

# Dokumentacija

## Faza I – Prikupljanje podataka

Grupa se sastoji od 5 članova, stoga je bilo neophodno pokriti 6 domena i prikupiti najmanje 5000 tokena po domenu. Domeni koji su izabrani su književni, pravno-administrativni, novinski, muzički, filmsko-pozorišni i tviter domen. Izabrani su jer u velikoj meri oslikavaju različite stilove i načine korišćenja jezika. Takođe prethodna analiza je pokazala da svaki od ovih domena potencijalno sadrži značajan broj entiteta relevantnih za klase od interesa. Za svaku od kategorija, svaki od članova tima prikupljao je najmanje po 1000 tokena. Po dogovoru, svaki pronađeni tekst iz različitog izvora je sačuvan u zaseban fajl.

### Književni domen

Tekstovi koji pripadaju književnom domenu su izabrani iz dela koja su prevedeni na srpski jezik kao i dela koja su u originalnom izdanju na srpskom jeziku. Takođe su za ovaj domen izabrani tekstovi moderne i klasične književnosti. Tekstovi birani za književni domen su uglavnom sa sajtova knjižara, na kojima je bila dostupna opcija delimičnog pregleda izabrane knjige. Izazov je bilo naći književne tekstove gde u prva dva poglavlja (pošto pregled knjige uglavnom sadrži svega par početnih poglavlja) postoji dovoljan broj entiteta od interesa. Tokom faze prikupljanja tekstova za književni domen, primećeno je da je jednostavnije naći tekstove sa potencijalno dovoljnim brojem različitih entiteta za PER anotaciju i LOC, ali ne toliko za ORG.

Tekstovi su pronađeni na sajtovima knjižara Laguna i Delfi zbog količine sadržaja i dostupnih tekstova. Primećeno je da najčešće dovoljan broj za sve tri anotacije sadrže najviše književni rodovi poput putopisa i istorijskih romana. Međutim radi pokrivanja što većeg spektra književnih tipova, podstaknut je dodatni trud za pretragom tekstova koji pripadaju klasičnoj književnosti, fantastici itd.

Broj sakupljenih fajlova za književni domen je 7.

### Novinski domen

Novinski domen se pokazao značajno manje izazovnim u odnosu na domene koji obuhvataju kulturne delatnosti. Izvori su mnogobrojni, a svake novine za sebe predstavljaju ogroman izvor tekstova. Takođe način pisanja je jako koncizan i takav da uglavnom se tekstovi, pogotovo za politiku i sport, sastoje od kraćih rečenica sa dosta informacija u kojima se nalaze uvek bar neki entitet a često i sva tri. Obuhvaćeni su različiti izvori, sa različitim stavovima. Takođe fokus nije bio samo na pronalaženje teksta vezano za domaću politiku i dešavanja, već je aktivno podstaknuta i pretraga vesti vezanih za inostrana dešavanja time obuhvatajući i mogućnost obrade stranih naziva.

Sajtovi sa kog su preuzeti novinski članci su:

1. Blic (<https://www.blic.rs>)
2. N1 info (<https://n1info.rs>)
3. Nova (<https://nova.rs>)
4. Sport Klub (<https://sportklub.n1info.rs>)
5. Nin (<https://www.nin.rs> )
6. Sportal u okviru Blic-a (<https://sportal.blic.rs>)
7. Politika (<https://www.politika.rs>)
8. Forbes u okviru N1 info (<https://forbes.n1info.rs>)
9. Rts (<https://rts.rs> )

Broj sakupljenih tekstova su raspoređeni u 23 fajla. Može se primetiti značajno veća količina fajlova u odnosu na domen kulture u svakom smislu, usled manjih tekstova u pojedinačnim vestima i člancima.

Broj tokenaje takođe veći. Što jasno ukazuje potrebu za celim tekstom usled njegove konciznosti bez rasparčavanja, kao i dostupnost tekstova koji zadovoljavaju kriterijume.

#### Pravno-administrativni domen

Tekstovi pravno-administrativnog domena, koji su izabrani, su presude Vrhovnog suda u Beogradu jer sadrže sve tri klase za NER anotaciju od značaja. Tekstovi pravno-administrativnog domena su svi imali problem slabe raznolikosti potencijalnih entiteta. Takođe često se koriste samo inicijali ukoliko se radi o osobama koje nisu sudije, advokati itd.. Zbog toga akcenat je postavljen da se traže tekstovi sa raznovrsnijim entitetima za svaku grupu, iako sam njihov broj je generalno mali procentualno u odnosu na ostatak teksta.

Sajt sa kog su uzeti tekstovi za pravni domen je sajt Vrhovnog suda (<https://www.vrh.sud.rs>).

Broj sakupljenih fajlova je 11.

#### Tviter domen

Tviter domen obuhvata tekstove informativnog karaktera, kao i one koji prikazuju svakodnevni život ljudi. Budući da je Tviter platforma sa ograničenjem broja karaktera po objavi, domen sadrži najveći broj pojedinačnih fajlova, usled manjeg obima tekstova. Zbog pretežno neformalnog stila komunikacije, process traženja tvitova koji sadrže entitete je bio vremenski zahtevniji.

Zajedničkim dogovorom dozvoljeno je odabir tvitova koji sadrže i po dva potencijalna entiteta, ali da u sveskupnom korpusu tekstova budu zastupljeni sva tri tipa entiteta. Često se dešava nemanja raznolikosti u okviru jednog tvita, međutim to je nadomešteno broju tvitova koji su nađeni. Ispostavilo se da tvitovi vezani za sport i politiku poput tekstova iz domena novosti najviše sadrže tražene entitete, stoga je akcenat pao na njih. Ali je podržano traženje i komentara na različite teme od korisnika.

Broj sakupljenih fajlova je 56.

## Film-pozorište domen

Tekstovi za domen pozoriste i filma su izabrani tako da obuhvataju kritike zasebnih pozorišnih izvođenja ili filmova kao i najave kulturnih dešavanja, poput raznoraznih festivala. Dosta poznatih, i manje poznatih novina u okviru sebe sadrže poseban deo za kulturu što je bio značajan izvor tekstova za biranje.

Tokom traženja tekstova primećeno je da ukoliko bi se uzimala kritika nekog dela, kako je akcenat stavljen na glumce, većina potencijalnih entiteta se odnosila na tip PER. Eventualno se ponekad našlo dovoljno entiteta za tip LOC, najčešće u kontekstu mesta izvedba ili premijera. S druge strane bilo teže naći tekst sa dovoljno različitih entiteta, koji bi potencijalno pripadali ORG. Više primera za ORG i LOC je pronađeno u tekstovima čiji su za cilj imali prenos i najavu nekog dešavanja i festivala. Ipak, priroda takvih tekstova podrazumeva ograničeno navođenje konkretnih imena — fokus je uglavnom na sadržaju događaja i samoj radnji, a ne na učesnicima ili organizatorima.

Sajtovi sa kog su prikupljeni tekstovi za konkretan domen su:

1. Sajt City Magazine, odeljak popkultura (<https://citymagazine.danas.rs/popkultura>)
2. Filmske radosti (<https://filmske-radosti.com>)
3. Sajt Doma omladine (<https://domomladine.org>)
4. Sajt Nova.rs, odeljak kultura (<https://nova.rs/kultura>)
5. Sajt Narodnog pozorišta ( <https://www.narodnopozoriste.rs>)
6. Univerzitetski odjek, odeljak kultura (<https://www.univerzitetskiodjek.com/kultura>)
7. SeeCult – Portal za kulturu Jugoslavije (<https://www.seecult.org/>)

Broj sakupljenih fajlova u domenu, koji obuhvata film i pozorište, je 11 fajlova.

## Muzički domen

Tekstovi prikupljeni za muzički domen birani su sa online dostupnih izvora posvećenih muzici što uključuje najavu muzičkih festivala, konkretne kritike vezane za albume, nastupe i uopšteno za bendove. Za tekstove iz ovog domena, nije bio problem nalaženje dovoljnog broja različitih entiteta za sve tri kategorije, već njihova generalna retkost pojavljivanja u tekstovima. Tekstovi su uglavnom bili dužeg formata, međutim akcenat ovih tekstova jeste opis i pisanje doživljaja sa konkretnog koncerta, što je smanjivalo potrebu navođenja bendova i mesta posle prvog pominjanja.

Kako u većini nađenih tekstova postojali su pasusi koji nisu sadržali ni jedan entitet, stoga sakupljanje teksta se ograničilo na njihove delove koji su sadržali zadovoljavajući broj entiteta. Drugim rečima, u korpus nisu uključivani čitavi tekstovi, već selektovani odlomci.

Naravno razmatran je spektar tekstova tako da se uključe različiti žanrovi muzike, što od metal muzike do nastupa klasične, čime se obezbedio veći broj različitih termina, usled raznovrsnosti samih tekstova i izražavanja.

Online sajtovi sa kog su prikupljeni tekstovi su:

1. Balkanrock (<https://balkanrock.com>)
2. Sajt Klasicni Mirko (<https://www.klasicnimirko.com/muzicke-kritike>)
3. Muzički Limbo (<https://muzickilimbo.rs>)

Broj sakupljenih fajlova za muzički domen jeste 15.

## Faza II – Anotacija podataka

Svako član tima je anotirao tekstove koje je sam pronašao tokom faze traženja tekstova. Na taj način je obezbeđeno da svi članovi ravnomerno učestvuju u procesu I da približno anotira isti broj tokena u okviru svakog domena.

### Anotacioni uzorak 10 procenata

U skladu sa dogovorom unutar tima, za potrebe koraka kalibracije anotatora, svaki član je iz svog skupa podataka izabrao okvirno 10% tekstova. Izabrani uzorci teksta su korišćeni za međusobno usklađivanje načina anotiranja I proveru doslednosti u primeni pravila.

Kod, kao i dobijeni izveštaji za određivanje saglasnosti između dva anotatora se nalaze u folderu Prodjenje anotiranih fajlova. Takođe u izveštajima su sačuvane i konkretne razlike nađenje u tekstovima tokom procesa anotiranja.

*Tabela 1. – Tabelarni prikaz stepena saglasnosti između anotatora*

| Anotator 1 | Anotator 2 | Broj anotiranih tokena | Broj razlika u anotiranju | Stepen saglasnosti svih tokena (%) | Stepen saglasnosti za anotirane tokene (%) |
|------------|------------|------------------------|---------------------------|------------------------------------|--|
| Aleksandar | Marija     | 1304                   | 155                       | 97,42                              | 88,11                                      |
| Aleksandar | Milena     | 1341                   | 177                       | 97,05                              | 86,80                                      |
| Aleksandar | Ognjen     | 1375                   | 203                       | 96,62                              | 85,24                                      |
| Aleksandar | Teodora    | 1303                   | 158                       | 97,37                              | 87,87                                      |
| Marija     | Milena     | 1246                   | 96                        | 98,40                              | 92,30                                      |
| Marija     | Ognjen     | 1294                   | 131                       | 97,82                              | 89,88                                      |
| Marija     | Teodora    | 1324                   | 155                       | 97,42                              | 88,29                                      |
| Milena     | Ognjen     | 1240                   | 85                        | 98,58                              | 93,15                                      |
| Milena     | Teodora    | 1318                   | 146                       | 97,57                              | 88,92                                      |
| Ognjen     | Teodora    | 1325                   | 149                       | 97,52                              | 88,75                                      |

#### - Globalna statistika za korak kalibracije anotatora

Ukupan broj tokena: 6005

Ukupan broj anotiranih tokena: 1811

Procenat slicnosti anotiranih tokena (svaki anotirani token ima vrednost izmedju 0 i 1 u zavisnosti od toga koliko anotatora se slaze): 68.29%

Procenat slicnosti svih tokena (svaki anotirani token ima vrednost izmedju 0 i 1 u zavisnosti od toga koliko anotatora se slaze): 90.44%

Procenat sličnosti anotiranih tokena (anotirani tokeni se uzimaju kao dobro anotirani samo ako se svi anotatori slažu): 44.06%

Procenat sličnosti svih tokena (anotirani tokeni se uzimaju kao dobro anotirani samo ako se svi anotatori slažu): 83.13%

## Anotaciona pravila

Pravila za anotaciju su definisana u rules.md dokumentu, koji se nalazi u okviru projekta. Na osnovu izveštaja o razlikama anotatora, sve nedoumice koje su se javile dodatno su razjašnjenje i obrazložene. Ukoliko se anotatori nisu složili za neke primere, konačna odluka se dobijala odlukom većine. Konačne odluke i objašnjenja su zabeleženi u sledećim pravilima anotacije:

Tabela 2. – Tabelarni prikaz anotacionih pravila

| Pravila  | Primer uz pojašnjenje   |
|--|---|
| 1. <i>Prisvojni pridev zavisi od konteksta</i>   | <ul style="list-style-type: none"> <li>- Janin film = B-PER O</li> <li>- Milenina kuca = B-LOC I-LOC</li> <li>- somborska regija = B-LOC I-LOC</li> <li>- Somborac Ernest Bosnjaku - Ernest Bosnjaku je B-PER, Somborac nije nista</li> <li>- Princ Hari vojvoda od Saseksa - Hari je B-PER, vojvoda nista, Saseksa – loc</li> <li>- Gradonacelnik Beograda Aleksandar Sapic - (O B-LOC B-PER I-PER), gradonacelnik je nista (titula), Beograd - loc</li> </ul> |
| 2. <i>Festival/fondacija/muz. bend je organizacija</i>                                     | <ul style="list-style-type: none"> <li>- Hills of Rock festivala - festival nije org, ostalo jeste</li> <li>- Nazivi casopisa - sve je org</li> </ul>   |
| 3. <i>Ako je deo naziva muzičkog dela ne anotira se</i>                                    |   |
| 4. <i>Ako pise grad Nis, da li je grad deo anotiranja ili ne</i>                           | <ul style="list-style-type: none"> <li>- kao organizaciona jedinica I grad je deo org</li> <li>- ako je naznačen kao lokacija onda nije deo</li> </ul>  |
| 5. <i>Ako ima titulu dr. Ana Peric, dr nije B-PER</i>                                      |   |
| 6. <i>Inicijali AA su B-PER</i>  |   |
| 7. <i>Šta je O</i>   | <ul style="list-style-type: none"> <li>- Ime naroda je nista</li> <li>- austrijska vladavina – nista</li> <li>- Zakon o informisanju - nista</li> </ul>   |
| 8. <i>Imena ljudi, mesta, ustanova... u nazivima filmova nisu ni jedan entitet</i>         | <ul style="list-style-type: none"> <li>- Npr Vanjin dnevnik, Vanjin nije B-PER</li> </ul>   |
| 9. <i>Ukoliko je lokacija deo naziva organizacije onda se beleži kao org</i>               | <ul style="list-style-type: none"> <li>- Filmski festival u Lokranu - B-ORG (sve je org, i mesto)</li> </ul>  |
| 10. <i>Imena naučnih institucija sa nazivima lokacije zavisi od konteksta (pravilo 9.)</i> | <ul style="list-style-type: none"> <li>- Institut u Vinci - sve je org, Institut nista ne znaci (izuzetak)</li> </ul>   |

|   |   |
|---|---|
|   | <ul style="list-style-type: none"> <li>- Naučni institut Vinča - onda je sve org</li> <li>- Institutu Kiri u Parizu - Institut Kiri org, Parizu – loc ( Pariz nije deo naziva instituta)</li> </ul>   |
| 11. Državne institucije   | <ul style="list-style-type: none"> <li>- Osnovni sud u Novom Sadu - sve je org, mora mesto ima vise osn. Sudova</li> <li>- Vlada je org</li> <li>- Ustavni sud - samo jedan je Ustavni</li> <li>- Ukrajinska vlada - sve org</li> <li>- Izraelska mornarica - sve je org</li> </ul> |
| 12. Naziv drzave je loc samo ako se refereise kao lokacija, inace je uvek org | <ul style="list-style-type: none"> <li>- Gradonacelnig Beograda - 0 B-LOC</li> <li>- predsendik Srbije - 0 B-ORG</li> </ul>   |
| 13. Nazivi nagrade - nije nista   | <ul style="list-style-type: none"> <li>- Miselinova zvezdica - isto nista</li> </ul>  |
| 14. Deo mesta lokacije - konkretno mesto je loc, delovi tog mesta nisu        | <ul style="list-style-type: none"> <li>- atrijumu Gradske kuće - ovo je loc, Gradska kuća nije organizacija kao naziv restorana ili kafića.. atrijum nista</li> </ul>   |
| 15. nazivi kafica   | <ul style="list-style-type: none"> <li>- zaposleni restorarana Usce - zap. nije org, ostalo jesto</li> <li>- basta restorana Usce - zavisno od konteksta, ukoliko se nesto desava u restoranu Usce - loc, inace je org</li> </ul>   |
| 16. Imena ljudi, mesta, ustanova u nazivima ulica su deo LOC.                 |   |

## Statistika tokena za skup podataka

Cilj sprovedene statističke analize je da se prikaže ukupna raspodela tokena po svim klasama (B-PER, I-PER, B-LOC, I-LOC, B-ORG, I-ORG i O), odnosno da se utvrdi učestalost pojavljivanja u skupu podataka. Analiza omogućava uvid u uravnoteženosti skupa podataka po klasama i proceni kvaliteta i reprezentativnosti podataka pre evaluacije modela za prepoznavanje entiteta.

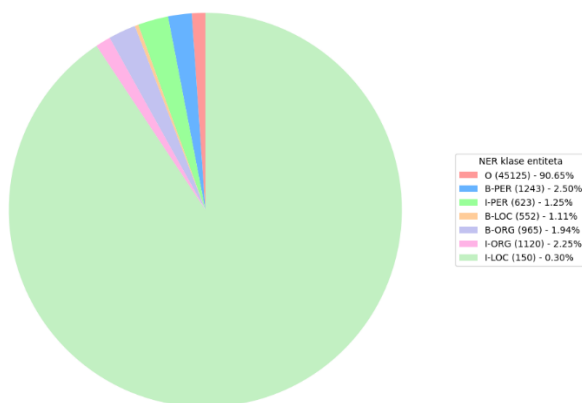
Za izračunavanje statistike korišćena je Python skriptatoken\_statistics.py, a dobijeni podaci su sačuvani u statistics.txt.

Analiza je sprovedena nad skupom od ukupno 49778 tokena u skupu podataka. Raspodela tokena po klasama od interesa je prikazana u tabeli 3., dok je vizuelni prikaz raspoređenosti podataka prikazan na slici 1.

Tabela 3. – Raspodela tokena po NER klasama

| Klasa               | Broj tokena | Procent (%) |
|---------------------|-------------|-------------|
| B-PER               | 1243        | 2,50        |
| I-PER               | 623         | 1,25        |
| <i>Ukupno (PER)</i> | 1866        | 3,75        |
| B-LOC               | 552         | 1,11        |
| I-LOC               | 150         | 0,30        |
| <i>Ukupno (LOC)</i> | 702         | 1,41        |
| B-ORG               | 965         | 1,94        |
| I-ORG               | 1120        | 2,25        |
| <i>Ukupno (ORG)</i> | 2085        | 4,19        |
| O                   | 45125       | 90,65       |
| <i>Ukupno</i>       | 49778       | 100         |

Distribucija tokena po klasama nad celim skupom podataka



Slika 1. Distribucija tokena po klasama nad celim skupom podataka

Može se primetiti da je 90,65% prikupljenih tekstova čine tokeni klafikovani kao O, dok svega 9,35% tokena predstavlja tokene koji su označeni labelama od interesa. Procentat NER tokena ukazuje na umerenu gustinu pojavljivanja označenih tokena, što je i očekivano u većini prirodnih tekstova.

Najzastupljenija kategorija označenih tokena jeste ORG (4,19% svih tokena), potom PER (3,75%) i LOC (1,41%), ukazujući da tekst nije



primarno orijentisan na lokacije, već organizacije i osobe. Što je razumljivo s obzirom na prirodu tekstova iz izabranih domena.

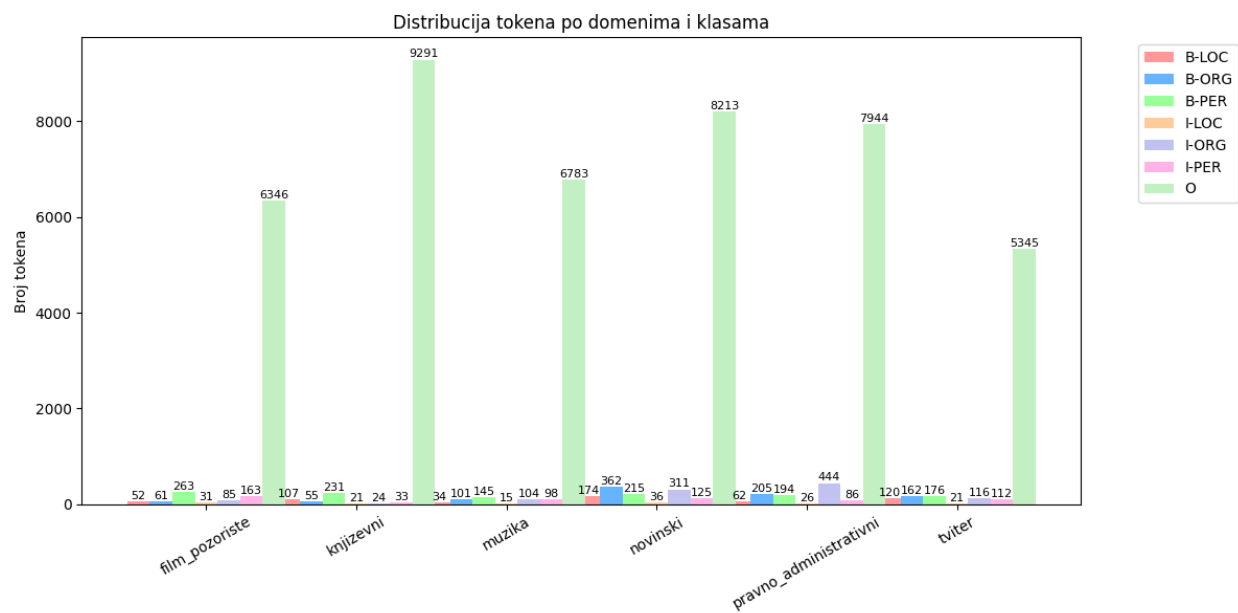
## Analiza po domenima

U nastavku izveštaja analiza je takođe izvršena nad celim skupom podatakam pri čemu su svi tokeni sa svojim labelama grupisani u tematske domene (književni, film-pozorište, muzika, novinski, tviter, pravno-administrativni). Na ovaj način je prikazana je zastupljenost klasa unutar pojedinačnih domena, kao i procena doprinosa širem kontekstu. Raspodela tokena po domenima, potom i klasama je prikazana na sledećim tabela.

*Tabela 4. Tabelarni prikaz raspoređenosti tokena grupisanih po domenu, potom po klasi*

| Domen          | Klasa        | Broj tokena | Procenat nad celim skupom(%) | Procenat u okviru domena (%) |
|----------------|--------------|-------------|------------------------------|------------------------------|
| Književni      | B-PER        | 231         | 0,46                         | 2,37                         |
|                | I-PER        | 33          | 0,07                         | 0,34                         |
|                | Ukupno (PER) | 264         | 0,53                         | 2,71                         |
|                | B-LOC        | 107         | 0,21                         | 1,10                         |
|                | I-LOC        | 21          | 0,04                         | 0,22                         |
|                | Ukupno (LOC) | 128         | 0,25                         | 1,32                         |
|                | B-ORG        | 55          | 0,11                         | 0,56                         |
|                | I-ORG        | 24          | 0,05                         | 0,25                         |
|                | Ukupno (ORG) | 79          | 0,16                         | 0,81                         |
|                | O            | 9291        | 18,66                        | 95,18                        |
|                | Ukupno       | 9762        | 19,61                        | 100,00                       |
| Film-pozorište | B-PER        | 263         | 0,53                         | 3,76                         |
|                | I-PER        | 163         | 0,33                         | 2,33                         |
|                | Ukupno (PER) | 426         | 0,83                         | 6,09                         |
|                | B-LOC        | 52          | 0,10                         | 0,74                         |
|                | I-LOC        | 31          | 0,06                         | 0,44                         |
|                | Ukupno (LOC) | 83          | 0,16                         | 1,18                         |
|                | B-ORG        | 61          | 0,12                         | 0,87                         |
|                | I-ORG        | 85          | 0,17                         | 1,21                         |
|                | Ukupno (ORG) | 146         | 0,29                         | 2,08                         |
|                | O            | 6346        | 12,75                        | 90,64                        |
|                | Ukupno       | 7001        | 14,06                        | 100,00                       |
| Muzika         | B-PER        | 145         | 0,29                         | 1,99                         |
|                | I-PER        | 98          | 0,20                         | 1,35                         |
|                | Ukupno (PER) | 243         | 0,49                         | 3,34                         |
|                | B-LOC        | 34          | 0,07                         | 0,47                         |
|                | I-LOC        | 15          | 0,03                         | 0,21                         |

|                               |              |      |       |        |
|-------------------------------|--------------|------|-------|--------|
|                               | Ukupno (LOC) | 49   | 0,10  | 0,68   |
|                               | B-ORG        | 101  | 0,20  | 1,39   |
|                               | I-ORG        | 104  | 0,21  | 1,43   |
|                               | Ukupno (ORG) | 205  | 0,41  | 2,82   |
|                               | O            | 6783 | 13,63 | 93,17  |
|                               | Ukupno       | 7280 | 14,61 | 100,00 |
| <b>Novinski</b>               | B-PER        | 215  | 0,43  | 2,28   |
|                               | I-PER        | 125  | 0,25  | 1,32   |
|                               | Ukupno (PER) | 340  | 0,68  | 3,60   |
|                               | B-LOC        | 174  | 0,35  | 1,84   |
|                               | I-LOC        | 36   | 0,07  | 0,38   |
|                               | Ukupno (LOC) | 210  | 0,42  | 2,22   |
|                               | B-ORG        | 362  | 0,73  | 3,84   |
|                               | I-ORG        | 311  | 0,62  | 3,30   |
|                               | Ukupno (ORG) | 673  | 1,35  | 7,14   |
|                               | O            | 8213 | 16,50 | 87,04  |
|                               | Ukupno       | 9436 | 18,96 | 100,00 |
| <b>Tviter</b>                 | B-PER        | 176  | 0,35  | 2,91   |
|                               | I-PER        | 112  | 0,22  | 1,85   |
|                               | Ukupno (PER) | 288  | 0,57  | 4,76   |
|                               | B-LOC        | 120  | 0,24  | 1,98   |
|                               | I-LOC        | 21   | 0,04  | 0,35   |
|                               | Ukupno (LOC) | 141  | 0,28  | 2,33   |
|                               | B-ORG        | 162  | 0,32  | 2,66   |
|                               | I-ORG        | 116  | 0,23  | 1,92   |
|                               | Ukupno (ORG) | 277  | 0,55  | 4,58   |
|                               | O            | 5345 | 10,74 | 88,33  |
|                               | Ukupno       | 6052 | 12,16 | 100,00 |
| <b>Pravno-administrativni</b> | B-PER        | 194  | 0,39  | 2,16   |
|                               | I-PER        | 86   | 0,17  | 0,96   |
|                               | Ukupno (PER) | 270  | 0,56  | 3,12   |
|                               | B-LOC        | 62   | 0,12  | 0,69   |
|                               | I-LOC        | 26   | 0,05  | 0,29   |
|                               | Ukupno (LOC) | 88   | 0,17  | 0,98   |
|                               | B-ORG        | 205  | 0,41  | 2,29   |
|                               | I-ORG        | 444  | 0,89  | 4,95   |
|                               | Ukupno (ORG) | 649  | 1,30  | 7,24   |
|                               | O            | 7944 | 15,96 | 88,65  |
|                               | Ukupno       | 8961 | 18,00 | 100,00 |



Distribucija tokena po domenima nad celim skupom podataka

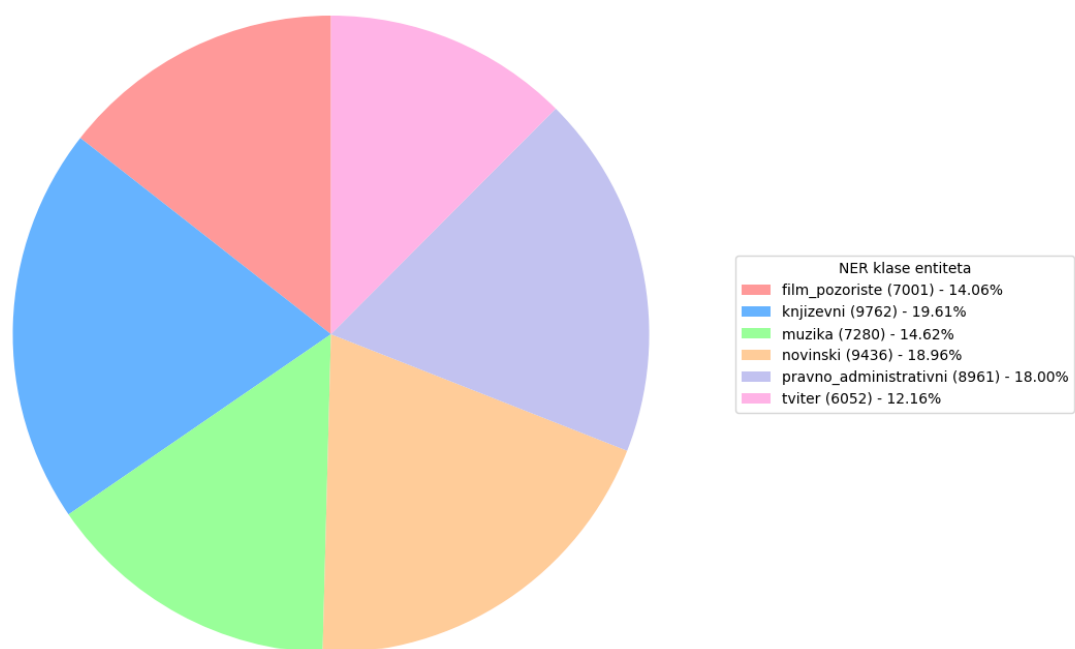
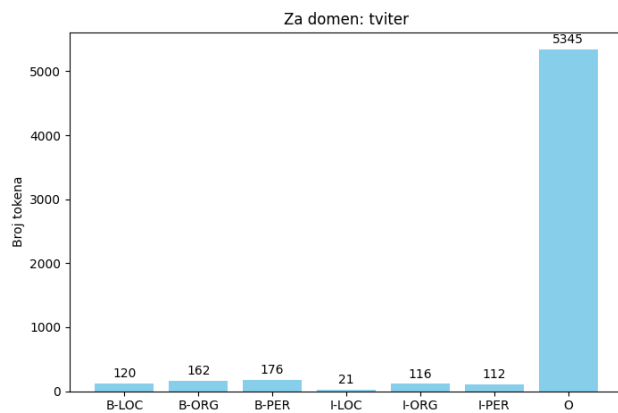
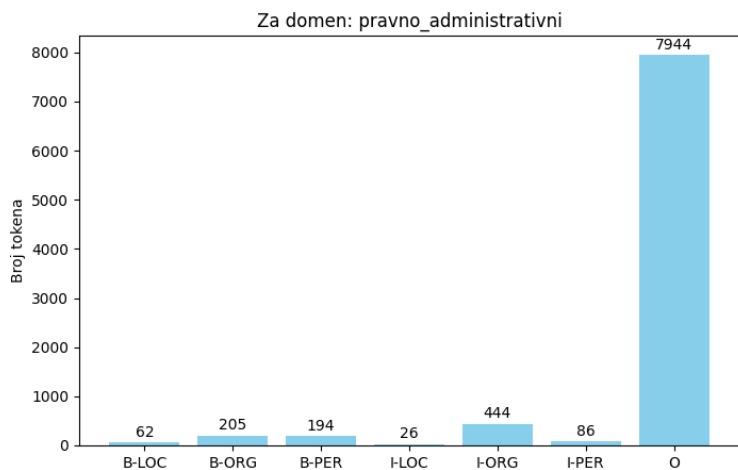
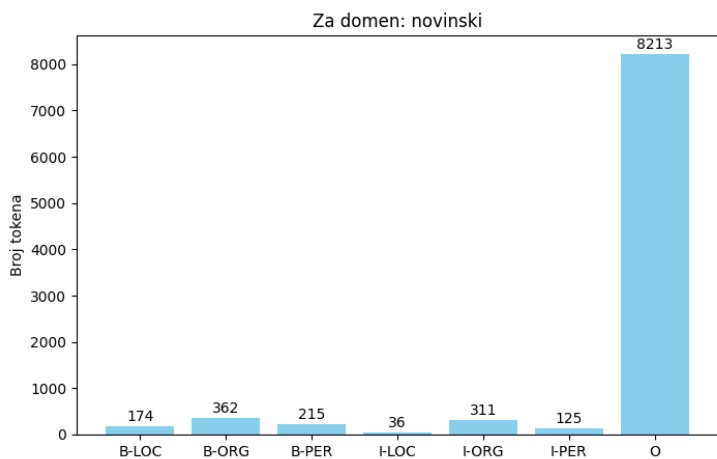
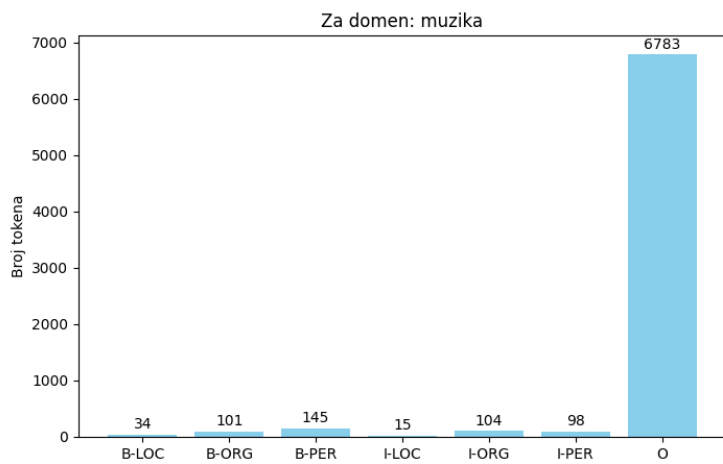
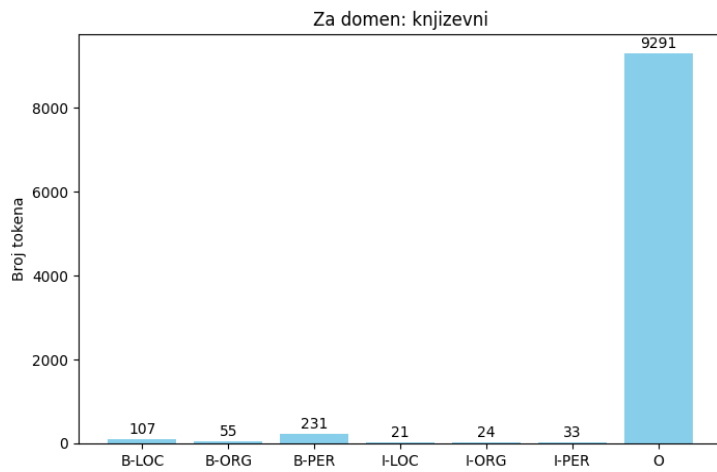
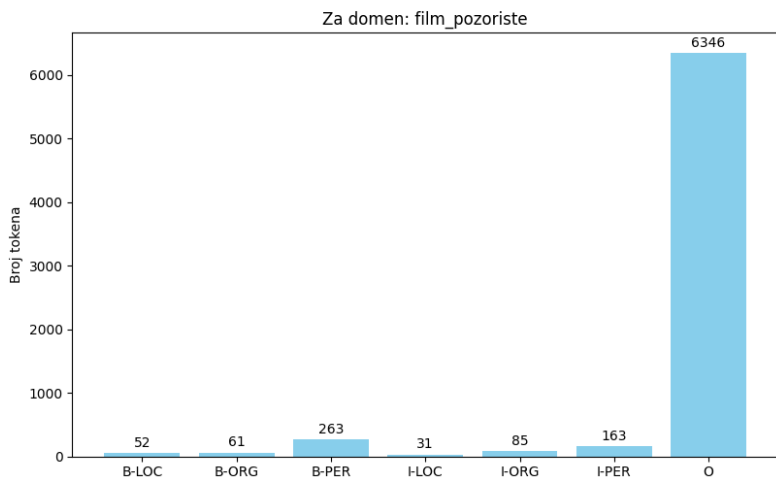


Tabela 4. – Tabelarni prikaz distribucije tokena unutar klasa grupisanih po domenima.

| Klase                     | Domen                  | Broj tokena | Procenat za ceo skup podataka(%) | Procenat za klasu (%) |
|---------------------------|------------------------|-------------|----------------------------------|-----------------------|
| <b>PER (B-PER, I-PER)</b> | Književni              | 264         | 0,53                             | 14,42                 |
|                           | Filmovi-pozorište      | 426         | 0,83                             | 23,27                 |
|                           | Muzika                 | 243         | 0,49                             | 13,27                 |
|                           | Novinski               | 340         | 0,68                             | 18,57                 |
|                           | Pravno-administrativni | 270         | 0,56                             | 14,75                 |
|                           | Tviter                 | 288         | 0,57                             | 15,73                 |
| <b>LOC (B-LOC, I-LOC)</b> | Književni              | 128         | 0,25                             | 18,31                 |
|                           | Filmovi-pozorište      | 83          | 0,16                             | 11,87                 |
|                           | Muzika                 | 49          | 0,10                             | 7,01                  |
|                           | Novinski               | 210         | 0,42                             | 30,04                 |
|                           | Pravno-administrativni | 88          | 0,17                             | 12,59                 |
|                           | Tviter                 | 141         | 0,28                             | 20,17                 |
| <b>ORG (B-ORG, I-ORG)</b> | Književni              | 79          | 0,16                             | 3,89                  |
|                           | Filmovi-pozorište      | 146         | 0,29                             | 7,20                  |
|                           | Muzika                 | 205         | 0,41                             | 10,10                 |
|                           | Novinski               | 673         | 1,35                             | 33,17                 |
|                           | Pravno-administrativni | 649         | 1,30                             | 31,99                 |
|                           | Tviter                 | 278         | 0,55                             | 13,65                 |
| <b>O</b>                  | Književni              | 9291        | 18,66                            | 21,15                 |
|                           | Filmovi-pozorište      | 6346        | 12,75                            | 14,45                 |
|                           | Muzika                 | 6783        | 13,63                            | 15,44                 |
|                           | Novinski               | 8213        | 16,50                            | 18,70                 |
|                           | Pravno-administrativni | 7944        | 15,96                            | 18,09                 |
|                           | Tviter                 | 5345        | 10,74                            | 12,17                 |

## Grafički prikazi broja tokena u zasebnim domenima

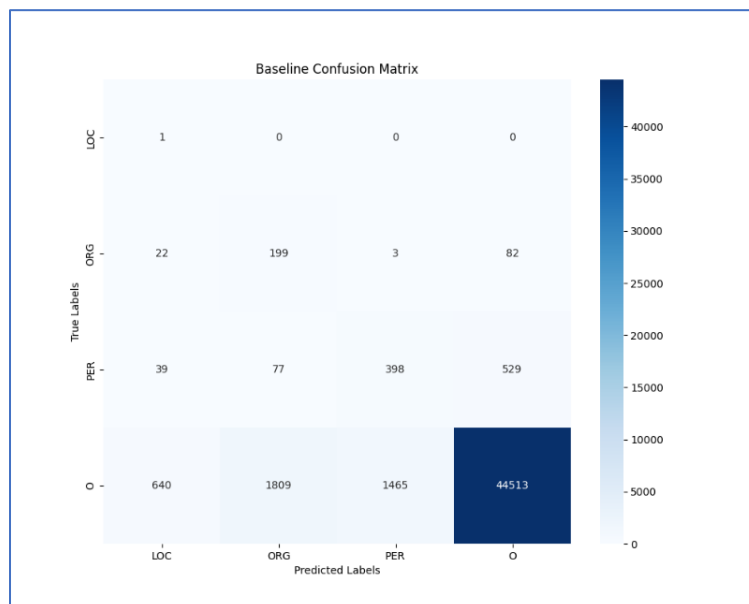


## Faza III – Evaluacija statističkih modela

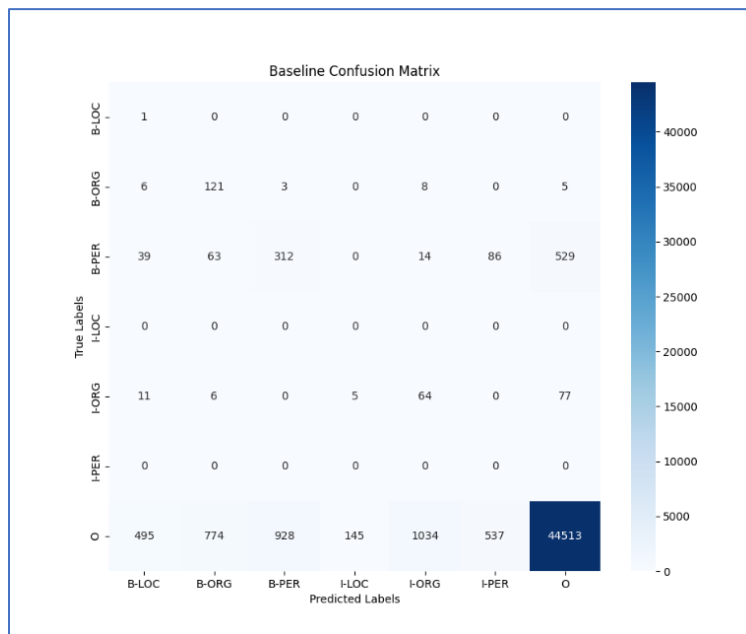
### Multinomijalni Bayes – Baseline model

Podela podataka na trening skup i testni skup urađena je preko unakrsne validacije u 10 slojeva pri čemu test set koji se koristi predstavlja kombinaciju 10 test setova, svaki od kojih predstavlja test set izabran u jednoj etapi unakrsne validacije. Za trening baseline modela korišćeni su vektori koji predstavljaju trenutni token, kapitalizaciju trenutnog tokena, prethodna 2 tokena (izabran broj tokena kako se povećavanjem broja nisu postizale bolje performanse), kapitalizacija prethodna 2 tokena kao i redni broj datog tokena u rečenici

Program i izveštaji u okviru projekta za model MB se nalazi u folderu faza-3/Baseline. Python skripta iskoristena za to je main.py. Izveštaj za model koji results.txt.



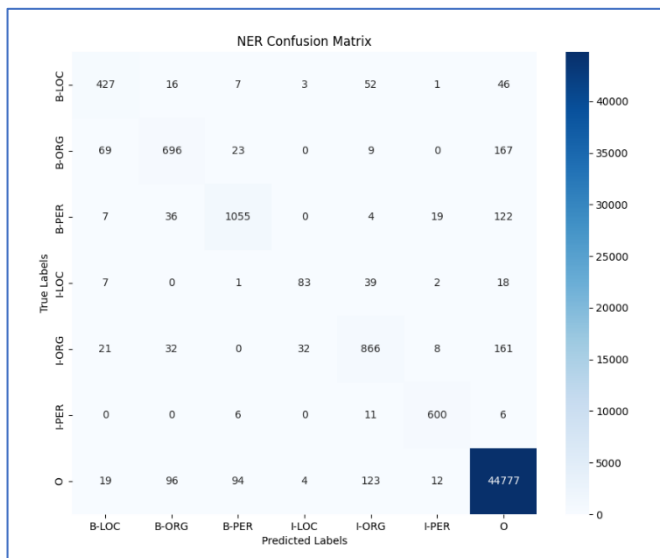
Slika 2. – MB matrica konfuzije za klase bez prefiksa



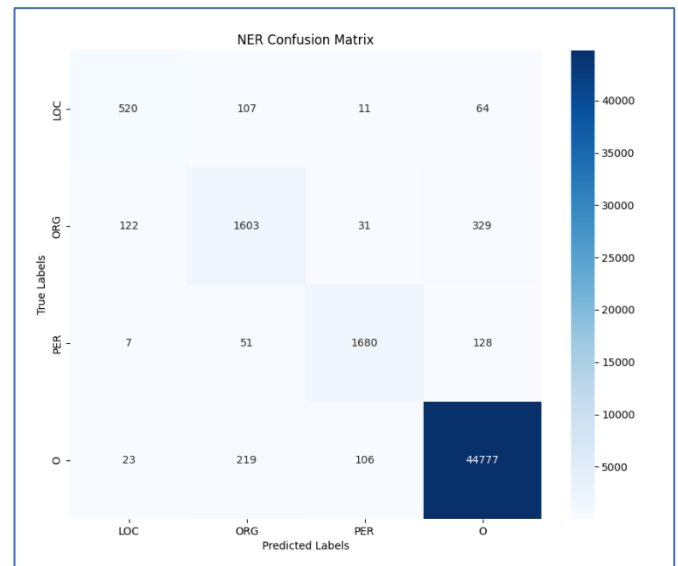
Slika 3. MB matrica konfuzije za klase sa prefiksima B- i I-

## BERTić NER

Program i izveštaji u okviru projekta za model BERTic se nalazi u folderu faza-3/bertic. Python skripta iskoritena za to je bertic.py. Izveštaj za model koji razmatra prefikse je classification\_report.txt a bez prefiksa classification\_report\_no\_prefix.txt.



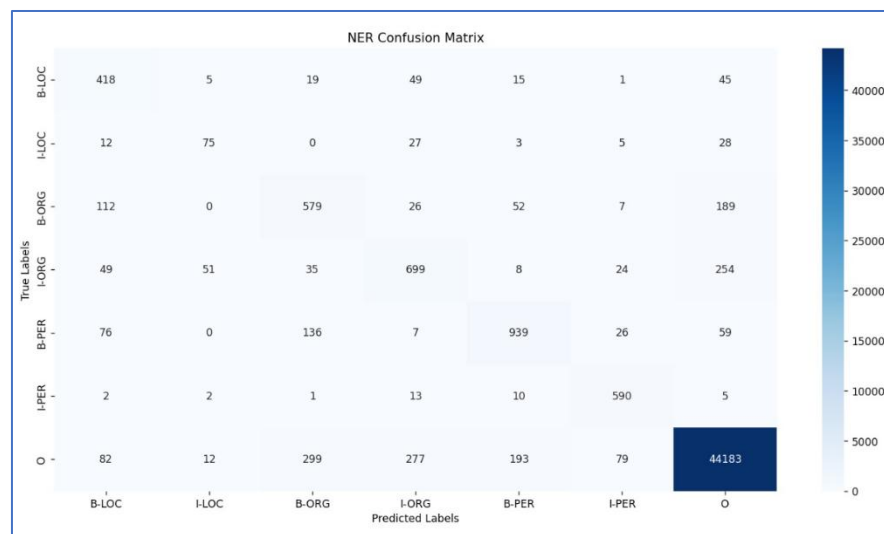
Slika 4. BERTić – matrica konfuzije za klase sa prefiksima B- i I-



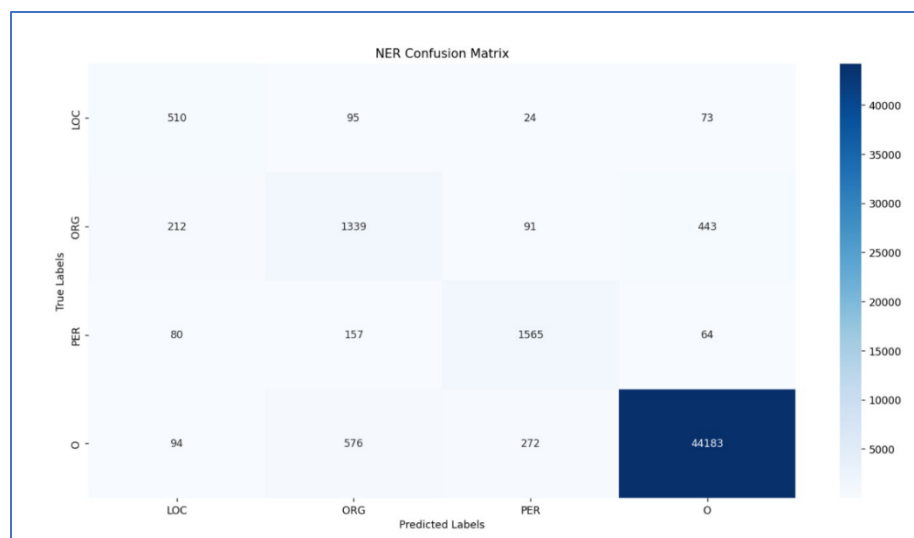
Slika 5. BERTić – matrica konfuzije za klase bez prefiksa

## CLASSLA – standardni (novinski) jezik

Program i izvestaji u okviru projekta za model CLASSLA za standardni jezik se nalazi u folderu faza-3/classla. Python skripta iskoritena za to je evaluate\_classla\_standard.py. Izveštaj za je classla/out/classla\_standard\_report\_mapped.txt.



Slika 6. CLASSLA std – matrica konfuzije za klase sa prefiksima B- i I-

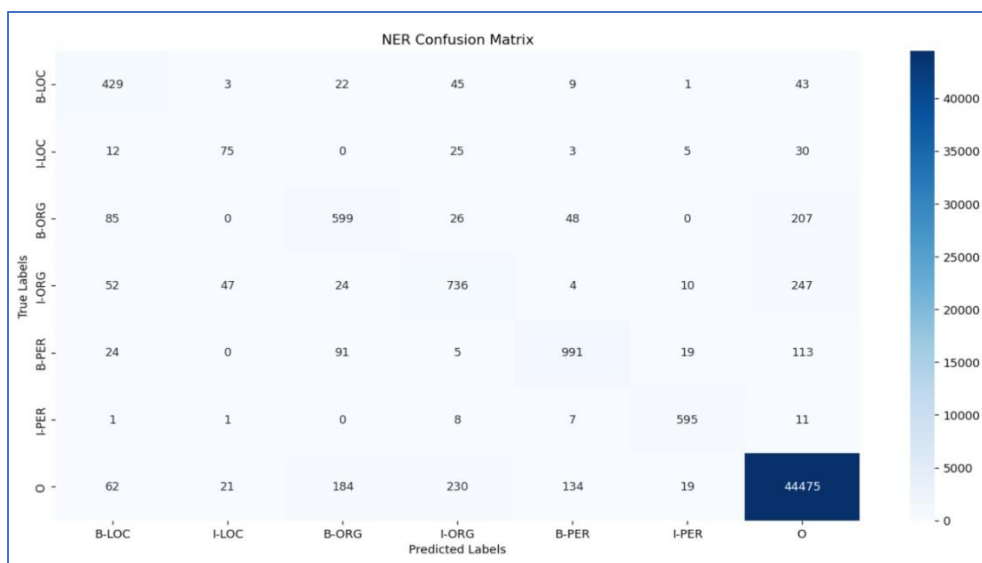


Slika 7. CLASSLA std – matrica konfuzije za klase bez prefiksa

## CLASSLA – nestandardni jezik (Tviter)

Program i izvestaji u okviru projekta za model CLASSLA za nestandardni jezik se nalazi u folderu faza-3/classla. Python skripta iskoritena za to je evaluate\_classla\_nonstandard.py. Izveštaj za je classla/out/classla\_nonstandard\_report\_mapped.txt.





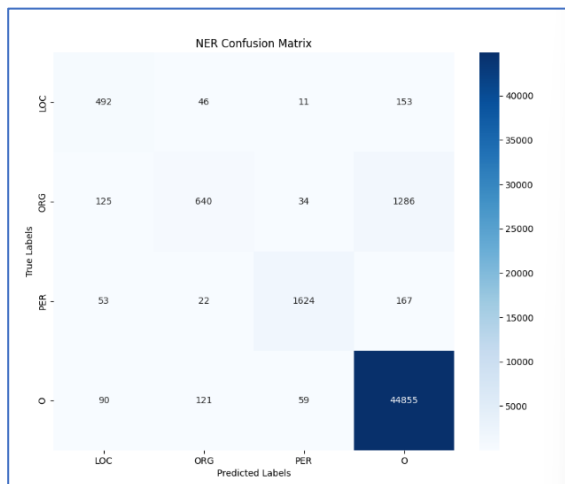
Slika 8. CLASSLA nonstd – matrica konfuzije za klase bez prefiksa



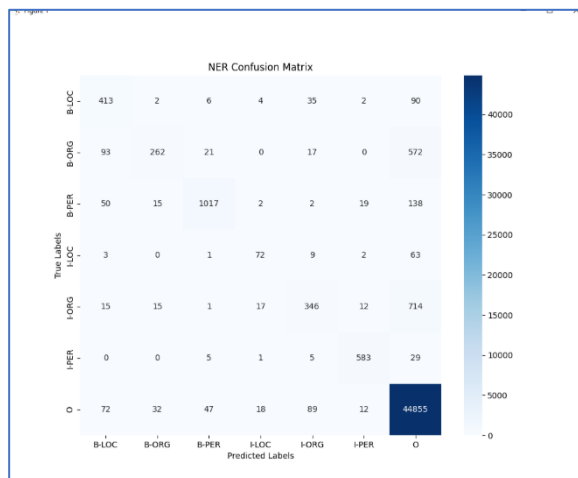
Slika 9. CLASSLA nonstd – matrica konfuzije za klase bez prefiksa

## COMtext.SR NER

Program i izvestaji u okviru projekta za model COMtext.SR se nalazi u folderu faza-3/COMtext.sr. Python skripta iskoritena za to je main2.py. Izveštaj za model koji razmatra prefikse je classification\_report.txt a bez prefiksa classification\_report\_no\_prefix.txt.



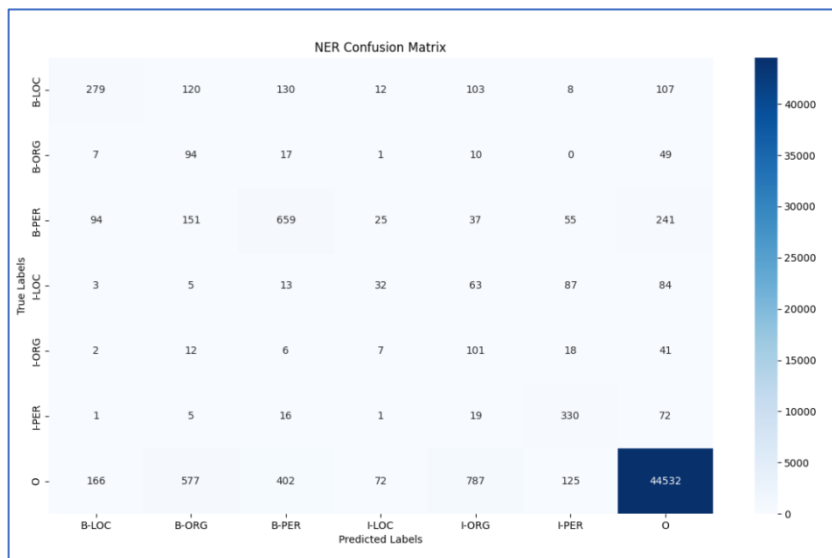
Slika 10. COMtext.SR – matrica konfuzije za klase sa prefiksima B- I I-



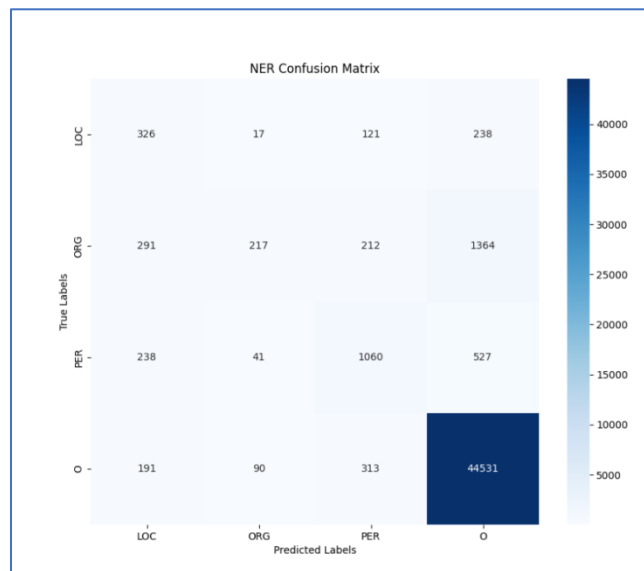
Slika 11. COMtext.SR – matrica konfuzije za klase bez prefiksa

## SrpCNNER

Program i izveštaji u okviru projekta za model SrpCNNER se nalazi u folderu faza-3/srbCNNER. Python skripta iskoristena za to je main.py. Izveštaj za model koji razmatra prefikse je classification\_report.txt a bez prefiksa classification\_report\_no\_prefix.txt.



Slika 12. SrpCNNER – matrica konfuzije za klase sa prefiksima B- I I-



Slika 13. SrpCNNER – matrica konfuzije za klase bez prefiksa

# Tabelarni prikaz rezultata – sa prefiksima

Tabela 5. Tabelarni prikaz grupisani po klasama

| Klasa        | Model                  | Precision | Recall | F1     |
|--------------|------------------------|-----------|--------|--------|
| <b>B-PER</b> | MB                     | 0,2991    | 0,2510 | 0,2730 |
|              | CLASSLA - standardni   | 0,7697    | 0,7554 | 0,7625 |
|              | CLASSLA - nestandardni | 0,8286    | 0,7973 | 0,8126 |
|              | BERTić-NER             | 0,8895    | 0,8488 | 0,8687 |
|              | COMtext.SR NER         | 0,9262    | 0,8182 | 0,8689 |
|              | SrpCANNER              | 0,5222    | 0,5302 | 0,5261 |
| <b>I-PER</b> | MB                     | 0,000     | 0,0000 | 0,0000 |
|              | CLASSLA - standardni   | 0,8060    | 0,9470 | 0,8708 |
|              | CLASSLA - nestandardni | 0,9168    | 0,9551 | 0,9355 |
|              | BERTić-NER             | 0,9346    | 0,9631 | 0,9486 |
|              | COMtext.SR NER         | 0,9262    | 0,9358 | 0,9306 |
|              | SrpCANNER              | 0,7432    | 0,5297 | 0,6186 |
| <b>B-ORG</b> | MB                     | 0,8462    | 0,1254 | 0,2184 |
|              | CLASSLA - standardni   | 0,5416    | 0,6000 | 0,5693 |
|              | CLASSLA - nestandardni | 0,6511    | 0,6207 | 0,6355 |
|              | BERTić-NER             | 0,7945    | 0,7220 | 0,7565 |
|              | COMtext.SR NER         | 0,8037    | 0,2715 | 0,4059 |
|              | SrpCANNER              | 0,5281    | 0,0975 | 0,1646 |
| <b>I-ORG</b> | MB                     | 0,3926    | 0,0571 | 0,0998 |
|              | CLASSLA - standardni   | 0,6366    | 0,6241 | 0,6303 |
|              | CLASSLA - nestandardni | 0,6847    | 0,6571 | 0,6706 |
|              | BERTić-NER             | 0,7844    | 0,7732 | 0,7788 |
|              | COMtext.SR NER         | 0,6879    | 0,3089 | 0,4264 |
|              | SrpCANNER              | 0,5401    | 0,0902 | 0,1546 |
| <b>B-LOC</b> | MB                     | 1,00      | 0,0018 | 0,0036 |
|              | CLASSLA - standardni   | 0,5566    | 0,7572 | 0,6416 |
|              | CLASSLA - nestandardni | 0,6451    | 0,7772 | 0,7050 |
|              | BERTić-NER             | 0,7764    | 0,7736 | 0,7750 |
|              | COMtext.SR             | 0,6393    | 0,7482 | 0,6895 |

|              |                        |        |        |        |
|--------------|------------------------|--------|--------|--------|
| <b>I-LOC</b> | NER                    |        |        |        |
|              | SrpCANNER              | 0,3676 | 0,5054 | 0,4256 |
|              | MB                     | 0,0000 | 0,0000 | 0,0000 |
|              | CLASSLA - standardni   | 0,5172 | 0,5000 | 0,5085 |
|              | CLASSLA - nestandardni | 0,5102 | 0,5000 | 0,5051 |
|              | BERTić-NER             | 0,6803 | 0,5533 | 0,6103 |
|              | COMtext.SR             | 0,6316 | 0,4800 | 0,5455 |
| <b>O</b>     | NER                    |        |        |        |
|              | SrpCANNER              | 0,1115 | 0,2133 | 0,1465 |
|              | MB                     | 0,9192 | 0,9865 | 0,9516 |
|              | CLASSLA - standardni   | 0,9870 | 0,9791 | 0,9831 |
|              | CLASSLA - nestandardni | 0,9856 | 0,9856 | 0,9856 |
|              | BERTić-NER             | 0,9885 | 0,9923 | 0,9904 |
|              | COMtext.SR             | 0,9654 | 0,9940 | 0,9795 |
|              | NER                    |        |        |        |
|              | SrpCANNER              | 0,9544 | 0,9868 | 0,9703 |

Tabela 6. Tabelarni prikaz macro rezultata za modele

| Model                  | Macro Precision | Macro Recall | Macro F1 | Accuracy |
|------------------------|-----------------|--------------|----------|----------|
| MB                     | 0,4321          | 0,1777       | 0,1933   | 0,9042   |
| CLASSLA - standardni   | 0,6878          | 0,7376       | 0,7094   | 0,9539   |
| CLASSLA - nestandardni | 0,7460          | 0,7561       | 0,7500   | 0,9623   |
| BERTić-NER             | 0,8355          | 0,8037       | 0,8183   | 0,9744   |
| COMtext.SR             | 0,7971          | 0,6509       | 0,6923   | 0,9552   |
| NER                    |                 |              |          |          |
| SrpCANNER              | 0,5382          | 0,4219       | 0,4295   | 0,9246   |

Tabelarni prikaz rezultatata – bez prefiksa

Tabela 7. Tabelarni prikaz rezultata grupisanih po klasama

| Klasa | Model                  | Precision | Recall | F1     |
|-------|------------------------|-----------|--------|--------|
| PER   | MB                     | 0,3816    | 0,2133 | 0,2736 |
|       | CLASSLA - standardni   | 0,8017    | 0,8387 | 0,8198 |
|       | CLASSLA - nestandardni | 0,8737    | 0,8639 | 0,8688 |
|       | BERTić-NER             | 0,9190    | 0,9003 | 0,9096 |
|       | COMtext.SR NER         | 0,9398    | 0,8703 | 0,9037 |
|       | SrpCANNER              | 0,6213    | 0,5681 | 0,5935 |
| ORG   | MB                     | 0,6503    | 0,0954 | 0,1665 |
|       | CLASSLA - standardni   | 0,6179    | 0,6422 | 0,6298 |
|       | CLASSLA - nestandardni | 0,6942    | 0,6643 | 0,6789 |
|       | BERTić-NER             | 0,8096    | 0,7688 | 0,7887 |
|       | COMtext.SR NER         | 0,7720    | 0,3070 | 0,4393 |
|       | SrpCANNER              | 0,5945    | 0,1041 | 0,1772 |
| LOC   | MB                     | 1,0000    | 0,0014 | 0,0028 |
|       | CLASSLA - standardni   | 0,5692    | 0,7265 | 0,6383 |
|       | CLASSLA - nestandardni | 0,6392    | 0,7393 | 0,6856 |
|       | BERTić-NER             | 0,7738    | 0,7407 | 0,7569 |
|       | COMtext.SR NER         | 0,6474    | 0,7009 | 0,4393 |
|       | SrpCANNER              | 0,3117    | 0,4644 | 0,3730 |
| O     | MB                     | 0,9192    | 0,9865 | 0,9516 |
|       | CLASSLA - standardni   | 0,9870    | 0,9791 | 0,9831 |
|       | CLASSLA - nestandardni | 0,9856    | 0,9856 | 0,9856 |
|       | BERTić-NER             | 0,9885    | 0,9923 | 0,9904 |
|       | COMtext.SR NER         | 0,9654    | 0,9940 | 0,9795 |
|       | SrpNNER                | 0,9544    | 0,9868 | 0,9703 |

Tabela 8. Tabelarni prikaz macro rezulata za modele

| Model | Macro Precision | Macro Recall | Macro F1 | Accuracy |
|-------|-----------------|--------------|----------|----------|
| MB    | 0,5902          | 0,2593       | 0,2789   | 0,9062   |

|                               |        |        |        |        |
|-------------------------------|--------|--------|--------|--------|
| <b>CLASSLA - standardni</b>   | 0,7440 | 0,7966 | 0,7677 | 0,9562 |
| <b>CLASSLA - nestandardni</b> | 0,7982 | 0,8133 | 0,8047 | 0,9641 |
| <b>BERTić-NER</b>             | 0,8727 | 0,8505 | 0,8614 | 0,9759 |
| <b>COMtext.SR NER</b>         | 0,8312 | 0,7180 | 0,7489 | 0,9565 |
| <b>SrpCANNER</b>              | 0,6205 | 0,5309 | 0,5285 | 0,9268 |

## Diskusija

U prethodnim tabelama prikaza su rezultati evaluacije različitih NER modela. Model MB je, pored evaluacije, treniran nad prikupljenim skupom podataka, i predstavlja baseline kao referentna tačka za merenje performansi ostalih modela. Svi modeli su evaluirani nad istom skupu podataka, sa i bez BIO prefiksa.

Metrike koje su korišćene obuhvataju Precision, Recall, F1-score za svaku klasu i makro-proseci, kao i Accuracy nad celim skupom podataka.

- Opšti pregled performansi

Opšti pregled performansi dat je u tabelama 6 i 8, gde su predstavljani makro rezultati svih modela. Time se dobija uvid u celokupne performanse modela nad svim klasama. U obe varijante evaluacije modela, model BERTić-NER sa vrednostima macro F1=0.8183 i macro F1 = 0.8614, sa i bez BIO prefiksa respektivno, pokazuje superiornost u odnosu na druge modele u zadatku razumevanja konteksta srpskog jezika. Modeli CLASSLA (nestandardni i standardni) takođe ostvaruju dobre rezultate (macro F1 u opsegu 0.75–0.80). Posebno je zanimljivo da ostvaruju konkurentne rezultate i pri radu sa tekstovima koji ne odgovaraju tipu jezika nad kojim je tip modela CLASSLA treniran. Model COMtext.SR NER pokazuje solidne performanse, pogotovo dobro izražene u prepoznavanju klase PER, ali sa nešto slabijim rezultatima kod organizacija i lokacija. Može se pretpostaviti, da model ima dobro kontekstualno shvatanje, kao i BERTić zbog transformer arhitekture, ali da ukazuje određena ograničenja u slučaju entiteta sa promenjivim dužinama i entitetima koji se mogu naći u više klasa. SrpCANNER beleži niže rezultate, ukazujući da konvolutivni pristup funkcioniše donekle ali ne dostiže sposobnost kontekstualnog shvatanja kao transformer modeli. MB baseline ostvaruju najslabije rezultate, s tim da baseline služi kao donja granica performansi.

Vrednosti accuracy-ja su genralno visoki kod svih modela, što ukazuje na dominaciju klase O u odnosu na druge klase, time ne dajući dobar uvid u tačnost i efikasnost modela modela.

- Analiza po klasama

Analiza po klasama data je u tabeli 7 i 9. Primećeno je da su modeli imali najbolje rezultate u prepoznavanju klase PER (lična imena). Najverovatnije zbog učestalosti i karakteristike kapitalizacije imena.

Klasa ORG (organizacije) rezultati su slabiji u odnosu na PER, naročito kod jednostavnijih modela. Ukazujući da je klasa teža za prepoznavanje, imajući u vidu da organizacije mogu sadržati različite dužine i strukture, često sa elementima koji se preklapaju sa imenima i lokacijama, što dodatno otežava situaciju. Kod klase LOC (lokacija), modeli imaju niži rezultat. Naime broj tokena za LOC je znatno manji u odnosu na broj tokena ostalih klasa, i ima dosta preklapanja sa klasom ORG, što je moguće da je dovelo do slabijih performansi generalno nad klasom LOC. O klasa svi modeli imaju visok F1 preko 0.95, što je razumljivo jer klasa O sadrži oko 90% sveukupnih podataka.

- Uticaj prefiksa (BIO oznaka)

Kod svih modela se vidi blago povećanje F1 kada se koristi bez prefiksa. Ukazuje na deo grešaka koji potiče iz netačne segmentacije entiteta. Međutim BERTić, CLASSLA i COMtext.SR najmanje gube performanse kad su prefiski uključeni, što znači da dobro razumeju strukturu entiteta.

- Zaključak

Na osnovu prikazanih rezultata, može se zaključiti da su transformer modeli (BERTić-NER) najefikasniji za izvršavanje zadatka prepoznavanja imenovanih entiteta na srpskom jeziku. CLASSLA modeli predstavljaju dobru alternative, naročito u uslovima nestandardnog jezika. COMText.SR je pouzdan model koji ima dobro kontekstualno shvatanje, međutim ne dostiže dobre performanse u svim klasama. Tradicionalni modeli i modeli zasnovani na konvoluciji, postižu zadovoljavajuće rezultate u osnovnim slučajevima, ali su manjkaju sposobnost generalizacije.