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

U okviru ovog predloga projekta dat je kratak opis teme i definicije rešavanog problema, kao i motivacija za implementaciju opisanog rešenja. Nakon motivacije sledi pregled vladajućih stavova i shvatanja u literaturi, a zatim opis korišćenog skupa podataka. Takođe, naveden je i softver koji će biti korišćen za realizaciju koraka opisanih u pregledu metodologije, kao i metod evaluacije. Na samom kraju dokumenta dat je plan rada na projektu.

1. Definicija problema

Cilj projekta je predikcija vremenskog perioda koje će video provesti u *YouTube trending*-u. Predikcija će se vršiti na osnovu numeričkih atributa skupa podataka (broj pregleda, lajkova, komentara i slično), leksičke analize naziva i opisa videa, kao i semantičke analize komentara. Takođe, planirano je i korišćenje *thumbnail*-a za predikciju, upotrebom konvolutivne neuronske mreže.

2. Motivacija

Poslednjih godina *YouTube* je postao najpopularnija platforma za širenje i prikaz video materijala. Broj ljudi koji provodi vreme na *YouTube*-u je u konstantnom porastu, što je dovelo do toga da je objavljivanje videa postalo izuzetno unosan posao. *YouTube* ima listu najpopularnijih videa za svaku državu, koja se naziva *Trending*. Svim velikim kanalima je u interesu da budu što duže na listi najpopularnijih, jer time zarađuju više novca. Zbog velike količine sadržaja koji se svakodnevno objavljuje, jasno je da to nije najjednostavniji zadatak, pa je neophodno detaljno planiranje izgleda sadržaja koji se objavljuje. Pod tim se, pored samog sadržaja videa, podrazumevaju i naziv videa, opis, *thumbnail* itd. Upravo u toj važnosti vremena provedenog u *Trending*-u krije se motivacija ovog rada. Cilj je kanalima omogućiti predikciju vremena koje će njihov video provesti na listi najpopularnijih, i to na osnovu podataka o videu dobijenih samo iz prvog dana boravka u *trending*-u.

3. Relevantna literatura

U ovom poglavlju dat je pregled literature relevantne za problem rešavan u ovom radu. Akcenat je na radovima koji se bave analizom sentimenta, jer je to deo projekta kom će biti posvećeno najviše pažnje, u cilju obogaćivanja početnog skupa podataka.

[1] Hanif Bhuiyan, Jinat Ara, Rajon Bardhan, Md. Rashedul Islam, “Retrieving YouTube Video by Sentiment Analysis on User Comment”, IEEE ICSIPA 2017, Malaysia, September 12-14, 2017.

- *Zadatak rada*
  Zadatak rada jeste automatizacija procesa pronalaženja adekvatnih videa na *YouTube*-u, upotrebom sentiment analize komentara korisnika. Ideja rada jeste poboljšavanje rezultata pretrage na *YouTube*-u adekvatnom analizom korisničkih komentara.
- *Metodologija*
  Metodologija opisana u radu sastoji se iz četiri koraka: prikupljanje podataka i pretprocesiranje, generisanje skupova podataka *NLP* tehnikama, računanje pozitivnosti i negativnosti sentimenta komentara upotrebom *SensiStrength*-a i računanje standardne devijacije za dobijanje konačnog rejtinga videa.

Prikupljanje komentara realizovano je upotrebom *YouTube API*-a. Nad tako prikupljenim komentarima izvršeno je pretprocesiranje, koje je obuhvatalo: izbacivanje datuma, linkova, specijalnih karaktera, znakova interpunkcije, kao i uklanjanje komentara koji nisu na engleskom jeziku.

Za svaki video za koji su se prikupljani komentari kreirana su po dva dataset-a. Prvi skup podataka formiran je izbacivanjem stop reči (*stop words*) iz komentara i prebacivanjem svih reči u jedninu. Drugi skup podataka formiran je izdvajanjem prideva iz komentara, upotrebom *Stanford Part-of-Speech tagger*-a (*POS tagger*). U okviru rada navedeno je da su empirijski i analizom podataka došli do zaključka da su pridevi najznačajniji identifikatori osećanja korisnika.

Za sentiment analizu komentara korišćen je *SentiStrength thesaurus*, nad oba skupa podataka. Svaka reč u okviru komentara dobija dve ocene sentimenta: -1 do -5 za negativan sentiment, i +1 do +5 za pozitivan. Na osnovu ocena pojedinačnih reči, komentar dobija konačnu ocenu sentimenta, za svaki od skupova podataka.

Poslednji korak metodologije opisane u radu jeste računanje dve vrednosti standardne devijacije za svaki video, po jedna za ocene sentimenta komentara iz oba skupa podataka. Aritmetička sredina te dve vrednosti predstavlja meru sličnosti sentimenta komentara za svaki video.

- <u>*Skup podataka*</u>
  Podaci su prikupljani pomoću *YouTube API*-a. Sakupljeno je po 1000 komentara za ukupno 1000 videa, raspoređenih u 10 različitih kategorija. Od prikupljenih komentara formirana su dva skupa podataka, koji su detaljnije opisani u okviru metodologije.
- <u>*Evaluacija rešenja*</u>
  Za evaluaciju predložene metode korišćena je mera tačnost, čija je vrednost bila oko 75%.
- <u>*Zaključak*</u>
  Planiramo da primenimo sličnu metodologiju za analizu sentimenta komentara kao u opisanom radu. Prikupljanje podataka ćemo vršiti na isti način, a za pretprocesiranje i obradu podataka *NLP* metodama planiramo da, pored ostalih tehnika koje ćemo primenjivati, isprobamo i navedene tehnike.

[2] Angel Iek, Hou Zhang, “Judging YouTube by its Covers”, Department of Computer Science and Engineering, University of California, San Diego, 2015.

- <u>*Zadatak rada*</u>
  Zadatak rada jeste predikcija popularnosti *YouTube* videa, tj. broja pregleda, na osnovu naziva videa.
- <u>*Metodologija*</u>
  Kreirano je šest rečnika, upotrebom prikupljenih naslova i opisa videa. Rečnici su se razlikovali po tome da li su kreirani za unigrame ili bigrame naslova i opisa videa, kao i po uslovu broja pojavljivanja reči (da li se reč pojavljuje više od jednom, tri ili pet puta u skupu naslova ili komentara). Pre kreiranja rečnika, vršeno je pretprocesiranje podataka, izbacivanjem uobičajenih reči koje imaju malu prediktivnu vrednost. Za svaki od rečnika kreirana je *sparse* matrica, čiji redovi predstavljaju naslovi/opisi videa, a vrednosti u 0 ili 1, zavisno od toga da li se određena reč iz rečnika nalazi u naslovu. Nad ovakvim podacima obučavana su tri klasifikatora: *Naive Bayes*, *SVM* i *kNN*, da bi se utvrdilo koji klasifikator je najbolji za date podatke.
- <u>*Skup podataka*</u>
  Podaci su prikupljani upotrebom *YouTube API*-a. Prikupljeni su podaci o skoro 53 hiljade videa, sa obeležjima: naziv videa, naziv kanala (profila) koji ga je objavio, opis, trajanje, broj lajkova, komentara i pregleda. Ciljna labela bila je broj pregleda, koji je diskretizovan, za potrebe obučavanja binarnog klasifikatora (visok/nizak broj pregleda). Za klasifikaciju su korišćeni samo naziv i opis videa.

- *Evaluacija rešenja*
  Za evaluaciju rešenja korišćena je mera tačnost, čija je vrednost bila oko 66%.
- *Zaključak*
  U navedenom radu prikazana je predikcija gledanosti videa pomoću naslova i opisa. Prikazano je da *Naive Bayes* model postiže bolje rezultate od *SVM* i *kNN* modela, za korišćene podatke. Najbolji rezultat tačnosti postignut je upotrebom upravo *Naive Bayes* modela, uz unigram rečnik kreiran od svih reči iz naslova videa. U radu je naglašeno da je za očekivati još bolje rezultate uključivanjem i *thumbnail* slike, što je uz korišćenje naslova i opisa svakako u planu u okviru našeg rada. Takođe, u radu je urađena dobra eksplorativna analiza podataka, što će biti od koristi za realizaciju našeg rada.

[3] William Hoiles, Anup Aprem, Vikram Krishnamurthy, "Engagement dynamics and sensitivity analysis of YouTube videos", IEEE Transaction on knowledge and data engineering, 2016.

- *Zadatak*
  U radu je sprovedeno istraživanje kako *meta-level feature*-i i interakcija *YouTube* kanala sa korisnicima utiče na broj pregleda videa, tj. njegovu popularnost. Takođe, vršena je i predikcija popularnosti *YouTube* videa na osnovu konstruisanih obeležja.
- *Metodologija*
  Metodologija opisana u radu sastoji se iz više koraka. Prvi korak je definisanje *meta-level feature*-a za glavne komponente svakog videa: *thumbnail*, naziv i ključne reči (*hashtags*). Potom je izvršena analiza osetljivosti (*sensitivity analysis*) konstruisanih obeležja nad ciljnom labelom, tj. brojem pregleda. Konačno, izvršena je predikcija broja pregleda na osnovu najbitnijih *meta-level feature*-a, upotrebom nekoliko različitih modela mašinskog učenja. Korišćene metode mašinskog učenja: *Extreme Learning Machine (ELM)*, *Feed-Forward Neural Network*, *Stacked Auto-Encoder Deep Neural Network*, *Elasticnet*, *Lasso*, *Relaxed Lasso*, *Quantile Regression with Lasso*, *Conditional Inference Random Forest (CIRF)*, itd.
- *Skup podataka*
  Rad je koristio skup podataka obezbeđen od strane *BBTV*-a (*BroadbandTV Corp*). Skup podataka sadrži dnevne informacije o *YouTube* videima na *BBTV* platformi od aprila 2007. do maja 2015. U okviru skupa podataka postoje podaci o oko 6 miliona videa na preko 25 hiljada kanala.
- *Evaluacija rešenja*
  Modeli su trenirani upotrebom unakrsne validacije sa 10 preklapanja (*10-folds cross validation*). Pošto se u rada predikcija popularnosti videa vršila regresijom, evaluacija rešenja vršena je pomoću mera *RMSE* (*Root Mean Squared Error*) i R^2.
- *Zaključak*
  U navedenom radu prikazan je veliki broj različitih regresionih modela mašinskog učenja, što će biti od velikog značaja za naš rad, u kom planiramo da koristimo regresiju za predviđanje vremena koje će video provesti u *trending*-u. Takođe, konstrukcija *feature*-a iz *thumbnail*-a, naslova videa i *hashtag*-ova detaljno je obrađena u ovom radu, i biće od velikog značaja za naš rad. U radu su pokazali da su ključni *meta-level feature*-i za predikciju popularnosti videa: broj pregleda tokom prvog dana, broj *subscriber*-a, kontrast *thumbnail*-a, *Google hits* (broj rezultata kada se u *Google* pretragu ukuca naslov videa) i broj ključnih reči (*hashtag*-ova).

4. Skup podataka

Skup podataka biće konstruisan upotrebom *YouTube API*-a, po uzoru na postojeći skup podataka koji se može preuzeti sa linka https://www.kaggle.com/datasnaek/youtube-new. Planirano je da se prikupljanje podataka vrši u dve faze. U prvoj fazi prikupljaju se podaci o samom videu, a u drugoj se prikupljaju komentari za odgovarajući video. Atributi (*feature*-i) koji će se prikupiti za svaki video u prvoj fazi su: ID videa, datum kog je video ušao u *trending*, naziv, naziv kanala koji je objavio video, kategorija videa, datum objavljivanja videa, lista tagova, broj pregleda, lajkova, dislajkova, komentara, link na *thumbnail*, opis videa, i binarne vrednosti da li su komentari i ocenjivanje omogućeni, kao i da li je video ukljonjen sa *YouTube*-a. Prvi korak u radu sa ovim podacima jeste formiranje ciljnog obeležja. *YouTube API* vraća podatke za isti video više puta, tako da svaka pojava određenog videa u skupu podataka predstavlja jedan dan proveden u *trending*-u. Prebrojavanjem broja pojavljivanja svakog videa u skupu podataka dobija se broj dana provedenih u *trending*-u za svaki video, tj. ciljno obeležje. Pošto je cilj rada predikcija vremena koje video provede na listi najpopularnijih, za svaki video će se posmatrati samo prvi red u skupu podataka. To znači da će se model obučavati tako da predviđa vreme koje će video provesti u *trending*-u samo na osnovu prvog dana kad je video ušao u *trending*. U okviru druge faze prikupljaju se komentari za svaki od videa u okviru prvog skupa podataka. Dva skupa podataka dobijena na ovaj način objedinjuju se u jedan pomoću obeležja ID videa, koje je jedinstveno za svaki video.

5. Metodologija

Prvi korak u realizaciji rada jeste prikupljanje podatka upotrebom *YouTube API*-a. Pošto *API* za svaki video vraća N redova, gde je N broj dana koje je video proveo u *trending*-u, potrebno je formiranje ciljne labele na osnovu broja pojavljivanja svakog videa. Potom se odbacuju svi podaci koji se ne odnose na prvi dan proveden u *trending*-u, jer je cilj rada predikcija vremena zadržavanja videa u *trending*-u, što ima smisla raditi čim video dospe na listu najpopularnijih. Takođe, za svaki video vrši se i prikupljanje komentara korisnika.

Kada su podaci prikupljeni, vršiće se eksplorativna analiza podataka uz adekvatnu vizuelizaciju, da bi saznali što više informacija o samim podacima. Eksplorativna analiza će biti neophodna za smislenu diskretizaciju vremena koje je video proveo u *trending*-u. Ovaj korak je neophodan, pošto ćemo predikciju vršiti i regresijom i klasifikacijom.

Nakon eksplorativne analize, sledeći korak je obučiti nekoliko različitih modela mašinskog učenja, pre vršenja bilo kakvih dodatnih sentiment analiza komentara, ekstrakcije *feature*-a iz *thumbnail*-a i naslova videa, itd. Ideja je krenuti od tih nekoliko modela obučenih nad „osnovnim“ skupom podataka, i potom iterativno obogaćivati podatke rezultatima analiza teksta, slike, itd. Predikcija vremena provedenog u *trending*-u biće realizovana regresionim i klasifikacionim modelima. Od regresionih modela, po uzoru na rad [3], isprobaće se *Lasso*, *ElasticNet*, *Ridge* i linearna regresija. Za klasifikaciju videa na osnovu vremena provedenog u *trending*-u planirano je da se isprobaju klasifikatori: *Random Forest*, *Support Vector Machine (SVM)*, *Gradient Boosting (XGBoost)*.

Sledeći korak jeste sentiment analiza korisničkih komentara, u cilju poboljšanja predikcije vremena provedenog u *trending*-u. Ideja je za svaki video pronaći koliko je bilo pozitivnih, neutralnih i negativnih komentara, i te podatke dodati u originalan skup podataka, da bi modeli mašinskog učenja za regresiju i klasifikaciju imali još obeležja kojima je moguće poboljšati predikciju. Da bi ovo bilo moguće, neophodno je izvršiti pretprocesiranje komentara, koje će obuhvatati: izbacivanje svih podataka koji nisu od velikog

značaja (datumi, brojevi, specijalni karakteri, komentari koji nisu na engleskom jeziku, itd.), izbacivanje svih stop reči i svođenje reči na njen koren (*stemming*). Od ovakvih podataka kreira se prvi *dataset*, koji sadrži pretprocesirane korisničke komentare. Drugi *dataset* formira se izdvajanjem prideva iz prvog *dataset*-a. Za izdvajanje prideva koristiće se *Stanford Part-of-Speech tagger*, kao što je urađeno u radu [1]. Ovim pristupom dobijamo po dva skupa podataka za svaki video. Sledeći korak je određivanje sentimenta komentara, pomoću dva kreirana skupa podataka. Za određivanje sentimenta prvo će se isprobati *SentiStrength* klasifikator, koji za svaku reč procenjuje jačinu pozitivnog i negativnog sentimenta. Sumiranjem vrednosti svih reči u okviru komentara, dobija se konačna vrednost sentimenta komentara, za svaki od skupova podataka. Pored ovog pristupa, u planu su da se isproba i upotreba *TextBlob*-a i *Vader*-a. Ukoliko pomenuta rešenja ne budu dovoljno kvalitetna, ručno ćemo labelirati komentare i obučiti klasifikatore: *SVM*, *Naive Bayes* i *Random Forest*, i videti sa kojim modelom postižemo najbolje rezultate.

Prilikom prikupljanja podataka, prikupljen je i *thumbnail* za svaki video. U planu je da se konvolutivnom neuronskom mrežom vrši ekstrakcija *feature*-a iz ovih slika, te da se i tako dobijeni podaci uključe u skup podataka. Za realizaciju ovog dela projekta verovatno će se koristiti konvolutivna mreža unapred obučena nad velikim skupom podataka (najverovatnije *VGG16*).

Još jedan način za obogaćivanje početnog skupa podataka jeste ekstrakcijom *feature*-a iz naslova videa. Ovaj deo rada biće realizovan *Bag-of-Words* pristupom, gde će se naslovi prvo pretprocesirati izbacivanjem stop reči i lematizacijom. Planirano je da se kreiraju unigram i bigram rečnici, i da se potom obuči klasifikacioni model koji će na osnovu *feature*-a iz naslova vršiti predikciju broja pregleda. Ovaj korak je neophodan da bi mogli da ustanovimo koji model je kvalitetniji, onaj koji je koristio unigrame ili bigrame, da bi potom na originalan skup podataka mogli da dodamo i *feature*-e dobijene uz adekvatan rečnik. Da bi mogli da vršimo klasifikaciju videa po broju pregleda, tu vrednost je neohodno diskretizovati, gde će prag biti medijan broja pregleda. Obučavaće se klasifikatori *SVM*, *Naive Bayes* i *Random Forest*.

6. Metod evaluacije

Konačni skup podataka biće podeljen na trening i test skup (trening 70% i test 30% od ukupnog broja). Podaci za test skup biće odabrani slučajno, neće se birati sa kraja ili početka originalnog skupa podataka. Kao mere evaluacije regresionih modela koristiće se *RMSE (Root Mean Squared Error)* i R^2. Za evaluaciju klasifikatora koristiće se preciznost, odziv i F1 mera.

7. Softver

Za implementaciju projekta koristiće se programski jezik *Python*. Manipulacija podacima i eksplorativna analiza biće realizovani upotrebom biblioteke *Pandas*. Za potrebe vizuelizacije koristiće se biblioteke *Matplotlib* i *Seaborn*. Pretprocesiranje tekstualnih podataka biće realizovano upotrebom *NLP* biblioteka *Nltk* i *SpaCy*. Za računanje sentimenta reči isprobaće se *TextBlob*, *Vader* i *SensiStrength*. Kreiranje i obučavanje različitih modela mašinskog učenja biće realizovano bibliotekama *Sklearn* i *Keras*.

8. Plan

- Prikupljanje podataka i formiranje skupa podataka
- Eksplorativna analiza podataka
- Obrada podataka, pretprocesiranje tekstualnih podataka
- Kreiranje modela
- Evaluacija modela

Članovi tima:

Marko Pejić E2 60/2019

Stefan Ruvčeski E2 64/2019

Mirko Ivić RA 47/2015