Универзитет у Београду

Електротехнички факултет



Пројекат на предмету обрада природних језика

|  |  |
| --- | --- |
| Ментор: | Кандидати: |
| Др Вук Батановић | Алекса Вучковић  Милан Боићевић  Милица Јевтовић |

Београд, Септембар 2025.

Садржај

[Садржај i](#_Toc209644335)

[1. Увод 1](#_Toc209644336)

[2. Прикупљање података 2](#_Toc209644337)

[2.1. Новински текстови 2](#_Toc209644338)

[2.2. Књижевни текстови 2](#_Toc209644339)

[2.3. Правно-административни текстови 3](#_Toc209644340)

[2.4. Текстови са твитера/X-a 3](#_Toc209644341)

[3. Анотација података 4](#_Toc209644342)

[3.1. Правила анотације 4](#_Toc209644343)

[4. Евалуација статистичких модела 5](#_Toc209644344)

[4.1. Baseline приступ – мултиномијални наивни Бајесов класификатор 5](#_Toc209644345)

[4.2. CLASSLО 5](#_Toc209644346)

[4.2.1. Стандардни језик 5](#_Toc209644347)

[4.2.2. Нестандардни језик 5](#_Toc209644348)

[4.3. BERTић 5](#_Toc209644349)

[4.4. COMtext.SR 5](#_Toc209644350)

1. Увод

Пројекат из предмета **Обрада природних језика** за школску 2024/2025. годину реализован је од стране групе студената:

* Алекса Вучковић, 2024/3040
* Милан Богићевић, 2024/3056
* Милица Јевтовић, 2024/3113

Тема пројекта јесте израда **baseline приступа и евалуација напреднијих модела за препознавање именованих ентитета (Named Entity Recognition – NER)** на српском језику. Фокус је на три основна типа ентитета – **особе (PER), локације (LOC) и организације (ORG)**, при чему је коришћен IOB2 систем означавања. Пројекат је имплементиран у програмском језику **Python**.

Израда пројекта одвија се кроз три главне фазе:

1. **Прикупљање података** – формирање скупа текстова из различитих тематских домена (новински, књижевни, правно-административни, твитер), њихово чишћење и припрема за даље кораке.
2. **Анотација података** – ручно означавање именованих ентитета уз формулисање јасних упутстава и спровођење калибрације и анализе квалитета анотације.
3. **Евалуација статистичких модела** – поређење перформанси baseline приступа са више постојећих NER модела (CLASSLA, BERTић-NER, COMtext.SR) и детаљна анализа добијених резултата.

Цео пројекат се може наћи на github-у на следећем [линку](https://github.com/MilanBogicevic8/Opj-projekat).

1. Прикупљање података

У оквиру пројекта било је неопходно формирати довољно велики и разноврстан корпус текстова на српском језику, који ће након анотације послужити за евалуацију различитих NER модела. Да би се обезбедила доменска разноликост и покривање различитих стилова писања, подаци су прикупљани из четири тематска домена:

1. **Новински текстови**
2. **Књижевни текстови**
3. **Правно-административни текстови**
4. **Поруке са друштвених мрежа (Твитер/X)**

Сваки извор је обрађен посебно, при чему је извршено аутоматско преузимање садржаја, чишћење текста и складиштење у унапред дефинисаној структури директоријума. Након тога сви текстови су пргледани мануелно и избачени су подаци који нису релевантни за даљу анализу. Ови текстови су даље коришћени за евалуацију различитих модела, који ће бити описани у четвртом поглављу, тако да се пазило да подаци који су коришћени за евалуацију не буду исти као они који су се користили за тренирање модела.

У тиму сваки члан је био задужен за ову фазу:

* **Милан** је сакупљао податке из Устава и закона за домен правних текстова, као и податке за новински и књижевни домен.
* **Милица** је сакупљала податке са различитих сајтова за правни домен и за правни и новински домен издвајала делове који ће бити коришћени,
* **Алекса** је прикупљао податке за твитер домен.
  1. Новински текстови

Први извор података били су новински портали. За потребе пројекта одабран је сајт **nova.rs**, из секције „Вести“. Коришћен је Python програм заснован на библиотекама requests и BeautifulSoup, који је имао две основне функције:

* **Сакупљање линкова** – програм је прво слао HTTP GET захтев ка страници са вестима и прикупљао све хипервезе које воде ка појединачним чланцима. Линкови су филтрирани тако да се задрже само они који припадају домену https://nova.rs/vesti/. Дупликати су уклоњени, а јединствени линкови су снимљени у посебан фајл.
* **Преузимање чланака** – затим је сваки од прикупљених линкова обрађен тако што су извучени текстуални садржаји пасуса (<p> тагови). Текст је очишћен од вишеструких празнина и специјалних знакова, а чланци краћи од 100 карактера су прескочени како би се избегли нерелевантни уноси. Сваки чланак је сачуван као појединачан .txt фајл у UTF-8 енкодингу.
* **Издвајање одабраних података** – преузети подаци су сачувани у фолдеру /data/newspapers/izvor, након чега су из њих извучени делови од значаја у фолдеру /data/newspapres/izvuceno. За одабир извучених делова, користе се они који садрже значајан број организација, особа и локација.

Уз сваки текст уметнут је и запис у метаподатке, који садржи име фајла и оригинални URL. На тај начин обезбеђена је трајна веза између извора и локално сачуваног текста.

* 1. Књижевни текстови

За књижевни домен коришћен је роман **„Ана Карењина” Лава Толстоја**, који је у целини јавно доступан у дигиталном формату. Текст је директно преузет и није вршена додатна модификација осим нормализације празних места и нових редова.

При обради је вођено рачуна да корпус садржи приближно **5000 токена.**

* 1. Правно-административни текстови

За домен правно-административних докумената прикупљени су подаци са портала **Paragraf** (<https://www.paragraf.rs>), који садржи законе, правилнике и друге прописе Републике Србије. Поред аутоматског преузимања већег броја докумената, као репрезентативан пример укључен је и **„Устав Републике Србије”**, који представља један од најзначајнијих правних текстова. Ови документи су сачувани у фолдеру /data/administrative\_data/paragraf\_rs.

Скрапинг је реализован помоћу библиотека **requests** и **BeautifulSoup**, при чему су издвајани пасуси из <p> елемената HTML страница. Сваки документ је сачуван као посебан .txt фајл у директоријуму, док су метаподацима (име фајла и изворни линк) смештени у датотеци metadata.txt.

Поред тога су ручно издвојени документи са сајтова Агенције за привредне регистре, Врховног суда, Министарства унутрашње и спољне трговине, Централног регистра обавезних социјалних осигурања, Правосудне академије и сличних. Сви линкови за коришћене документе се налазе у метаподацима. Поједини документи за које би линк аутоматски сачувао фајл се налазе фолдеру /data/administrative\_data/izvor.

Из наведених извора су извучени значајни делови фајлова и они се налазе у фолдеру /data/administrative\_data/izvuceno. Како су поједини фајлови писани у ћирилици, конвертовани су у латиницу и они се налазе у фоледру /data/administrative\_data/izvuceno\_latinica.

* 1. Текстови са твитера/X-a

Текстови са твитера/X-a Подаци су прикупљени претрагом на страници x.com/search, користећи филтере lang, since, until, min\_retweets, а подаци добављени у json формату су затим парсирани како би се издвојиле основне информације – текстуални садржај објаве и одговарајућа URL адреса. Садржај поруке је филтриран тако да су најпре избачени сувишни карактери (са уникод вредношћу већом или једнаком 0x10000), а затим одбачене објаве са мање од четрнаест речи, на основу размака. Подаци су смештени у data/twitter/metadata.json.

1. Анотација података

У овој фази било је потребно сваки од прикупљених текстова ручно обележити у погледу присуства основних именованих ентитета (особе – **PER**, локације – **LOC**, организације – **ORG**) по систему **IOB2**.

У нашем тиму сваки члан је анотирао један тематски домен:

* **Милан** је анотирао домен правних текстова,
* **Милица** новинске чланке,
* **Алекса** твитер постове и књижевни домен.

Поред тога, књижевне текстове смо сви заједно анотирали, тако што је сваки члан узео половину од укупног скупа, што износи око **2500 токена** по особи. На основу тих података извршена је **калибрација**, која ће бити детаљније описана у подпоглављу о калибрацији.

Током израде смо се међусобно договарали и дефинисали правила за анотацију у случају двосмислених ситуација или случајева у којима је било могуће поступити на више начина. На тај начин је постигнута већа доследност у генерисаним ознакама.

У наредном потпоглављу ће бити описана правила за анотирање података, а у наредном подпоглављу ће бити представљен поступак калибрације података.

* 1. Правила анотације
     1. Скраћенице

Уколико се наиђе на податак који представља **скраћеницу**, она ће бити анотирана у зависности од тога шта представља.

* Пример: **Bgd.** → LOC
* Пример: **UN** → ORG
  + 1. Град и адреса као јединствена локација

Уколико се **град налази уз адресу** и сматра се делом ње, онда се **цела целина анотира као једна локација**.

* Пример: „**Краља Петра 12, Београд**“ → B-LOC I-LOC I-LOC
* Пример: „**Немањина 4, Ниш**“ → B-LOC I-LOC
  + 1. Називи улица и бројеви

**Назив улице и број улице** увек се сматрају делом **локације** и заједно се анотирају као једна целина.

* Пример: „**Булевар Краља Александра 73**“ → B-LOC I-LOC I-LOC I-LOC
* Пример: „**Немањина 4**“ → B-LOC I-LOC
  + 1. Ентитети који се протежу преко више редова

Уколико се **назив улице или неки други ентитет протеже у више редова**, а токенизован је у различитим редовима, онда се **тип ентитета наставља у следећем реду**.

* Пример:

A white sheet of paper with black and red lines

AI-generated content may be incorrect.

* + 1. Реч „улица“ уз назив

Реч **„улица“** када се налази уз назив улице, сматра се делом **локације**, **без обзира да ли је написана малим или великим словом**.

* Пример:
  + „**Булевар Краља Александра**“ → B-LOC I-LOC I-LOC
  + „**улица Немањина**“ → B-LOC I-LOC
  + „**Улица Кнеза Милоша**“ → B-LOC I-LOC I-LOC
    1. Општи појмови који не означавају конкретну локацију

Речи или изрази који **не означавају конкретну локацију** анотирају се као **O**.

* Примери:
  + „**novobeogradske ulice**“ → O O
  + „**na raskrsnici**“ → O O
    1. Називи организација

Речи које се **односе на конкретну организацију** увек се анотирају као **ORG**.

* Примери:
  + „**policija**“ → B-ORG
  + „**vlada**“ → B-ORG
  + „**Univerzitet u Beogradu**“ → B-ORG I-ORG I-ORG
    1. Oпшти појмови који означавау организације

Речи или изрази који **не означавају конкретну организацију** анотирају се као **O.**

* Пример:
  + „**srpsko tužilaštvo**“ → O O
    1. Имена особа у називу организације

Ако се **назив особе појави у имену организације**, та целина се и даље анотира као **ORG**, а не као PER.

* Пример:
  + „**Univerzitet Nikola Tesla**“ → B-ORG I-ORG I-ORG
  + „**Fondacija Marko Marković**“ → B-ORG I-ORG I-ORG
    1. Назив места после организације

Уколико се након **назива организације** нађе и **назив места**, цела конструкција се анотира као **једна организација (ORG)**.

* Примери:
  + „**Policija Beograd**“ → B-ORG I-ORG
  + „**Vlada Republike Srbije**“ → B-ORG I-ORG I-ORG
  + „**Univerzitet u Nišu**“ → B-ORG I-ORG I-ORG
    1. Придеви уз назив организације

**Придеви који се налазе уз назив организације** се **не сматрају делом организације** и анотирају се као **O**.

* Пример:
* „**srpskoj Vojsci** “ → O B-ORG
  + 1. ZПоодсеци организација
* Уколико се **назив организације и њен пододсек** појаве заједно (нпр. „ETF – Одсек за рачунарство“), **анотирају се спојено као једна организација (ORG)**.
* Примери:
  + „**EТФ – Одсек за рачунарство**“ → B-ORG О I-ORG I-ORG I-ORG
    1. Скраћеница након назива организације

Ако се **скраћеница појави након пуне форме назива организације**, скраћеница се анотира као **I-ORG**.

* Примери:
  + „**Elektrotehnički fakultet – ETF**“ → B-ORG I-ORG O I-ORG
  + „**Univerzitet u Beogradu – UB**“ → B-ORG I-ORG I-ORG O I-ORG
  + „**Ministarstvo unutrašnjih poslova – MUP**“ → B-ORG I-ORG I-ORG O I-ORG
    1. Наводници у називу

Ако су **наводници саставни део назива** (нпр. назив организације или књиге), **не анотирају се као O**, већ се третирају као део назива.

* Пример:
  + „**Универзитет Никола Тесла“**“ → О B-ORG I-ORG I-ORG O
    1. Место после организације и ближа одредница

1. Ако је **назив места одвојен зарезом** од саме организације, **не улази у састав организације** и анотира се као **место (LOC)**.
2. Ако је за место дата **ближа одредница у заградама**, она се третира као **посебно место** и започиње са **B-LOC**.

* Примери:
  + „**Policija, Beograd**“ → B-ORG O B-LOC
  + „**Vlada, Novi Sad (Petrovaradin)**“ → B-ORG O B-LOC I-LOC O B-LOC O
    1. Именице у различитим предлошко-падежним конструкцијама

Ако се ради о **деривату имена који означава особу**, тај дериват се анотира као **PER**.

* Примери:
  + „**Vučićevi**“ → PER
    1. Иницијали особе

Ако се наведу **само иницијали особе**, они се анотирају као **PER**.

* Примери:
* „**M. B.**“ → B-PER O I-PER
* „**J. K.**“ → B-PER O I-PER
  + 1. Уметничка имена и титуле

1. **Уметничка имена** се анотирају као **PER**.

* Примери:
  + „**DJ Žeks**“ → B-PER I-PER
  + „**Lady Gaga**“ → B-PER I-PER

1. **Титуле и звања** која претходе имену особе **не улазе у PER анотацију**.

* Примери:
  + „**председник Србије Александар Вучић**“ → O B-LOC B-PER I-PER
  + „**др Јован Јовановић**“ → O B-PER I-PER
    1. Називи бендова и група

Називи **музичких бендова или група** анотирају се као **ORG**.

* Примери:
  + „**Beogradski sindikat**“ → B-ORG I-ORG
  + „**Riblja Čorba**“ → B-ORG I-ORG
    1. Име особе са надимком

Ако се име особе појави са **надимком**, цела конструкција се анотира као **један PER ентитет**.

* Пример:
  + „**Aleksandar Vučić - Aca**“ → B-PER I-PER I-PER
    1. Образовне установе

Називи **образовних установа** анотирају се као **ORG**.

* Примери:
  + „**Факултет техничких наука**“ → B-ORG I-ORG I-ORG
  + „**Универзитет у Београду**“ → B-ORG I-ORG I-ORG
    1. Називи штампаних новина

Називи **штампаних новина** анотирају се као **ORG**.

* Примери:
  + „**Politika**“ → B-ORG
  + „**Blic**“ → B-ORG
    1. Називи мора и океана

Називи **мора и океана** анотирају се као **LOC**.

* Примери:
  + „**Атлантски океан**“ → B-LOC I-LOC
  + „**Јадранско море**“ → B-LOC I-LOC
    1. Називи у хаштаговима

Ако хаштаг садржи **назив локације, особе или организације**, он се анотира **према одговарајућем типу ентитета**.

* Примери:
  + „**#Beograd**“ → B-LOC
  + „**#Blic**“ → B-ORG
    1. Кориснички називи профила

Називи **корисничких профила на друштвеним мрежама** **се не анотирају**.

* Примери:
  + @avucic → O
  + @BlicOnline → O
  + @RTSinfo → O

Детаљнији примери се налазе на следећем [линку](https://docs.google.com/document/d/1pfpEGfowzCBZrc-sHqYqL8C5sd60J1b2J7TR5PA34Z0/edit?tab=t.0).

* 1. Калибрација

1. Евалуација статистичких модела

У овој фази пројекта фокус је на евалуацији различитих *NER* модела, почевши од базног приступа заснованог на статистичким методама, па све до напреднијих дубоких модела. Циљ је да се анализирају перформансе појединих модела, идентификују њихове предности и ограничења и изврши упоредна анализа резултата.

Прво се разматра ***baseline* модел** заснован на мултиномијалном наивном Бајесовом класификатору, где се сваки токен класификује независно. За ову евалуацију користи се 10-слојна унакрсна валидација, а као основне одлике модела укључују one-hot репрезентацију тренутног и претходних неколико токена, капитализацију токена и позицију токена у реченици.

Након тога, евалуација се проширује на напредније *NER* моделе као што су **CLASSLО, BERTić** и **COMtext.SR**, уз коришћење одговарајућих интерфејса и правилно мапирање излазних ознака на три основне класе именованих ентитета: **PER** (особе), **LOC** (локације) и **ORG** (организације). Поред тога, сваки модел се анализира у две варијанте: једној у којој се B- и I- ознаке третирају као одвојене класе и другој у којој се посматрају само типови ентитета, без разликовања B/I префикса.

* 1. Baseline приступ – мултиномијални наивни Бајесов класификатор

*Baseline* модел спроводи **класификацију појединачних токена** у оквиру *Named Entity Recognition* (енгл. *NER*) задатка користећи **Multinomial Naive Bayes**. Сваки токен се класификује **независно**, при чему се секвенцијалне зависности узимају у обзир само преко претходна два токена као додатних карактеристика.

* + 1. Карактеристике (енгл. features)

За сваки токен модел користи:

* **token\_lower** – токен у малим словима (корисно за нормализацију и смањење варијанти).
* **is\_capitalized** – буловска вредност која показује да ли је прво слово велико (често указује на имена и локације).
* **has\_hyphen** – буловска вредност која показује да ли токен садржи цртицу.
* **prefix\_1** – прво слово токена (може помоћи у идентификацији облика речи или ентитета).
* **suffix\_1** – последње слово токена (често користи за препознавање скраћеница или типова речи).
* **position** – редни број токена у реченици, што омогућава моделу да уочи да ли се ентитети чешће појављују на почетку или крају контекста.
* **prev\_token** – претходни токен у реченици (ако постоји), што омогућава делимично хватање локалног контекста.
* **prev2\_token** – токен два места пре тренутног, такође ради делимичне репрезентације контекста.
  + 1. Тренирање и евалуација модела
* Скуп података се прво дели на **80% података за cross-validation и 20% података за тестирање**, уз коришћење **стратификације** да би се очувала расподела класа.
* **10-fold cross-validation**: скуп података је подељен на 10 једнаких делова (енгл. fold). Модел се тренира на 9 делова и тестира на 10. делу. Овај процес се понавља 10 пута, сваки пут користећи други *fold* за тестирање.
* **Метрике за мерење перформанси**:
  + **Accuracy (тачност)**: проценат правилно класификованих токена.
  + **Confusion matrix (матрица конфузије)**: омогућава преглед где модел најчешће греши, које класе се мешају.
  + **Precision:** колики део токена предвиђених као одређена класа је заправо припадао тој класи.
  + **Recall:** колики део стварних токена одређене класе је модел правилно предвидео.
  + **F1-score:** хармонијска средина између precision и recall. Корисно је за класе са мањим бројем примера, где accuracy може бити превише обмањујућа.
  + **Macro F1-score / Macro avg:** просечан F1-score по свим класама без узимања у обзир броја примера по класи. Свака класа се третира једнако, па ова метрика добро показује перформансе на мање заступљеним класама.
  + **Micro avg:** узима у обзир све токене као једну групу и рачуна укупни F1-score (или precision/recall) преко свих токена, што тежи класама са више примера.
    1. Варијанте евалуације

Евалуација је урађена на скупу за тестирање који се добија стратификацијом 20% података из почетног скупа добијеног анотирањем.

**Одвојене класе (B-/I- тагови одвојено)**

* Класе: B-PER, I-PER, B-LOC, I-LOC, B-ORG, I-ORG, O
* 10-fold CV резултат: просечна тачност ≈ **0.839**
* Пример предикције:

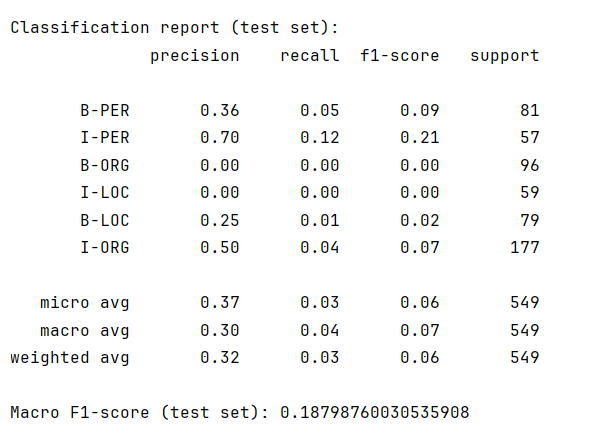
Beograd -> I-LOC

vlada -> I-LOC

Marko -> I-LOC

ETF -> B-ORG

У наставку су приказани резултати и анализа различитих метрика које су искоришћене на тест скупу података.

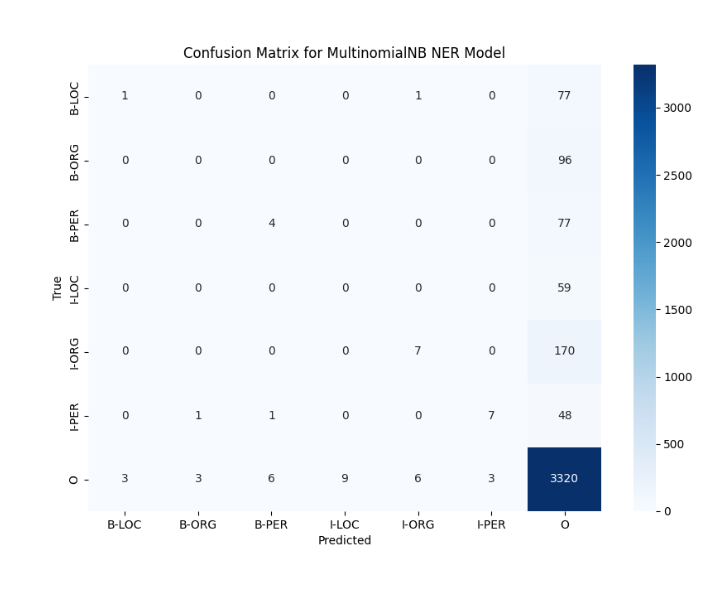


Када се разликују B- и I- тагови, резултати показују да модел има добру тачност на тест скупу (0.839), али ова тачност је већим делом последица доминантне класе O. Анализа извештаја класификације на тест скупу показује да модел веома слабо препознаје мање заступљене класе: почетак PER ентитета (B-PER) има низак recall (0.05) и F1-score (0.09), док I-PER има бољи precision (0.70) али и даље низак recall (0.12). Остале класе као што су B-ORG и I-LOC практично нису предвиђене, а B-LOC и I-ORG имају минималне вредности F1-score.

Метрике микро и макро просека показују сличан тренд: микро F1-score (0.06) узима све токене као једну групу и указује на лоше перформансе за мање заступљене ентитете, док макро F1-score (0.07) прави просек по свим класама и јасно показује да модел није способан да правилно класификује мање учестале B-/I- токене. Weighted F1-score такође остаје низак (0.06), јер доминантна класа O не компензује слабости на мање заступљеним класама.

Закључак је да модел добро препознаје доминантну класу O, али веома слабо идентификује почетке и наставке ентитета (B-/I-).

У наставку је приказана матрице конфузије које приказују резултате предикције модела.



Када су **B-** и **I-** класе одвојене, предвиђања модела су знатно лошија. Главни разлог је тај што број података у свакој ентитетској класи постаје веома мали у односу на доминантну класу **O**. Као резултат тога, модел често погрешно категорише токене који припадају **B/I** класама као O, што се види у матрици конфузије, где већина предвиђања пада у редове и колоне класе O. Ово илуструје колико дисбаланс у подацима утиче на *NER* модел и показује да раздвајање **B** и **I** класа може погоршати тачност предвиђања за стварне ентитете.

**Спојене класе (B-/I- префикси игнорисани)**

* Класе: PER, LOC, ORG, O
* 10-fold CV резултат: просечна тачност ≈ **0.847**
* Пример предикције:

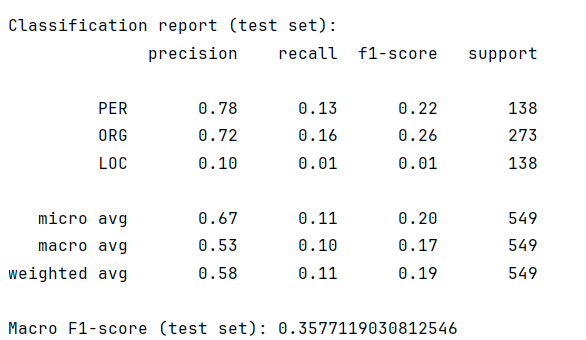
Beograd -> LOC

vlada -> ORG

Marko -> LOC

ETF -> ORG

У наставку су приказани резултати и анализа различитих метрика које су искоришћене на тест скупу података.



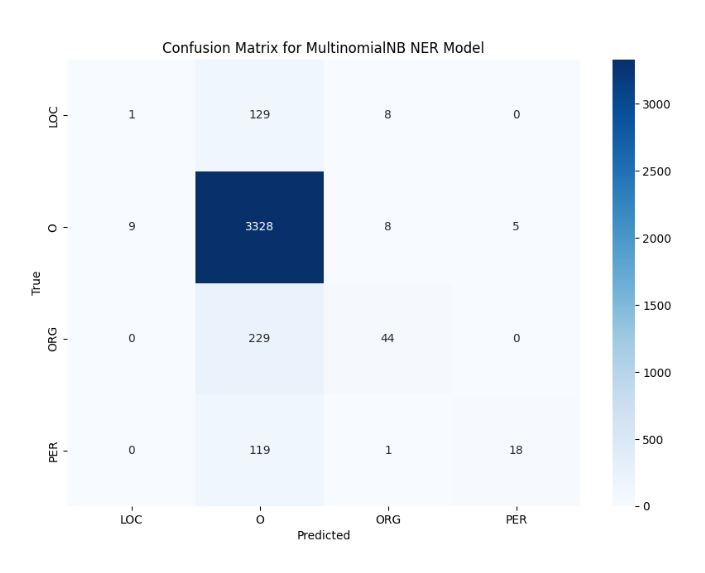
Када се B- и I- ознаке третирају као једна класа, модел постиже релативно боље резултате у односу на случај када се ове ознаке раздвоје као и у претходном случају на тест скупу, са релативно високом просечном тачношћу од 0.847. Ово је делимично последица доминације класе **O**, која чини већину токена у корпусу, па модел лако препознаје ове токене.

Међутим, евалуација на тест скупу и даље показује значајне слабости у детекцији стварних ентитета. За класе PER и ORG, модел постиже високу прецизност (редом 0.78 и 0.72), али веома низак recall (0.13 и 0.16), што указује да велики број ентитета остаје непримећен. Класа LOC постиже практично неупотребљиве резултате (precision 0.10, recall 0.01), што се може објаснити малим бројем примера за локације и доминацијом O токена.

Укупни macro F1-score од 0.358 и weighted average F1-score од 0.19 показују да, иако модел повремено исправно класификује ентитете, његова способност за идентификацију мање заступљених ентитетских класа је ограничена. Ови резултати илуструју да чак и када се B и I ознаке споје, наивни Бајесов приступ не даје добре резултате.

Спајање B и I ознака у једну класу је у овом случају повољније за наивни Бајесов модел јер омогућава бољу идентификацију ентитета уз минималне податке и смањује негативан утицај дисбаланса класа. Међутим, чак и у овом случају, модел има ограничену способност препознавања мање заступљених ентитета, што указује на потребу за коришћењем сложенијих *NER* приступа за практичне задатке.

У наставку је приказана матрице конфузије које приказују резултате предикције модела.

****

У овој матрици се види да је предикција за модел код кога су **B-** и **I-** класе спојене доста добра уколико се посматра главна дијагонала, што указује на то да модел правилно класификује већину токена у исправне категорије. Међутим, већина података припада класи **O**, што је типично за NER задаке, јер већина речи у тексту није ентитет. То може изазвати дисбаланс у класификацији, где модел има тенденцију да предвиђа класу O чешће него друге класе.

* + 1. Закључак
* *Baseline* модел пружа једноставну и брзу *NER* класификацију.
* Спојене класе олакшавају генерализацију и смањују сложеност.
* Резултати служе као основа за поређење са напреднијим моделима који ће бити описани у наредним потпоглављима.
  1. CLASSLО
     1. Стандардни језик
     2. Нестандардни језик
  2. BERTић
  3. COMtext.SR