UDC: 004.8 Info M str. 28-39

PRIMJENA MAŠINSKOG UČENJA ZA OPISIVANJE SLIKA POMOĆU TEKSTA
USING MACHINE LEARNING FOR IMAGE CAPTIONING

Aleksije Mićić, prof. dr Zoran Đurić

REZIME: Moderni informacioni sistemi za e-trgovinu sve češće integrišu sisteme za preporuku proizvoda kako bi unaprijedili korisničko iskustvo. U ovom radu su analizirane primjene opisivanja slika pomoću tehnika mašinskog učenja, s ciljem generisanja boljih preporuka za e-trgovine koje se bave prodajom odjevnih predmeta. Poseban fokus stavljen je na ograničenja tradicionalnih pristupa koji se oslanjaju isključivo na vizuelnu sličnost među proizvodima. U praktičnom dijelu rada evaluirani su opisi generisani pomoću osam različitih modela, pri čemu su se najbolje pokazali opisi generisani sa velikim jezičkim modelima. Nakon toga je evaluirano predloženo rješenje koje kombinuje tradicionalni pristup sa analizom sličnosti opisa na prethodno opisanim problemima. Dobijeni rezultati ukazuju na opravdanost ovog pristupa. Na kraju rada dati su mogući pravci daljeg istraživanja.

KLJUČNE REČI: Sistemi za preporuku, Opisivanje slika pomoću teksta, Pretraga unutar prodavnice, Pretraga od potrošača do prodavnice, Veliki jezički modeli

ABSTRACT: Modern information systems for e-commerce increasingly integrate product recommendation engines to enhance the user experience. This paper analyzes the application of image captioning techniques based on machine learning, with the goal of generating better recommendations for e-commerce platforms focused on selling clothing. Special attention is given to the limitations of traditional approaches that rely solely on visual similarity between products. Image captions generated by eight different models were evaluated, with large language models performing the best. Subsequently, the proposed solution, which combines the traditional visual similarity approach with semantic similarity analysis of the generated descriptions, was evaluated across previously defined problems. The results demonstrate the effectiveness and justification of this new approach. Finally, the paper outlines potential directions for future research.

KEY WORDS: Recommendation systems, Image captioning, Shop-to-Shop retrieval, Consumer-to-shop retrieval, Large language models

1. UVOD

Kupovina proizvoda putem informacionih sistema za e-trgovinu je u stalnom porastu. Ovakav vid kupovine omogućava generisanje preporuka za proizvode na osnovu korisničkih akcija, što može pozitivno uticati na zaradu prodavača [1]–[3] S druge strane, korisnička očekivanja, vezana za kvalitet preporuka, su sa godinama samo porasla, a od kvaliteta preporuka uveliko zavisi cjelokupno korisničko iskustvo, naročito u e-trgovinama koje se bave prodajom odjevnih predmeta.

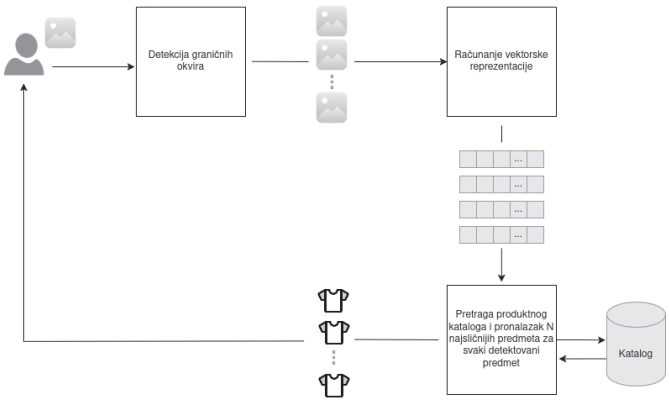
Za pronalazak sličnih i povezanih proizvoda, posebno za pretrage unutar prodavnice (eng. *shop-to-shop retrieval*) i od potrošača do prodavnice (eng. *consumer-to-shop retrieval*), u daljnjem tekstu pretraga prodavnice, koristi se analiza sličnosti slika [4], [5]. Kao ključ pretrage uzima se slika na kojoj su prikazani odjevni predmeti od interesa, a kao rezultat pretrage dobijaju se dostupni slični i povezani odjevni predmeti iz kataloga proizvoda e-trgovine. Ovaj pristup veoma brzo generiše preporuke koje su, u zavisnosti od složenosti upita, vizuelno slične predmetima koji su dio ključa pretrage. Takođe, memorijski zahtjevi za čuvanje vektorske reprezentacije slika svih proizvoda su relativno male, u odnosu na same slike i obično staju u radnu memoriju bilo kog modernog računara, čak kada je riječ i o velikim katalozima sa milionima proizvoda. Ipak ovakav pristup ima i svoje mane. Proces prikupljanja i labelisanja podataka koji će se koristiti tokom obučavanja modela je mukotrpan i iziskuje mnogo vremena, ali i zahtjeva obučene anotatore, zbog čega može biti skup. Takođe nije dovoljno samo obučiti model koji vrši analizu sličnosti slika odjevnih predmeta, jer dati model kao ulaz podrazumijevano dobija pojedinačne predmete. Zbog čega je potrebno i obučiti model koji detektuje pojedinačne predmete, tako što određuje njihov granični okvir (eng. *bounding box*) i dodijeli im odgovarajuću klasu [4]–[6].

U zavisnosti od složenosti ključa pretrage, može doći do detekcije mnogo lažno pozitivnih predmeta, npr. kada je ključ majica koja prikazuje ljude koji na sebi imaju nove odjevne predmete za koje se onda takođe traže slični proizvodi. Takođe, ako je ključ pretrage složeniji, odnosno ako je na slici osoba koja nosi višeslojnu garderobu, tada je često nemoguće identifikovati sve odjevne predmete. Kada je detekcija predmeta uspješna, problem ponekad predstavlja i to što model pronalazi vizuelno sličnu odjeću, bez razumijevanja konteksta, ako je npr. ključ pretrage košulja koju nosi neka osoba, onda će košulje iz kataloga koje su na slici savijene imati manju sličnost, čak i kada je riječ o identičnom predmetu. Ovaj problem je još više izražen, kada je kao ključ pretrage data majica sa nekim likom iz popularne kulture, npr. lik iz filma, jer će se kao rezultat dobiti majice koje prikazuju druge, vizuelno slične likove, ne nužno povezane sa datim likom, gdje bi uobičajeno bilo adekvatnije prikazati majice sa drugim likovima povezanim sa datim likom, ili čak majice sa natpisima vezani za istu franšizu.

Za prevazilaženje ovih mana predloženi su pristupi koji kombinuju analizu sličnosti slika sa opisima slika generisanim pomoću tehnika mašinskog učenja. Ovaj rad se bavi primjenom opisivanja slika pomoću tehnika mašinskog učenja za poboljšanje korisničkih preporuka, posebno za problem pretraživanja prodavnice. U drugom poglavlju je analiziran tradicionalni pristup za pretragu prodavnice baziran na detekciji objekata i analizi sličnosti slika, kao i problemi sa ovakvim pristupom. U trećem poglavlju je dat opisa rješenja koje se temelji na kombinovanju starog pristupa sa generisanjem opisa. Detaljno je objašnjen postupak generisanja opisa, te je dat pregled najrelevantnijih vektorskih reprezentacija teksta, koje su korištene u eksperimentalnom dijelu rada. U četvrtom poglavlju, opisane su ukratko metrike za evaluaciju kvaliteta opisa, te su dati rezultati za sve predložene modele. Nakon toga, analizirani su rezultati primjene generisanja opisa za probleme navedene u drugom poglavlju. Peto poglavlje sadrži zaključak.

2. OPIS PROBLEMA

Primjer tradicionalnog pristupa za pretragu prodavnice je ilustrovan na slici 1. Prvi korak uključuje detekciju graničnih okvira na ulaznoj slici *I*, gdje se koristi napredni algoritam detekcije objekata, poput YOLO[1] ili Mask R-CNN, da identifikuje skup graničnih okvira $B = \{b_1, b_2, \ldots, b_k\}$ pri čemu svaki okvir b_i definiše koordinate (x, y, w, h) [7], [8]. Nakon identifikacije graničnih okvira, za svaki okvir b_i računa se vektorska reprezentacija primjenom konvolucione neuronske mreže (eng. *convolutional neural network* - CNN), rezultujući skupom vektora $V = \{v_1, v_2, \ldots, v_k\}$ gdje je $v_i \in \mathbb{R}^d$. Sljedeći korak uključuje pretragu kataloga proizvoda $C = \{c_1, c_2, \ldots, c_m\}$, gdje svaki proizvod c_j ima svoju vektorsku reprezentaciju. Za svaki vektor v_i izračunava se sličnost sa proizvodima c_j korišćenjem metričke funkcije, kao što su kosinusna sličnost ili Euklidova udaljenost, čime se identifikuje skup N najsličnijih proizvoda $S_i = \{s_{i1}, s_{i2}, \ldots, s_{iN}\}$. Konačno, korisniku se prikazuje skup preporuka $S = \{S_1, S_2, \ldots, S_k\}$ [4], [5].



Slika 1. *Šematski prikaz arhitekture za tradicionalni pristup pretrage prodavnice*

[1] https://docs.ultralytics.com/

2.1. *Obučavanje modela za detekciju objekata*

Neka je dat skup podataka $D = \{(x_i, y_i)\}_{(i=1)}^{N}$, gdje je $x_i \in \mathbb{Z}^{H \times W \times K}$ slika, a y_i pripadajuće oznake za tu sliku. Svaka oznaka y_i sastoji se od skupa graničnih okvira i odgovarajućih klasa za tu sliku. Ako se na slici x_i nalazi n_i objekata garderobe, tada je:

$$yi = \left\{ \left(b_{i1}, c_{i1}\right), \left(b_{i2}, c_{i2}\right), \ldots, \left(b_{in_i}, c_{in_i}\right) \right\}$$

gdje je $b_i j = (x_{\min}^{ij}, y_{\min}^{ij}, x_{\max}^{ij}, y_{\max}^{ij})$ skup koordinata koje definišu granični okvir objekta j na slici i, a $c_{ij} \in C$ je klasa tog objekta (npr., "majica", "haljina", "pantalone", itd.). Cilj je obučiti model detekcije objekata f_θ, koji mapira ulaznu sliku x_i u skup predviđenih graničnih okvira i klasa:

$$\widehat{y_i} = \{(\widehat{b_{ik}}, \widehat{c_{ik}})\}_{k=1}^{\widehat{n_i}},$$

gdje je $\widehat{n_i}$ broj predviđenih objekata na slici x_i, $\widehat{b_{ik}}$ su predviđeni granični okviri, a $\widehat{c_{ik}}$ su predviđene klase. Model f_θ se trenira da minimizuje ukupnu funkciju gubitka $L(\theta)$, koja kvantifikuje razliku između predviđanja modela i stvarnih oznaka. Funkcija gubitka sastoji se iz dva dijela:

- **Gubitak klasifikacije** (L_{cls}): Mjera greške u predviđanju klasa objekata. Obično se koristi unakrsna entropija (eng. *cross-entropy loss*) za više klasa.
- **Gubitak regresije** (): Mjera greške u predviđanju koordinata graničnih okvira. Često se koristi *Smooth* L_1 gubitak ili L_2 norma za regresiju koordinata

Ukupna funkcija gubitka data je izrazom:

$$L(\theta) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{n_i} \left(L_{\text{cls}} \left(c_{ij}, \widehat{c_{ij}} \right) + \lambda L_{\text{reg}} \left(b_{ij}, \widehat{b_{ij}} \right) \right),$$

gdje je N broj slika u skupu podataka, n_i broj stvarnih objekata na slici x_i, λ je hiperparametar koji balansira uticaj između gubitka klasifikacije i gubitka regresije, c_{ij} i $\widehat{c_{ij}}$ su stvarne i predviđene klase, a b_{ij} i $\widehat{b_{ij}}$ su stvarni i predviđeni granični okviri.

2.2. *Obučavanje modela za analizu sličnosti slika*

Nakon detekcije objekata i izdvajanja pojedinačnih predmeta garderobe sa slika, cilj je obučiti model koji može da mapira ove predmete u vektorski prostor gdje su više slični predmeti bliži jedni drugima, a manje slični predmeti udaljeniji jedni od drugih. Neka je dat model g_ϕ, koji mapira ulaznu sliku predmeta x_{ij} (određenu graničnim okvirom) u vektorsku reprezentaciju b_{ij}, odnosno $z_{ij} = g_\phi(x_{ij})$.

Obučavanje modela g_ϕ zasniva se na korišćenju funkcije gubitka tipa *triplet loss* zajedno sa strategijom *hard mining* [5]. Formiraju se trojke (a, p, n), gdje su:

- a sidro (eng. *anchor*), slika nekog predmeta,
- p pozitivan uzorak, slika istog predmeta,
- n negativni uzorak, slika različitog predmeta.

Nakon što se svaka od tih slika proslijedi kroz model g_ϕ, dobijaju se vektorske reprezentacije $z_a = g_\phi(a)$, $z_p = g_\phi(p)$ i $z_n = g_\phi(n)$. *Triplet loss* funkcija definiše se kao:

$$L_{\text{triplet}}(\phi) = \sum_{(a,p,n)} \left[\left| z_a - z_p \right|_2^2 - \left| z_a - z_n \right|_2^2 + \alpha \right]_+ ,$$

gdje je $\alpha > 0$ margina koja definiše minimalnu razliku između pozitivnih i negativnih parova, kako ne bi bili suviše blizu jedan drugom, a $[x]_+ = \max(0, x)$ se koristi u funkciji gubitka kako bi se osiguralo da samo pozitivni tj. problematični slučajevi doprinesu gubitku, dok se negativni zanemaruju.

Hard mining strategija podrazumijeva izbor najtežih pozitivnih i negativnih uzoraka tokom obučavanja. Za svako sidro *a*, bira se pozitivni primjer *p* koji je najudaljeniji od *a* u trenutnom vektorskom prostoru, što znači da ga model teže razlikuje od negativnih. Za svako sidro *a*, bira se negativni primjer *n* koji je najbliži *a*, tj. onaj koji je za model najteže razlikovati od pozitivnih. Kroz iterativno ažuriranje parametara ϕ, model uči vektorske reprezentacije koje bolje odražavaju semantičku sličnost između predmeta. Na ovaj način, model je sposoban da kodira predmete garderobe tako da su slični predmeti (npr., ista majica fotografisana u različitim kontekstima) blizu u vektorskom prostoru, dok su različiti predmeti udaljeni.

Za potrebe rada dotrenirana (eng. *fine-tuned*) su tri modela za detekciju objekata: model baziran na Detectron2 Mask-R-CNN, YOLOv8 i YOLOv11 nad *DeepFashion2*[2] *– DF2* skupu podataka. Pored toga, obučen je jedan model za potrebe analize sličnosti slika baziran na ResNet50 CNN nad istim skupom podataka. DF2 je skup podataka namijenjen za istraživanje u oblasti računarskog vida i modne industrije. Sadrži slike odjevnih predmeta s detaljnim oznakama, uključujući granične okvire, maske segmentacije (eng. *segmentation masks*), atribute odjeće i informacije o katego-rijama.

2.3. *Mane tradicionalnog pristupa za pretragu proizvoda*

Razlikujemo dvije vrste nedostataka kod tradicionalnog pristupa za pretragu proizvoda, one koji potiču od modela za detekciju objekata i one koje potiču od modela za analizu sličnosti slika.

2.3.1. Nedostaci modela za detekciju objekata

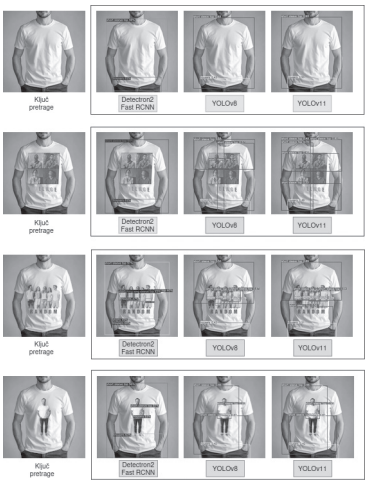
Pošto je prvi korak u sistemu pretrage prodavnice, detekcija odjevnih predmeta i njihovih graničnih okvira, greške u ovoj fazi mogu nepovratno poremetiti čitav proces. Najčešće greške su [9]:

- Pogrešno određen granični okvir. Dešava se kada model pogrešno odredi granice okvira za neki odjevni predmet, pa se može desiti da se ne pronađu dovoljno slični predmeti.
- Pogrešna klasifikacija graničnog okvira. Dešava se kada model dodijeli pogrešnu klasu nekom graničnom okviru. Moguće je da je okvir pravilno određen, ali da se zbog pogrešne klase ne pronađu dovoljno slični predmeti.

[2] https://github.com/switchablenorms/DeepFashion2

Prethodna dva problema se mogu uglavnom prevazići analizom pogrešnih uzoraka i proširivanjem trening skupa. Ipak postoje i greške koje se ne mogu prevazići na ovakav način:

- Detekcija lažno pozitivnih uzoraka. Ako je ključ pretrage majica na kojoj su naslikani ljudi, model će detektovati granične okvire za garderobu koju ti naslikani ljudi nose, što rezultuje velikim brojem lažno pozitivnih uzoraka. Primjer detekcije lažno pozitivnih uzoraka prikazan je na slici 2.



Slika 2. Primjer detekcije lažno pozitivnih uzoraka, kada je na majici prikazan neki broj ljudi

- Detekcija slojevite odjeće. Slojevita garderoba predstavlja složen ključ pretrage i predstavlja veliki problem za modele za detekciju. Kako su određeni slojevi odjeće samo djelimično vidljivi, model ih "ne vidi", obično se u takvim situacijama samo detektuje posljednji sloj garderobe koji je i najvećim dijelom vidljiv. Ovo je problem u situacijama kada potrošač želi u potpunosti da rekreira stil garderobe iz ključa pretrage. Primjer problema sa slojevitom garderobom je prikazan na slici 3, gdje od vidljive garderobe ni na jednoj osobi nije detektovan džemper.

Slika 3. *Primjer problema sa detekcijom sve odjeće kod slojevite garderobe*

2.3.2. Nedostaci modela za analizu sličnosti slika

Ako su pravilno detektovani granični okviri sljedeći korak je da se za njih kreiraju vektorske reprezentacije i da se pronađu najsličniji odjevni predmeti iz kataloga proizvoda. Ipak i u ovom koraku postoje određeni problemi:

- Nemogućnost razumijevanja konteksta iz ključa pretrage. Pošto je model obučen da pronalazi vizuelno slične slike, on nije u stanju da pravilno odredi sličnost odjevnih predmeta koji možda nisu toliko vizuelno slični, ali pripadaju zajedničkom kontekstu. Npr. ako je ključ pretrage majica sa likom iz neke serije, a katalog proizvoda ima samo majice sa natpisima iz te serije, tada katalog vjerovatno neće odabrati te majice među najsličnijima, a važi i obrnuto. Takođe, ako je ključ pretrage majica sa nekim likom specifičnog izgleda iz određene franšize, sistem će kao najsličnije da vrati majice sa likovima sličnog izgleda, a koji su možda iz potpuno drugih franšiza, a koje uopšte ne interesuju potrošača. Na slici 4 je ilustrovan dati problem, gdje majica sa likom čarobnjaka i majica sa engleskom riječi za čarobnjaka nisu uopšte vizuelno slične pa se ne bi pojavili u rezultatima pretrage jedna za drugu.

Slika 4. *Primjer problema sa nerazumijevanjem konteksta kod pretrage majica u prodavnici*

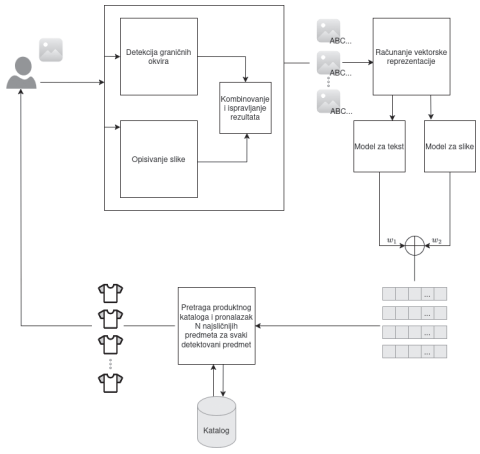
- Nedovoljna invarijantnost na različite načine prikaza nekih odjevnih predmeta. Kako bi se postigla invarijantnost u pretrazi prodavnice, potrebno je da se u trening skupu podataka nađe dovoljan broj uzoraka koji su prikazani pod različitim orijentacijama i u različitim okolnostima. Ipak model ima tendenciju da preferira vizuelno slične odjevne predmete i potpunu invarijantnost je nemoguće postići. Ako je ključ pretrage košulja prikazana na nekoj osobi, a u katalogu proizvoda prodavača se nalazi ista ta košulja, ali na slici na kojoj je savijena, tada se ona vjerovatno neće naći među najsličnijim košuljama koje su vraćene kao rezultat pretrage, ova situacija je ilustrovana na slici 5.

Slika 5. *Primjer problema sa nerazumijevanjem konteksta i nedovoljne invarijantnosti kod pretrage košulja u prodavnici*

- Poteškoće u pronalasku povezanih proizvoda. Sistemi za pretragu bazirani na analizi vizuelne sličnosti nisu u mogućnosti identifikovati proizvode koji su funkcionalno ili kontekstualno povezani, ali vizuelno različiti. Na primjer, korisnik koji pretražuje duks određene robne marke može biti zainteresovan za komplementarne odjevne predmete istog brenda, poput trenerke ili majice iz iste kolekcije. Međutim, ovi predmeti se vizuelno značajno razlikuju i stoga algoritam za pretragu sličnosti slika ih neće prepoznati kao povezane. Kako sistem nema implementirano razumijevanje konteksta ni semantičkih veza između predmeta, on nije u stanju adekvatno identifikovati komplementarne proizvode koji bi mogli biti relevantni za korisnika.

3. PRIJEDLOG RJEŠENJA

Za rješavanje prethodno navedenih problema moguće je primijeniti opisivanje slika pomoću tehnika mašinskog učenja, pomoću kojeg se može simulirati razumijevanje konteksta. Na slici 6 je prikazana arhitekturalna šema takvog jednog sistema.



Slika 6. *Šematski prikaz arhitekture sistema za pretragu prodavnice koja kombinuje tradicionalni pristup sa opisivanjem slika*

Slično kao i kod sistema prikazanog na slici 1. prvi korak u sistemu uključuje detekciju graničnih okvira na ulaznoj slici I, kako bi se identifikovao skup $B = \{b_1, b_2, \ldots, b_k\}$. Paralelno s tim, koristi se model za generisanje opisa slike koji analizira ulaznu sliku i generiše skup tekstualnih opisa $O = \{o_1, o_2, \ldots, o_l\}$, od kojih svaki opisuje jedan od detektovanih odjevnih predmeta. Sljedeći korak uključuje kombinovanje i ispravljanje rezultata, gdje se rješavaju prethodno opisane mane. Problem generisanja lažno pozitivnih uzoraka rješava se tako što se, u slučaju kada model detektuje više okvira $|B| > 1$, a generiše samo jedan opis $|O| = 1$, zadržava samo najveći okvir $b_{\max}$, definisan površinom $w \cdot h$. Problem slojevite odjeće rješava se tako što se, u slučaju da model detektuje manje okvira $|B|$ nego što ima generisanih opisa $|O| > |B|$, opisi i okviri uparuju prema kategorijama (npr. "jakna", "majica"), dok se za opise koji ostanu neupareni generišu samo tekstualni vektori. Za svaki par (b_i, o_i), vektor slike v_i^s se računa pomoću CNN, dok se vektor opisa v_i^t generiše korišćenjem NLP modela kao što su Word2Vec, BERT ili sl. Rezultujući vektor v_i se dobija kao linearna kombinacija: $v_i = w_1 v_i^s + w_2 v_i^t$, gdje su w_1 i w_2 hiperparametri koji se određuju eksperimentalno. U slučaju da vektori v_i^s i v_i^t nisu iste dimenzije, vektor veće dimenzije se redukuje na dimenziju manjeg vektora. Nakon što su svi parovi (b_i, o_i) obrađeni, rezultujući vektori se upoređuju sa vektorskim reprezentacijama proizvoda iz kataloga $C = \{c_1, c_2, \ldots, c_m\}$. Slični proizvodi se pronalaze na isti način kao i ranije.

Problem razumijevanja konteksta se rješava tako što generisani opisi sadrže informacije koje su od značaja za neki odjevni predmet, opisujući ujedno i neke kontekstualne informacije vezane za taj predmet, npr. naziv franšize i lika koji je naslikan na majici, ili da je riječ o savijenoj košulji.

3.1. *Generisanje opisa slika*

Opisivanje slika pomoću teksta (eng. *image captioning* - IC) je zadatak generisanja deskriptivnih tekstualnih opisa za slike koristeći ML tehnike. Ovaj zadatak spada u multimodalne probleme jer integriše vizuelni i jezički domen, zahtijevajući efikasno premošćivanje semantičkog jaza između njih. Cilj ovog postupka je razviti model koji, datu sliku $I \in \Im$, mapira na tekstualnu sekvencu $S = \{w_1, w_2, \ldots, w_T\}$, gdje je svaka riječ w_t element predefinisanog vokabulara V. Formalno gledano, cilj je modelovati uslovnu vjerovatnoću $P(S \mid I)$, odnosno pronaći sekvencu riječi S koja maksimizira ovu vjerovatnoću. Stoga, cilj je optimizacija sljedećeg izraza:

$$S^* = \arg\max_S P(S \mid I)$$

gdje S^* predstavlja optimalnu sekvencu riječi koja najbolje opisuje sliku I. Da bi se riješio ovaj zadatak, sistem IC tipično se sastoji od dvije glavne komponente: enkodera i dekodera.

Enkoder je komponenta koja obrađuje ulaznu sliku I i ekstrahuje njene značajne vizuelne karakteristike $F \in \mathbb{R}^d$, gdje predstavlja dimenzionalnost vektora karakteristika. Tipično, enkoder je implementiran korišćenjem CNN ili vizuelnih transformatora, koji su prethodno trenirani na velikim skupovima podataka za zadatke klasifikacije slika. Formalno, enkoder može biti predstavljen kao funkcija:

$$f_{\text{enc}} : \Im \to \mathbb{R}^d, \; F = f_{\text{enc}}(I),$$

gdje F predstavlja vektorsku reprezentaciju slike I u prostoru karakteristika. Dekoder, s druge strane, ima zadatak da generiše tekstualni opis $S = \{w_1, w_2, \ldots, w_T\}$ na osnovu vektorske reprezentacije F. Dekoder se obično implementira korišćenjem LSTM ili transformatora. Dekoder modeluje uslovnu vjerovatnoću $P(S \mid I)$ kao produkt uslovnih vjerovatnoća svake riječi u sekvenci:

$$P(S \mid I) = \prod_{t=1}^{T} P\left(w_t \mid w_1, w_2, \ldots, w_{t-1}, F\right),$$

gdje w_t predstavlja t-tu riječ u sekvenci. Da bi dekoder pravilno obradio izlaz enkodera, često je neophodno uskladiti dimenzije vektorskih reprezentacija enkodera sa očekivanim ulaznim dimenzijama dekodera. Zbog toga se u praksi, posebno kod multimodalnih LLM-ova, primjenjuje linearna projekcija na izlaz enkodera:

$$F' = W \times F + b,$$

gdje je $F \in \mathbb{R}^d$ vektorska reprezentacija slike dobijena enkoderom, $W \in \mathbb{R}^{ddec \times d}$ matrica težina b je vektor pomaka (bias). Kao rezultat transformacije dekoder dobija ulaz $F' \in \mathbb{R}^{ddec}$ koji je dimenziono usklađen. Na ovaj način se sprječava problem da dekoder prima vektor neodgovarajuće dimenzije, što bi onemogućilo normalnu propagaciju signala tokom treninga ili generisanja opisa. Dekoder se formalno može predstaviti kao funkcija:

$$f_{\text{xe}}^{\text{c}}: (\mathbb{R}^d, w_1, w_2, \ldots, w_{t-1}) \to P(w_t \mid w_1, w_2, \ldots, w_{t-1}, F).$$

Kombinovanjem enkodera i dekodera, cjelokupan model može se posmatrati kao funkcija:

$$\text{f}_{\text{IC}}: I \to S, \quad S = f_{\text{IC}}(I),$$

gdje S predstavlja prostor svih mogućih tekstualnih sekvenci. Trening modela za IC uključuje optimizaciju parametara enkodera i dekodera kako bi se maksimizovala vjerovatnoća generisanja tačnih opisa za slike u skupu za obuku. Ovo se formalno izražava kao minimizacija negativnog logaritamskog vjerovatnosnog gubitka (eng. *negative log-likelihood loss*):

$$L = -\sum_{(I,S)\in D} \sum_{t=1}^{T} \log P\left(w_t \mid w_1, w_2, \ldots, w_{t-1}, F\right),$$

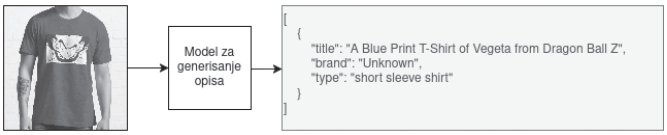
gdje D predstavlja skup podataka za obuku, a $F = f_{\text{enc}}(I)$. Za potrebe rada analizirani su razni IC modeli, ali su se najbolje pokazali multimodalni LLM jer su u stanju da identifikuju i detaljno opišu više objekata nezavisno jedan od drugih, dok se ostali modeli mogu koristiti za jednostavne ključeve pretrage gdje je na slici prikazan jedan odjevni predmet.

3.2. *Analizirani modeli za generisanje opisa slika*

Modeli za IC koji su prethodili multimodalnim LLM nisu bili direktno namijenjeni za detaljno opisivanje svih predmeta koji su prikazani na slici, već je fokus bio na generisanju uopštenog opisa za cjelokupnu sliku. Ipak određeni modeli poput Florence-2 su u stanju da generišu relativno detaljne opise, koji obuhvataju većinu predmeta na slici, ali ni ti modeli ne

generišu struktuiran opis, gdje se jasno izdvajaju pojedinačni predmeti [10] Zbog toga su u radu stariji modeli isključivo analizirani, zajedno sa multimodalnim LLM, za problem razumijevanja konteksta gdje je kao ključ pretrage data majica, odnosno gdje je majica glavni predmet za koji se opis generiše.

Primjenom multimodalnih LLM moguće je opisati sve predmete na slici, jer se kao odgovor može generisati strukturirani izlaz, npr. u JSON formatu, gdje je moguće na lak način izdvojiti opise za pojedinačne odjevne predmete [11], [12]. Na slici 7 su prikazani opisi za neki ključ pretrage generisani pomoću **gpt-4o-mini-2024-07-18** modela.



Slika 7. *Primjer opisa generisanog pomoću multimodalnog LLM modela u JSON formatu*

U tabeli 1. su dati svi analizirani modeli.

Model	Primjena u radu	Opis
GenerativeImage2Text - GIT_BASE	Razumijevanje konteksta	Dotreniran za opisivanje majica i košulja
Florence-2 - Florence2-base		
BLIP-2 - BLIP-2 Opt. 2.7B		
gpt-4o-mini-2024-07-18	Svi navedeni problemi	Pretreniran komercijalni model
gpt-4o-2024-08-06		
claude-3-5-sonnet-20241022		
claude-3-haiku-20240307		
llama3.2-vision-11b		Pretreniran javno dostupan model

Tabela 1. *Analizirani modeli za generisanje opisa slika*

3.3. *Analizirane vektorske reprezentacije teksta*

Obrada prirodnog jezika (eng. *natural language processing* - NLP) je skup metoda koje imaju za cilj da ljudski (prirodni) jezik učine razumljivim računarima. Pomoću NLP-a i ML tehnika se rješavaju mnogi zadaci kao što je analiza sličnosti teksta, koju je potrebno sprovesti da bi se izvršila pretraga kataloga proizvoda na osnovu generisanih opisa proizvoda [13].

Tekst je vremenski nezavisan sadržaj sa linearnom strukturom, a u računarima je kodovan u binarnom obliku [14]. Kako bi se tekstualni sadržaj tumačio pomoću NLP metoda, neophodno je kreirati odgovarajuću vektorsku reprezentaciju. Što je reprezentacija bogatija informacijama, to se efikasnije može primijeniti u različitim NLP zadacima. Reprezentacije variraju od jednostavnih, koje su bazirane na frekvenciji riječi, do onih složenijih, baziranih na ugrađivanju riječi (eng. *word embeddings*) ili na kontekstualnim jezičkim modelima (eng. *contextual language models*) koje trenutno predstavljaju najnapredniji pristup. Pošto su vektori predloženih tekstualnih reprezentacija veće dimenzije nego vektori korištene vizuelne reprezentacije, vršena je redukcija dimenzionalnosti, prije računanja linearne kombinacije.

3.3.1. Reprezentacije bazirane na ugrađivanju riječi

Reprezentacije bazirane na ugrađivanju riječi omogućavaju kodiranje semantičkih i sintaksičkih odnosa između riječi. Ugrađivanje riječi je tehnika koja svakom pojmu u jeziku pridružuje gusto popunjeni vektor niske dimenzionalnosti, čime se omogućava zadržavanje informacija o sličnosti između riječi, kao i odnosa među njima. Cilj ovih metoda je smanjiti dimenzionalnost podataka uz očuvanje ključnih informacija o značenju i kontekstu. Tehnike ugrađivanja riječi bazirane su na principu da se slične riječi, s obzirom na njihovo značenje i kontekst upotrebe, moraju nalaziti blizu jedna druge u prostoru vektorskih reprezentacija. Zajednički nedostatak ovih reprezentacije je njihova nemogućnost da riješe problem polisemije, jer dodjeljuju isti vektor višeznačnim riječima bez obzira na kontekst [15].

3.3.1.1. Word2Vec

Word2Vec koristi neuronske mreže za učenje vektorskih reprezentacija riječi na osnovu njihovog konteksta u tekstualnim podacima. Postoje dvije osnovne arhitekture koje koristi Word2Vec za učenje ovih vektora *Continuous Bag of Words* (CBOW) i *Skip-Gram* (SG). CBOW model predviđa ciljnu riječ koristeći njene susjedne riječi. Nasuprot tome, SG koristi ciljnu riječ da bi predvidio njene susjedne riječi. Word2Vec se pokazao kao efikasan model za generisanje vektorskih reprezentacija riječi [16]. Međutim, Word2Vec ima i nekoliko značajnih nedostataka. Prvo, koristi lokalni kontekstni prozor, što znači da zanemaruje globalne korelacije između udaljenih riječi u tekstu, čime se gube važne semantičke informacije. Ovaj problem djelimično rješava GloVe [17], [18]. Drugo značajno ograničenje Word2Vec-a je njegova nemogućnost rada sa nepoznatim riječima, odnosno riječima koje nisu bile prisutne u trening skupu. Ovo je posebno problematično u kontekstu jezika sa bogatom morfologijom, kao što je srpski. Ovaj problem je efikasno riješen sa FastText [17], [19].

U ovom radu korišćen je Google-ov Word2Vec model[3], koji je treniran na dijelu skupa podataka iz *Google News* korpusa, ukupno obuhvatajući oko 100 milijardi riječi. Model sadrži vektorske reprezentacije za oko 3 miliona jedinstvenih riječi i fraza, koje su predstavljene kao vektori dimenzije 300.

3.3.1.2. FastText

FastText uvodi koncept podriječi (eng. subwords), što omogućava da se riječi predstavljaju kao skup slovnih n-grama, pružajući fleksibilniju i informativniju vektorsku reprezentaciju. Na taj način, vektorska reprezentacija riječi je dobijena kao zbir vektora njenih n-gramova. Na primjer, riječ “bašta” može biti podijeljena na n-gramove dužine 3, kao što su “<ba”, “baš”, “ašt”, “šta” i “ta>”. Ovdje su dodani specijalni simboli “<” i “>” kako bi se označio početak i kraj riječi, čime se omogućava modelu da razlikuje prefikse i sufikse od unutrašnjih dijelova riječi. Ovakav pristup omogućava FastText-u da efikasno generalizuje i na riječi koje nisu prisutne u trening skupu, jer se vektorska reprezentacija nepoznate riječi može formira-

[3] https://code.google.com/archive/p/word2vec/

ti na osnovu n-gramova koje dijeli sa poznatim riječima. FastText koristi istu arhitekturu kao Word2Vec, odnosno podržava i CBOW i SG. U CBOW varijanti, model koristi n-gramove podriječi susjednih riječi kako bi predvidio ciljnu riječ, dok u SG varijanti model koristi n-gramove podriječi ciljne riječi kako bi predvidio susjedne riječi [17], [19].

U ovom radu korišćen je FastText model treniran na skupu podataka Common Crawl[4], sa ukupno oko 600 milijardi riječi, koji sadrži dva miliona vektorskih reprezentacija riječi dimenzije 300.

3.3.1.3. GloVe

GloVe koristi matričku faktorizaciju i globalne informacije o distribuciji riječi u vidu matrice ko-pojavljivanja riječi, što rezultuje vektorskim reprezentacijama koje reflektuju kako lokalne, tako i globalne odnose u tekstu.GloVe koristi cijeli korpus kako bi izgradio svoju matricu ko-pojavljivanja. Ovaj globalni pristup omogućava da se uhvate odnose između riječi koje se možda nikada ne pojavljuju zajedno u istom lokalnom kontekstu, ali koje dijele slične obrasce ko-pojavljivanja sa drugim riječima. Na primjer, riječi poput "kralj" i "kraljica" mogu imati slične obrasce ko-pojavljivanja sa riječima poput "kruna", "prijestolje" ili "monarhija", iako se možda ne pojavljuju zajedno u istom rečenici. GloVe takođe omogućava da se u vektorskom prostoru modeluju linearni odnosi između riječi. Na primjer, poznata je sposobnost GloVe-a da uhvati analogije poput "kralj" : "kraljica" = "muškarac" : "žena". Ovi linearni odnosi proizlaze iz činjenice da su vektorske reprezentacije riječi u GloVe-u organizovane tako da razlike između vektora reflektuju semantičke razlike između riječi. Na primjer, vektor razlike između riječi "kralj" i "kraljica" može biti sličan vektoru razlike između riječi "muškarac" i "žena", što omogućava modelu da prepozna analogne odnose između različitih parova riječi [17], [18].

U ovom radu korišćen je GloVe model treniran na skupu podataka Wikipedia 2014 + Gigaword 5[5], koji sadrži ukupno oko šest milijardi riječi i 400 hiljada vektorskih reprezentacija riječi. Za potrebe rada korišćena je varijanta modela dimenzije 200.

3.3.2. Reprezentacije bazirane na kontekstualnim jezičkim modelima

Reprezentacije bazirane na kontekstualnim jezičkim modelima uzimaju u obzir kontekst u kojem se riječ pojavljuje, što znači da generišu različite vektorske reprezentacije za istu riječ u različitim kontekstima, čime se efikasno rješava problem polisemije i sinonimije (različite riječi sa sličnim značenjem). Ovi modeli koriste arhitekture zasnovane na dubokim neuronskim mrežama. Posebno se izdvajaju transformatorske arhitekture zbog svoje sposobnosti da analiziraju cijeli tekstualni kontekst. Njihova efikasnost proističe iz treniranja na velikim količinama podataka.

3.3.2.1. ELMo

ELMo uvodi koncept dinamičkih, kontekstualnih reprezentacija. Osnovna arhitektura ELMo modela zasniva se na dvosmjernim LSTM mrežama (eng. *bidirectional* LSTM - BiLSTM). BiLSTM omogućava modelu da obrađuje tekst u oba smjera. Ova dvosmjerna analiza omogućava modelu da uhvati bogatije kontekstualne informacije jer svaka riječ može biti interpretirana u odnosu na riječi koje joj prethode, ali i na riječi koje dolaze poslije nje. ELMo koristi slovne, odnosno znakovne CNN za generisanje inicijalnih vektorskih reprezentacija riječi na nivou znakova, čime se efikasno nosi sa nepoznatim i rijetkim riječima. ELMo koristi višeslojnu arhitekturu, gdje se vektorske reprezentacije riječi generišu iz različitih slojeva mreže. Svaki sloj modela uči različite nivoe jezičkih informacija – niži slojevi hvataju osnovne morfološke i sintaksičke informacije, dok viši slojevi obuhvataju složenije semantičke odnose. Konačna vektorska reprezentacija svake riječi u ELMo-u se dobija kao linearna kombinacija vektora iz svih slojeva modela. Ova kombinacija je ponderisana na način da se težine mogu prilagoditi specifičnom NLP zadatku, što omogućava fleksibilnost i prilagodljivost modela. ELMo je pokazao poboljšanja u performansama na širokom spektru NLP zadataka u odnosu na prijašnjem modele [20]. Međutim, iako je ELMo bio revolucionaran u trenutku svog predstavljanja, ubrzo su se pojavili još napredniji modeli zasnovani na transformatorskim arhitekturama, kao što je BERT, koji je imao bolju sposobnost da razumiju kontekstualne informacije.

U ovom radu korišten je ELMo model treniran na Wikipedia skupu podataka[6], koji sadrži ukupno oko pet milijardi riječi i vektore dimenzije 1024 [21].

3.3.2.2. BERT

BERT koristi transformator arhitekturu koja omogućava simultano uzimanje u obzir svih riječi u rečenici, bez obzira na njihov položaj. Ova sposobnost omogućava modelu da razumije kontekst svake riječi u odnosu na sve ostale riječi u tekstu, što ga čini posebno efikasnim za zadatke koji zahtijevaju duboko razumijevanje jezika. Osnovna arhitektura BERT-a zasniva se na mehanizmu pažnje (eng. *attention mechanism*), koji omogućava modelu da dinamički dodjeljuje različite težine različitim riječima u rečenici u zavisnosti od njihovog značaja za analizu konteksta. Ključni element ovog mehanizma je koncept samopažnje (eng. *self-attention*), gdje model izračunava relacije svake riječi sa svim ostalim riječima u rečenici, omogućavajući tako precizno hvatanje semantičkih i sintaktičkih odnosa. BERT generiše reprezentacije riječi koje zavise od njihovog konteksta u tekstu, pa se ista riječ može predstaviti različitim vektorima u zavisnosti od njenog značenja u datom okruženju [15], [22]. U ovom radu korišćen je BERT model treniran na skupu podataka *BooksCorpus* i English Wikipedia[7], ukupno obuhvatajući oko 3,3 milijarde riječi. Korišćena je varijanta modela *BERT-Base*, koja generiše vektorske reprezentacije dimenzije 768.

[4] https://fasttext.cc/docs/en/english-vectors.html

[5] https://nlp.stanford.edu/projects/glove/

[6] http://vectors.nlpl.eu/repository/

[7] https://huggingface.co/google-bert/bert-base-uncased

3.3.2.3. Modeli sa MTEB rangiranja

MTEB rangiranje (*Massive Text Embedding Benchmark leaderboard*) predstavlja rang-listu koja upoređuje performanse različitih modela za generisanje vektorskih reprezentacija teksta kroz skup standardizovanih zadataka, kao što su pretraga klasifikacija i klasterizacija [23]as various models are constantly being proposed without proper evaluation. To solve this problem, we introduce the Massive Text Embedding Benchmark (MTEB. U okviru ovog rada korišćeni su modeli koji su ostvarili dobre rezultate na MTEB rang-listi, posebno u zadacima pretrage. Primijenjeni su OpenAI komercijalni modeli *text-embedding-3-large* i *text-embedding-3-small*, bazirani na transformatorskoj arhitekturi, koji su se pokazali veoma efikasnim u raznim NLP zadacima, iako njihova tačna implementacija nije javno dostupna zbog komercijalne prirode proizvoda. Pored OpenAI modela, korišćene su i otvorene alternative koje takođe bilježe visoke performanse na MTEB rangiranju za zadatke pretrage: *snowflake-arctic-embed-l-v2.0*, *nomic-embed-text-v1.5* i *all-mpnet-base-v2*.

4. EKSPERIMENTALNI DIO

Za potrebe obučavanja modela za detekciju objekata i modela za analizu sličnosti slika korišten je DF2 skup podataka. U pitanju je skup podataka koji sadrži veliki broj slika odjevnih predmeta iz trinaest kategorija, od kojih su neke uslikane od strane potrošača, a druge su profesionalne slike iz prodavnica odjeće, na slici 8 su prikazane osnovne karakteristike DF2 trening skupa podataka, a slične su raspodjele i za validacioni i testni skup.

```
# Trening skup:
Ukupan broj slika: 191961
Ukupan broj odjevnih predmeta: 312186
Ukupan broj graničnih okvira: 312186
Broj graničnih okvira po kategoriji:
  long sleeve outwear: 13457
  trousers: 55387
  skirt: 30835
  short sleeve top: 71645
  vest dress: 17949
  shorts: 36616
  long sleeve top: 36064
  vest: 16095
  short sleeve dress: 17211
  sling dress: 6492
  long sleeve dress: 7907
  short sleeve outwear: 543
  sling: 1985
Ukupan broj parova prodavnica-potrošač: 14555
Ukupan broj odjevnih predmeta koji se ne mogu upariti: 94408
Raspodjela prema izvoru slike:
  potrošač: 59804
  prodavnica: 132157
```

Slika 8. *Osnovne karakteristike DF2 trening skupa podataka*

Slike su uparene, tako da slike na kojima su prikazani ista garderoba iz prodavnice i one uslikane od strane potrošača imaju isti identifikator *pair_id*, pri čemu pojedinačni predmeti imaju svoj identifikator *style* kako bi se pravilno uparili, a oni predmeti koji ne mogu da se upare imaju *style* vrijednost 0. Prilikom obučavanja kao sidra su birani samo oni odjevni predmeti koji se mogu upariti.

4.1. *Metrike za evaluaciju kvaliteta generisanih opisa slika*

Za evaluaciju generisanih opisa korištene su metrike: BLEU, METEOR, ROUGE-L, CIDEr, SPICE, SPIDER. Navedene metrike su odabrane jer su to najčešće korištene metrike u radovima koji se bave problemom generisanja opisa pomoću tehnika mašinskog učenja, kako bi bili uporedljivi sa rezultatima drugih istraživanja. BLEU mjeri sličnost između generisanog opisa i referentnog opisa koristeći n-gram preciznost i kaznu za dužinu. Definiše se na sljedeći način:

$$BLEU = BP * \exp\left(\sum_{n=1}^{N} w_n \cdot \ln p_n\right)$$

Gdje je BP, kazna za dužinu (eng. *brevity penalty*) koja osigurava da generisani opisi budu dovoljno dugi da prenesu cjelokupni smisao referentnog opisa, p_n je modifikovana *n*-gram preciznost, w_n su težine za svaki n-gram (obično je $w_n = \frac{1}{N}$), *c* je dužina generisanog opisa, a *r* je dužina referentnog opisa [25].

ROUGE-L koristi najdužu zajedničku podnisku (eng. *longest common subsequence* - LCS) između generisanog i referentnog opisa, te predstavlja mjeru baziranu na LCS. Definiše se na sljedeći način:

$$F_{\text{ROUGE-L}} = \frac{\left(1+\beta^2\right)\cdot R_{\text{LCS}} \cdot P_{\text{LCS}}}{R_{\text{LCS}} + \beta^2 \cdot P_{\text{LCS}}}$$

Gdje je

$$P_{\text{LCS}} = \frac{LCS(X,Y)}{\text{du⁓ina kandidatskog opisa}},$$

$$R_{\text{LCS}} = \frac{LCS(X,Y)}{\text{duzina referentnog opisa}},$$

LCS(X,Y) je dužina najduže zajedničke podniske između kandidatskog opisa X i referentnog opisa Y, a β je parametar koji balansira značaj preciznosti i odziva i često se uzima $\beta = \frac{P_{\text{LCS}}}{R_{\text{LCS}}}$ [26].

METEOR uzima u obzir unigram podudaranja, preciznost, odziv i kaznu za nepodudaranje reda riječi. METEOR skor se definiše kao:

$$\text{METEOR} = \text{F}_{\text{mean}}(1 - \text{K})$$

Gdje je $\text{F}_{\text{mean}} = \frac{10 \cdot P \cdot RP}{9 \cdot R}$, P je unigram preciznost:

$$P = \frac{\text{broj podudarnih unigrama}}{\text{ukupan broj unigrama u kanidatskom opisu}},$$

R je unigram odziv:

$$R = \frac{\text{broj podudarnih unigrama}}{\text{ukupan broj unigrama u referentnom opisu}},$$

K je kazna za nepodudaranje reda riječi:

$$\text{K} = 0{,}5\left(\frac{C}{M}\right)^3$$

Gdje je *M* broj podudaranja (eng. *matches*) i *C* broj segmenata (eng. *chunks*), odnosno grupa uzastopnih podudaranja [27].

CIDEr procjenjuje kvalitet opisa koristeći TF-IDF ponderisane n-grame i računa kosinusnu sličnost između kandidatskog opisa i skupa referentnih opisa. CIDEr se definiše kao:

$$\text{CIDEr} = \frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{N} w_n \cdot \text{sim}_n\left(c, r_m\right)$$

Gdje je *M* broj referentnih opisa, *N* maksimalna dužina n-grama (obično je $N = 4$), w_n težine za svaki n-gram (obično je $w_n = \frac{1}{N}$), $\text{sim}_n(c, r_m)$ kosinusna sličnost između TF-IDF vektora n-grama kandidatskog opisa *c* i referentnog opisa r_m za n-gram dužine *n* [28].

SPICE ocjenjuje semantički sadržaj opisa poredeći semantičke grafove izdvojene iz kandidatskog i referentnog opisa. SPICE metrika se definiše kao F1-mjera između semantičkih predikata:

$$\text{SPICE} = \frac{2 \cdot |S_c \cap S_r|}{|S_c| + |S_r|}$$

Gdje je S_c skup semantičkih predikata iz kandidatskog opisa, S_r skup semantičkih predikata iz referentnog opisa, a $|S_c \cap S_r|$ broj zajedničkih semantičkih predikata između kandidatskog i referentnog opisa [29]. SPIDEr je kombinovana metrika koja predstavlja prosječnu vrijednost CIDEr i SPICE metrika. SPIDEr se definiše kao [30]:

$$\text{SPIDEr} = \frac{\text{CIDEr} + \text{SPICE}}{2}.$$

Za potrebe dotreniravanja transformatorskih modela iskorišten je skup podataka sa opisima odjevnih predmeta formiran iz *Eureka-Attr* kataloga proizvoda, čije su karakteristike prikazane u tabeli 2.

	Kategorija	Broj odjevnih predmeta
Trening	*t-shirts*	100.000
	shirts	10.000
Validacioni	*t-shirts*	20.000
	shirts	2.000

Tabela 2. *Raspodjela Eureka-Attr skupa podataka sa opisima slika za odjevne predmete*

Za potrebe evaluacije generisanih opisa, detekcije slojevite odjeće, kao i za validiranje kvaliteta preporuka korištena su tri druga skupa podataka (disjunktni u odnosu na prethodne), a svi oni predstavljaju podskup *Eureka-PC* kataloga proizvoda. Osnovne karakteristike ovih skupova podataka su date u tabeli 3.

Kategorija	Broj odjevnih predmeta	Broj jednoznačnih predmeta	Problemi
t-shirts	100.000	80.870	Detekcija lažno pozitivnih graničnih okvira, (ne)razumijevanje konteksta
shirts	100.000	50.440	
layered	10.000	0	Problem sa detekcijom slojevite odjeće

Tabela 3. *Raspodjela podskupa Eureka-PC kataloga proizvoda korištenog za analizu rješenja za četiri vrste opisanih problema*

U tabeli 4. su dati rezultati evaluacije različitih modela korištenjem opisanih metrika, na skupu od 10k slika od čega je 5k iz kategorije *t-shirts* i 5k iz kategorije *shirts*, nasumično odabranih iz prethodno opisanog skupa.

Model	BLEU	METEOR	ROUGE-L F1	CIDEr	SPICE	SPIDEr
Dotreniran GIT_BASE	18.45	50.78	54.67	1.876	0.278	1.077
Dotreniran Florence2-base	20.12	53.34	57.23	2.101	0.301	1.201
Dotreniran BLIP-2 Opt. 2.7B	23.01	56.89	60.34	2.678	0.345	1.511
gpt-4o-mini-2024-07-18	34.57	78.84	79.19	4.997	0.578	2.788
gpt-4o-2024- 08-06	36.91	81.23	81.34	5.123	0.593	2.858
claude-3-5-sonnet-20241022	33.12	76.45	77.56	4.678	0.541	2.609
claude-3-haiku-20240307	31.45	73.67	74.89	4.321	0.512	2.416
llama3.2-vision-11b	29.78	70.89	72.34	4.012	0.478	2.245

Tabela 4. *Rezultati evaluacije analiziranih modela na Eureka-Attr skupu podataka*

Najbolje rezultate ostvaruju OpenAI ChatGPT modeli, ali su dobri i rezultati ostvareni od strane Antrophic Claude modela i LLaMa modela. Može se učiniti kako su dobijeni rezultati niži nego rezultati dobijeni na nekim drugim studijama, ali u pitanju je testiranje generisanih opisa na vrlo specifičnom domenu, pri čemu su svi referentni opisi detaljni, zbog čega su ocjene nešto niže nego kada bi se generisali uopšteniji opisi.

4.2. *Analiza rezultata za predloženo rješenje*

Rezultati evaluacije tri predložena modela za detekciju graničnih okvira i objekata, kao i pet predloženih LLM modela su dati u tabeli 5. Korišten je podskup proizvoda iz *Eureka-PC* kataloga, gdje je analizirano po pet hiljada proizvoda iz kategorije *t-shirts* i pet hiljada iz kategorije *shirts*, na svakoj slici je prikazan tačno jedan odjevni predmet, ali neki od odjevnih predmeta imaju na sebi naslikane ljudske likove.

Model	Prosječan broj detektovanih predmeta		Tačnost	
	t-shirts	*shirts*	*t-shirts*	*shirts*
GPT-4o	1.003	1.002	99.60%	99.98%
GPT-4o-mini	1.005	1.0076	99.58%	99.46%

Claude-3.5 Sonnet	1.003	1.0004	99.60%	99.60%
Claude-3 Haiku	1.007	1.0083	99.30%	99.30%
LlaMa-3.2	1.007	1.0013	99.30%	99.30%
YOLOv8	1.275	1.1712	77.90%	95.60%
YOLOv11	1.272	1.164	78.34%	95.60%
Mask R-CNN	1.28	1.1698	77.50%	77.50%

Tabela 5. *Rezultati evaluacije predloženih modela na problemu detekcije lažno pozitivnih uzoraka*

Najbolji rezultati su ostvareni korištenjem GPT-4o modela, ali i svi ostali predloženi modeli su ostvarili visoku tačnost na datom testnom skupu.

Rezultati evaluacije predloženih modela na problemu detekcije slojevite odjeće su dati u tabeli 6. Korišten je podskup iz *Eureka-PC* kataloga proizvoda, od hiljadu proizvoda, slike proizvoda prikazuju ljude koji nose slojevitu garderobu, gdje je ručno određen tačan broj predmeta na svakoj slici iz trinaest kategorija koje su obrađene u DF2 skupu podataka, dato ograničenje je uvedeno zbog poređenja sa modelima za detekciju koji su obučeni na DF2 skupu i samo mogu da detektuju navedene kategorije odjeće, dok u praksi LLM modeli mogu da detektuju i predmete iz drugih kategorija. Pravilna detekcija je slučaj kada neki model detektuje pravilan broj odjevnih predmeta, ali i pravilne kategorije za date predmete, dakle nije dovoljno samo pogoditi broj predmeta.

Model	**Tačnost**
GPT-4o	70.80%
GPT-4o-mini	71.00%
3.5 Sonnet	71.00%
3 Haiku	68.70%
LlaMa-3.2-V	67.10%
YOLOv8	35.40%
YOLOv11	36.40%
Mask R-CNN	41.40%

Tabela 6. *Rezultati evaluacije predloženih modela na problemu detekcije slojevite odjeće*

Najbolji rezultati su ostvareni pomoću GPT-4o modela, dok modeli za detekciju graničnih okvira znatno zaostaju za LLM modelima.

Za potrebe analize razumijevanja konteksta, kreiran je skup od 100 ključeva pretrage, pri čemu polovinu čine majice s printom, odnosno sa dodatnim kontekstom, a drugu polovinu obične majice bez printa. Skup je pažljivo izbalansiran kako bi rezultati analize bili validni. Naime, da je analiza vršena isključivo na ključevima pretrage s print majicama, dobiveni rezultati bi bili preprilagođeni specifičnom problemu (eng. *overfitted*). S druge strane, za eksperiment je korišten katalog proizvoda od 10.000 majica, formiran na osnovu *Eureka-PC* kataloga proizvoda, gdje je za svaki ključ pretrage određeno 20 najsličnijih majica. Poređeni su rezultati pretrage proizvoda u dva scenarija:

1. Korišćenje samo vizuelne reprezentacije slike.
2. Korišćenje linearna kombinacija vizuelne reprezentacije v_1 i tekstualne reprezentacije generisanog opisa v_2.

Eksperiment je uključivao poređenje dva pristupa: YOLOv11 + ResNet50 kao najboljeg predstavnika starog pristupa i ChatGPT-4o modela kao najboljeg predstavnika LLM modela. Za svaki ključ pretrage pronalazilo se dvadeset najsličnijih predmeta pomoću oba pristupa, a rezultati su upoređeni kako bi se utvrdilo koji model vraća više relevantnih predmeta. Kako bi se postigla optimalna linearna kombinacija vektora v_1 i v_2, težine w_1 i w_2 su određene tako da zadovoljavaju uslov $w_1 + w_2 = 1$. Vrijednosti težina su ispitivane unutar intervala [0,1] s korakom EPS = 0.01. Na taj način, za svaku kombinaciju težina testirani su rezultati kako bi se pronašao optimalan omjer. Eksperiment je pokazao da se bolji rezultati postižu korištenjem linearne kombinacije obje vektorske reprezentacije (v_1 i v_2) u odnosu na korištenje samo jedne od njih. U tabeli 7. su prikazani rezultati poređenja novog pristupa sa starim za različite modele vektorske reprezentacije opisa.

Model	**% Bolji rezultati starog modela**	**% Isti**	**% Bolji rezultati novog modela**	**Optimalan omjer $w_1 : w_2$**
text-embedding-3-small	15	10	75	14 : 86
Word2Vec	12	26	62	52 : 48
FastText	9	27	64	41 : 59
GloVe	11	26	63	398 : 62
ELMo	8	27	65	43 : 57
BERT	14	16	70	34 : 66
snowflake-arctic-embed-l-v2.0	8	10	82	27 : 73
text-embedding-3-large	10	13	77	20 : 80
nomic-embed-text-v1.5	18	20	62	27 : 73
all-mpnet-base-v2	5	28	67	22 : 78

Tabela 7. *Rezultati evaluacije za problem razumijevanje konteksta nad majicama*

Najbolje performanse postignute su s modelom *snowflake-arctic-embed-l-v2.0*, pri čemu je optimalan omjer težina bio $w_1 : w_2 = 27 : 73$.

Analizirana je i primjena na problemu razumijevanja konteksta i nedovoljne invarijantnosti kad su ključevi pretrage košulje. Za potrebe analize, kreiran je skup od 50 ključeva pretrage, od čega polovinu čine savijene košulje, a drugu polovinu košulje na ljudima. Cilj je bio ispitati performanse različitih pristupa u pronalaženju sličnih predmeta unutar kataloga koji sadrži 10.000 predmeta, takođe formiranog na osnovu *Eureka-PC* kataloga proizvoda. Važno je napomenuti da ovaj katalog nije isti kao u prethodnom primjeru za majice, ali postoji određeno preklapanje kada su u pitanju nerelevantni predmeti. Za svaki ključ pretrage identifikovano je deset najsličnijih predmeta. U analizi su upoređena dva pristupa:

1. Korišćenje samo vizuelne reprezentacije slike.
2. Korišćenje linearna kombinacija vizuelne reprezentacije v_1 i tekstualne reprezentacije generisanog opisa v_2.

Testirani su sljedeći modeli: YOLOv11+ResNet50 i ChatGPT-4o. Jedan od izazova u analizi je bio taj što detekcija graničnih okvira kod starog pristupa, nije uvijek dobro funkcionisala na savijenim košuljama. Uzrok ovog problema vjerovatno leži u nedostatku primjera sa savijenim košuljama u trening skupu podataka, zbog čega detekcija nije uvijek moguća, ili se loše odredi granični okvir. Iz tog razloga je, prilikom analize u slučaju da nije došlo do detekcije graničnog okvira, korištena cijela ulazna slika bez dodatnog isjecanja. Na ovaj način osigurano je da se ne donose zaključci isključivo na osnovu problema sa detekcijom graničnih okvira, već da se analizira i širi kontekst performansi modela. Kao i u prethodnom primjeru, korišćenje linearne kombinacije vizuelne reprezentacije v_1 i reprezentacije generisanog teksta v_2 dalo je bolje rezultate u poređenju sa korišćenjem samo vizuelne reprezentacije. Optimalna linearna kombinacija težina w_1 i w_2 određena je tako da zadovoljava uslov $w_1 + w_2 = 1$, pri čemu su težine ispitivane u intervalu [0,1] s korakom EPS = 0.01. U tabeli 8. su prikazani rezultati poređenja novog pristupa sa starim za različite modele vektorske reprezentacije opisa.

Model	**% Bolji rezultati starog modela**	**% Isti**	**% Bolji rezultati novog modela**	**Optimalan omjer $w_1 : w_2$**
text-embedding-3-small	20	10	70	73 : 27
Word2Vec	16	26	58	86 : 14
FastText	12	26	62	82 : 18
GloVe	14	26	60	83 : 17
ELMo	14	24	62	84 : 16
BERT	14	22	64	79 : 21
snowflake-arctic-embed-l-v2.0	18	12	70	72 : 28
text-embedding-3-large	16	14	70	70 : 30
nomic-embed-text-v1.5	24	20	56	69 : 31
all-mpnet-base-v2	22	18	60	79 : 21

Tabela 8. *Rezultati evaluacije za problem razumijevanje konteksta i nedovoljne invarijantnosti nad košuljama*

Najbolji rezultati postignuti su s modelom *text-embedding-3-large*, pri čemu je optimalan omjer bio $w_1 : w_2 = 70 : 30$. Međutim, analiza je pokazala zanimljiv obrazac:

- Kod savijenih košulja, kombinacija vizuelne i tekstualne reprezentacije značajno je nadmašila pristupe bazirane samo na vizuelnim modelima. Ovo je očekivano jer tekstualni opisi omogućavaju bolji kontekstualni uvid u karakteristike savijenih košulja.
- Kod košulja na ljudima, stariji pristup (YOLO+ResNet50) pokazao se boljim u određenim slučajevima. Razlog za to je što je teško riječima precizno opisati šaru (eng. *pattern*) na košulji, dok vizuelni modeli poput ResNet50 bolje prepoznaju te suptilne vizuelne detalje. Zbog toga je omjer takav da mnogo više ide u korist vizuelne reprezentacije.

Konačno, analizirana je i primjena modela na problemu pronalaženja povezanih proizvoda, što predstavlja specifičan slučaj u kojem nije cilj pronaći slične artikle u klasičnom smislu, već identifikovati proizvode koji su konceptualno komplementarni s obzirom na ključ pretrage. Za potrebe eksperimenta, kreiran je skup od 50 ključeva pretrage, pri čemu su svi ključevi predstavljali slike dukseva. Nasuprot tome, katalog proizvoda je sadržavao 5.000 slika donjih dijelova trenerki, formiranih na osnovu Eureka-PC kataloga proizvoda. Za svaki ključ pretrage ručno je određeno tačno jedno komplementarno odgovarajuće uparivanje iz kataloga, koje je služilo kao referenca za evaluaciju.

U skladu s prethodnim eksperimentima, i ovde su upoređena dva pristupa: korišćenje isključivo vizuelne reprezentacije slike i kombinovana reprezentacija slike i generisanog teksta. Testirani su sljedeći modeli: YOLOv11+ResNet50 i ChatGPT-4o. U ovom eksperimentu uzeti su u obzir prva dva rezultata (top-2) iz povratnog skupa. Na ovaj način, omogućena je veća tolerancija u evaluaciji, s obzirom na to da povezani proizvodi mogu imati varijabilniju vizuelnu i deskriptivnu reprezentaciju, što otežava tačno pozicioniranje u top-1 rezultatu. Kao i ranije optimalna linearna kombinacija težina w_1 i w_2 određena je tako da zadovoljava uslov $w_1 + w_2 = 1$, pri čemu su težine ispitivane u intervalu [0,1] s korakom EPS = 0.01. U tabeli 9. su prikazani rezultati poređenja novog pristupa sa starim za tri najbolja modela vektorske reprezentacije teksta iz prethodnih eksperimenata.

Model	**% Bolji rezultati starog modela**	**% Isti**	**% Bolji rezultati novog modela**	**Optimalan omjer $w_1 : w_2$**
text-embedding-3-small	0	32	68	52 : 48
snowflake-arctic-embed-l-v2.0	0	32	68	49 : 51
text-embedding-3-large	0	26	74	41: 59

Tabela 9. *Rezultati evaluacije za problem razumijevanje konteksta pri pronalasku povezanih proizvoda*

Najbolji rezultati postignuti su s modelom *text-embedding-3-large*, pri čemu je optimalan omjer bio $w_1 : w_2 = 41 : 59$. Rezultati su u skladu sa očekivanjima jer stari pristup nema dobre mogućnosti pronalaska povezanih proizvoda zbog nerazumijevanja konteksta.

5. PRAVCI DALJEG ISTRAŽIVANJA

Mogući pravci daljeg istraživanja obuhvataju dotreniravanje LLM-ova za dati specifični problem generisanja opisa za pretragu proizvoda, kao i primjena temeljnog *prompt engineering*-a u cilju dobijanja što boljih opisa koji bi detaljno opisali sve karakteristike odjevnih predmeta, kao i testiranje sve većeg broja dostupnih LLM-ova. Takođe od interesa je i primjena generisanih opisa za kvalitetniju kategorizaciju i razvrstavanje nekog kataloga proizvoda, npr. prema materijalima, dužini, okolnostima kada se neka garderoba nosi, sezonalnosti, trendu kojem pripada itd.

6. ZAKLJUČAK

U ovom radu opisan je prijedlog rješenja za primjenu opisa generisanih pomoću tehnika mašinskog učenja za problem pretrage proizvoda. Ovim radom pokazano je da je za uspješniju pretragu bolje kombinovati model za analizu sličnosti slika zajedno sa modelom za opisivanje slika, odnosno linearno kombinovanje vektorske reprezentacije dobijene na osnovu vizuel-

nih karakteristika sa vektorskom reprezentacijom generisanog opisa. Posebna pažnja je posvećena problemima sa klasičnim pristupom za pretragu proizvoda koji se oslanja samo na analizu sličnost slika, gdje problemi potiču ili od modela za detekciju koji se nalazi na početku takvog sistema, ili od modela za analizu sličnosti koji se nalazi na kraju tog sistema.

LITERATURA

[1] D.-C. Pahonțu i E.-Ștefania Enache, „An Overview of AI-driven Recommendation Systems: Enhancing Personalization & User Experience (Qualitative Study)“.

[2] D. G. Balasubramanian, „THE ROLE OF AI IN ENHANCING PERSONALIZATION IN ECOMMERCE: A STUDY ON CUSTOMER ENGAGEMENT AND SATISFACTION“, 2024.

[3] „Assessing the role of artificial intelligence in enhancing customer personalization: A study of ethical and privacy implications in digital marketing“. Pristupljeno: 14. prosinac 2024. [Na internetu]. Dostupno na: https://www.researchgate.net/publication/386078486_Assessing_the_role_of_artificial_intelligence_in_enhancing_customer_personalization_A_study_of_ethical_and_privacy_implications_in_digital_marketing

[4] Y. Ge, R. Zhang, L. Wu, X. Wang, X. Tang, i P. Luo, „DeepFashion2: A Versatile Benchmark for Detection, Pose Estimation, Segmentation and Re-Identification of Clothing Images“, 23. siječanj 2019., *arXiv*: arXiv:1901.07973. doi: 10.48550/arXiv.1901.07973.

[5] D. Morelli, M. Cornia, i R. Cucchiara, „FashionSearch++: Improving Consumer-to-Shop Clothes Retrieval with Hard Negatives“.

[6] P. Alirezazadeh, F. Dornaika, i A. Moujahid, „Deep Learning with Discriminative Margin Loss for Cross-Domain Consumer-to-Shop Clothes Retrieval“, *Sensors*, sv. 22, izd. 7, Art. izd. 7, sij. 2022, doi: 10.3390/s22072660.

[7] J. Redmon, S. Divvala, R. Girshick, i A. Farhadi, „You Only Look Once: Unified, Real-Time Object Detection“, 09. svibanj 2016., *arXiv*: arXiv:1506.02640. doi: 10.48550/arXiv.1506.02640.

[8] K. He, G. Gkioxari, P. Dollár, i R. Girshick, „Mask R-CNN“, 24. siječanj 2018., *arXiv*: arXiv:1703.06870. doi: 10.48550/arXiv.1703.06870.

[9] D. Hoiem, Y. Chodpathumwan, i Q. Dai, „Diagnosing Error in Object Detectors“, u *Computer Vision – ECCV 2012*, sv. 7574, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, i C. Schmid, Ur., u Lecture Notes in Computer Science, vol. 7574. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, str. 340–353. doi: 10.1007/978-3-642-33712-3_25.

[10] B. Xiao *i ostali*, „Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks“, 10. studeni 2023., *arXiv*: arXiv:2311.06242. doi: 10.48550/arXiv.2311.06242.

[11] „Introducing Structured Outputs in the API“. Pristupljeno: 12. siječanj 2025. [Na internetu]. Dostupno na: https://openai.com/index/introducing-structured-outputs-in-the-api/

[12] „Increase output consistency (JSON mode)“, Anthropic. Pristupljeno: 12. siječanj 2025. [Na internetu]. Dostupno na: https://docs.anthropic.com/en/docs/test-and-evaluate/strengthen-guardrails/increase-consistency

[13] J. Eisenstein, „Natural Language Processing“.

[14] Risojević, Vladimir, *Multimedijalni Sistemi*. Banja Luka: Univerzitet u Banjoj Luci Elektrotehnički fakultet, 2018.

[15] „Transformer: A Novel Neural Network Architecture for Language Understanding“. Pristupljeno: 19. studeni 2024. [Na internetu]. Dostupno na: http://research.google/blog/transformer-a-novel-neural-network-architecture-for-language-understanding/

[16] T. Mikolov, K. Chen, G. Corrado, i J. Dean, „Efficient Estimation of Word Representations in Vector Space“, 07. rujan 2013., *arXiv*: arXiv:1301.3781. Pristupljeno: 04. studeni 2024. [Na internetu]. Dostupno na: http://arxiv.org/abs/1301.3781

[17] A. CR, „Word Embeddings in NLP | Word2Vec | GloVe | fastText“, Analytics Vidhya. Pristupljeno: 04. studeni 2024. [Na internetu]. Dostupno na: https://medium.com/analytics-vidhya/word-embeddings-in-nlp-word2vec-glove-fasttext-24d4d4286a73

[18] J. Pennington, R. Socher, i C. Manning, „GloVe: Global Vectors for Word Representation“, u *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, i W. Daelemans, Ur., Doha, Qatar: Association for Computational Linguistics, lis. 2014, str. 1532–1543. doi: 10.3115/v1/D14-1162.

[19] A. Joulin, E. Grave, P. Bojanowski, i T. Mikolov, „Bag of Tricks for Efficient Text Classification“, 09. kolovoz 2016., *arXiv*: arXiv:1607.01759. Pristupljeno: 11. studeni 2024. [Na internetu]. Dostupno na: http://arxiv.org/abs/1607.01759

[20] M. E. Peters *i ostali*, „Deep contextualized word representations“, 22. ožujak 2018., *arXiv*: arXiv:1802.05365. doi: 10.48550/arXiv.1802.05365.

[21] M. Fares, A. Kutuzov, S. Oepen, i E. Velldal, „Word vectors, reuse, and replicability: Towards a community repository of large-text resources“, u *Proceedings of the 21st nordic conference on computational linguistics*, Gothenburg, Sweden: Association for Computational Linguistics, svi. 2017, str. 271–276. [Na internetu]. Dostupno na: http://www.aclweb.org/anthology/W17-0237

[22] J. Devlin, M.-W. Chang, K. Lee, i K. Toutanova, „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding“, 24. svibanj 2019., *arXiv*: arXiv:1810.04805. doi: 10.48550/arXiv.1810.04805.

[23] N. Muennighoff, N. Tazi, L. Magne, i N. Reimers, „MTEB: Massive Text Embedding Benchmark“, 19. ožujak 2023., *arXiv*: arXiv:2210.07316. doi: 10.48550/arXiv.2210.07316.

[24] „OpenAI Platform - emb dok“. Pristupljeno: 12. siječanj 2025. [Na internetu]. Dostupno na: https://platform.openai.com

[25] K. Papineni, S. Roukos, T. Ward, i W.-J. Zhu, „Bleu: a Method for Automatic Evaluation of Machine Translation“, u *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, P. Isabelle, E. Charniak, i D. Lin, Ur., Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, srp. 2002, str. 311–318. doi: 10.3115/1073083.1073135.

[26] C.-Y. Lin, „ROUGE: A Package for Automatic Evaluation of Summaries“, u *Text Summarization Branches Out*, Barcelona, Spain: Association for Computational Linguistics, srp. 2004, str. 74–81. Pristupljeno: 09. lipanj 2024. [Na internetu]. Dostupno na: https://aclanthology.org/W04-1013

[27] S. Banerjee i A. Lavie, „METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments“, u *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, J. Goldstein, A. Lavie, C.-Y. Lin, i C. Voss, Ur., Ann Arbor, Michigan: Association for Computational Linguistics, lip. 2005, str. 65–72. Pristupljeno: 16. lipanj 2024. [Na internetu]. Dostupno na: https://aclanthology.org/W05-0909

[28] R. Vedantam, C. L. Zitnick, i D. Parikh, „CIDEr: Consensus-based Image Description Evaluation“, 03. lipanj 2015., *arXiv*: arXiv:1411.5726. doi: 10.48550/arXiv.1411.5726.

[29] P. Anderson, B. Fernando, M. Johnson, i S. Gould, „SPICE: Semantic Propositional Image Caption Evaluation“, 29. srpanj 2016., *arXiv*: arXiv:1607.08822. doi: 10.48550/arXiv.1607.08822.

[30] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, i K. Murphy, „Improved Image Captioning via Policy Gradient optimization of SPIDEr“, u *2017 IEEE International Conference on Computer Vision (ICCV)*, lis. 2017, str. 873–881. doi: 10.1109/ICCV.2017.100.

Aleksije Mićić, dipl. ing Bravo Systems d.o.o. Banja Luka, RS, BiH / Elektrotehnički fakultet, Univerzitet u Banjoj Luci, RS, BiH
Kontakt: micic@oddbytes.com
Oblasti interesovanja: mašinsko učenje, data science, sistemi za preporuke, objektno-orijentisano programiranje i dizajn, informacioni sistemi, web servisi

prof. dr Zoran Đurić, Elektrotehnički fakultet, Univerzitet u Banjoj Luci, RS, BiH
Kontakt: zoran.djuric@etf.unibl.org
Oblasti interesovanja: sigurnost, kriptografija, PKI, platni sistemi i protokoli, formalna verifikacija, mašinsko učenje, data science, objektno-orijentisano programiranje i modelovanje, Internet programiranje, razvoj mobilnih aplikacija, XML-bazirana međuoperativnost, web servisi, računarske mreže, penetration testing, sistem integracija