Зборник радова Факултета техничких наука, Нови Сад

UDK: 007:34
DOI: https://doi.org/10.24867/35BE12Roljic

Компаративна анализа *NER* модела за пресуде на црногорском језику

Comparative Analysis of NER Models for Judgements in Montenegrin

Бранислав Рољић, Стеван Гостојић, *Факултет техничких наука, Нови Сад*

Област – ЕЛЕКТРОТЕХНИЧКО И РАЧУНАРСКО ИНЖЕЊЕРСТВО

Кратак садржај – *Препознавање именованих ентитета у језицима са ограниченим ресурсима отежано је малим бројем доступних анотираних података и доменски специфичних модела. У овом раду се евалуирају трансформерски и генеративни NER приступи на ручно анотираном корпусу црногорских правосудних пресуда. Методологија спаја лингвистичке увиде и технике дубоког учења како би се модели прилагодили правном домену. Спроведени експерименти анализирају перформансе по типовима ентитета и показују кључне предности и ограничења сваког приступа. Резултати доприносе развоју правно-језичких NLP алата и указују на могуће правце даљег унапређења NER система у условима ограничених ресурса.*

Кључне речи: *NER, правни NER, BERT, LLM*

Abstract – *Named entity recognition in under-resourced languages is hindered by limited annotated data and the absence of domain-adapted models. This study evaluates transformer-based and generative NER approaches on a manually annotated corpus of Montenegrin legal texts. The methodology integrates linguistic insight with deep learning techniques to adapt models to the legal domain. Through comprehensive experiments, we examine performance across entity types, assess generalization, and identify key strengths and limitations of each method. The results support the development of NLP tools for Montenegrin legal language and highlight directions for advancing NER in low-resource legal settings.*

Keywords: *NER, legal NER, BERT, LLM*

НАПОМЕНА: Овај рад проистекао је из мастер рада чији ментор је био др Стеван Гостојић, ред. проф.

1. УВОД

Препознавање именованих ентитета (енг. *named entity recognition*, *NER*) у правним текстовима представља изазован задатак, посебно за језике са ограниченим дигиталним ресурсима као што је црногорски. Правни документи карактеришу се формалним стилом, комплексном синтаксом и специфичном терминологијом, што отежава поуздану аутоматску идентификацију ентитета као што су судови, странке и правни акти. Постојећи *NER* модели, иако веома успешни у општим доменима, обично не обезбеђују задовољавајуће резултате у правним корпусима са мањим бројем анотација.

Циљ овог рада је компаративна анализа више приступа заснованих на модерним језичким моделима прилагођеним јужнословенским језицима, као и модела оптимизованих за рад у условима ограничених ресурса. Истраживање обухвата прикупљање и ручну анотацију корпуса црногорских пресуда, дефинисање доменски специфичних категорија ентитета и евалуацију више архитектура.

Вредност овог рада огледа се у систематском поређењу различитих архитектура на истом корпусу, чиме се добија поуздан увид у њихову стабилност, ограничења и применљивост у правном домену. Практична мотивација проистиче из растуће потребе за аутоматизованом анализом правних докумената у Црној Гори, док научни допринос лежи у развоју иницијалних језичких ресурса и метода за *NER* на мање заступљеним језицима.

2. СРОДНА ИСТРАЖИВАЊА

Истраживања у области *NER*-а еволуирала су од раних метода заснованих на лингвистичким правилима и статистичким моделима ка дубоким неуронским мрежама и трансформер архитектурама. Рани *BiLSTM* и *CNN* приступи омогућили су моделовање контекстуалних зависности, али је појава *BERT*-а представљала кључни заокрет захваљујући бидирекционом механизму самопажње и богатим контекстуализованим представама токена. Његови деривати, као што су *DistilBERT* и *RoBERTa*, даље су унапредили резултате у различитим *NER* задацима [1]. Бројне студије показују да трансформери надмашују класичне секвенцијалне моделе у доменски специфичним контекстима. *LegalBERT* значајно премашује *BiLSTM-CRF* у правним текстовима [2], што потврђује предност трансформерске архитектуре у задацима обраде правних докумената. За јужнословенске језике посебно је значајно истраживање [3], које показује да доменски прилагођена токенизација може побољшати *NER* перформансе за 4,5–54,6%, што указује на важност квалитетног сегментирања у морфолошки богатим језицима.

Структурни аспекти *NER*-а такође су предмет анализе: избор анотационе шеме значајно утиче на коначне метрике – једноставни ентитети боље се препознају у *IO* формату, док су *IOBES* шеме погоднije за сложеније структуре [4]. Шири преглед модерних приступа дат је у [1], који истиче тренд ка интеграцији

трансформера, *LLM* модела и техника структурисаног предвиђања као што је *CRF*.
За јужнословенски простор, студија [5] показује да *bcms-BERTuh* и *XLM-R-BERTuh* постижу највише вредности *F1*-мере (до 0,942), што потврђује ефикасност мултијезичне *BCMS* обуке. На правосудним текстовима, рад [6] показује да фино подешавање *BERTuha* на доменски специфичним пресудама даје изузетно високе резултате ($F1 \approx 0.96$), чак и у условима ограничених ресурса.

3. МЕТОД

3.1 Претпроцесирање и анализа података

Како не постоји јавно доступан, анотиран корпус црногорских правних докумената за *NER*, у овом истраживању је конструисан нови корпус од 225 пресуда (7201 ентитет). Документи су прикупљени аутоматизованим *scraping*-ом са портала црногорских судова [7], коришћењем *Playwright* библиотеке [8] ради обраде динамичког једностраничног садржаја (енг. *single-page application*, *SPA*). Након *scraping*-а, вршено је уклањање *HTML*/*CSS* артефаката и нормализација текста у *plain-text* формат погодан за анотацију.
Анотација је извршена у *LabelStudio* [9] окружењу, уз дефинисање 14 ентитетских категорија специфичних за правни домен, заснованих на постојећим правним *NER* системима [2], [6]. Обухваћени су следећи ентитети: назив суда, датум пресуде, број предмета, судија, записничар, тужилац, окривљени, кривично дело, одлука, врста санкције, износ/трајање санкције, материјалноправне одредбе, процесноправне одредбе и трошкови поступка.
Током анотације уочени су бројни изазови, укључујући варијабилне формате ентитета (посебно код бројева предмета и датума), различите обрасце анонимизације и изражену структурну недоследност међу документима.

3.2 Припрема података за моделе

Дужина пресуда у великом броју случајева премашује ограничење од 512 токена које намеће трансформер архитектуре, те је била неопходна примена механизма клизног прозора (енг. *sliding window*) – сегменти дужине 512 токена са вредношћу помераја (енг. *stride*) од 128, што омогућава очување контекста око граница сегмената.
Токенизатор дели речи на подречи, при чему први подтокен добија *BIO* ознаку, а сви остали вредност -100, да би били игнорисани у функцији губитка. У циљу избегавања неважећих *BIO* секвенци, I-ознаке на почетку прозора конвертују се у B-ознаке. Сваки сегмент се затим попуњава (енг. *padding*) и допуњује [*CLS*]/[*SEP*] токенима.
За објективну процену модела и избегавање претренирања примењена је стратификована петострука унакрсна валидација (енг. *stratified 5-fold cross-validation*), уз очување пропорционалне заступљености ентитетских класа у сваком *fold*-у. У свакој итерацији око 180 докумената коришћено је за обуку, док је приближно 45 докумената служило као валидациони скуп.

3.3 NER модели

Циљ методологије је обухватање више комплементарних приступа, од класичних трансформера до генеративних *zero-shot* и *few-shot* модела.

3.3.1 Трансформер модели

BERTuh, као први велики *BCMS* трансформер, је претрениран користећи приступ детекције замењених токена (енг. *replaced token detection*, *RTD*). Овакав приступ омогућава ефикасније учење у нискоресурсним условима, што је посебно важно за црногорски правни домен.
***BERTuh* са тежинама класа** користи модификовану *CrossEntropyLoss* функцију са тежинама пропорционалним учесталости класа, како би се ублажио изражен дисбаланс у корпусу и побољшало препознавање слабо заступљених ентитета.
***BERTuh* са фокалним губитком** – примењује фокални губитак (енг. *focal loss*) како би се смањио утицај лако класификованих примера и појачала осетљивост на тешке и варијабилне ентитете, што доводи до стабилнијих перформанси у правним текстовима.
XLM-R-BERTuh је мултијезичка варијанта *XLM-RoBERTa* модела, додатно обучена на *BCMS* подацима. Претходна истраживања [5] показују да овај модел постиже највише вредности *F1*-мере на *NER* задацима за јужнословенске језике, нарочито код језичких варијација и флективних структура.
BERTuh-CRF **–** ова архитектура комбинује контекстуализоване енкодинге *BERTuh*-а са *CRF* слојем који моделује зависности између ознака у секвенци. *BERTuh* обезбеђује богате представе токена у сложеним правним реченицама, док *CRF* осигурава структурно конзистентне *BIO* секвенце и прецизније обележавање дугачких, вишечланих ентитета, који су чести у правним текстовима.
***BERTuh* са *DAPT-MLM* –** ради бољег хватања специфичне правне терминологије и стилских образаца, модел *BERTuh* је додатно претрениран *MLM* техником на корпусу од 2600 пресуда. Овај поступак омогућава моделовање доменски релевантних дистрибуција речи и побољшава квалитет токенских репрезентација. Као што показује рад [3], *DAPT* доследно унапређује перформансе трансформер модела у задацима *NER*-а специфичним за одређени домен.

3.3.2 Zero-shot и Few-shot модели

GLiNER третира *NER* као *span-level* класификацију и омогућава *zero-shot* препознавање ентитета на основу њихових текстуалних описа, без употребе *BIO* шеме. Модел користи ембединг простор који истовремено представља текст и дефиниције ентитета, што му омогућава да препозна и категорије које нису биле део тренинг скупа, што је посебно значајно у нискоресурсним доменима као што је правни.

PromptNER формулише *NER* као генеративни задатак, где *LLM* добија упутство за издвајање ентитета и производи структуриран *JSON* излаз. Примењени су *zero-shot* и *few-shot* режими, како би се испитала способност *LLM* модела да без додатног обучавања, или уз минималан број примера, обрађују ентитете у специјализованом правном домену.

3.4 Евалуација

Евалуација је изведена на нивоу ентитета, јер делимично препознати ентитети немају практичну вредност у правном домену. За *BIO* секвенце коришћена је библиотека *seqeval* [10], а за *GLiNER* и *PromptNER* прилагођена *span-based* логика.
Мерене су метрике прецизност, одзив и *F1*-мера по класама, као и макро и пондерисани *F1* агрегати.

4. ИМПЛЕМЕНТАЦИЈА

Обука свих модела спроведена је по јединственој методологији која обезбеђује фер поређење, контролисане услове и репродуктивност резултата.

4.1. Рачунарска инфраструктура

Експерименти су реализовани на *cloud* инфраструктури са **NVIDIA RTX A4000 (16GB)**, користећи **Python 3.11**, **PyTorch 2.1** и **HuggingFace Transformers 4.36**. Конзистентна *GPU* конфигурација обезбедила је стабилност и поновљивост обуке свих модела.

4.2. Заједнички хиперпараметри

За све моделе примењени су идентични хиперпараметри: *learning rate*: $\mathbf{3\times10^{-5}}$ (*AdamW*, *weight decay* 0.01), *warmup ratio*: **0.01** (линеарни *schedule*), број епоха: **8** (осим *DAPT-MLM* фазе: 2 епохе), *batch* size: **4**, уз **акумулацију 4 градијента** (ефективни *batch* 16), *mixed precision* (***fp16***) обука, *sliding window*: **512 / stride 128**

4.3. Динамика обуке и конвергенција

Тренинг и валидациони губитак праћени су током обуке, заједно са макро/микро метрикама на сваких 100 корака. Механизам раног заустављања је активиран након три узастопне стагнације *F1*-мере, што је спречило претренирање и обезбедило стабилну конвергенцију. Сви модели су конвергирали у опсегу између шесте и осме епохе.

4.4. Састав тренинг скупа

Применом *sliding window* токенизације добијено је 11.330-12.464 тренинг примера по *fold*-у, док су валидациони скупови садржали 2.747-2.993 примера. Ова расподела одговара 80/20 подели докумената (приближно 180 за тренинг, 45 за валидацију). Корпус је показао изражен дисбаланс класа.

4.5. Стратификована расподела

Стратификација је заснована на комбинацијама ентитетских образаца (16 доминантних шаблона), што је омогућило да сваки *fold* садржи пропорционалан број примера свих типова ентитета.

5. РЕЗУЛТАТИ И ДИСКУСИЈА

У табели 1 су приказани резултати за све варијанте *BERTuh* модела, као и за мултијезички *XLM-R-BERTuh*, те *zero-shot* и *few-shot* приступе.

Табела 1 Поређење резултата различитих NER модела на корпусу црногорских правних пресуда

Модел	Прецизност	Одзив	*F1*-мера
BERTuh	0.8020	0.7656	0.7628
BERTuh + Class Weights	0.4790	0.9213	0.5951
BERTuh + Focal Loss	0.7874	0.7345	0.7319
BERTuh + CRF	0.8109	0.8201	0.8049
BERTuh + DAPT-MLM	0.7980	0.7872	0.7768
XLM-R-BERTuh	**0.8942**	**0.8842**	**0.8858**
GLiNER Large	0.05	0.08	0.07
GLiNER Large v2	0.09	0.07	0.07
GLiNER bi-large (knowledge)	0.19	0.12	0.14
GLiNER multi v2.1	0.19	0.12	0.14
PromptNER Zero-shot	0.5643	0.2910	0.3567
PromptNER Few-shot	0.4848	0.4466	0.4397

Основни *BERTuh* модел представља поуздану референтну основу за *NER* у пресудама. Макро-*F1* износи 0,76, што показује да *BCMS* модели адекватно обрађују синтаксичке и семантичке обрасце карактеристичне за правни језик, упркос ограниченој величини корпуса. Најбоље резултате модел постиже за учестале категорије, док ређе класе остају изазов због ограничене заступљености у подацима.
Примена тежина класа довела је до значајног погоршања резултата. Иако је одзив био веома висок, прецизност је драстично опала, што је резултовало макро-*F1* вредношћу од 0,59. Ово указује да инверзно пондерисање класа није ефикасна стратегија за правне пресуде: због великог дисбаланса ентитета, модел генерише превише лажних позитивних предвиђања.
С друге стране, варијанта *BERTuh*-а са фокалним губитком се показала знатно стабилнијом. Модел је успео да боље контролише утицај тешких примера, уз макро-*F1* $\approx$ 0,73. Фокални губитак се показао као ефикаснија стратегија, јер је селективно појачавао учење на примерима са већим степеном неизвесности, без нарушавања стабилности процеса тренинга.
Увођење *CRF* слоја довело је до побољшања у структурној доследности предвиђања. Модел *BERTuh-CRF* остварује макро-*F1* од 0,80, што указује да експлицитно моделовање транзиција између ознака ублажава грешке на границама ентитета. Ово је нарочито важно за правни домен, где су ентитети често дугачки и вишечлани.
Примена *DAPT-MLM* преттренирања показала је да доменска адаптација има значајну вредност. Модел *BERTuh* са *DAPT-MLM*-ом постиже макро-*F1* од 0,78. Иако ово није највиши резултат међу трансформерима,

доменско преттренирање побољшава репрезентације правних текстова и доприноси стабилнијој конвергенцији у односу на основни модел.

Најбоље резултате остварује *XLM-R-BERTuћ*, са макро-*F1* оценом од 0,89. Мултијезичко *BCMS* преттренирање и снажнија енкодер архитектура *XLM-R* омогућили су моделу да ефикасно обухвати морфолошке и синтаксичке варијације присутне у правним документима. Модел показује ниску варијансу између *fold*-ова и највиши ниво робусности.

Zero-shot приступи су тестирани ради испитивања применљивости модела у условима без анотираних података. *GLiNER* варијанте остариле су веома ниске макро-*F1* вредности (0,07–0,14), са релативно бољим одзивом, али недовољном прецизношћу за практичну употребу у правном домену. Ово је очекивано, јер модели нису прилагођени језичким структурама правосудних пресуда.

PromptNER показује боље перформансе од *GLiNER*-а: *zero-shot* варијанта постиже макро-*F1* од 0,36, а *few-shot* варијанта 0,44. Иако ови резултати и даље знатно заостају за фино подешеним трансформерима, они указују да *LLM* модели имају потенцијал у сценаријима ограничених ресурса, али тренутно не могу заменити трансформерске моделе обучене на доменски специфичном корпусу.

6. ЗАКЉУЧАК

У раду је представљена компаративна анализа више приступа препознавању именованих ентитета у црногорским пресудама, заснована на новоформираном, ручно анотираном корпусу и доменски релевантним категоријама ентитета. Испитани су трансформерски модели *BERTuћ* и *XLM-R-BERTuћ*, као и њихове варијанте са *CRF* слојем, класним тежинама, фокалним губитком и доменском адаптацијом. Резултати показују да најстабилније и најпрецизније решење представља мултијезички *XLM-R-BERTuћ* (≈0.89 макро-*F1*), док *CRF* додатно побољшава конзистентност секвенцијалног означавања. Модел *BERTuћ* са *DAPT-MLM* преттренирањем такође је остварио унапређење у односу на основну варијанту, што потврђује значај доменске адаптације у условима ограничених ресурса. Ови резултати су упоредиви са перформансама достигнутим за српски правни *NER*, што потврђује ефикасност трансформера и доменске адаптације и за мање језике.

Главна ограничења рада односе се на величину корпуса и изостанак подршке за угнежђене ентитете. *Zero-shot* и *few-shot* приступи показују потенцијал, али и даље знатно заостају за моделима обученим на доменским подацима.

Будућа истраживања треба усмерити ка проширивању корпуса (укључујући *active learning* и синтетичке податке), моделовању односа ентитета на нивоу целог документа, подршци за хијерархијске и угнежђене ознаке, као и тестирању архитектура за дуге секвенце. Посебно перспективан смер представља истраживање хибридних модела који би спојили високу прецизност трансформера у означавању секвенци са напредним контекстуалним резоновањем *LLM* модела.

7. ЛИТЕРАТУРА

[1] I. Keraghel, S. Morbieu, и M. Nadif, „Recent Advances in Named Entity Recognition: A Comprehensive Survey and Comparative Study“, 20. Децембар 2024., arXiv: arXiv:2401.10825. doi: 10.48550/arXiv.2401.10825.

[2] H. Darji, J. Mitrović, и M. Granitzer, „German BERT Model for Legal Named Entity Recognition“, у Proceedings of the 15th International Conference on Agents and Artificial Intelligence, 2023, стр. 723–728. doi: 10.5220/0011749400003393.

[3] M. Bogdanović, M. Frtunić Gligorijević, J. Kocić, и L. Stoimenov, „An Analysis of the Training Data Impact for Domain-Adapted Tokenizer Performances—The Case of Serbian Legal Domain Adaptation“, Appl. Sci., том 15, изд. 13, стр. 7491, Јули 2025, doi: 10.3390/app15137491.

[4] I. Ait Talghalit, H. Alami, и S. O. El Alaoui, „Exploring Different Annotation Schemes for Single and Consecutive Named Entity Recognition in the Arabic Biomedical Domain using Transformer Models and Contextual Semantic Embeddings“, Eng. Technol. Appl. Sci. Res., том 15, изд. 2, стр. 21854–21860, Апр. 2025, doi: 10.48084/etasr.10019.

[5] M. Škorić, „New Language Models for Serbian“, Infotheca, том 24, изд. 1, стр. 7–28, 2024, doi: 10.18485/infotheca.2024.24.1.1.

[6] V. Kalušev и B. Brkljač, „Named entity recognition for Serbian legal documents: Design, methodology and dataset development“, 14. Фебруар 2025., arXiv: arXiv:2502.10582. doi: 10.48550/arXiv.2502.10582.

[7] „Sudovi Crne Gore“. Приступљено: 28. Октобар 2025. [На Интернету]. Available at: https://sudovi.me/sdvi/odluke

[8] Playwright. [На Интернету]. Приступљено: 28. Октобар 2025. Available at: https://playwright.dev/

[9] LabelStudio. [На Интернету]. Приступљено: 28. Октобар 2025. Available at: https://labelstud.io/

[10] seqeval. [На Интернету]. Приступљено: 28. Октобар 2025. Available at: https://huggingface.co/spaces/evaluate-metric/seqeval

Кратка биографија:

Бранислав Рољић рођен је 1999. године у Бањалуци. Основне академске студије завршио је 2022. године на Електротехничком факултету у Бањалуци, а Мастер академске студије на Факултету техничких наука у Новом Саду 2025. године.
Контакт:branislavroljic99@gmail.com