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](#_Toc210085591)

[1. Увод 1](#_Toc210085592)

[2. Прикупљање података 2](#_Toc210085593)

[2.1. Новински текстови 2](#_Toc210085594)

[2.2. Књижевни текстови 3](#_Toc210085595)

[2.3. Правно-административни текстови 3](#_Toc210085596)

[2.4. Текстови са твитера/X-a 3](#_Toc210085597)

[3. Анотација података 4](#_Toc210085598)

[3.1. Правила анотације 4](#_Toc210085599)

[3.1.1. Скраћенице 4](#_Toc210085600)

[3.1.2. Град и адреса као јединствена локација 4](#_Toc210085601)

[3.1.3. Називи улица и бројеви 5](#_Toc210085602)

[3.1.4. Ентитети који се протежу преко више редова 5](#_Toc210085603)

[3.1.5. Реч „улица“ уз назив 5](#_Toc210085604)

[3.1.6. Општи појмови који не означавају конкретну локацију 5](#_Toc210085605)

[3.1.7. Називи организација 5](#_Toc210085606)

[3.1.8. Oпшти појмови који означавау организације 6](#_Toc210085607)

[3.1.9. Имена особа у називу организације 6](#_Toc210085608)

[3.1.10. Назив места после организације 6](#_Toc210085609)

[3.1.11. Придеви уз назив организације 6](#_Toc210085610)

[3.1.12. ZПоодсеци организација 6](#_Toc210085611)

[3.1.13. Скраћеница након назива организације 6](#_Toc210085612)

[3.1.14. Наводници у називу 7](#_Toc210085613)

[3.1.15. Место после организације и ближа одредница 7](#_Toc210085614)

[3.1.16. Именице у различитим предлошко-падежним конструкцијама 7](#_Toc210085615)

[3.1.17. Иницијали особе 7](#_Toc210085616)

[3.1.18. Уметничка имена и титуле 7](#_Toc210085617)

[3.1.19. Називи бендова и група 7](#_Toc210085618)

[3.1.20. Име особе са надимком 8](#_Toc210085619)

[3.1.21. Образовне установе 8](#_Toc210085620)

[3.1.22. Називи штампаних новина 8](#_Toc210085621)

[3.1.23. Називи мора и океана 8](#_Toc210085622)

[3.1.24. Називи у хаштаговима 8](#_Toc210085623)

[3.1.25. Кориснички називи профила 8](#_Toc210085624)

[3.2. Калибрација 9](#_Toc210085625)

[3.3. Анализа анотације 10](#_Toc210085626)

[4. Евалуација статистичких модела 12](#_Toc210085627)

[4.1. Baseline приступ – мултиномијални наивни Бајесов класификатор 12](#_Toc210085628)

[4.1.1. Карактеристике (енгл. features) 12](#_Toc210085629)

[4.1.2. Тренирање и евалуација модела 13](#_Toc210085630)

[4.1.3. Варијанте евалуације 13](#_Toc210085631)

[4.1.4. Закључак 17](#_Toc210085632)

[4.2. CLASSLA 17](#_Toc210085633)

[4.2.1. Стандардни језик 18](#_Toc210085634)

[4.2.2. Нестандардни језик 18](#_Toc210085635)

[4.2.3. Анализа 19](#_Toc210085636)

[4.3. BERTић 20](#_Toc210085637)

[4.3.1. Резултати евалуације 20](#_Toc210085638)

[4.3.2. Анализа резултата 23](#_Toc210085639)

[4.4. COMtext.SR 24](#_Toc210085640)

[4.4.1. Резултати евалуације 24](#_Toc210085641)

[4.4.2. Анализа резултата 26](#_Toc210085642)

[4.5. Закључак 27](#_Toc210085643)

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, 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.

Током анотације исправљене су грешке токенизовања које су настале на појединим местима услед коришћења неодговарајућег кода латиничиног слова ћ.

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

За процес калибрације одабрана је прва половина књижевног домена. Помоћу скрипте су израчунати бинарни степени сагласности. Милан и Милица су у сагласности 99.004%, Милан и Алекса 99.210%, док су Милица и Алекса у сагласности 99.313%. Просек бинарних сагласности износи 99.176%. Грешке које су се појављивале су издвојене у табели.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **broj linije** | **token** | **Milan** | **Milica** | **Aleksa** | **većinska odluka** | **objašnjenje** |
| 57 | guvernantom | B-PER | O | O | O | Није име особе. |
| 58 | Francuskinjom | I-PER | O | O | O | Није име особе. |
| 206 | ekonomkom |  | O | O | O | Није анотирано. |
| 257 | Stepan | B-PER | O | B-PER | B-PER | Јесте особа. |
| 258 | Arkadijevič | I-PER | B-PER | I-PER | I-PER | Грешка је последица претходне грешке. |
| 261 | Stiva | I-PER | I-PER | B-PER | I-PER | Надимак након имена особе сматрамо истим ентитетом. |
| 493 | sećao |  | O | O | O | Није анотирано. |
| 948 | Doli | O | B-PER | O | O | Јесте особа. |
| 1041 | S |  | O | O | O | Није анотирано. |
| 1209 | Doli | B-PER | O | O | O | Јесте особа. |
| 1255 | Stepan | O | B-PER | B-PER | B-PER | Јесте особа. |
| 1256 | Arkadijevič | B-PER | I-PER | I-PER | I-PER | Грешка је последица претходне грешке. |
| 1257 | . | I-PER | O | O | O | Грешка је последица претходне грешке. |
| 1730 | m-lle | B-PER | O | O | O | Не сматрамо делом имена. |
| 1731 | Roland | I-PER | B-PER | B-PER | B-PER | Грешка је последица претходне грешке. |
| 2046 | upita | O | B-PER | O | O | Није особа. |
| 2047 | Stepan | B-PER | I-PER | B-PER | B-PER | Грешка је последица претходне грешке. |
| 2048 | Arkadijevič | I-PER | O | I-PER | I-PER | Грешка је последица претходне грешке. |
| 2142 | Stepana | B-PER | I-PER | B-PER | B-PER | Означено као унутрашњост ентитета али је почетак. |
| 2143 | Arkadijeviča | I-PER | O | I-PER | I-PER | Јесте особа. |
| 2165 | Matvej | B-PER | O | B-PER | B-PER | Јесте особа. |
| 2519 | Matvej | B-PER | B-PER | O | B-PER | Јесте особа. |
| 2638 | Stepan | B-PER | B-PER | O | B-PER | Јесте особа. |
| 2699 | Matvej | B-PER | B-PER | B-PER | B-PER | Размак на почетку анотације. |
| 3365 | Stepan | I-PER | B-PER | B-PER | B-PER | Јесте особа. |
| 3366 | Arkadijevič | O | I-PER | I-PER | I-PER | Грешка је последица претходне грешке. |
| 3574 | Ako | O |  | O | O | Није анотирано. |
| 3644 | Liberalna | B-ORG | O | B-ORG | B-ORG | Јесте организација. |
| 3645 | je | I-ORG | O | O | O | Није део организације. |
| 3646 | partija | O | O | I-ORG | O | Јесте организација. |
| 3663 | Arkadijevič | I-PER | B-PER | I-PER | I-PER | Означено као почетак ентитета али је унутрашњост. |
| 3741 | Liberalna | B-PER | B-ORG | B-ORG | B-ORG | Јесте организација. |
| 3743 | partija | I-PER | I-ORG | I-ORG | I-ORG | Јесте организација. |
| 3831 | Arkadijevič | I-PER | B-PER | I-PER | I-PER | Означено као почетак ентитета али је унутрашњост. |
| 3872 | Ruriku | B-PER | B-PER | B-LOC | B-PER | Јесте особа. |
| 3983 | vlada | O | O | B-ORG | O | Није организација - не односи се на конкретну. |
|  |  |  |  |  |  |  |

* 1. Анализа анотације

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **domen** | **O** | **I-PER** | **B-PER** | **I-ORG** | **B-ORG** | **I-LOC** | **B-LOC** |
| administrative\_texts | 3967 | 65 | 65 | 595 | 221 | 149 | 120 |
| newspapers | 4160 | 108 | 149 | 258 | 198 | 112 | 157 |
| twitter | 4670 | 69 | 109 | 29 | 58 | 32 | 111 |
| literature | 4856 | 46 | 90 | 3 | 4 | 0 | 7 |

Резултати за процес анализе анотације су приказани табеларно.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **domen** | **PER** | **ORG** | **LOC** | **ukupno entiteta** | **ukupno tokena** |
| administrative\_texts | 130 | 816 | 269 | 1215 | 5182 |
| newspapers | 257 | 456 | 269 | 982 | 5142 |
| twitter | 178 | 87 | 143 | 408 | 5078 |
| literature | 136 | 7 | 7 | 150 | 5006 |

На основу анализе можемо запазити да се од именованих ентитета у административном и новинском домену највише појављују организације, а у књижевном и твитер домену особе.

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **domen** | **PER entiteta** | **ORG entiteta** | **LOC entiteta** | **entiteta u tokenima** |
| administrative\_texts | 10.70% | 67.16% | 22.14% | 23.45% |
| newspapers | 26.17% | 46.44% | 27.39% | 19.10% |
| twitter | 43.63% | 21.32% | 35.05% | 8.03% |
| literature | 90.67% | 4.67% | 4.67% | 3.00% |
| prosek | 42.79% | 34.90% | 22.31% | 13.39% |

При анализи пронађене су грешке на које треба обратити пажњу при процесу тренирања и евалуације:

* + - у новинском скупу података на линији 164 недостаје анотација
    - у твитер скупу података на линији 2699 постоји размак пре анотације „ B-PER“

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 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
* 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 ≈ **0.845**
* Пример предикције:

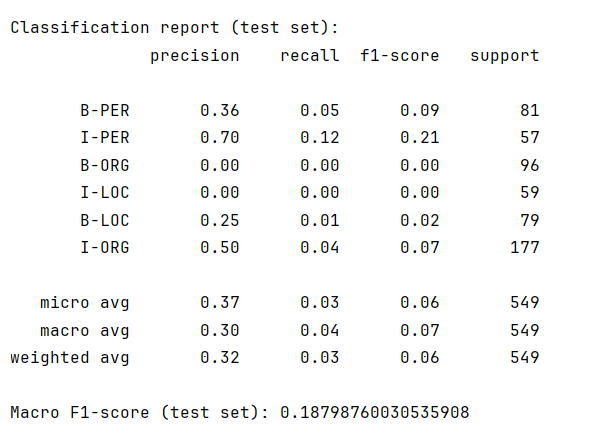
Beograd -> I-LOC

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.



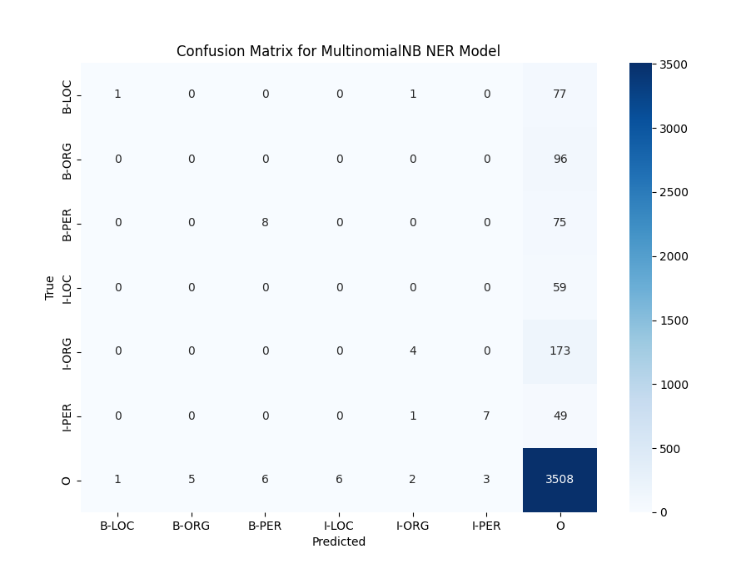
Резултати показују да је **просечна тачност на скупу за тренирање** била задовољавајућа, али на тест скупу модел има видљива ограничења.

За ентитете типа **B-PER** и **I-PER** добијени су нешто бољи резултати у односу на друге класе (прецизност 0.57 и 0.70), али је **осетљивост ниска** (0.10 и 0.12), што значи да модел ретко препознаје све појаве ових ентитета. Код класа **I-ORG**, **B-LOC** и нарочито **I-LOC** и **B-ORG**, резултати су веома слаби — модел их готово уопште не препознаје (recall ≈ 0.00–0.02).

Гледано на нивоу целог скупа, **micro F1-score** износи 0.07, што указује да модел укупно има слабу способност предвиђања када се све класе узму заједно. Са друге стране, **macro F1-score** је 0.1956, што потврђује да је просечна перформанса по класама ниска, а неравнотежа у бројности класа значајно утиче на укупан резултат.

Ови налази показују да **Multinomial Naive Bayes** као базни модел није довољно ефикасан за прецизно разликовање између ентитета, посебно у условима велике неравнотеже између класа и сложенијих зависности између токена.

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:

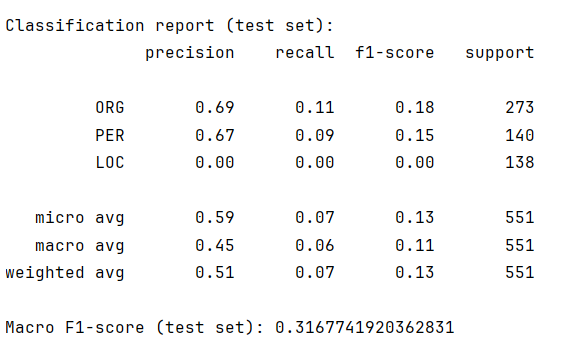
Beograd -> LOC

vlada -> ORG

Marko -> LOC

ETF -> ORG

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



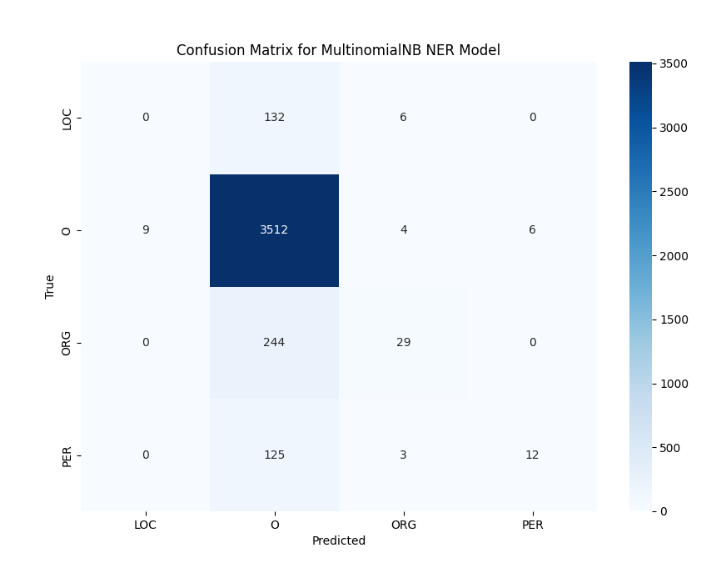
Када се **B-** и **I-** ознаке третирају као једна класа, модел постиже нешто боље резултате у односу на случај када су ове ознаке раздвојене, са релативно високом просечном тачношћу на тренинг скупу од око 0.847. Ово је делимично последица доминације класе **O**, која чини већину токена у корпусу, па модел најчешће правилно класификује баш ове токене.

Ипак, резултати на тест скупу и даље показују значајне слабости у детекцији стварних ентитета. За класе **ORG** и **PER** добијена је релативно висока прецизност (0.69 и 0.67), али уз веома низак recall (0.11 и 0.09), што указује да модел препознаје мали број стварних ентитета и пропушта већину њих. Класа **LOC** постиже практично неупотребљиве резултате (precision 0.00, recall 0.00), што се може објаснити малим бројем примера за локације у корпусу и великом доминацијом **O** токена.

Укупни **macro F1-score** износи 0.317, док је **weighted F1-score** 0.13, што указује да, иако модел понекад исправно класификује ентитете, његова способност за поуздано препознавање мање заступљених класа је веома ограничена.

Спајање B и I ознака у једну класу се у овом случају показало повољнијим за Наивни Бајесов модел јер смањује негативан утицај дисбаланса класа и поједностављује задатак. Ипак, чак и у овом приступу, модел има велика ограничења у препознавању ентитета, што јасно указује на потребу за применом сложенијих *NER* модела како би се постигли употребљиви резултати у пракси.

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. CLASSLA

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

Статистика целог скупа података Статистика за скраћени облик анотације

|  |  |  |  |
| --- | --- | --- | --- |
| **Tag** | **Precision** | **Recall** | **F1 score** |
| B-LOC | 0.77 | 0.62 | 0.68 |
| B-ORG | 0.69 | 0.48 | 0.57 |
| B-PER | 0.82 | 0.6 | 0.69 |
| I-LOC | 0.42 | 0.61 | 0.5 |
| I-ORG | 0.72 | 0.66 | 0.69 |
| I-PER | 0.91 | 0.79 | 0.85 |
| macro avg | 0.72 | 0.63 | 0.66 |
| micro avg | 0.62 | 0.73 | 0.67 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Tag** | **Precision** | **Recall** | **F1 score** |
| LOC | 0.68 | 0.67 | 0.68 |
| ORG | 0.78 | 0.64 | 0.7 |
| PER | 0.89 | 0.7 | 0.79 |
| macro avg | 0.78 | 0.67 | 0.72 |
| micro avg | 0.66 | 0.78 | 0.72 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Tag** | **Precision** | **Recall** | **F1 score** |
| LOC | 0.84 | 0.59 | 0.69 |
| ORG | 0.51 | 0.2 | 0.28 |
| PER | 0.88 | 0.54 | 0.67 |
| macro avg | 0.74 | 0.44 | 0.55 |
| micro avg | 0.45 | 0.79 | 0.57 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tag** | **Precision** | | **Recall** | **F1 score** |
| LOC | | 0.48 | 0.68 | 0.57 |
| ORG | | 0.82 | 0.65 | 0.73 |
| PER | | 0.78 | 0.83 | 0.8 |
| macro avg | | 0.69 | 0.72 | 0.7 |
| micro avg | | 0.67 | 0.74 | 0.7 |

**Статистика твитер података Статистика админстративних података**

* + 1. Нестандардни језик

Статистика целог скупа података Статистика за скраћени облик анотације

|  |  |  |  |
| --- | --- | --- | --- |
| **Tag** | **Precision** | **Recall** | **F1 score** |
| B-LOC | 0.82 | 0.71 | 0.76 |
| B-ORG | 0.69 | 0.73 | 0.71 |
| B-PER | 0.86 | 0.8 | 0.83 |
| I-LOC | 0.53 | 0.69 | 0.6 |
| I-ORG | 0.68 | 0.8 | 0.73 |
| I-PER | 0.91 | 0.91 | 0.91 |
| macro avg | 0.75 | 0.77 | 0.76 |
| micro avg | 0.77 | 0.74 | 0.75 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Tag** | **Precision** | **Recall** | **F1 score** |
| LOC | 0.77 | 0.78 | 0.77 |
| ORG | 0.71 | 0.8 | 0.75 |
| PER | 0.9 | 0.86 | 0.88 |
| macro avg | 0.79 | 0.81 | 0.8 |
| micro avg | 0.81 | 0.77 | 0.79 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Tag** | **Precision** | **Recall** | **F1 score** |
| LOC | 0.83 | 0.74 | 0.79 |
| ORG | 0.52 | 0.57 | 0.54 |
| PER | 0.92 | 0.74 | 0.82 |
| macro avg | 0.76 | 0.68 | 0.72 |
| micro avg | 0.71 | 0.8 | 0.75 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Tag** | **Precision** | **Recall** | **F1 score** |
| LOC | 0.62 | 0.72 | 0.67 |
| ORG | 0.69 | 0.8 | 0.74 |
| PER | 0.81 | 0.89 | 0.85 |
| macro avg | 0.71 | 0.8 | 0.75 |
| micro avg | 0.79 | 0.69 | 0.73 |
|  |  |  |  |

**Статистика твитер података Статистика админстративних података**

* + 1. Анализа

На твитер скупу података јасно се примећује предност нестандардног модела у анализи неформалних текстова. Сад друге стране, административни текстови имау сличан резултат за оба модела. Ниска прецизност у препознавању локација може се објаснити начином на који су ручно анотиране организације које садрже називе локација, а које се често појављују у тексту:

|  |  |  |
| --- | --- | --- |
| **Token** | **Expected** | **Result** |
| IZMEĐU | O | I-ORG |
| VLADE | B-ORG | B-ORG |
| UJEDINJENOG | I-ORG | I-ORG |
| KRALJEVSTVA | I-ORG | B-LOC |
| VELIKE | I-ORG | I-LOC |
| BRITANIJE | I-ORG | I-LOC |
| I | I-ORG | O |
| SEVERNE | I-ORG | B-LOC |
| IRSKE | I-ORG | I-LOC |

Низак *recall* може се објаснити погрешним анотирањем локација и адреса, где се често појављују ORG ознаке:

|  |  |  |
| --- | --- | --- |
| **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 |

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 свих модела по анотацији, за све евалуиране моделе.

* 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:

A screenshot of a computer screen

AI-generated content may be incorrect.

Новински домен:

A screenshot of a computer screen

AI-generated content may be incorrect.

Твитер домен:

A screenshot of a computer screen

AI-generated content may be incorrect.

Књижевни домен:

A screenshot of a computer screen

AI-generated content may be incorrect.

Резултати за све домене:

A screenshot of a computer screen

AI-generated content may be incorrect.

Креирана је матрица конфузије.

A graph with numbers and a blue square

AI-generated content may be incorrect.

* + 1. Анализа резултата

Фајлови у којима су приказани анотирани токени са одговарајућим резултатима предикције се налазе у фолдеру /*evaluation/bertic/predictions*.

На основу резултата можемо приметити грешке евалуираног модела:

|  |  |  |  |
| --- | --- | --- | --- |
| **Token** | **Predikcija** | **Konvertovana predikcija** | **Očekivano** |
| Agenciji | I-MISC | O | B-ORG |
| za | I-ORG | I-ORG | I-ORG |
| privredne | I-ORG | I-ORG | I-ORG |
| registre | I-ORG | I-ORG | I-ORG |

Као и грешке процеса анотације:

|  |  |  |  |
| --- | --- | --- | --- |
| **Token** | **Predikcija** | **Konvertovana predikcija** | **Očekivano** |
| Agenciji | B-ORG | B-ORG | B-LOC |
| za | I-ORG | I-ORG | I-LOC |
| privredne | O | O | I-LOC |
| registre | O | O | I-LOC |

Док поједине грешке могу да потичу из различитих правила анотације:

|  |  |  |  |
| --- | --- | --- | --- |
| **Token** | **Predikcija** | **Konvertovana predikcija** | **Očekivano** |
| Hrvatima | B-PER | B-PER | O |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Token** | **Predikcija** | **Konvertovana predikcija** | **Očekivano** |  |
| Vučićevi | О | O | B-PER |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Token** | **Predikcija** | **Konvertovana predikcija** | **Očekivano** |
| TC | B-ORG | B-ORG | B-LOC |
| “ | O | O | I-LOC |
| Ušće | I-ORG | I-ORG | I-LOC |
| “ | O | O | I-LOC |

Модел предикције често не укључује крај наводника у састав ентитета, док почетак углавном укључује. На основу правила анотације наводници су укључени у ентитет ако се налазе у средини њега, што доводи до додатне несагласности.

* 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:

A screenshot of a computer screen

AI-generated content may be incorrect.

Новински домен:

A screenshot of a computer screen

AI-generated content may be incorrect.

Твитер домен:

A screenshot of a computer screen

AI-generated content may be incorrect.

Књижевни домен:

A screenshot of a computer screen

AI-generated content may be incorrect.

Резултати за све домене:

A screenshot of a computer screen

AI-generated content may be incorrect.

Креирана је матрица конфузије.

A graph with numbers and a blue square

AI-generated content may be incorrect.

* + 1. Анализа резултата

Фајлови у којима су приказани анотирани токени са одговарајућим резултатима предикције се налазе у фолдеру /*evaluation/comtext/predictions*.

У резултатима се јављају грешке у предикцији модела:

|  |  |  |  |
| --- | --- | --- | --- |
| **Token** | **Predikcija** | **Konvertovana predikcija** | **Očekivano** |
| EU | B-TOP | B-LOC | B-ORG |

Модел предикције подржава већи број тагова због чега поједини ентитети губе своје ознаке. Да оба подржавају исте тагове, сагласност би била већа:

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

Грешке су примећене на сличним примерима као за модел БЕРТић. Иако су резултати евалуације лошији, модел се понаша боље у одређеним ситуацијама:

|  |  |  |  |
| --- | --- | --- | --- |
| **Token** | **Predikcija** | **Konvertovana predikcija** | **Očekivano** |
| Agencije | B-INST | B-ORG | B-ORG |
| za | I-INST | I-ORG | I-ORG |
| privredne | I-INST | I-ORG | I-ORG |
| registre | O | O | I-ORG |

|  |  |  |  |
| --- | --- | --- | --- |
| **Token** | **Predikcija** | **Konvertovana predikcija** | **Očekivano** |
| Hrvatima | O | O | O |

Као БЕРТић, модел не укључује наводнике у састав ентитета, што такође доводи до несагласности.

* 1. Закључак

Дата је табела за поређење ***мacro F1-score* резултата свих модела, по домену.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Домен** | **CLASSLA - standardni** | **CLASSLA - nestandardni** | **BERTić** | **COMtext.SR** |
| Твитер | 0.55 | 0.72 | 0.84 | 0.79 |
| Правно-административни | 0.70 | 0.75 | 0.87 | 0.85 |
| Новински | 0.83 | 0.88 | 0.94 | 0.89 |
| Књижевни | 0.37 | 0.55 | 0.92 | 0.88 |
| Сви домени | 0.72 | 0.80 | 0.90 | 0.86 |

BERTić се показао као најбољи модел за све домене, након њега COMtext.SR, па CLASSLA – нестандардни и на крају CLASSLA – стандардни. Сви модели су најбоље евалуирали новински домен, затим књижевни, правно-административни, па твитер домен. Ово може бити последица комплексности ентитета у појединим доменима.

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