačeni outlayeri. Upotrebom gotovog rešenja za analizu sentimenta opisa posta i sentimenta komentara vezanih za post (StanfordCoreNLP, Textblob….). Za analizu podataka koristićemo i metode eksplorativne analize. Takođe radiće se i analiza slika pomoću neuronske mreže, tj prepoznavaće se

objekti na slikama, kao i kakav je kvalitet slike. Radiće se i klasifikacija korisnika na više grupa po broju pratilaca(po broju ljudi koji oni prate i broj ljudi koji ih prati). Takođe radiće se i klasifikacija postova u zavisnosti od vremena objavljivanja (jutro, podne, veče, itd.) Na osnovu ovih parametara biće kreirana dodatna obeležja koja će se koristiti za predikciju broja lajkova. Za predikciju broja lajkova biće upotrebljeni Random Forest, Linear Regression, Gradient Boosted Trees, neuronske mreže, a problem ćemo po potrebi pokušati rešiti i nekim drugim metodama koje otkrijemo u relevantnoj literaturi.
s

6. Metod evaluacije

Skup podataka ćemo podeliti na trening i test, pri čemu će u skupu za testiranje biti 20% postova od svakog korisnika. Za evaluaciju ćemo koristiti metrike koje smo pronašli u relevantnoj literaturi, kao što su R2, Mean Square Error, Mean Absoulte Error i druge. Dobijene vrednost će biti izračunate za sve modele koje budemo koristili i biće izvršeno njihovo poređenje i analiza.