Универзитет у Београду

Електротехнички факултет



Пројекат на предмету обрада природних језика

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

Београд, Септембар 2025.

Садржај

[Садржај i](#_Toc210370308)

[1. Увод 1](#_Toc210370309)

[2. Прикупљање података 2](#_Toc210370310)

[2.1. Новински текстови 2](#_Toc210370311)

[2.2. Књижевни текстови 3](#_Toc210370312)

[2.3. Правно-административни текстови 3](#_Toc210370313)

[2.4. Текстови са твитера/X-a 3](#_Toc210370314)

[3. Анотација података 4](#_Toc210370315)

[3.1. Правила анотације 4](#_Toc210370316)

[3.1.1. Скраћенице 4](#_Toc210370317)

[3.1.2. Град и адреса као јединствена локација 4](#_Toc210370318)

[3.1.3. Називи улица и бројеви 4](#_Toc210370319)

[3.1.4. Ентитети који се протежу преко више редова 5](#_Toc210370320)

[3.1.5. Реч „улица“ уз назив 5](#_Toc210370321)

[3.1.6. Општи појмови који не означавају конкретну локацију 5](#_Toc210370322)

[3.1.7. Називи организација 5](#_Toc210370323)

[3.1.8. Oпшти појмови који означавау организације 5](#_Toc210370324)

[3.1.9. Имена особа у називу организације 6](#_Toc210370325)

[3.1.10. Назив места после организације 6](#_Toc210370326)

[3.1.11. Придеви уз назив организације 6](#_Toc210370327)

[3.1.12. Пододсеци организација 6](#_Toc210370328)

[3.1.13. Скраћеница након назива организације 6](#_Toc210370329)

[3.1.14. Наводници у називу 7](#_Toc210370330)

[3.1.15. Место после организације и ближа одредница 7](#_Toc210370331)

[3.1.16. Именице у различитим предлошко-падежним конструкцијама 7](#_Toc210370332)

[3.1.17. Иницијали особе 7](#_Toc210370333)

[3.1.18. Уметничка имена и титуле 7](#_Toc210370334)

[3.1.19. Називи бендова и група 7](#_Toc210370335)

[3.1.20. Име особе са надимком 8](#_Toc210370336)

[3.1.21. Образовне установе 8](#_Toc210370337)

[3.1.22. Називи установа које се баве услужном делатношћу 8](#_Toc210370338)

[3.1.23. Називи штампаних новина 8](#_Toc210370339)

[3.1.24. Националности 8](#_Toc210370340)

[3.1.25. Називи мора и океана 9](#_Toc210370341)

[3.1.26. Називи у хаштаговима 9](#_Toc210370342)

[3.1.27. Кориснички називи профила 9](#_Toc210370343)

[3.2. Калибрација 9](#_Toc210370344)

[3.3. Анализа анотације 11](#_Toc210370345)

[4. Евалуација статистичких модела 12](#_Toc210370346)

[4.1. Baseline приступ – мултиномијални наивни Бајесов класификатор 12](#_Toc210370347)

[4.1.1. Карактеристике (енгл. features) 12](#_Toc210370348)

[4.1.2. Тренирање и евалуација модела 13](#_Toc210370349)

[4.1.3. Варијанте евалуације 13](#_Toc210370350)

[4.1.4. Закључак 17](#_Toc210370351)

[4.2. CLASSLА 18](#_Toc210370352)

[4.2.1. Стандардни језик 18](#_Toc210370353)

[4.2.2. Нестандардни језик 18](#_Toc210370354)

[4.2.3. Анализа 19](#_Toc210370355)

[4.3. BERTић 21](#_Toc210370356)

[4.3.1. Резултати евалуације 21](#_Toc210370357)

[4.4. COMtext.SR 22](#_Toc210370358)

[4.4.1. Резултати евалуације 22](#_Toc210370359)

[4.5. SrpCNNER 23](#_Toc210370360)

[4.5.1. Резултати евалуације 23](#_Toc210370361)

[4.6. Mатрице конфузије 24](#_Toc210370362)

[4.7. Анализа резултата 25](#_Toc210370363)

[4.8. Закључак 26](#_Toc210370364)

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 SrpCNNER) 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, елиминишу се текстови који се понављају и користе се делови које 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, ручно су 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: „**Kralja Petra 12, Beograd**“ → B-LOC I-LOC I-LOC I-LOC
* Пример: „**Nemanjina 4, Niš**“ → B-LOC I-LOC I-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: „**Bulevar Kralja Aleksandra 73**“ → B-LOC I-LOC I-LOC I-LOC
* Пример: „**Nemanjina 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~~**~~.~~ У најновијим изменама је избегнуто преламање једне реченице у више редова, па овај случај више није могућ.

* 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:
  + „**Bulevar Kralja Aleksandra**“ → B-LOC I-LOC I-LOC
  + „**ulica Nemanjina**“ → B-LOC I-LOC
  + „**Ulica Kneza Miloša**“ → 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**. Док се речи које означавају опште појмове који могу бити организација анотирају са О.

* Примери:
  + „**policija**“ → О
  + „**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. Пододсеци организација
* Уколико се **назив организације и њен пододсек** појаве заједно (нпр. „ETF – Одсек за рачунарство“), **анотирају се спојено као једна организација (ORG)**. Интерпункцијски знакови улазе у назив организације.
* Примери:
  + „**ETF – Odsek za računarstvo**“ → B-ORG I-ORG I-ORG I-ORG I-ORG
  + „**Stanica - servisa za savremeni ples**“→ B-ORG I-ORG I-ORG I-ORG I-ORG
    1. Скраћеница након назива организације

Ако се **скраћеница појави након пуне форме назива организације**, скраћеница се анотира као **B-ORG**, уколико су раздвојене неким интерпункцијским знаком.

* 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:
  + „**Univerzitet Nikola Tesla“** → B-ORG I-ORG I-ORG I-ORG I-ORG
    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 I-PER
* „**J. K.**“ → B-PER 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:
  + „**predsednik Srbije Aleksandar Vučić**“ → O B-LOC B-PER I-PER
  + „**dr Jovan Jovanović**“ → 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** уколико нема интерпункцијских знакова, али уколико се јави интерпункцијски знак, онда се надимак посмтра као одвојен ентитет.

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

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

* Примери:
  + „**Fakultet tehničkih nauka**“ → B-ORG I-ORG I-ORG
  + „**Univerzitet u Beogradu**“ → B-ORG I-ORG I-ORG
    1. Називи установа које се баве услужном делатношћу

За овакве типове установа са LOC ће бити анотирано све што улази у сам назив установе. Док ће ближе одреднице типа установе бити анотиране са О.

* Примери:
  + „**Tržni centar Ušće**“ → B-LOC I-LOC I-LOC
  + „**hotel Hajat**“→ O B-LOC
    1. Називи штампаних новина

Називи **штампаних новина** анотирају се као **ORG**. Уколико се јави нека ближа одредница за тип новина, она се анотира са О.

* Примери:
  + „**Politika**“ → B-ORG
  + „**Blic**“ → B-ORG
  + „**nedeljnik** **Vreme**“ → O B-ORG
    1. Националности

Све националности су анотиране са О.

* Примери:
  + „**Srbi**“ → O
  + „**Hrvati**“ → O
    1. Називи мора и океана

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

* Примери:
  + „**Atlantski okean**“ → B-LOC I-LOC
  + „**Jadransko more**“ → 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

За процес калибрације одабрана је прва половина књижевног домена. Помоћу скрипте су израчунати бинарни степени сагласности. Милан и Милица су у сагласности 99.141%, Милан и Алекса 99.313%, док су Милица и Алекса у сагласности 99.347%. Просек бинарних сагласности износи 99.267%. Грешке које су се појављивале су издвојене у табели.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Број линије** | **Токен** | **Милан** | **Милица** | **Алекса** | **Већинска одлука** | **Објашњење** |
| 48 | guvernantom | B-PER | O | O | O | Није особа. |
| 49 | Francuskinjom | I-PER | O | O | O |
| 211 | Stepan | B-PER | O | B-PER | B-PER | Пропуштен почетак ентитета. |
| 212 | Arkadijevič | I-PER | B-PER | I-PER | I-PER |
| 215 | Stiva | I-PER | I-PER | B-PER | I-PER | Надимак након "-" не сматрамо истим ентитетом. |
| 792 | Doli | O | B-PER | O | O | Јесте особа. |
| 1013 | Doli | B-PER | O | O | O | Јесте особа. |
| 1054 | Stepan | O | B-PER | B-PER | B-PER | Анотирани погрешни редови. |
| 1055 | Arkadijevič | B-PER | I-PER | I-PER | I-PER |
| 1056 | . | I-PER | O | O | O |
| 1460 | m-lle | B-PER | O | O | O | Титулу не сматрамо делом имена. |
| 1461 | Roland | I-PER | B-PER | B-PER | B-PER |
| 1722 | upita | O | B-PER | O | O | Анотирани погрешни редови. |
| 1723 | Stepan | B-PER | I-PER | B-PER | B-PER |
| 1724 | Arkadijevič | I-PER | O | I-PER | I-PER |
| 1800 | Stepana | B-PER | I-PER | B-PER | B-PER | Почетак ентитета анотиран са тагом I-, а унутрашњост прескочена. |
| 1801 | Arkadijeviča | I-PER | O | I-PER | I-PER |
| 1822 | Matvej | B-PER | O | B-PER | B-PER | Јесте особа. |
| 2126 | Matvej | B-PER | B-PER | O | B-PER | Јесте особа. |
| 2227 | Stepan | B-PER | B-PER | O | B-PER | Јесте особа. |
| 2278 | Matvej | B-PER | B-PER | B-PER | B-PER | Размак на почетку анотације. |
| 2843 | Stepan | I-PER | B-PER | B-PER | B-PER | Почетак ентитета анотиран са тагом I-, а унутрашњост прескочена. |
| 2844 | Arkadijevič | O | I-PER | I-PER | I-PER |
| 3069 | Liberalna | B-ORG | O | B-ORG | B-ORG | Реч "је" не припада ентитету. |
| 3070 | je | I-ORG | O | O | O |
| 3071 | partija | O | O | I-ORG | O |
| 3084 | Arkadijevič | I-PER | B-PER | I-PER | I-PER | Означено као почетак ентитета али је унутрашњост. |
| 3149 | Liberalna | B-PER | B-ORG | B-ORG | B-ORG | Јесте организација. |
| 3151 | partija | I-PER | I-ORG | I-ORG | I-ORG |
| 3219 | Arkadijevič | I-PER | B-PER | I-PER | I-PER | Означено као почетак ентитета али је унутрашњост. |
| 3252 | Ruriku | B-PER | B-PER | B-LOC | B-PER | Јесте особа. |
| 3342 | vlada | O | O | B-ORG | O | Није организација - не односи се на конкретну. |

* 1. Анализа анотације

Резултати појављивања ознака приликом анотације су приказани табеларно.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Домен** | **O** | **I-PER** | **B-PER** | **I-ORG** | **B-ORG** | **I-LOC** | **B-LOC** |
| Правни | 4001 | 65 | 65 | 569 | 229 | 115 | 138 |
| Новински | 4177 | 110 | 150 | 243 | 198 | 105 | 159 |
| Твитер | 4628 | 69 | 109 | 29 | 55 | 32 | 111 |
| Књижевни | 4852 | 48 | 95 | 3 | 3 | 0 | 6 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Домен** | **PER** | **ORG** | **LOC** | **Укупно ентитета** | | **Укупно токена** | |
| Правни | 130 | 798 | 253 | 1181 | 5182 | |
| Новински | 260 | 441 | 264 | 965 | 5142 | |
| Твитер | 178 | 84 | 143 | 405 | 5033 | |
| Књижевни | 143 | 6 | 6 | 155 | 5007 | |

У наредној табели је приказано процентуално појављивање сваког засебног ентитета међу свим ентитетима као и проценат именованих ентитета међу свим токенима.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Домен** | **PER ентитета** | **ORG ентитета** | **LOC ентитета** | **Ентитета у токенима** |
| Правни | 11.01% | 67.57% | 21.42% | 22.79% |
| Новински | 26.94% | 45.70% | 27.36% | 18.77% |
| Твитер | 43.95% | 20.74% | 35.31% | 8.05% |
| Књижевни | 92.26% | 3.87% | 3.87% | 3.10% |
| Просечно | 43.54% | 34.47% | 21.99% | 13.18% |

На основу анализе можемо запазити да у правном и новинском домену приближно петина токена представљају ентитете. Међу њима се највише појављују организације, док се у књижевном и Твитер домену највише појављују особе.

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 **CLASSLA, BERTić, COMtext.SR и SrpCNNER**, 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 додатнa карактеристика.

* + 1. Карактеристике (енгл. features)

За сваки токен модел користи:

* **token\_lower** – токен записан 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
* Цео скуп података се користи **за cross-validation података**.
* **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 мења у свакој итерацији **cross-validation**, и тако 10 пута заредом.

**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 ≈ **0.8513**
* Пример предикције:

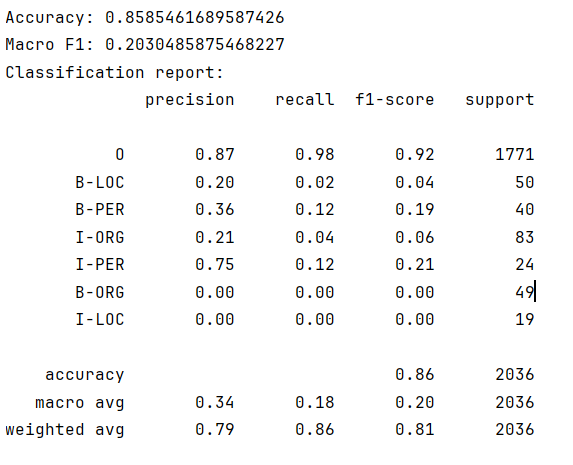
Beograd -> I-LOC

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.



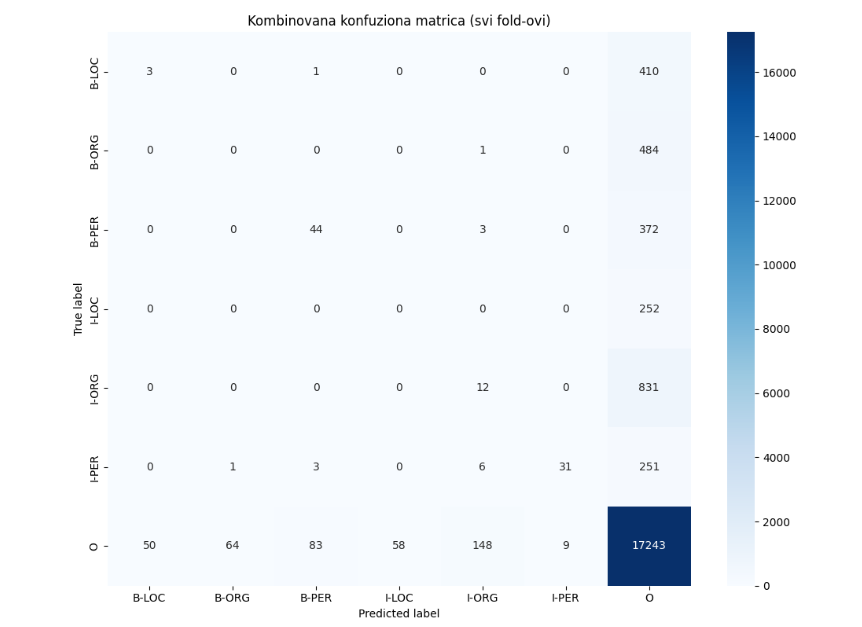
Резултати показују да је укупна тачност модела на тест скупу релативно висока (**0.859**), што је углавном последица доминације класе **O**, за коју је прецизност 0.87, recall 0.98 и F1-score 0.92. Међутим, када је реч о препознавању стварних ентитета, модел има изразите слабости.

Класе **B-PER** и **I-PER** постижу релативно бољу прецизност (**0.36 и 0.75**), али ниску осетљивост (**0.12**), што значи да модел препознаје само мали део појава ових ентитета. Остале класе као што су **I-ORG**, **B-LOC**, **B-ORG** и нарочито **I-LOC** имају практично непостојеће резултате (recall ≈ 0.00–0.04), што указује да модел готово уопште не успева да идентификује ове ентитете.

Укупни **macro F1-score** износи **0.203**, што показује да је просечна перформанса по класама веома ниска, док **weighted F1-score 0.81** више одражава добру класификацију доминантне класе **O**. Ови резултати јасно потврђују да Multinomial Naive Bayes као базни модел није довољно ефикасан за прецизно разликовање ентитета, посебно у условима велике неравнотеже између класа и сложених контекста у тексту.

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

У овој матрици су приказани збирни резултати **cross-validation** тренирања података.



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 ≈ **0.8590**
* Пример предикције:

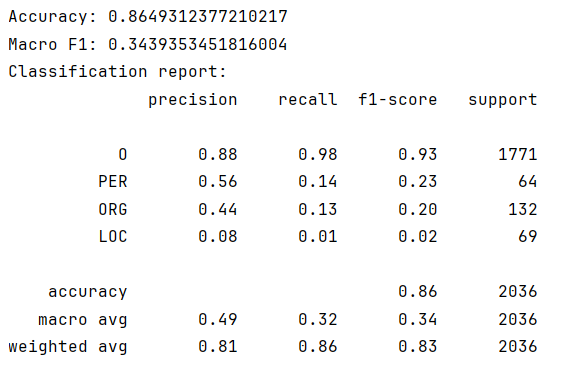
Beograd -> LOC

vlada -> ORG

Marko -> LOC

ETF -> ORG

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

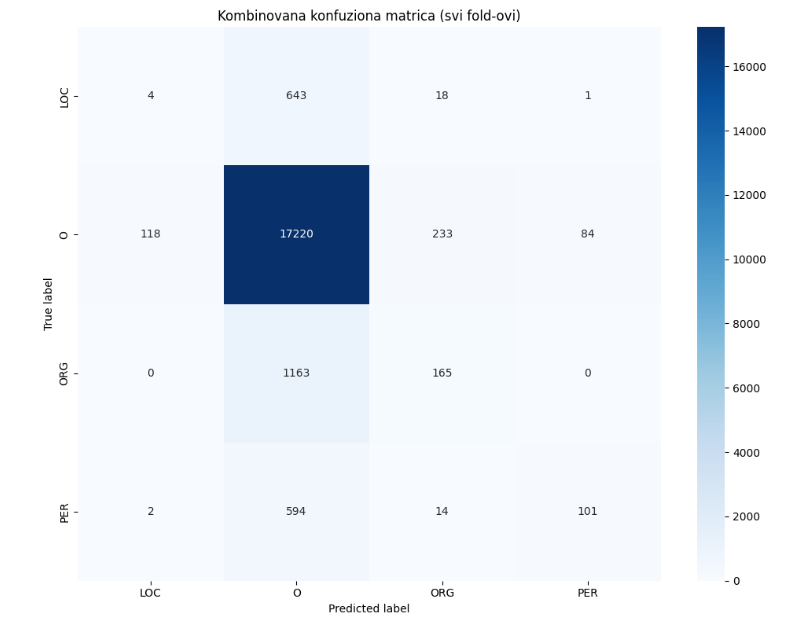


Када се **B-** и **I-** ознаке третирају као једна класа, модел постиже нешто боље резултате у односу на случај када су ове ознаке раздвојене. Модел постиже релативно високу укупну тачност од **0.865**, што је пре свега последица доминације класе **O**, која чини већину токена у корпусу. За ову класу добијени су добри резултати – **прецизност 0.88, *recall* 0.98 и F1-score 0.93**, што показује да модел скоро увек исправно препознаје токене који не припадају ниједном ентитету.

Међутим, када је у питању препознавање стварних ентитета, модел има значајне слабости. За класу **PER** добијена је умерена прецизност (**0.56**), али веома низак *recall* (**0.14**), што значи да модел препознаје само мали број појављивања особа, док већину пропушта. Слична ситуација је и са класом **ORG**, где је прецизност **0.44**, али recall свега **0.13**. Ово указује на то да модел повремено тачно класификује неке организације, али није у стању да доследно препозна већину њихових појављивања у тексту. Најслабији резултати постигнути су за класу **LOC**, са **прецизношћу 0.08 и *recall*-ом 0.01**, што практично чини ову класу неупотребљивом. Ово се може објаснити малим бројем примера за локације и израженим дисбалансом класа.

Укупни **macro F1-score** износи **0.344**, што показује да је просечна ефикасност модела на свим класама ниска, док **weighted F1-score 0.83** више одражава добру перформансу на већинској класи **O**.

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. CLASSLА

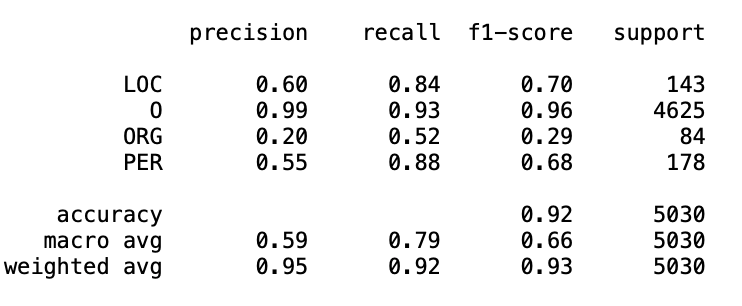
За евалуацију *classla* модела коришћене су *standard* и *nonstandard* варијанте модела. Као улазни податак коришћен је унапред токенизован текст из свих домена. У наставку су издвојени резултати у виду статистичких података.

* + 1. Стандардни језик

A screenshot of a graph

AI-generated content may be incorrect.A number of numbers in a row

AI-generated content may be incorrect.Статистика целог скупа податакa

**A screenshot of a graph

AI-generated content may be incorrect.Статистика твитер података Статистика админстративних података**

* + 1. Нестандардни језик

A number of numbers on a white background

AI-generated content may be incorrect.Статистика целог скупа података

A number of numbers on a white background

AI-generated content may be incorrect.A screenshot of a graph

AI-generated content may be incorrect.A table of numbers with black text

AI-generated content may be incorrect.**Статистика твитер података Статистика админстративних података**

* + 1. Анализа

|  |  |  |
| --- | --- | --- |
| **Token** | **Expected** | **Result** |
| BULEVAR | B-LOC | B-ORG |
| JAŠE | I-LOC | B-PER |
| TOMIĆA | I-LOC | I-PER |
| 23 | I-LOC | O |
| , | I-LOC | O |
| sprat | I-LOC | O |
| 5 | I-LOC | O |
| , | I-LOC | O |
| stan | I-LOC | O |
| 2 | I-LOC | O |
| , | I-LOC | O |
| NOVI | I-LOC | B-ORG |
| SAD | I-LOC | I-ORG |
| , | I-LOC | O |
| Srbija | I-LOC | B-LOC |

|  |  |  |
| --- | --- | --- |
| **Token** | **Expected** | **Result** |
| Poslovno | O | B-ORG |
| ime | O | I-ORG |
| obveznika | O | I-ORG |
| OD | B-ORG | O |
| SRCA | I-ORG | B-ORG |
| humanitarno | O | B-ORG |
| udruženje | O | I-ORG |
| zaposlenih | O | I-ORG |

На твитер скупу података јасно се примећује предност нестандардног модела у анализи неформалних текстова. Са друге стране, административни текстови имају сличан резултат за оба модела. У наставку су дати примери неусаглашености анотација који доприносе ниском *recall* резултату за LOC и ниској прецизности за ORG анотације у административном домену, за стандардни модел.

Оба модела имају значајно боље резултате од изнад описаног *baseline* модела.

Фајлови са статистиком и разликама у анотацији за све домене и моделе налазе се у фолдеру *pythonProject/evaluation/classla*, а на крају овог документа је представљено поређење са осталим моделима. У наставку су дате матрице конфузије.

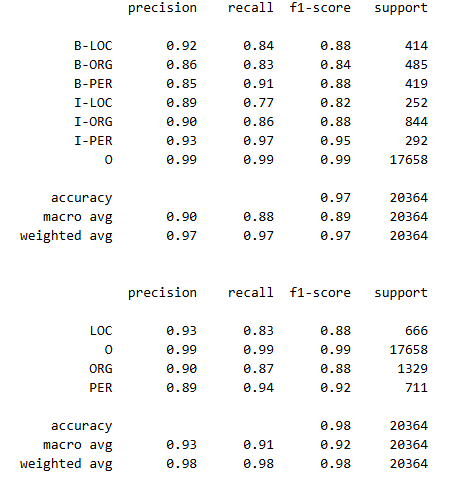
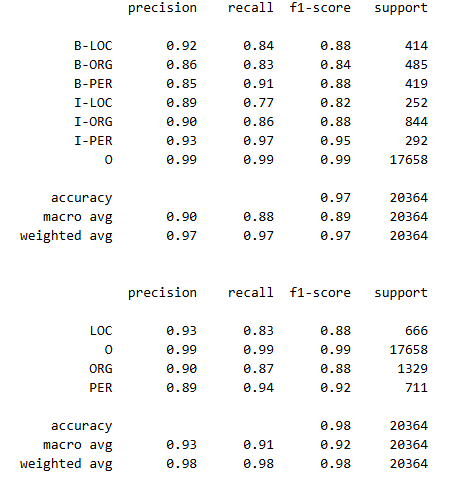
|  |  |
| --- | --- |
| CLASSLA - Standard | A screenshot of a computer  AI-generated content may be incorrect. |
| CLASSLA - Nonstandard | A screenshot of a computer screen  AI-generated content may be incorrect. |

* 1. BERTић
     1. Резултати евалуације

Евалуација је одрађена по доменима и за све домене заједно. Комплетни резултати се налазе у фолдеру /*evaluation/bertic/output*. За евалуацију су се користили пакети *sklearn* (са и без B- i I- тагова) i *seqeval* (у *default* и *strict* режиму рада). У наставку су приказани резултати добијени помоћу *sklearn* пакета за појединачне домене игноришући B- i I- тагове, као и резултати за све домене заједно користећи B- i I- тагове и без њих.

|  |  |
| --- | --- |
| Правно-административни домен | Твитер домен |
| A screenshot of a computer screen  AI-generated content may be incorrect. | A screenshot of a graph  AI-generated content may be incorrect. |
| Новински домен | Књижевни домен |
| A screenshot of a graph  AI-generated content may be incorrect. | A screenshot of a graph  AI-generated content may be incorrect. |

Резултати за све домене:

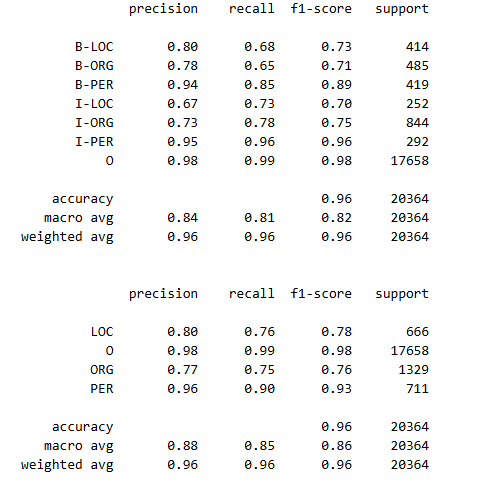
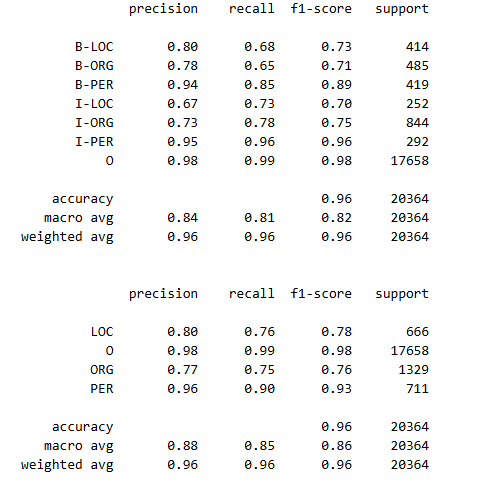


* 1. COMtext.SR
     1. Резултати евалуације

Евалуација је одрађена по доменима и за све домене заједно. Комплетни резултати се налазе у фолдеру /*evaluation/comtext/output*. За евалуацију су се користили пакети *sklearn* (са и без B- i I- тагова) i *seqeval* (у *default* и *strict* режиму рада). У наставку су приказани резултати добијени помоћу *sklearn* пакета за појединачне домене игноришући B- i I- тагове, као и резултати за све домене заједно користећи B- i I- тагове и без њих.

|  |  |
| --- | --- |
| Правно-административни домен | Твитер домен |
| A screenshot of a computer screen  AI-generated content may be incorrect. | A screenshot of a graph  AI-generated content may be incorrect. |
| Новински домен | Књижевни домен |
| A screenshot of a computer screen  AI-generated content may be incorrect. | A screenshot of a graph  AI-generated content may be incorrect. |

Резултати за све домене:

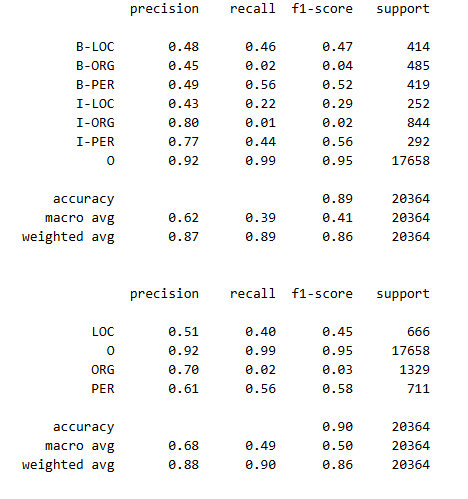
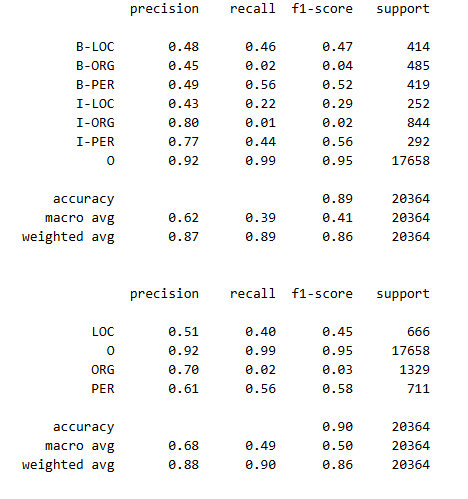


* 1. SrpCNNER
     1. Резултати евалуације

Евалуација је одрађена по доменима и за све домене заједно. Комплетни резултати се налазе у фолдеру /*evaluation/srpcnner/output*. За евалуацију су се користили пакети *sklearn* (са и без B- i I- тагова) i *seqeval* (у *default* и *strict* режиму рада). У наставку су приказани резултати добијени помоћу *sklearn* пакета за појединачне домене игноришући B- i I- тагове, као и резултати за све домене заједно користећи B- i I- тагове и без њих.

|  |  |
| --- | --- |
| Правно-административни домен | Твитер домен |
| A screenshot of a graph  AI-generated content may be incorrect. | A screenshot of a computer screen  AI-generated content may be incorrect. |
| Новински домен | Књижевни домен |
| A screenshot of a computer screen  AI-generated content may be incorrect. | A screenshot of a computer screen  AI-generated content may be incorrect. |

Резултати за све домене:



* 1. Mатрице конфузије

|  |  |
| --- | --- |
| BERTić | A graph with numbers and a blue square  AI-generated content may be incorrect. |
| COMtext | A graph with numbers and a blue square  AI-generated content may be incorrect. |
| SrpCNNER | A graph with numbers and a blue square  AI-generated content may be incorrect. |
|  |  |
|  |  |

Анализа резултата

Табеларно је приказано неколико издвојених грешака у предикцији модела, док се комплетни резултати налазе у фолдеру /*evaluation /predictions*.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Пр.** | **Токен** | **Очекивано** | **BERTic** | **COMtext** | **SrpCNNER** | **Образложење** |
| 1 | EU | B-ORG | B-ORG | B-LOC | B-PER | Грешке у класификацији организације. |
| 2 | Jutjub | B-ORG | B-ORG | O | B-PER |
| 3 | Orbanu | B-PER | B-PER | B-LOC | O | Грешке у класификацији особе. |
| 4 | Aleksandar | B-PER | B-PER | I-ORG | B-PER |
| Šapić | I-PER | I-PER | I-ORG | I-PER |
| 5 | Rusi | O | B-PER | B-LOC | O | Различито третирање националних назива. |
| 6 | Predsedništva | B-ORG | B-ORG | O | O | Пропуштен таг који означава почетак ентитета. |
| BiH | I-ORG | I-ORG | I-ORG | O |
| 7 | Republike | B-LOC | O | B-LOC | B-LOC |
| Srbije | I-LOC | I-LOC | I-LOC | I-LOC |
| 8 | Merkatora | B-LOC | B-ORG | B-ORG | B-LOC | Разлика у класификовању услужних објеката и третирању наводника. |
| 9 | Tržnog | B-LOC | B-ORG | B-ORG | O |
| centra | I-LOC | I-ORG | I-ORG | O |
| „ | I-LOC | I-ORG | O | O |
| Ušće | I-LOC | I-ORG | I-ORG | O |
| “ | I-LOC | O | O | O |
| 10 | Srpske | B-ORG | B-ORG | B-ORG | O | Разликују се одлуке за укључвање скраћенице у састав ентитета. |
| napredne | I-ORG | I-ORG | I-ORG | O |
| stranke | I-ORG | I-ORG | I-ORG | O |
| ( | O | O | I-ORG | O |
| SNS | B-ORG | B-ORG | I-ORG | O |
| ) | O | O | I-ORG | O |

Поједини модели подржавају већи број ознака због чега се ентитети могу изгубити конверзијом.

|  |  |  |  |
| --- | --- | --- | --- |
| **Token** | **Očekivano** | **Predikcija** | **Konvertovana predikcija** |
| Službeni | B-ORG | I-LAW | O |
| glasnik | I-ORG | I-LAW | O |
| RS | I-ORG | I-LAW | O |

* 1. Закључак

Приказана је табела која поређује ***мacro F1-score* свих модела без коришћења** B- i I- тагова**.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Домен** | **CLASSLA - standardni** | **CLASSLA - nestandardni** | **BERTić** | **COMtext** | **SrpCNNER** |
| Твитер | 0.66 | 0.78 | 0.85 | 0.80 | 0.54 |
| Правно-административни | 0.77 | 0.82 | 0.90 | 0.85 | 0.46 |
| Новински | 0.88 | 0.91 | 0.95 | 0.89 | 0.51 |
| Књижевни | 0.52 | 0.65 | 0.92 | 0.92 | 0.45 |
| Сви домени | 0.79 | 0.85 | 0.92 | 0.86 | 0.50 |

BERTić се показао као најбољи модел за све домене, након њега COMtext, CLASSLA – нестандардни, CLASSLA – стандардни и на крају SrpCNNER. CLASSLA и BERTić постижу најбоље резултате у новинском домену, COMtext у књижевном, док SrpCNNER даје најбоље резултате у Твитер домену. SrpCNNER, иако је трениран на скупу књижевних текстова, на нашем корпусу даје најлошије резултате за овај домен. Исти домен је проблематичан и за модел CLASSLA, док је за BERTić и COMtext то Твитер домен.

Поређење *F1-score* резултата по врсти анотације је дато на следећем графикону: