Predikcija popularnosti YouTube video snimka

Nela Jović
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Novi Sad, Srbija
jovic.e222.2023@uns.ac.rs

Filip Ilić
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Novi Sad, Srbija
ilic.e225.2023@uns.ac.rs

***Apstrakt*—U današnje doba, koje odlikuje razvoj tehnologija, sve je veći broj takozvanih internet zanimanja. Iako se na početku na ove poslove gledalo neozbiljno, trenutno je moguće živeti, a i dobro zaraditi, od ovakvih zanimanja. Jedno od takvih zanimanja je kreiranje video sadržaja i njihovo postavljanje na internet. *YouTube* predstavlja najpopularniju platformu za ovakvu svrhu, i odavno su ljudi prepoznali njegov potencijal. Kako količina novca koja može da se zaradi zavisi od popularnosti video snimka, osobe koje postavljaju snimke na razne načine pokušavaju da privuku publiku. Upravo rešavanjem ovog problema se bavi ovaj rad. Logično je da sadržaj samog snimka utiče na njegovu popularnost, ali su za ovaj rad od interesa ostali atributi snimka, kao i njihov, pozitivan ili negativan, uticaj na popularnost snimka. Analiza ovih atributa obuhvata više stvari. Pre svega je upotrebom dubokog učenje analiziran uticaj slike i tekstualnih atributa na popularnost snimka. Za tu svrhu korišćeni su *ResNet* konvolutivna neurnoska mreža i *fine-tuning BERT* jezičkog modela. Nakon toga su primenjene razne tehnike pretprocesiranja podataka, kako bi se oni pripremili za treniranje. Na kraju su modeli trenirani u poređeni su rezultati. U obzir su uzeti sledeći modeli: linearna regresija, *XGBoost* model, *CatBoost* model, *SVR* model. Evaluacija modela i poređenje rezultata je odrađena na osnovu R2 mere i korena srednje vrednosti kvadrata greške (*RMSE*). Kao najbolji se pokazao *XGBoost* model koji je ima R2 meru od 0.98 i *RMSE* od 0.375. Daljom analizom, uočen je šablon kod podataka koji prouzrokuju veliku grešku, i njihovim uklanjanjem rezultati su dodatno poboljšani. Na kraju je *XGBoost* model ima R2 meru od 0.98, a vrednost *RMSE* se smanjila na 0.344.**

Ključne reči—YouTube; popularnost; mašinsko učenje; duboko učenje; regresija; predikcija

I. UVOD

U današnje vreme, najpopularniji sadržaj na internetu predstavlja video sadržaj. Ovo se može ogledati u sve većem broju platforma za objavljivanje video snimaka. Jedna od prvih, a trenutno verovatno i najpoplarnijih platforma za ovu svrhu je *YouTube*. Glavni indikator ovoga je broj osoba kojima je postavljanje snimaka na *YouTube*-u posao, kao i količina novca koja se može na ovaj način zaraditi.

Kako bi učinili svoj video što uspešnijim, osobe koje postavljaju snimke na YouTube-u (u nastavku *YouTube*-eri) na razne načine pokušavaju da privuku publiku. Jasno je da sam sadržaj snimka u velikoj meri utiče na njegovu popularnost i uspešnost, međutim da bi se saznao sadržaj potrebno je odgledati snimak, pa je potrebno privući gledaoce na alternativne načine. *YouTube*-eri to najčešće postižu zadavanjem privlačnih naslova i slika za svoje snimke. Naslov video snimka u velikoj meri utiče na to da li će ga korisnici odgledati. Dobar naslov će korisnicima sugerisati sadržaj snimka i pomoći će im da nasulte šta mogu da vide u samom snimku. Analogno naslovu, dobra slika video snimka će korisnicima ili prikazati neki ključan deo snimka ili na neki drugi način sugerisati o čemu se u samom snimku radi. *YouTube*-eri su svesni ovoga i pokušavaju da zloupotrebe naslov i sliku. To se postiže postavljanjem obmanjujućih naslova i slika koje nemaju veze sa sadržajem snimka, već imaju zadatak da privuku publiku i povećaju broj pregleda. Međutim, ovo će se negativno odraziti na broj lajkova, pa je logičnije posmatrati broj lajkova, umesto broj pregleda, kao meru popularnosti video snimka.

Upravo je cilj ovog rada rešavanje problema predikcije popularnosti YouTube snimka. Postoji dosta radova koji su rešavali ovaj problem, međutim svim tim radovima je slično to što su rešavali klasifikacioni problem, odnosno što su klasifikovali video snimke po kategorijama popularnosti. Ovaj rad će se baviti rešavanjem regresionog problema, pa će se na osnovu atributa video snimka predviđati broj lajkova koje će taj snimak imati. Za analizu teksta i slika koristiće se duboko učenje, pa će rezultati tih modela, zajedno sa određenim ostalim atributima, kao što su broj pregleda, komentara i dr., biti ulaz regresionog modela koji će predviđati broj lajkova.

U narednom poglavlju će biti izložen detaljan pregled relevatne literature i postojećih radova. Treće poglavlje će se baviti opisom skupa podataka. U četvrtom poglavlju će biti objašnjeno kako je korišćeno duboko učenje za analizu teksta i slike. Poglavlje pet će sadržati detaljan opis metoda koje su korišćene za rešavanje regresionog problema, kao i opis postignutih rezultata. Šesto poglavlje će biti posvećeno analizi grešaka . Poslednje poglavlje će služiti za izvođenje zaključka, kao i predstavljanje predloga za poboljšanje rešenja.

II. PREGLED RELEVANTNE LITERATURE

Radi boljeg upoznavanja sa problemom, izvršeno je istraživanje postojećih radova koji su rešavali isti ili sličan problem. Pronađeno je nekoliko takvih radova, čiji se zaključci i zapažanja mogu primeniti na ovaj rad, pa će u nastavku biti opisani neki od tih radova.

U radu [1] rešavan je problem predviđanja popularnosti video snimaka korišćenjem dubokog učenja i klasifikacije sekvence. Korišćen je skup od 200 video snimaka različitih dužina, od 3 sekundi pa do preko 30 minuta. Prvo je rađena ekstrakcija atributa za svaki snimak, gde se za svaki *frame* video

snimka računala njegova *RGB* vrednost, i na osnovu svih vrednosti računala se srednja vrednost za snimak. Pored ove vrednosti, ulaz u model bili su i broj pregleda i broj lajkova, dok je izlaz modela ukazivao na to da li je video snimak popularan ili ne. Model je predstavljen dubokom neuronskom mrežom sa samo jednim skrivenim slojem, gde je aktivaciona funkcija *ReLU*. Za računanje gubitka korišćen je *Binary Cross Entropy*, dok je za optimizaciju korišćen *Adam* algoritam. Podaci su podeljeni u skup za trening i testiranje u odnosu 65:35, gde se za evaluaciju koristila tačnost. U radu nije navedena postignuta vrednost za tačnost, ali je vrednost funkcije gubitka približna nuli. Iz ovog rada se može zaključiti da i relativno jednostavni modeli mogu da daju dobre rezultate ukoliko se dobro odradi ekstrakcija atributa.

Rad [2] se bavi klasifikacijom video snimaka na jednu od 4 klasa popularnosti: nepopularni video snimci, neutralni video snimci, video snimci sa pretežno pohvalama i video snimci sa pretežno lošim ocenama. Skup podataka je preuzet sa sajta *kaggle* [3], i sastoji se od 29089 podataka. Nakon izbacivanja podataka o video snimcima koji više nisu dostupni, ostaje 24726 podataka koji se dele na podatke za trening, validaciju i test u odnosu 70:20:10. Na početku je korišćen *backwards search* algoritam kako bi se odredila važnost atributa. Od polaznih 6 atributa (naslov, vremenski razmak, kategorija, oznake, opis i trajanje) algoritam je odabrao vremenski razmak, kategoriju i opis kao najvažnije karakteristike. Nakon toga su se trenirali modeli zasnovani na različitim algoritmima i poredili rezultati. Za ovu svrhu korišćeni su sledeći algoritmi: *stochastic gradient descent, multilayer perceptron neuron network, decision trees, random forest, gradient boosting and extreme gradient boosting.* Za poređenje rezultata različitih modela koristi se F1 mera, i najbolje rezultate postigao je model zasnovan na *extreme gradient boosting* algoritma, sa F1 merom od 0.736. Značaj ovog rada ogleda se u poređenju velikog broja algoritma za višeklasnu klasifikaciju. Najbolje se pokazao *extreme gradient boosting* algoritam, pa će se on razmatrati i u ovom radu.

U radu [4] je opisano rešavanje problema klasifikacije video snimaka na popularne i na one koji to nisu. Korišćen je skup od 5000 podataka, koji se dele na 3500 podataka za trening i 1500 podataka za validaciju. U ovom radu su rađeni selekcija, ekstrakcija i spajanje atributa, kako bi se što više atributa uzelo u obzir, i od njih selektovali samo oni atributi koji su u velikoj korelaciji sa popularnosti video snimka. Nakon navedenog pretprocesiranja, ulaz u model predstavljaju kategorija video snimka, kvalitet video snimka, trajanje, broj pregleda i rezultat (dobijen spajanjem broja komentara, broja lajkova i broja dislajkova). Za klasifikaciju su korišćeni model zasnovan na *decision trees* algoritmu, model zasnovan na *extreme gradient boosting* algoritmu, dok je poslednji model takođe zasnovan na *extreme gradient boosting* algoritmu, samo što su parametri modela optimizovani. Za poređenje rezultata su korišćene tačnost, preciznost, odziv i F1 mera. Najbolje rezultate postigao je model zasnovan na *extreme gradient boosting* algoritmu sa optimizovanim parametrima, i on je imao tačnost od 88%, preciznost od 86%, odziv od 67% i F1 meru od 75%. Značaj ovog rada ogleda se u pretprocesiranju podataka, kao i u optimizovanju parametara modela kako bi se postigli što bolji rezultati.

Rad [5] daje značaj detekciji autlajera, odnosno rešava problem predviđanja da li je korisnik stvarno hteo da odgleda neki video. Postoje slučajevi kada korisnici neželjno otvore neki video, na primer kada video ima obmanjujuć naslov ili sliku, ukoliko korisnik greškom klikne na neki video ili ukoliko neko koristi tuđi nalog za gledanje, pa je cilj rada razvoj algoritma za detekciju autlajera, koji uzima u obzir sledeće atribute: kvalitet snimka, dužinu snimka, kategoriju snimka, identifikator kanala, odnos broja komentara i broja pregleda, odnost broj lajkova i dislajkova i ukupan broj pregleda. Podaci nad kojima je isproban algoritam su prikupljeni na osnovu YouTube istorije 10 korisnika, gde su se za svakog korisnika posmatrala bar 150 poslednjih odgledanih video snimka. Pored istorije, korsinici su za svaki snimak vrednovali značaj atributa koje algoritam uzima u obzir, kao i da li su zaista hteli da odgledaju taj snimak. Nakon toga algoritam je odredio autlajere, i rezultati su poređeni sa tačnim vrednostima. Pokazalo se da je algoritam u 63% slučajeva ispravno prepoznao video snimak koji korisnici nisu nameravali da odgledaju.

Rad [6] je jedan od retkih radova koji rešava regresioni problem. Cilj rada je predviđanje broja pregleda koje će video snimak imati na osnovu ostalih atributa. Korišćen je skup podataka sa sajta *kaggle* [7], a atributi koji su uzeti u obzir su broj lajkova, broj dislajkova, broj komentara i broj pregleda. Razmatrani su modeli zasnovani na algoritmima *Ordinary Least Square Method* i *Stohastic Gradient Descent* i poređeni su rezultati. Model zasnovan na *Stohastic Gradient Descent* algoritmu je dao bolje rezultate, međutim i model zasnovan na algoritmu *Ordinary Least Square Method* je pokazao nadprosečne rezultate. Zaključak je da selekcijom većeg broja atributa i pretprocesiranjem podataka mogće je da model zasnovan na linearnoj regresiji daje dobre rezultate.

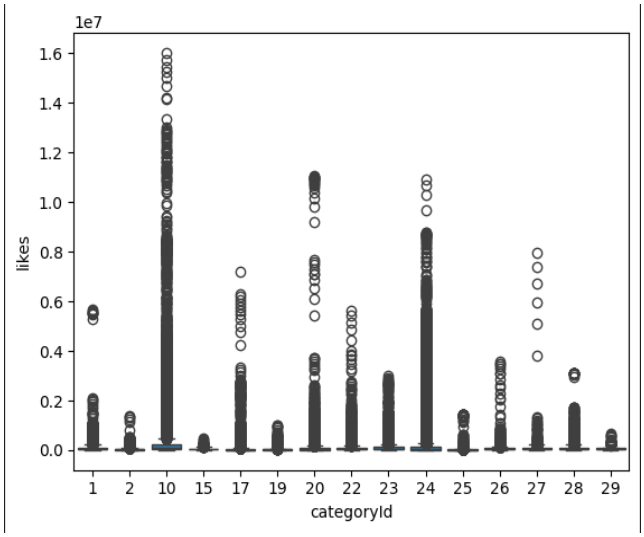
III. Skup podataka

U ovom poglavlju će biti opisan skup podataka koji je korišćen. Polazni skup podataka je preuzet sa sajta *kaggle* [8], i on sadrži informacije o trending YouTube snimcima, počev od 11.08.2020., u raznim zemljama. Za ovaj rad su od interesa snimci vezani sa Sjedinjene Američke Države, i kada je skup podataka preuzet 16.12.2023. on je sadržao 245987 podataka. Svaki podatak ima 16 atributa, od kojih su korišćeni: naslov, datum objave, identifikator kanala, naziv kanala, identifikator kategorije, trending datum, oznake, broj pregleda, broj lajkova, broj dislajkova, broj komentara, link do slike i opis. Atributi koji nisu korišćeni su identifikator kanala, oznaka da li su komentari onemogućeni i oznaka da li su lajkovi.

Polazni skup podataka je proširen na više načina. Prvo se na osnovu identifikatora kanala dobavio datum kreiranja kanala. U slučaju da je kanal obrisan, ova vrednost se ne može dobaviti pa je to tretirano kao nedostajuća vrednost. Na osnovu datuma kreiranja kanala i trending datuma za snimak računala se starost kanala u trenutku kada su za video prikupljeni podaci. Na osnovu trending datuma i datuma objavljivanja snimka računala se starost samog snimka kada su za njega prikupljeni podaci. Skup je takođe proširen podacima koji predstavljaju dužinu naslova, dužinu opisa i broj oznaka koje video snimak ima. Na osnovu linka dobavljene su slike za sve video snimke.

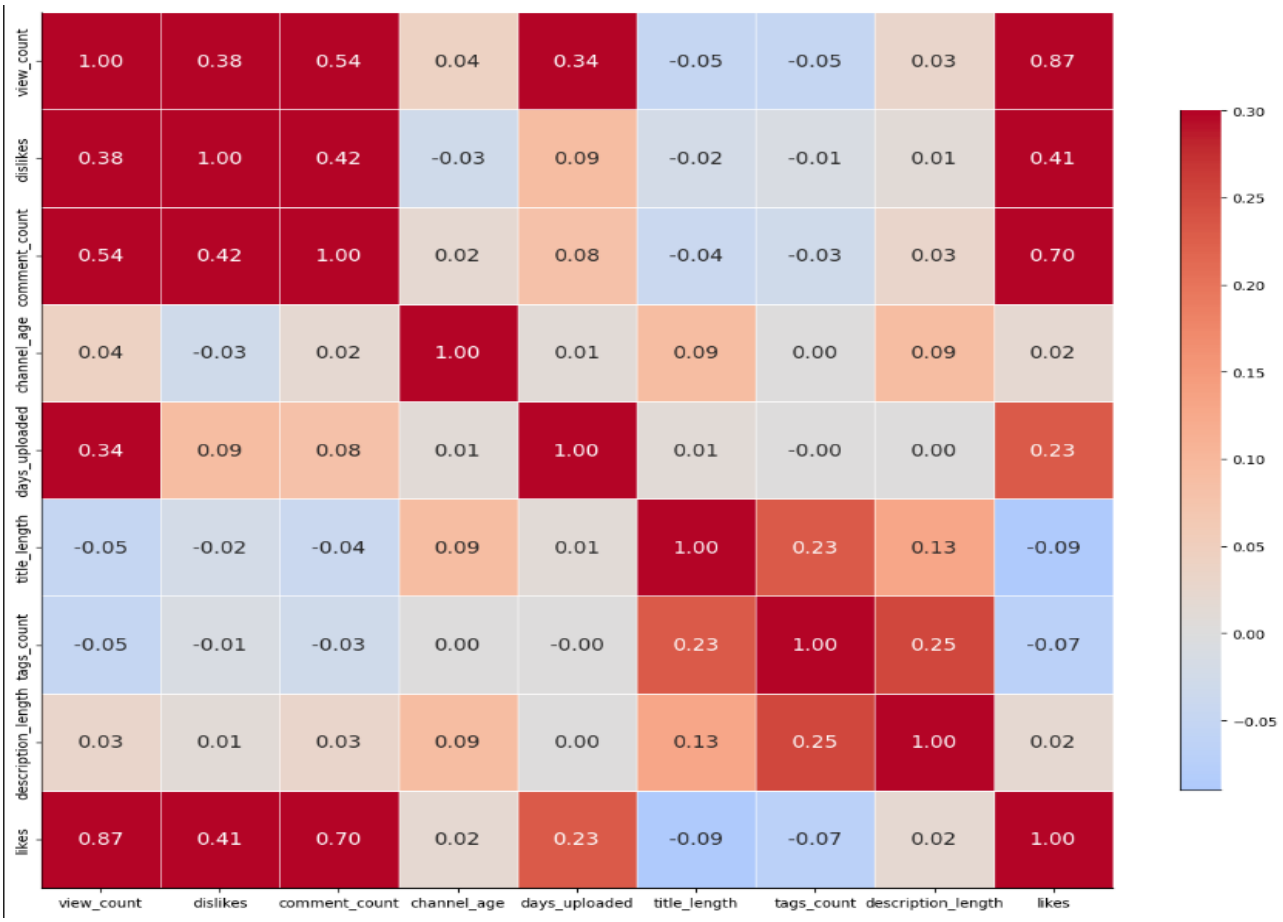
Sledeći korak je analiza podataka. Prvo je posmatran skup vrednosti koje imaju atributi i tu je uočeno nekoliko zanimljivih stvari. Prva takva stvar je veliki broj nedostajućih vrednosti za broj dislajkova, čak 147916, a razlog za to je činjenica da od 13.12.2021. broj dislajkova za video snimak nije javno dostupan. Osim za broj dislajkova, broj nedostajućih vrednosti za ostale attribute je relativno mali: 76 za broj pregleda, 1446 za broj lajkova i 677 za starost kanala. Analiziranjem dužine naslova video snimka, uočeno je da postoje snimci sa predugačkim naslovima, odnosno prosečna dužina naslova je 20, a maksimalna 100. Ista stvar je uočena i za opis, gde je prosečna dužina 900, dok postoje snimci i sa dužinom opisa od 4998 karaktera. Što se tiče broja oznaka, što više oznaka ima snimak to je veća šansa da će se pojaviti u pretrazi, međutim u nekim slučajevima se i sa ovim preteralo. Tako je prosečna vrednost za broj oznaka 168, dok postoje snimci i sa 500 oznaka

Kako je ciljno obeležje broj lajkova, analiziran je odnos ostalih atributa sa brojem lajkova. Posmatranjem kategorije video snimka, uočeno je da najviše snimaka pripada kategorijama Zabava (19.99%), Gejming(19.88%) i Muzika(16.39%). Takođe je upčeno da snimci koji pripadaju određenim kategorijama imaju značajno manji broj lajkova u odnosu druge kategorije. Slika 1. prikazuje broj lajkova po kategorijama video snimaka, i tu se može uočiti da kategorije Automobili i vozila(2), Životinje i ljubimci(15), Putovanja i događaji(19), Vesti i politika(25) i Aktivizam(29) imaju manji broj lajkova u odnosu na druge snimke.



Slika 1. Broj lajkova u odnosu na kategoriju snimka

Kako bi se ispitala korelacija između broja lajkova i ostalih numeričkih atributa, napravljena je matrica korelacija. Slika 2. pokazuje da postoji velika korelacija broja lajkova sa brojem pregleda, brojem dislajkova i brojem komentara, dok broj lajkova je u veoma slaboj korelaciji sa starošću kanala i sa dužinom opisa, pa se ova 2 atributa neće razmatrati u nastavku.



Slika 2. Matrica korelacija

IV. DUBOKO UČENJE

Pre nego što je rađeno predviđanje broja lajkova, dubokim učenjem je odrađena analiza uticaja određenih atributa na popularnost video snimka. Konkretno su posmatrani slika i tekstualni atributi video snimka(naslov, naziv kanala, opis i oznake), i za tu svrhu su kreirana i trenirana 2 zasebna modela. Ideja je da se izlazi ovih modela koriste kao ulazi u regresioni model koji predviđa broj lajkova, pa će u nastavku detaljno biti opisano: prikupljanje i anotiranje podataka, priprema podataka za treniranje, kreiranje i treniranje modela i rezultati.

Prvo će biti opisana analiza uticaja slika na popularnost video snimka. Za ovu svrhu, slike su dobavljane na osnovu linkova koji se nalaze u skupu podataka [3] . Pošto se u skupu jedan snimak može pojaviti više puta, na ovaj način je dobavljeno 42729 različitih slika. Za anotaciju slika koristio se odnos broja lajkova i broja pregleda za video snimak, i na ovaj način slike su podeljene u 4 kategorije:

- Nepopularne slike, koje imaju odnos manji od 2% (8657 slika)
- Srednje popularne slike, koje imaju odnos veći ili jednak 2% a manji od 5% (16595 slika)
- Popularne slike, koje imaju odnos veći ili jednak 5% a manji od 10% (14392 slika)
- Veoma popularne slike, koje imaju odnos veći ili jednak 10% (3085 slika)

Kako bi se skup podataka izbalansirao rađene su augmentacije slika. Najviše su augmentovane slike iz veoma popularne kategorije, i nad njima su vršene sledeće transformacije: rotacija, treperenje boja, horizontalno okretanje, vertikalno okretanje, zamućenje, promena perspective, isecanje slike, invertovanje boje i dve različite promene kontraskta. Nad slikama ostalih kategorija primenjene su samo neke od nabrojanih transformacija, tako da je na kraju skup podataka sadržao 34628 nepopularnih slika, 33190 srednje popularnih slika, 28784 popularnih slika i 33935 popularnih slika. Pre kreiranja modela, slike su podeljene u skup za treniniranje i testiranje u odnosu 80:20.

Model za klasifikaciju slika nije pravljen od nule, već je korišćen princip transfera učenja nad jednim od pretreniranih *state of the art* modela. Odabrana je *ResNet* arhitektura, s pretpostavkom da će svojom dubinom uspeti da izvuče ključne osobine za klasifikaciju. Isprobani su modeli različitih veličina, a kao najbolji se pokazao model od 25557032 parametara, jer je najmanje *overfit*-ovao podatke i imao je najbolje rezultate. Što se tiče slojeva koji su trenirani, i ovde su isprobane razne

kombinacije. Prvo je treniran samo klasifikator, međutim ovakav pristup je imao dosta loše rezultate. Zatim je pored klasifikatora treniran i veći broj slojeva za ekstrakciju osobina, ali je ovaj pristup veoma brzo krenuo da *overfit*-uje podatke. Najmanje *overfit*-ovanje i najbolje rezultate imao je pristup gde se pored klasifikatora trenirao samo još poslednji *bottleneck* sloj. Za optimizaciju je korišćen *Adam* algoritam, dok je za funkciju gubitka odabrana *Cross Entropy Loss*, kojoj se prosleđuju težine kako bi u obzir uzela nebalansiranost skupa podataka. Formula (1) prikazuje kako su računate težine po klasama.

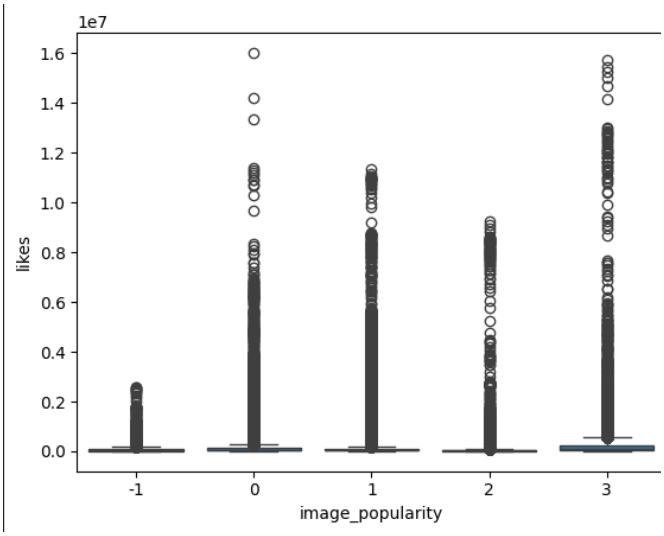
$$weight_{class} = \frac{total}{samples_{class}} \qquad (1)$$

Za hiperparametre *batch_size*, *learning_rate* i *epochs* su isprobane razne vrednosti, a najbolji rezultat je postignu za vrednosti 32, 0.00001 i 12 respektivno. Nakon treniranja model je nad trening podacima imao tačnost od 74%, dok je model nad podacima za test imao tačnost od 66%. Vrednosti po klasama za preciznost, odziv i F1 meru prikazani u Tabeli 1.

	preciznost	**odziv**	**F1**
Popularne	0.61	0.71	0.65
Srednje popularne	0.58	0.64	0.60
Nepopularne	0.74	0.68	0.71
Veoma popularne	0.73	0.60	0.66

Tabela 1. Rezultati klasifikacije

Na kraju je istreniran model iskorišćen kako bi se skup podataka za regresiju proširio još jednim atributom koji označava kategoriju slike. Najviše snimaka je imalo srednje popularnu sliku (39.04%), zatim popularnu sliku (38.43%), dok ostatak čine nepopularne i veoma popularne slike. Za 2.64% video snimaka nije bilo moguće dobaviti sliku. U tom slučaju je za kategoriju slike postavljena vrednost -1, što se dalje tretiralo kao nedostajuća vrednost. Slika 3. prikazuje broj lajkova po kategorijama. Može se zaključiti da snimci za koje se slika nije mogla dobaviti imaju manji broj lajkova u odnosu na druge kategorije.



Slika 3. Broj lajkova po kategoriji slike

Tekstualni atributi koji su se našli u skupu podataka, odnosno naziv kanala, naziv video snimka i opis pretvoreni su u jednu tekstualnu kategoriju koja je predstavljala ulaz u *BERT* model. Tekstualni atributi su na osnovu procenta koji je izračunat kao odnos broja dislajkova u odnosu na broj lajkova svrstani u četri kategorije:

- Veoma pozitivni video snimci, procenat broja dislajkova u odnosu na lajkove je manji od 2%
- Pozitivni, procenat broja dislajkova u odnosu na lajkove je između 2 i 5%.
- Negativni, procenat broja dislajkova u odnosu na lajkove je između 5 i 10%
- Veoma negativni, procenat broja dislajkova u odnosu na lajkove je preko 10%

Nad ulaznim podacima radi se tokenizacija, stemovanje i *padding*.

Za obučavanje modela se koristio princip *fine-tuning* modela, koji je zasnovan na dodatnom treniranju prethodno obučenog modela, tako da se model specijalizuje za rad sa podacima nad kojima je treniran. Kako se ne bi isti video snimci ponavljali izdvojeni su svi sa jednistvenim video_id atributom. Nakon eleminasanja duplikata, skup podataka je imao 44009 redova. Od toga 10466 je veoma negativnih video snimaka, 12358 negativnih, 9949 pozitivnih i 11236 veoma pozitivnih. Nakon 4 epohe treniranja tačnost nad testnim podacima bila je 55%. Na osnovu *recall* mere uviđeno je da model slabo prepoznaje pozitivne i negativne video snimke. Iz tog razloga skup podataka je proširen dodatnim skupovima podataka koji se nalaze na sajtu kaggle [3]. Nakon toga je broj podataka porastao na 67249 redova. Ovo se pokazala kao dobra odluka jer je nakon toga tačnost modela nad testnim podacima iznosila 68%, dok se *recall* mera za pozitivnu i negativnu kategoriju popravila. U Tabeli 2. se mogu videti rezultati klasifikacije po kategorijama u toku treninga.

	preciznost	**odziv**	**F1**
Veoma pozitivni	0.80	0.81	0.80
Pozitivni	0.81	0.85	0.83
Negativni	0.83	0.78	0.81
Veoma negativni	0.88	0.87	0.88

Tabela 2. Rezultati klasifikacije nad trening podacima

Nad podacima pokušana je augumentacija korišćenjem *ContextualWordEmbsAug* augumentatora, međutim ovde se to pokazala kao loša praksa jer je tačnost opala na 35%.

Što se tiče parametara, *learning rate* daje isti rezultat za vrednosti 1e-5 i 2e-5, dok sve veće vrednosti daju lošiji rezultat. Zbog velike količine podataka korišćen je *AdamW* optimizator ali je testirano i sa *Stochastic gradient descendent* koji je dao duplo gori rezultat. *Batch size* je uticao sasvim malo skoro nikako na tačnost modela, razlika je nekad u 1% ako se koristi veći *batch size*. U nastavku je prikazana tabela (Tabela 3.) sa rezultatima za različite parametre.

Learning rate	Batch size	Optimizator	Postignuta tačnost
1e-5	8	AdamW	0.67
1e-5	16	AdamW	0.67
1e-5	32	AdamW	0.67
2e-5	8	AdamW	0.67
2e-5	16	AdamW	0.67
2e-5	32	AdamW	0.67
3e-5	8	AdamW	0.48
1e-5	8	SGD	0.38

Tabela 3. Rezultati treniranja

Nakon obučavanja, ovaj *BERT* model iskorišćen je da nad skupom podataka na osnovu tekstualnih atributa za svaki red kaže koja je procena popularnosti tekstualnih atributa. Time je dobijena novi atribut koji zamenjuje sve ostale tekstualne atribute.

V. METODOLOGIJA I REZULTATI

Cilj ovog poglavlja je opis pripreme podataka za treniranje, kao i treniranje i optimizacija modela za predviđanje broja lajkova.

Pošto podaci imaju nedostajuće vrednosti za određene atribute, prvo se pristupilo zameni tih vrednosti. Za broj pregleda, broj lajkova i starost kanala nedostajuće vrednosti su zamenjene srednjom vrednošću. Što se tiče kategorije slike, snimci za koje se nije mogla dobaviti slika su smešteni u odgovarajuću kategoriju na osnovu odnosa broja lajkova i broja pregleda. Za zamenu nedostajućih vrednosti broja dislajkova razmatrano je više opcija, a najefektivnije se pokazao sledeći metod: prvo se izračuna srednja vrednost u procentima odnosa broja dislajkova i broja lajkova, kao i broja dislajkova i broja pregleda, pa se nedostajuće vrednosti zamene manjom od dobijenih vrednosti.

Od pretprocesiranja odrađene su još dve stvari. Prvo je skup podataka proširen dodatnim atributima na sledeći način: za svaki kanal je izračunata srednja, minimalna i maksimalna vrednost broja pregleda za snimke koje pripadaju tom kanalu, pa su se podaci o tim snimcima proširili dobijenim vrednostima. Isto je odradjeno i za broj komentara, pa se skup podataka proširio sa dodatnih 6 atributa. Nakon toga je rađena normalizacija podataka. Prvo je isprobana normalizacija upotrebom *StandardScaler* klase iz biblioteke *sklearn*, međutim dobijeni rezultati su bili nezadovoljavajući pa se normalizacija radila logaritmovanjem za osnovu 2. Konačan ulaz u modeli čine sledeći atributi: broj pregleda, broj dislajkova, broj komentara, starts video snimka, dužina naslova, broj oznaka, kategorija video snimka, kategorija slike, kategorija tekstualnih atributa, srednja vrednost broja pregleda za kanal, minimalna vrednost broja pregleda za kanal, maksimalna vrednost broja pregleda za kanal, srednja vrednost broja komentara za kanal, minimalna vrednost broja komentara za kanal i maksimalna vrednost broja komentara za kanal.

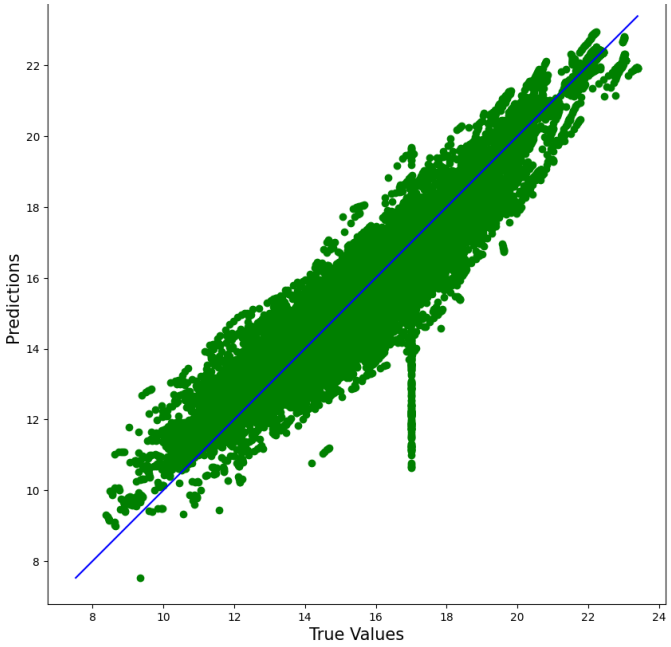
Nakon što su podaci pripremljeni, sledeći korak je treniranje modela. U ovom radu razmatrani su sledeći regresioni modeli:

- Linearna regresija
- *XGBoost*
- *CatBoost*
- *Support Vector Regression*

Za sve modele, osim za model zasnovan na linearnoj regresiji, je rađena optimizacija parametara kako bi se postigli što bolji rezultati. Za evaluaciju modela su se koristile R2 mera i koren srednje vrednosti kvadrata greške (*RMSE*) Tabela 4. prikazuje rezultate koje je imao model zasnovan na linearnoj regresiji, dok Slika 4. grafički ilustruje rezultate na podacima za testiranje.

Trening RMSE	**Test RMSE**	**Trening R2**	**Test R2**
0.884	0.710	0.89	0.93

Tabela 4. Evaluacija modela zasnovanog na linearnoj regresiji

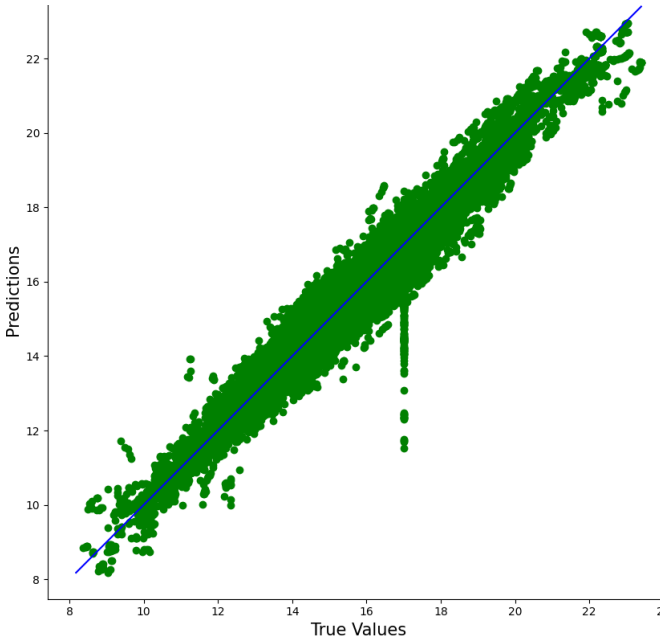


Slika 4. Rezultati linearne regresije

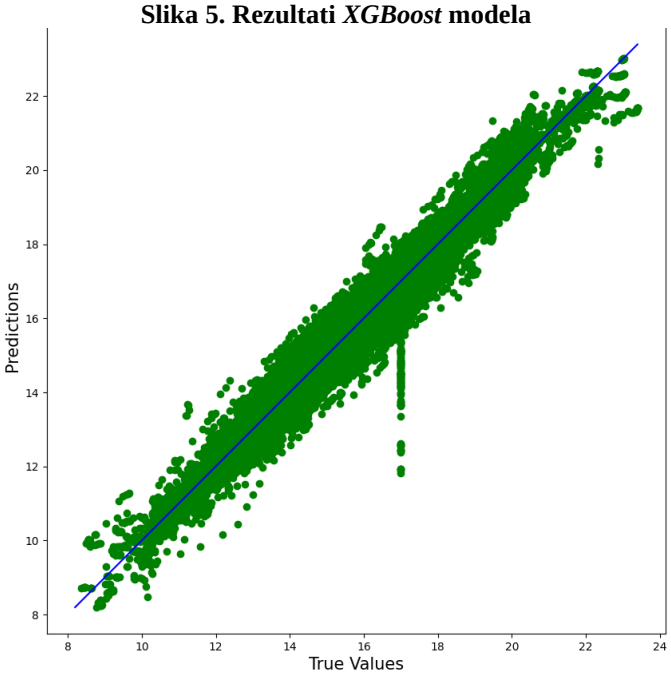
Za *XGBoost* model su optimizovani sledeći parametri: *n_estimators*, *learning_rate*, *max_depth* i *colsample_bytree*. Parametri *n_estimators* i *learning_rate* su optimizovani zajedno, kako bi se postigli najbolji rezultati sa najboljim performansama. Za *n_estimators* su isprobane vrednosti 1000, 5000, 10000, 20000 i 30000, dok su za *learning_rate* uzete u obzir vrednosti 0.1, 0.01 i 0.001. Najbolji rezultati dobijeni su za vrednosti 20000 i 0.01. Za parameter *max_depth* je isprobano 10,8 i 6, a najbolje se pokazala vrednost 8, dok su se za *colsamples_bytree* probale vrednosti 0.4, 0.5, 0.6 i 0.7, a najbolje se pokazala vrednost 0.6. Tabela 5. prikazuje rezultate *XGBoost* modela sa optimizovanim parametrima, a Slika 5. predstavlja grafički prikaz rezultata nad podacima za testiranje.

Trening RMSE	Test RMSE	Trening R2	Test R2
0.081	0.375	0.99	0.98

Tabela 5. Evaluacija *XGBoost* modela



Slika 5. Rezultati *XGBoost* modela



Slika 6. Rezultati *CatBoost* modela

Što se tiče *Suport Vector Regression* algoritma, kod njega je moguće promeniti nekoliko *krenel*-a. Isprobani su *linear, poly i rbf*. Zaključeno je da je problem linearan jer je *poly kernel* davao dobre rezultate samo ako je stepen polinoma jedan, što predstavlja linearan problem. Što se tiče ostalih parametara menjana je vrednost parametra C i najbolje rezultate algoritam daje kad ostane predefinisana vrednost 1.0 . Za broj iteracija primenjene su vrednosti 10000, 50000, 100000, 200000 i 300000. Vrednost 200000 se pokazala kao najbolja jer 300000 nije promenilo ništa u odnosu na 200000 dok manji broj iteracija daje lošiji rezultat. Promena *tol* (tolerancija) parametra se uopšte nije osetila. Tabela 7. prikazuje rezultate modela, dok se na Slici 7. može videti grafički prikaz nad podacima za testiranje.

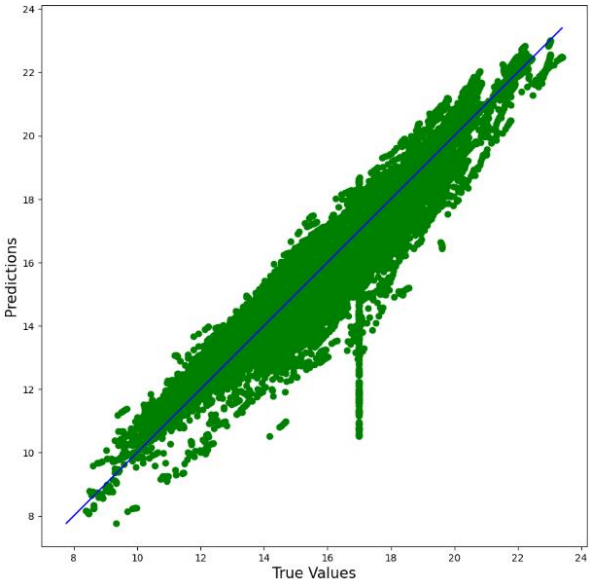
CatBoost model je optimizovan na sličan način kao *XGBoost model.* Za *n_estimators* su isprobane vrednosti 10000, 20000, 30000, 40000, 50000 i 60000, a kao najbolje je odabrana vrednost 50000, uz vrednost 0.01 za parameter *learning_rate.* Za parameter *depth* su isprobane vrednosti 6, 8 i 10, a najmanju grešku daje vrednost 8, dok je za parametar *rsm*(analogan parametru *colsamples_bytree* kod *XGBoost* modela) najmanju grešku dala vrednost 0.6, od isprobanih 0.4, 0.5 i 0.6. Tabela 6. prikazuje rezultate modela, dok se na Slici 6. može videti grafički prikaz nad podacima za testiranje.

Trening RMSE	Test RMSE	Trening R2	Test R2
0.152	0.376	0.99	0.98

Tabela 6. Evaluacija *CatBoost* modela

Trening RMSE	Test RMSE	Trening R2	Test R2
0.94	0.62	0.91	0.78

Tabela 7. Evaluacija *Support Vector Regression* modela



Slika 7. Rezultat Suport Vector Regression modela

VI. Analiza grešaka

Fokus ovog poglavlja je analiza grešaka modela, kako bi se rezultati dodatno poboljšali. To podrazumeva posmatranje podataka za koje model pravi velike greške, nalaženje nekih šablona u tim podacima, kao i ponovno treniranje modela nad skupom podataka koji ne sadrži podakte koji prouzrokuju grešku. Pošto se *XGBoost* model pokazao kao najbolji, analiza grešaka je odrađena nad njim.

Od 49181 podataka za test, izdvojeni su oni podaci kod kojih je greška veća od 50% vrednosti za očekivan broj lajkova, i postoji 2798 takvih podataka. Analizom kategoričkih atributa (kategorija snimka, kategorija slika, kategorija tekstualnih atributa), uočeno je da najviše podataka pripadaju onim kategorijama koje su najzastupljenije, odnosno nije pronađena veza između neke od kategorija i greške. Nakon toga su posmatrani numerički podaci, i uočeno je da podaci koji imaju veliku grešku uglavnom imaju velike vrednosti za broj pregleda, broj dislajkova, broj komentara, dužinu naslova i broj oznaka. Uklanjanjem podataka sa velikim vrednostima za ove atribute, skup podataka se sa 245903 podataka smanjio na 204829. *XGBoost* model je ponovo treniran, i Tabela 8. prikazuje dobijene rezultate.

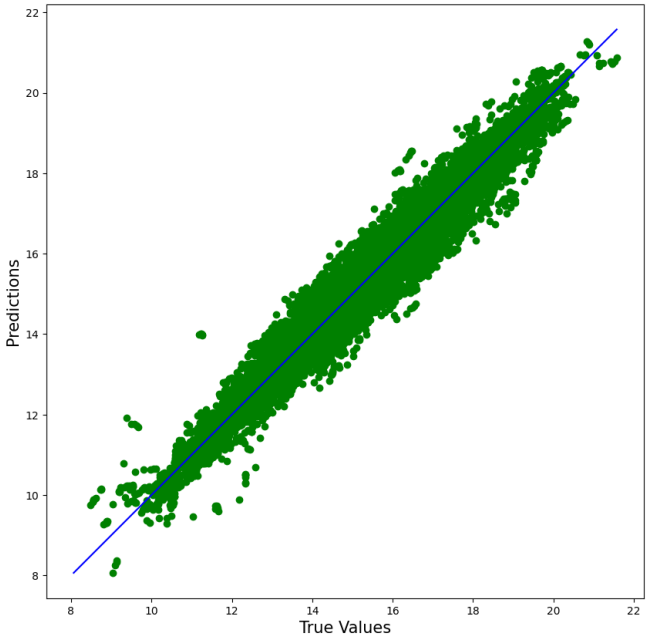
Trening RMSE	Test RMSE	Trening R2	Test R2
0.069	0.372	0.99	0.98

Tabela 8. Evaluacija *XGBoost* modela nad novim skupom podataka

I pored boljih rezultata i manje greške, analiza grešaka je nastavljena. Ponovo su posmatrani podaci koji prouzrokuju velike greške, i uočeno je da gotovo svi imaju istu vrednost za broj lajkova: 131122.742079866. Naime, određen broj video snimaka je imao nedostajuću vrednost za broj lajkova, pa je ta vrednost zamenjena srednjom vrednošću, što je baš vrednost broja lajkova koju imaju podaci sa velikom greškom. Izbacivanjem podataka sa ovom vrednošću iz skupa, skup podataka se dodatno smanjio na 203555 podataka. Nakon ponovnog treniranja modela, dobijenu su još bolji rezultati, a Tabela 9. poredi rezultate pre i posle izbacivanja podataka sa velikom greškom. Slika 8. predstavlja grafički prikaz novih rezultata.

	Trening RMSE	Test RMSE	Trening R2	Test R2
Nad svim podacima	0.081	0.375	0.99	0.98
Nad redukovanim skupom podataka	0.065	0.344	0.99	0.98

Tabela 9. Poređenje rezultata *XGBoost* modela



Slika 8. Rezultati *XGBoost* modela nad redukovanim skupom podataka

VII. Zaključak

Ovaj rad se bavio rešavanjem problema predikcije popularnosti YouTube snimka, gde se kao mera popularnosti koristio broj lajkova. U današnje vreme, kada sve više i više ljudi bira *online* zanimanja, kao što su kreiranje video sadržaja, bitno je znati kako se istaći, kao i koji su to faktori koji utiču na popularnost i uspeh kreiranog sadržaja.

Prvi korak u rešavanju ovog problema je kreiranje odgovarajućeg skupa podataka. Skup podataka [8] je proširen dodatnim atributima kako bi se povećala tačnost modela. Zatim je korišćenjem dubokog učenja odrađena analizia slike i tekstualnih atributa video snimka, čime se polazni skup podatak dodatno proširio. Za analizu slika je korišćena *ResNet* konvolutivna neuronska mreža i transfer učenja, dok se za analizu teksta radio *fine-tuning BERT* jezičkog modela. Na kraju su primenjene razne tehnike za pretprocesiranje podataka kako bi se modeli što bolje istrenirali nad njima.

Nakon što su podaci pripremljeni, pristupilo se treniranju modela. Ovaj rad je razmatrao 4 modela i poredio njihove rezultate. U obzir su uzeti: linearna regresija, *XGBoost* model, *CatBoost* model, *SVR* model. Modeli su evaluirani i njihovi rezultati su poređenji na osnovu R2 mere i korena srednje vrednosti kvadrata greške (*RMSE*). Kao najbolji se pokazao *XGBoost* model, koji je imao R2 vrednost od 0.98 i RMSE od 0.375. Kako bi se rezultati dodatno poboljšali urađena je analiza grešaka, čime su se iz skupa podataka izbacili podaci koji uzrokuju veliku grešku. Nakon treniranja *XGBoost* modela nad novim skupom, R2 mera je ostala ista, dok se RMSE vrednost smanjila na 0.344.

Dalji razvoj i usavršavanje ovog rešenja obuhvata proširenje skupa podataka većim brojem atributa, kao i primena još nekih tehnika pretprocesiranja. Takođe je moguće uzeti u razmatranje još neke modele koji nisu razmatrani u ovom radu.

LITERATURA

[1] Deshak Bhatnagar & Siddhaling Urolagin. YouTube Video Popularity Analysis& Prediction Using Deep Learning. International Journal of Computational Intelligence in Control 13, 2021

[2] Yuping Li, Kent Eng, Liqian Zhang Department of Civil and Environmental Engineering, YouTube Videos Prediction: Will this video be popular?, Stanford University, Stanford, CA 94305

[3] YouTube Video Statistics - https://www.kaggle.com/datasets/datasnaek/youtube-new

[4] Nisa M.U., Mahmood D., Ahmed G., Khan S., Mohammed M.A., Damaševicius R. Optimizing Prediction of YouTube Video Popularity Using XGBoost. Electronics 2021, 10, 2962.

[5] Braun, P., Cuzzocrea, A., Doan, L. M. V., Kim, S., Leung, C. K., Matundan, J. F. A., & Robby Singh, R. (2017). *Enhanced Prediction of User-Preferred YouTube Videos Based on Cleaned Viewing Pattern History*. *Procedia Computer Science, 112, 2230–2239.* doi:10.1016/j.procs.2017.08.129

[6] Rui, Lau & Afif, Zehan & Saedudin, Rd & Mustapha, Aida & Razali, Nazim. (2019). A regression approach for prediction of Youtube views. Bulletin of Electrical Engineering and Informatics. 8. 10.11591/eei.v8i4.1630.

[7] Trending YouTuve Video Statistics, https://www.kaggle.com/datasnaek/youtube-new

[8] YouTube Trending Video Dataset(updated daily), https://www.kaggle.com/datasets/rsrishav/youtube-trending-video-dataset/data