# Predlog projekta

## Definicija projekta

Predviđanje popularnosti okačenog sadržaja na sajtu *9gag* (*9gag post*) na osnovu analize okačene slike ili upotrebljenog ranije ustanovljenog šablona, broja komentara koji su drugi korisnici sajta ostavili na *post-u* i njihovog značenja, ukupan balans između korisnika koji su označili da im se sadržaj dopada i onih kojima se sadržaj nije svideo. Ideja ovog projekta bila bi analiziranje svih pomenutih parametara koji prate jedan *9gag post*, odnosno utvrđivanje da li i u kojoj meri oni utiču na njegovu eventualnu popularnost, kao i korišćenje stečenih informacija prilikom predviđanja popularnosti novopostavljenog sadržaja.

## Motivacija

Svaki *9gag post* ima određeni stepen popularnosti koji zavisi od brojnih parametara koji ga prate. Parametri okačenog sadržaja koji mogu biti od značaja su činjenica da li je okačena slika ili je korišćen ranije utvrđen šablon sadržaja. Takođe, bitna stavka prilikom predviđanja popularnosti bio bi i tekst *post*-a, kako prateći, tj. onaj koji se nalazi na samoj slici ili šablonu, tako i sam naslov. Kako *9gag* dozvoljava pisanje komentara na sadržaje, potrebno je osvrnuti se i na komentare, odnosno utvrditi u kojoj meri oni utiču na eventualnu popularnost. Pretpostavka je da, što su komentari pozitivniji, to i sam *post* ima veću popularnost. Stoga je potrebno i izvršiti analizu sadržaja komentara. Prilikom kačenja svakog *9gag post*-a, korisnik se nada da postigne maksimalnu moguću popularnost. Ova analiza može umnogome doprineti razumevanju svih kriterijuma koje treba zadovoljiti da bi se postigla optimalna popularnost.

## Relevantna literatura

[1] *Meghawat, Mayank, et al. "A multimodal approach to predict social media popularity." 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). IEEE, 2018.* https://arxiv.org/pdf/1807.05959.pdf

Cilj ovog rada je predviđanje popularnosti objava na društvenoj mreži *Flickr* na osnovu multimedijalnog sadržaja, informacija o kontekstu i društvenih informacija. Opisani pristup koristi vizuelna obeležja, koja se dobijaju analizom slike, zatim obeležja dobijena analizom teksta objave i društvena obeležja, poput prosečnog broja pregleda onoga čija je objava. Korišćen je SMP-TI skup podataka koji je originalno sadržao 432 hiljade objava sa društvene mreže *Flickr*. U ovom skupu podataka postoje dodatne kontekstne informacije u koje spadaju broj komentara, informacija o tome da li se na slici nalaze ljudi, dužina naslova, dužina opisa, broj tagova itd. Autori su dodatno proširili skup obeležja bitnim tekstualnim informacijama u koje spadaju naslov, tagovi i opis fotografije. Za analizu kontekstnih i društvenih informacija korišćeni su *Random forest* algoritam i namenski kreirana konvolutivna neuronska mreža sa šest slojeva. Za analizu slike korišćena je modifikovana verzija duboke konvolutivne neuronske mreže *InceptionResnetV2* arhitekture. Analiza naslova i tagova se bazirala na upotrebi rečnika, a sentiment opisa je određen pomoću *Stanford CoreNLP* biblioteke. Dobijena obeležja (njih 15) se normalizuju i pomoću konvolutivnog modela se vrši predviđanje popularnosti objave. Za evaluaciju rešenja korišćeni su Spirmanov koeficijent korelacije, srednja apsolutna greška i srednja kvadratna greška. Primena multimodalnog pristupa se pokazala kao opravdana i zbog toga bi i mi u našem radu mogli da na takav način pristupimo rešavanju problema. Razlika u odnosu na opisani rad bi bila u tome što bismo prilikom analize slike vršili detekciju objekata koji pripadaju unapred određenim klasama i dodatno bi bila vršena analiza sentimenta komentara. Analiza kontekstnih i tekstualnih informacija, kao i konačna predikcija bi

mogle da se rade na sličan način. Evaluacija bi mogla da se radi takođe upotrebom Spirmanovog koeficijenta korelacije.

[2] *Mazloom, Masoud, et al. "Multimodal popularity prediction of brand-related social media posts." Proceedings of the 24th ACM international conference on Multimedia. ACM, 2016.* https://staff.fnwi.uva.nl/m.mazloom/Papers/mazloom-ACM-MM-2016.pdf

Rad predstavlja težnju autora da identifikuju koji aspekti u objavama koji se odnose na brendove brze hrane na društvenoj mreži Instagram utiču na to da objave budu popularnije. Za to se koristi devet obeležja koje su autori nazvali *engagement parameters*. 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* metode. Koncepti na slici su detektovani pomoću konvolutivne neuronske mreže trenirane na *ImageNet* skupu podatak. Problem popularnosti objave se posmatra kao problem rangiranja. Nad dobijenim vrednostima je primenjena L2 regularizacija i za predviđanje je korišćena SVR metoda. Podaci su podeljeni na trening i test skup, a za pronalazak optimalne vrednosti za regularizacioni parametar C za metodu SVM korišćena je unakrsna validacija sa pet podela. Skup podataka je dobijen skupljanjem podataka 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. U našem projektu analiza koncepata na slikama bi mogla da se radi na sličan način, samo sa manjim brojem koncepata koji se prepoznaju, pri čemu bi neki od tih koncepata svakako bile i osobe na slici. Takođe bi mogla da se vrši analiza sentimenta *hashtag*-ova, naslova i komentara na objavama, kao u opisanom radu.

[3] *Yang, Diyi, et al. "Humor recognition and humor anchor extraction." Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015.* https://www.aclweb.org/anthology/D15-1284.pdf

Cilj ovog rada je kreiranje sistema koji je u stanju da prepozna humor u rečenicama prirodnog jezika. Zbog toga su posmatrane četiri skrivene semantičke osobine za koje se smatra da dovode do postojanja humora, a to su neskladnost (engl. *incongruity*), dvosmislenost, međuljudski efekat i fonetski stil. Za ove četiri osobine su kreirani posebni ekstraktori obeležja. Neskladnost je određivana uz pomoć *Word2Vec* metode. Dvosmislenost je određivana pomoću *WordNet*-a kao leksičkog resursa. Međuljudski efekat je određen analizom sentimenta. Fonetski stil je meren korišćenjem *CMU Pronouncing Dictionary*-a, na način da je analizirana aliteracija i rima u tekstu. Korišćeni skup podataka se sastojao iz dva izvora koji su sadržali tekstove sa humorom (jedan je sa veb sajta *Pun of the Day*, a drugi *16000 One-Liner* skup podataka) i nekoliko izvora za koje se smatralo da ne sadrže smešne rečenice. Eksperiment je vršen kao za klasičan problem klasifikacije teksta. *Random Forest* algoritam je korišćen za unakrsnu validaciju sa deset podela. U metodi nazvanoj *Human Centric Features* (skraćeno HCF) su pored četiri ranije navedene osobine korišćena obeležja dobijena primenom KNN klasifikatora koji koristi klase pet najbližih rečenica po značenju. Implementiran je i tkz. *SaC* ansambl koji koristi stilske osobine rečenice, poput aliteracije, autonomije i žargona. Kao mera evaluacije korišćeni su preciznost, odziv, tačnost i F1 mera. Najbolje rezultate je dao pristup koji je kombinovao *Word2Vec* i HCF metode, što potvrđuje značaj semantičkih osobina koje je bilo cilj analizirati. Ovakav zaključak bi u našem radu mogao da se iskoristi tako što bi neka od obeležja mogla da sadrže informaciju o tome da li naslov, i eventualno tekst na slici,

predstavljaju rečenice u kojima je prepoznato postojanje humora, što bi trebalo da utiče na popularnost objave.

## Skup podataka

Za ovaj projekat potrebno je obezbediti skup podataka nad kojim će se vršiti dalje analize. Planirano je da se skup podataka samostalno formira koristeći API *9gag-a* kojim se svaki *post* predstavlja u JSON formatu. Primer jednog takvog *post-a* bio bi:

https://9gag.com/v1/group-posts/group/default/type/hot?after=aqge97Y&fbclid=IwAR0rUHXaNQs06_igsmnj3nxI0AH0OAuJEBuQGf6iaCFc4UPkiiTfsf6Jwtk .

Odavde je moguće preuzeti sve potrebne informacije, moguće je preuzeti samu sliku, naslov, broj *downvote* i *upvote* glasova, tagove koji prate sadržaj itd. Ono što je ovde posebno korisno je što u ovom skupu informacija postoji i neka vrsta pokazivača ka komentarima, kao i pokazivači ka sledećem *post*-u i onom koji mu je prethodio.

Korišćenjem ovog API-a moguće je razdvojiti sadržaje koji se nalaze u kategorijama “HOT”, “Trending” i “Fresh”, kao i utvrđivanje kojoj podkategoriji oni pripadaju.

## Metodologija

U ovom projektu podaci koje ćemo koristiti za analizu su slika, tekst na objavi, naslov objave, *hashtag*-ovi i komentari vezani za svaku objavu. Dakle, biće rađena analiza određenih koncepata, odnosno prepoznavanje objekata na slikama, upotrebom konvolutivne neuronske mreže primenom metode transfera učenja sa parametrima dobijenim treniranjem na *ImageNet* skupu podataka i uz korisćenje neke poznate arhitekture konvolutivnih neuronskih mreža. Zatim, biće rađena analiza sentimenta tekstova, naslova i komentara upotrebom *SentiStrength* algoritma koji je pogodan za analizu kratkih *web* tekstova. Dok će se za svaku objavu prisustvo *hashtag*-a zapisivati u formi vektora i u tom formatu koristiti za određivanje sentimenta. Dodatno, biće analizirano postojanje humora u tekstovima na objavama i naslovima putem *Word2Vec* i *HCF* metoda. Iz prethodno navedenih parametara bi dobili određena obeležja na osnovu kojih bi vršili predikciju popularnosti objave pomoću neke neuronske mreže ili algoritama poput SVM (tačnije SVR) i Random Forest algoritma.

**Commented [JS1]:** Hashtag-ove možda možete koristiti i tako što napravite rečnih čestih hashtag-ove i svaki hashtag je jedno obeležje sa 1/0 vrednošću (post ga sadrži ili ne)

**Commented [JS2]:** Ovde bih razmislila i o drugim modelima. NN ne mora raditi najbolje kada su u pitanju sređeni numerički podaci, možda biste imali više sreće sa SVM ili Random Forest.

Ono što možete razmisliti je i glasanje različitih klasifikatora gde je npr. jedan klasifikator je treniran na obeležjima slike, a drugi na komentarima, treći na tekstu i naslovu post-a. četvrti na hashtag-ovima.

## Metod evaluacije

Krajnji rezultat, dobijen pomoću neke od navedenih metoda, biće predikcija popularnosti objave u vidu *score*-a, koji u ovom slučaju predstavlja odnos *upvote*/*downote*. Taj dobijeni *score* će se porediti sa stvarnim *score*-om te objave pomoću Spirmanovog koeficijenta korelacije kako bi dobili tačnost. Dodatno, za određivanje uspešnosti predikcije bi bile korišćene i mere R2 (c*oefficient of determination*) i MSE (*mean squared error*). Podaci će biti podeljeni na *train* i *test* skup, gde će se na *train* podacima obučavati model, dok će se na *test* podacima vršiti verifikacija modela.

**Commented [JS3]:** U kom se opsegu kreće score?

Možda razmotrite i problem klasifikacije – da li će neki post završiti u hot, trending ili fresh (ili definišite kategorije po score-u, ali se trudite da imate neko prirodno opravdane za njih). Jer možda nije toliko bitan tačan broj glasova koliko da li je post jako popularan/umereno popularan ili nepopularan. Pazite samo da tu informaciju o kategoriji fresh/trending/hot ne koristite kao obeležja u modelu jer sadrže deo informacije o ciljnoj labeli.

Ne bi bilo loše da sprovedete i analizu grešaka modela. Možete recimo razmatrati da li vaš model greši na slikama koje nisu šabloni ili je najnesigurniji za post-ove fresh kategorije i slično.

## Softver

Za izradu projekta će biti korišćeni *PyCharm* radno okruženje i *python* programski jezik. Za skladištenje podataka će se koristiti *MongoDB* baza podataka.

## Plan

Plan izrade projekta može da se podeli na sledeće bitne celine:

- Prikupljanje podataka
- Obrada podataka
- Kreiranje modela
- Verifikacija modela
- Analiza i prezentacija rezultata