SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Primjena postupka optičkog prepoznavanja znakova i velikih jezičnih modela na staru arhivsku dokumentaciju

Lovro Magdić
Voditelj: *doc. dr. sc. Juraj Petrović*

Zagreb, svibanj 2025.

SADRŽAJ

PRIMJERI

POPIS TABLICA

1. Uvod

U digitalnom dobu, arhive sadrže neprocjenjive povijesne podatke, no njihova obrada često je otežana zbog rukom pisanih ili pisaćom mašinom ispisanih dokumenata. Za efikasnu digitalizaciju i pretraživanje takvih materijala, ključna je primjena metoda optičkog prepoznavanja znakova (OCR), koje pretvaraju tekst sa slika u strojno čitljiv oblik.

Ovaj seminarski rad se fokusira na primjenu velikih jezičnih modela (engl. Large Language Models, LLMs) za automatsku korekciju pogrešno prepoznatih riječi i prilagodbu teksta hrvatskom jeziku. Koristit će se unaprijed trenirani modeli prilagođeni hrvatskom, slovenskom i engleskom, kao i fino podešavanje modela ByT5 na sintetički generiranim podacima. Cilj je razviti sustav koji digitalizira arhivske dokumente i poboljšava kvalitetu prepoznatog teksta kroz naprednu semantičku analizu. Ovakav pristup omogućava lakši pristup, pretraživanje i analiza vrijednih povijesnih informacija.

2. Opis problema

Ovaj seminarski rada krenuo je kao ideja poboljšanja završnog rada [3] potaknuta znanstvenim radovima koji se bave sličnom problematikom.[2][5]. Zadatak ovog seminarskog rada je provesti poboljšanje rezultate postupka optičkog prepoznavanja znakova na staroj arhivskoj dokumentaciji korištenjem velikih jezičnih modela u svrhu postizanja poboljšanja izdvojenog teksta.

Dokumentacija sadrži 3052 slika skeniranih dokumenata na crnoj pozadini. U većini dokumentaciju čine oštećeni i poderani papiri, različitih rotacija, razina osvjetljenja i dimenzija te ne cjelovitih dijelova teksta. Tekst dokumenata je pisano hrvatsko-srpskim jezikom na pisaćoj mašini. Većinski kroz cjelovitu dokumentaciju prisutne su ne konzistencije razine otiska pojedinih znakova i problematične metode ispravka zatipaka. Zatipci su ispravljeni na način da se preko pogrešno otisnutog znaka otiskuje ispravan znak i time smanjuje konačna interpretacija znaka.

Navedeni problemi su obrađeni i većinski ispravljeni u završnom radu [3] te se u sklopu ovog seminarskog rada koristi krajnji rezultat navedenog završnog rada što su tekstualne datoteke izdvojenog teksta. Na navedenim tekstualnim datotekama provodi se postupak poboljšanja uz primjenu predtreniranih i fino podešenih velikih jezičnih modela. Model je fino podešavan na sintetičko generiranom skupu podataka parova neispravnih i ispravnih riječi na hrvatskom jeziku. Kroz navedeni postupak model je naučio predvidjeti ispravnu riječ na ulaz riječi s jednom ili više pogrešaka.

3. Korištene tehnologije

3.1. Razvojno okruženje i jezik

Ovaj seminarski rad razvijen je u razvojnom okruženju "Visual Studio Code" u programskom jeziku "Python" verziji 3.11.1. Tijekom razvoja rada korištene su razvojne biblioteke: PyTorch i Hugging Face.

PyTorch je razvojna biblioteka dubokog učenja korištena za dizajniranje i treniranje neuronskih mreža. Koristi se u širokom području istraživačkog i razvojnog okruženja raznih primjena strojnog učenja. Hugging Face je razvojna biblioteka dizajnirana za rad s vrhunskim velikim jezičnim modelima(BERT, GPT i T5). Najčešće korištena u raznim obradama prirodnog jezika kao klasifikacija, prevođenje i rezimiranju teksta. Biblioteka omogućuje alate za tokeniziranje i fino podešavanje koji su ključni za obradu prirodnog jezika.

4. Opis rada sustava

Cjevovod sustava čine pripremljeni ulazni podatci koji se sastoje od tekstualnih datoteka dobivenih provedbom optičkog prepoznavanja podataka i tekstualnih datoteka koje predstavljaju temeljnu istinu koje se koriste pri analizi točnosti. Ulazni podatci su dobiveni kao rezultat završnog rada [3]. Navedeni ulazni podatci prolaze kroz dodatna dva cjevovoda obrade u kojima se koriste predtrenirani model ByT5 koji je fino podešen na 5000, 10000, 65000 i 100000 primjera. Također se koristi predtrenirani model crosloengual. Kombinacijom ova dva modela postiže se krajnji rezultat sustava koji u pravilu predstavlja ispravljen, interpretabilan i točan primjer ulaza.

4.1. Ulazni podaci

Podatci korišteni u ovom seminarskom radu su rezultat završnog rada [3] u kojem je obrađena stara arhivska dokumentacija i provedeno je optičko prepoznavanje znakova. Ulazni podatci modela korištenih u seminarskom radu su rečenice sadržane u tekstualnim datotekama koje predstavljaju tekst optički prepoznat iz skeniranih dokumenata stare arhivske dokumentacije. Rečenice su pisana na hrvatskom-srpskom jeziku s učestalim znakovnim pogreškama u riječima što je uzrok primjene optičkog prepoznavanja znakova.

Primjer 4.1: Primjer tekstualne datoteke dobivene primjenom metode optičkog prepoznavanja znakova

```
tae a SE’s
fe eh
, SUP ZACRED STHOCO POVIERL TIVO?
UPRAVA DIZAVNE HEZIJZEDNOSTI Samo aa litre informacijy Wye >
 TL Odjetjenjo . Odtajek ~~
: TI  7 - Prawn yo UNE ES
Geo canes eet Srennnnnes Ot * asvrsen YN, 8, -
Satta UM Fapured is
i . Abert
Akos jo "Nin"
SPECIJALNA INFORMACHIA
Boop: 7.
PREDMET: Saradnik Gustav otputo~
Yeo ponovne na red uv Mn~
chen. =
FRIMJEDDA
Veze spec. inf. br.258 of 26.XII 1 br.262 Od 30.XII
1966. godine __
Gustev" je u daljojem razgovoru Gao infor-
maciju da je u Minchenu povezen sa HiO-om 1 da je
Sesto u druStvu 98 Prpi Wikoloa neki Bruco, rodom
iz Zegrebe, Gustav kate da je za Bracu uapio saz~
nati da Je stari exigront, star oko Jo godina, ca
Je 2@ vrijone rate imeo svoju gostiony ispod Sije -
mena 1 de stalno #ivi u MOnchenu, adje ime evoju
eradjevinsku radnju. Zotio da isa svoj sutomobil 1
da avude vozi Frpie, jer da Pprpi nema automobila,
Ze Jeli Dranke Quetev" kete da o@ u zed -
nje vrijeme vozi u crnom Mercedesu", 8 da je rani~
Je dolazio u Augsburg u Volkawagenu.
 pripremags emigrecije za izvrienje atente
| ta na Predsjednike Republike prilikos posjete Aust-
. P4ji, koje ee trebala obaviti u proljece 1966. godi
| ne, Guetev kete do je o tome razgoverso aa Prpi ~
| em 1 Cori Ivenom u III mjesecu ove godine i da je
+ tom prilikom Prpi izjevio de je nakon objave posje
| ta, u SR Wjomatku ubaten iz renlje, veliki broj cr-
venih Spijuna, kojima da je uspjelo saznati nemsere
i emigracije, te da oe xbog toga morao odgoditi pos -
(| @jet.
```

4.2. Model ByT5

Model ByT5[6] je varijanta T5 modela kojeg razvija Google i oslanja se na procesiranje neobrađenog teksta na razini UTF-8 bajtova. Ne koristi tokeniziranje kao slični modeli već se oslanja na usporedbu bajtova što ga čini neovisnim o jeziku odnosno nije vezan za jezična pravila i vokabular. Primjer obrade na razini bajtova za riječ ("računalo") bi se sastojao od rastavljanja riječi na UTF-8 enkodirane znakove (["r","a","č","u","n","a","l","o"]) odnosno u heksadekadski oblik (["72","61","c4 8d","8d","6e","61","6c","6f"]). Navedeni postupak se također koristi i tijekom finog podešavanje modela. U ovom seminarskom radu model ByT5 je fino podešen na 5000, 10000, 65000 i 100000 sintetičkih primjera pogrešaka optičkog prepoznavanja znakova. Primjeri su zapisani u obliku ".jsonl" datoteke u kojoj svaki red predstavlja jedan primjer, oznaka `input_text` označava ulaz, a `target_text` ispravan izlaz. Kreirani su tako da su izdvojene rečenice iz skupa podataka starih novinskih članaka[1] na hrvatskom jeziku te su rečenice potom razdvojene na riječi. Za svaku izdvojenu riječ uvodimo od jedne do nekoliko zamjena znakova.

Primjer 4.2: Primjer podataka korištenih za fino podešavanje modela

```
{"input_text": "Dubroanuk,", "target_text": "Dubrovnik,"}
{"input_text": "bikc", "target_text": "Kiki"}
{"input_text": "obnce", "target_text": "dance"}
{"input_text": "plvsatice", "target_text": "plesačice"}
{"input_text": "Žućj", "target_text": "Župe"}
{"input_text": "Čučrovačke,", "target_text": "dubrovačke,"}
{"input_text": "pljsga", "target_text": "plesna"}
{"input_text": "ggŠpa", "target_text": "grupa"}
{"input_text": "Ćiwmička", "target_text": "Ritmička"}
{"input_text": "greva", "target_text": "grupa"}
{"input_text": "brš", "target_text": "još"}
{"input_text": "nemih", "target_text": "nekih"}
{"input_text": "sših", "target_text": "Svih"}
{"input_text": "zabavlČatm", "target_text": "zabavljati"}
{"input_text": "Vjevđzica,", "target_text": "Vjeverica,"}
```

4.3. Model CroSloEngual BERT

CroSloEngual BERT[4] je višejezični model baziran na BERT arhitekturi. Sastoji se od 12 blokova transformera i svaki blok obrađuje podatke sekvencijalno. Model ima

110 milijuna parametara što uključuje težine i pristranost u unaprijednim mrežama i slojevima pažnje. Navedeni parametri su trenirani na velikom tekstovnom korpusu: Wikipedia, OpenSubtitles i news corpora; na hrvatskom, slovenskom i engleskom jeziku. Model omogućuje NER(Named Entity Recognition), klasifikaciju, semantičku analizu teksta i POS označavanje (Part Of Speech tagging). Razvijen je u sklopu projekta EMBEDDIA s ciljem podržavanja europskih jezika s malo dostupnih resursa.

4.4. Korištene metode

U obradi ulaznih podataka koristimo dvije metode obrade preciznije dva različita cjevovoda podataka. Metode se razlikuju u usporedbi brojčanih vrijednosti razlike dva znakovna niza odnosno kada je razlika prihvatljiva i kada nije. Metode rade na način da ulaznu datoteku dijele na rečenice što je ekvivalentno jednom redu odnosno do znaka novog reda. Svaka riječ te rečenice prolazi kroz model crosloengual i uspoređuje se s izlazom modela u kontekstu sličnosti korištenjem metrike levenshtein-ove udaljenosti(engl. levenshtein distance). Rezultat metrike je brojčana vrijednost koja označava razliku dva znakovna niza.

$$D(i,j) = \begin{cases} \max(i,j) & \text{ako } \min(i,j) = 0, \\ \min \begin{cases} D(i-1,j)+1, \\ D(i,j-1)+1, \\ D(i-1,j-1)+\delta(s_i,t_j) \end{cases} & \text{inače,} \end{cases}$$

Prva metoda uspoređuje ako je razlike između dvije riječi veća od razlike duljine ulazne riječi i polovine duljine ulazne riječi. Onda se ulazna riječ ne zamjenjuje procijenjenom riječi modela odnosno druga metoda ako je razlika dvije riječi veća od polovine duljine ulazne riječi. Usporedba ovih dvaju brojčanih vrijednosti je rezultat metoda pokušaja i pogrešaka dok nije postignuto očekivano ponašanje. U obje metode ako je Levenshtein-ova udaljenost veća smatramo da procjena modela nema kontekstualnog smisla s ostatkom rečenice. Ovo je najčešće slučaj za riječi koje predstavljaju imena osoba, gradove i slično. Ulazna riječ koja nije zamijenjena procijenjenom riječju modela ostaje dio rečenice inače ulazna riječ zamijenjena je procijenjenom riječju i koristi se dalje u procjeni ostalih riječi te rečenice. Nakon prolaska kroz model crosloengual rečenica prolazi model ByT5 koji za svaku riječ procjenjuje pravopisno točnu riječ i time postižemo rečenice u kojima su ispravljene kontekstualne pogreške modelom crosloengual i pravopisne pogreške modelom ByT5.

Primjer 4.3: Primjer programskog koda logike prve metode

```
# corrector == model crosloengual
# model == model ByT5
for sentence in file:
    sentence=sentence.replace("\n", "")
    sentence_list=sentence.split(" ")
    for word in sentence_list:
        wrong_word=word
        result=corrector.correct(sentence=sentence,
        wrong_word=wrong_word, topk=5)
        levenshtein=Levenshtein.distance(result, wrong_word)
        if levenshtein>(len(wrong_word)-(len(wrong_word)/2)):
            corrected_wrong_word=model.correct(wrong_word)
            corrected_sentence.append(corrected_wrong_word)
        else:
            corrected_sentence.append(result)
```

Primjer 4.4: Primjer programskog koda logike druge metode

```
# corrector == model crosloengual
# model == model ByT5
for word in words:
    try:
        prediction=corrector.correct(sentence=sentence,
        wrong_word=word, topk=5)
    except ValueError:
        prediction=word
    if Levenshtein.distance(word, prediction)>len(word)//2:
        corrected_word=model.correct(word)
        corrected_words.append(corrected_word)
    else:
        corrected_words.append(prediction)
corrected_sentence=" ".join(corrected_words)
corrected_lines.append(corrected_sentence)
```

5. Rezultati

Model ByT5 fino je podešen na 5000, 10000, 65000 i 100000 primjera, a svaki od modela nazvan je redom: ByT5-5k, ByT5-small, ByT5-v2 i ByT5-v3. Model crosloengual isti je u svim iteracijama rezultata. Rezultati su prikazani decimalnim brojevima od 0 do 1, gdje 0 označava potpunu nepodudarnost, a 1 označava dvije identične datoteke. Rezultati su prikazani tablično gdje svaka tablica prikazuje rezultate jednog od modela ByT5. U stupcima su prikazne sličnosti teksta obrađenog velikim jezičnim modelom. i temeljne istine, sličnosti teksta obrađenog optičkim prepoznavanjem i temeljne istine i usporedba odnosa teksta obrađenog optičkim prepoznavanjem i velikim jezičnim modelom. U većini slučajeva veću točnost pronalazimo u usporedbi optičkog prepoznavanja i temeljne istine što ukazuje da obrada velikim jezičnim modelima pogoršava rezultate. Promatranjem dobivenih brojčanih vrijednosti zaključili bi da optičko prepoznavanje rezultira boljim rezultatima dok obrada velikim jezičnim modelima pogoršava ali ručnom provjerom rezultata veliki jezični modeli rezultiraju interpretabilniji i kontekstualno točnijim rezultatima. Za fino podešavanje ByT5 modela korišteni su podatci iz novijih novinskih članaka na hrvatskom jeziku dok je tekst obrađene dokumentacije na hrvatsko-srpskom jeziku što je uzrok pojedinih pogrešaka kao i ispravljanje pravopisnih grešaka na imenima auta i slično. Model nije fino podešen za takve primjere i navedenu kombinaciju jezika pa su greške bile očekivane. Zadnji stupac tablice prikazuje odnos teksta obrađenog metodom optičkog prepoznavanja podataka i velikim jezičnim modelom. S obzirom na to da su tekstovi obrađeni OCR-om ulazni podatci za velike jezične modele navedeni omjer prikazuje Koliko teksta je ostalo ne promijenjeno nakon obrade LLM-om. Zanimljivo je za uočit da ByT5 modeli fino podešeni na manjem broju primjera kreiraju manje promjena na ulaznim podatcima ali i postižu bolji odnos s temeljnom istinom. To je prikaz modela koji su naučeni postizati bolje rezultate uvođenjem niti kakvih promjena.

Primjer 5.1: Primjer rezultata primjenom velikih jezičnih modela

```
taj a SE’S
“Je je
“, SVP ZAGREB STVORO POVJERA TISU?
UPRAVA DIZAJNE REZIJJEDNOSTI Samo za litre informacija Web >
« će Odjeljenje –. Ostajek (.
: će » 7 – Pravo je ALI će
Gdj Manes već Srenander Od »« osvrsen TN, 1, –
Sasta će Fabured” je
i . Abert
Akon je "Nis" `
SPECIJALNA INFORMACIJA
Book: 1.
PREDMET: Saradnik “Gustav” otputo,
Yak ponovne je jer je Ma,
chen. =
PRIMJERNA
Veze sveg. ina. br.258 od 26.XII 1 br.262 od 20.XII
1996. godine ".
“Gustov" je u daljnjem razgovoru Gao infor-
maniju da je u Minchenu povezen sa HDZ-om 1 da je
Sesto u DruStvu 10 Prvi¢ Wikolić neki Bruko, godin
je Zagrebe, “Gustav” kaže da je za Braću napio sam,
naše je je stari exigrant, star oko Je godina, je
Je 2. vrijeme kaže imao svoju gostitno ispod Svje –
među 1 se stalno živi u MONCHENU, koje ima svoju
gradjevinsku radnju. Zatim je ima svoj sutomobil 1
je svoje koji France, jer je Porpi¢ nema automobila,
Za Jedin Dragom “Queter" kaže je N. u jer –
nje vrijeme vozi u prvom “Mercedesu", 8 je e radi,
Ja došao u Augustusa u Volkanagenu.
» pripremala emigracije za izvršenje atente
u je za Predsjednike Republike prilikom posjete Aust-
. Puli, koje je trebalo dobiti u proljeće 1996. godine
u je, “Gueter” kaže da je o koje razgovorio je Prvi .
u ¢em 1 Coris Ivanom u ILI mjesecu ove godine i je je
+ tom prilikom Prvi¢e izjavio se je nakon objave posto
u je, u NA Wromatku ubaten je reklje, veliki broj je-
većih Spijena, kojima je je uspjelo saznati nemjere
i emigracije, te da je zbog koja morao raditi pod –
(U "jer.
```

Primjer 5.2: Primjer rezultata primjenom optičkog prepoznavanja znakova

```
tae a SE’s
fe eh
, SUP ZACRED STHOCO POVIERL TIVO?
UPRAVA DIZAVNE HEZIJZEDNOSTI Samo aa litre informacijy Wye >
 TL Odjetjenjo . Odtajek ~~
: TI  7 - Prawn yo UNE ES
Geo canes eet Srennnnnes Ot * asvrsen YN, 8, -
Satta UM Fapured is
i . Abert
Akos jo "Nin"
SPECIJALNA INFORMACHIA
Boop: 7.
PREDMET: Saradnik Gustav otputo~
Yeo ponovne na red uv Mn~
chen. =
FRIMJEDDA
Veze spec. inf. br.258 of 26.XII 1 br.262 Od 30.XII
1966. godine __
Gustev" je u daljojem razgovoru Gao infor-
maciju da je u Minchenu povezen sa HiO-om 1 da je
Sesto u druStvu 98 Prpi Wikoloa neki Bruco, rodom
iz Zegrebe, Gustav kate da je za Bracu uapio saz~
nati da Je stari exigront, star oko Jo godina, ca
Je 2@ vrijone rate imeo svoju gostiony ispod Sije -
mena 1 de stalno #ivi u MOnchenu, adje ime evoju
eradjevinsku radnju. Zotio da isa svoj sutomobil 1
da avude vozi Frpie, jer da Pprpi nema automobila,
Ze Jeli Dranke Quetev" kete da o@ u zed -
nje vrijeme vozi u crnom Mercedesu", 8 da je rani~
Je dolazio u Augsburg u Volkawagenu.
 pripremags emigrecije za izvrienje atente
| ta na Predsjednike Republike prilikos posjete Aust-
. P4ji, koje ee trebala obaviti u proljece 1966. godi
| ne, Guetev kete do je o tome razgoverso aa Prpi ~
| em 1 Cori Ivenom u III mjesecu ove godine i da je
+ tom prilikom Prpi izjevio de je nakon objave posje
| ta, u SR Wjomatku ubaten iz renlje, veliki broj cr-
venih Spijuna, kojima da je uspjelo saznati nemsere
i emigracije, te da oe xbog toga morao odgoditi pos -
(| @jet.
```

Primjer 5.3: Primjer temeljne istine

```
specijalna informacija
saradnik "Gustav" otputovao ponovno na rad u Munchen

veza spec. inf. br.258 od 28.xii i br.262
od 30.xii 1966. godine
"Gustav" je u daljnjem razgovoru dao informaciju
da je u munchenu povezan sa hno-om i da je često u društvu
sa prpićem nikolom neki braco, rodom iz Zagreba
"Gustav" kaže da je za bracu uspio saznati
da je stari emigrant, star oko 50 godina, da
je za vrijeme rata imao svoju gostionu ispod sljemena
i da stalno živi u munchenu, gdje ima svoju
gradjevinsku radnju, zatim da ima svoj
automobil i da svuda vozi prpića,
jer da prpić nema automobil.
za jelić branka "gustav" kaže da se u zadnje
vrijeme vozi u crnom "mercedesu", a da je rani
je dolazio u augsburg u volkswagenu
o pripremama emigracije za izvršenje atentata
na predsjenika republike prilikom posjete aust riji
koja se trebala obaviti u proljeće 1966. godine
"gustav" kaže da je o tome razgovarao sa prpićem
i ćorić ivanom u iii mjesecu ove godine i da je
tom prilikom prpić izjavio da je nakon objave posjeta u
sr njemačku ubačen iz zemlje, velik broj crvenih špijuna
kojima da je uspjelo saznati namjere
emigracije, te da se zbog toga morao odgoditi posjet.
```

Tablica 5.1: Rezultati za model ByT5-v3 korištenjem prve metode

Naziv	LLM vs Truth	OCR vs Truth	OCR vs LLM
Z05353396	0.4648	0.5233	0.7648
Z05353401	0.5457	0.5676	0.7597
Z05353422	0.5858	0.6506	0.7987
Z05353424	0.6961	0.7390	0.8200
Z05353668	0.5920	0.6516	0.7484
Z05353673	0.5673	0.5747	0.7811
Z05353721	0.2632	0.3030	0.7799
Z05353758	0.5955	0.6416	0.7963
Z05353778	0.6529	0.6934	0.7519
Z05353836	0.6642	0.6924	0.7863
Prosjek	**0.5627**	**0.5537**	**0.7787**

Tablica 5.2: Rezultati za model ByT5-v2 korištenjem prve metode

Naziv	LLM vs Truth	OCR vs Truth	OCR vs LLM
Z05353396	0.4777	0.5233	0.7748
Z05353401	0.5419	0.5676	0.7753
Z05353422	0.5812	0.6506	0.8080
Z05353424	0.7064	0.7390	0.8428
Z05353668	0.5936	0.6516	0.7647
Z05353673	0.5680	0.5747	0.7915
Z05353721	0.2679	0.3030	0.7799
Z05353758	0.5984	0.6416	0.8071
Z05353778	0.6483	0.6934	0.7661
Z05353836	0.6789	0.6924	0.8042
Prosjek	**0.5362**	**0.5537**	**0.7830**

Tablica 5.3: Rezultati za model ByT5-small korištenjem prve metode

Naziv	LLM vs Truth	OCR vs Truth	OCR vs LLM
Z05353396	0.4937	0.5233	0.8570
Z05353401	0.5576	0.5676	0.8703
Z05353422	0.6183	0.6506	0.8685
Z05353424	0.7281	0.7390	0.9055
Z05353668	0.6185	0.6516	0.8544
Z05353673	0.5668	0.5747	0.8613
Z05353721	0.2654	0.3030	0.7630
Z05353758	0.5961	0.6416	0.8493
Z05353778	0.6737	0.6934	0.8746
Z05353836	0.6783	0.6924	0.8907
Prosjek	**0.5816**	**0.5537**	**0.8554**

Tablica 5.4: Rezultati za model ByT5-5k korištenjem prve metode

Naziv	LLM vs Truth	OCR vs Truth	OCR vs LLM
Z05353396	0.4922	0.5233	0.8385
Z05353401	0.5659	0.5676	0.8558
Z05353422	0.6128	0.6506	0.8379
Z05353424	0.7106	0.7390	0.8789
Z05353668	0.6068	0.6516	0.8377
Z05353673	0.5455	0.5747	0.8188
Z05353721	0.2714	0.3030	0.6810
Z05353758	0.5998	0.6416	0.8602
Z05353778	0.6553	0.6934	0.8532
Z05353836	0.6505	0.6924	0.8667
Prosjek	**0.5712**	**0.5537**	**0.8229**

Tablica 5.5: Rezultati za model ByT5-v3 korištenjem druge metode

Naziv	LLM vs Truth	OCR vs Truth	OCR vs LLM
Z05353396	0.4741	0.5233	0.7881
Z05353401	0.5541	0.5676	0.7834
Z05353422	0.5940	0.6506	0.8197
Z05353424	0.6991	0.7390	0.8407
Z05353668	0.6013	0.6516	0.7674
Z05353673	0.5787	0.5747	0.8054
Z05353721	0.2806	0.3030	0.8283
Z05353758	0.6134	0.6416	0.8230
Z05353778	0.6652	0.6934	0.7712
Z05353836	0.6688	0.6924	0.8111
Prosjek	**0.5729**	**0.6545**	**0.8038**

Tablica 5.6: Rezultati za model ByT5-v2 korištenjem druge metode

Naziv	LLM vs Truth	OCR vs Truth	OCR vs LLM
Z05353396	0.4874	0.5233	0.7981
Z05353401	0.5511	0.5676	0.7991
Z05353422	0.5893	0.6506	0.8289
Z05353424	0.7097	0.7390	0.8632
Z05353668	0.6035	0.6516	0.7836
Z05353673	0.5794	0.5747	0.8164
Z05353721	0.2857	0.3030	0.8283
Z05353758	0.6165	0.6416	0.8342
Z05353778	0.6605	0.6934	0.7857
Z05353836	0.6840	0.6924	0.8295
Prosjek	**0.5777**	**0.6537**	**0.8177**

Tablica 5.7: Rezultati za model ByT5-small korištenjem druge metode

Naziv	LLM vs Truth	OCR vs Truth	OCR vs LLM
Z05353396	0.5039	0.5233	0.8819
Z05353401	0.5671	0.5676	0.8954
Z05353422	0.6267	0.6506	0.8910
Z05353424	0.7319	0.7390	0.9279
Z05353668	0.6295	0.6516	0.8738
Z05353673	0.5782	0.5747	0.8881
Z05353721	0.2828	0.3030	0.8081
Z05353758	0.6141	0.6416	0.8778
Z05353778	0.6868	0.6934	0.8974
Z05353836	0.6846	0.6924	0.9196
Prosjek	**0.5626**	**0.6537**	**0.8761**

Tablica 5.8: Rezultati za model ByT5-5k korištenjem druge metode

Naziv	LLM vs Truth	OCR vs Truth	OCR vs LLM
Z05353396	0.5025	0.5233	0.8640
Z05353401	0.5757	0.5676	0.8805
Z05353422	0.6198	0.6506	0.8573
Z05353424	0.7134	0.7390	0.8990
Z05353668	0.6176	0.6516	0.8576
Z05353673	0.5574	0.5747	0.8444
Z05353721	0.2893	0.3030	0.7172
Z05353758	0.6174	0.6416	0.8903
Z05353778	0.6673	0.6934	0.8765
Z05353836	0.6537	0.6924	0.8946
Prosjek	**0.5714**	**0.6537**	**0.8537**

6. Zaključak

Glavni cilj ovog seminarskog rada bio je razviti metodu obrade optičko prepoznatog teksta sa stare arhivske dokumentacije nakon obrade primjenom metoda računalnog vida. Metoda je razvijena korištenjem velikih jezični modela ByT5 i crosloengual za ispravljanje pravopisnih pogrešaka hrvatskog jezika odnosno kontekstualnu analizu riječi u rečenici na hrvatskom jeziku. Preciznije razvijene su dvije metode obrade ulaznih podataka koje se razlikuju u pragu prihvaćanja riječi procijenjenih velikim jezičnim modelom. Rezultat obrade su tekstualne datoteke obrađene velikim jezičnim modelima nad tekstualnim datotekama dobivenih provedbom optičkog prepoznavanja stare arhivske dokumentacije. Razvijene metode, uz manje pogreške, kreiraju kontekstualno i pravopisno točnije podatke od same primjene optičkog prepoznavanja, iako krajnji rezultat nije uvijek bliži stvarnoj istini. Daljnje unaprjeđenje metoda zahtijevalo bi kreiranje prilagođenog skupa podataka za hrvatsko-srpski jezik, s karakterističnim pogreškama za sustave optičkog prepoznavanja znakova.

7. Sažetak

Veliki jezični model (engl. Large Language Model) je model dubokog učenja koji se sastoji od neuronske mreže trenirane upotrebom samonadziranog učenja. Koriste se za kreiranje, raspoznavanje i semantičku analizu teksta. U sklopu ovog seminara koristit će se za prepoznavanje i korekciju krivo prepoznatih riječi odnosno riječi koje nisu u sklopu hrvatskog jezika. Zbog ograničenih podatkovnih i računalnih resursa koristit ćemo unaprijed trenirane jezične modele na hrvatskom jeziku: T5 i CroSloEngual BERT.

Optičko prepoznavanje znakova (engl. Optical Character Recognition, OCR) uključuje tehnike pripreme i obrade digitalnih slika na kojima je potrebno prepoznati tekst. U okviru seminara rada potrebno je istražiti, prilagoditi i primijeniti tehnike optičkog prepoznavanja znakova za što bolje prepoznavanje teksta dostupnog na fotografijama stare arhivske građe na hrvatskom jeziku.

8. Literatura

[1] Hans Christensen i Liling Tan. HC Corpora newspapers. https://www.kaggle.com/datasets/alvations/old-newspapers.

[2] Jenna Kanerva, Cassandra Ledins, Siiri Käpyaho, i Filip Ginter. Ocr error post-correction with llms in historical documents: No free lunches. https://arxiv.org/pdf/2502.01205.

[3] Lovro Magdić. Primjena postupka optičkog prepoznavanja podataka na staru arhivsku dokumentaciju. https://repozitorij.fer.unizg.hr/islandora/object/fer%3A12514.

[4] M. Ulčar i M. Robnik-Šikonja. FinEst BERT and CroSloEngual BERT: less is more in multilingual models. https://doi.org/10.1007/978-3-030-58323-1_11.

[5] Martijn Veninga. Llms for ocr post-correction. https://essay.utwente.nl/102117/1/Veninga_MA_EEMCS.pdf.

[6] Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, i Colin Raffel. Byt5: Towards a token-free future with pre-trained byte-to-byte models. https://arxiv.org/pdf/2105.13626.