Компаративна анализа NER модела за пресуде на црногорском језику

Бранислав Рољић
Департман за рачунарство и аутоматику
Факултет техничких наука Универзитета у Новом Саду
Нови Сад, Република Србија
branislavroljic99@gmail.com

Стеван Гостојић
Департман за рачунарство и аутоматику
Факултет техничких наука Универзитета у Новом Саду
Нови Сад, Република Србија
gostojic@uns.ac.rs

***Сажетак* — Препознавање именованих ентитета у језицима са ограниченим ресурсима представља бројне изазове, пре свега због недостатка анотираних корпуса и специјализованих модела. У раду се евалуирају трансформерски и генеративни *NER* модели на ручно анотираном корпусу правних текстова, креираном у оквиру истраживања. Методологија комбинује лингвистичке увиде и технике дубоког учења како би се модели прилагодили специфичностима правног језика у Црној Гори. Кроз опсежне експерименте анализирају се перформансе модела по типовима ентитета, процењује се генерализација и утврђују снаге и слабости различитих приступа. Резултати доприносе развоју алата за обраду природног језика у правном домену Црне Горе и указују на правце даљег унапређења *NER* система у условима ограничених ресурса.**

Кључне речи— NER, правни NER, BERT, LLM

I. 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, процесни учесници 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 судских 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.

У другом поглављу дат је сажет преглед релевантних истраживања из области препознавања именованих ентитета, са нагласком на правни домен и језике са ограниченим ресурсима. Треће поглавље описује процес изградње и ручне анотације корпуса, као и дефинисање доменски специфичних категорија ентитета. У четвртом поглављу приказан је развој и прилагођавање анализираних модела, док пето поглавље представља методологију евалуације и резултате компаративне анализе. На крају, шесто поглавље сумира кључне закључке и предлаже потенцијалне правце будућег истраживања.

II. Претходна решења

Истраживања у области *NER*-а прошла су кроз значајан развој – од метода заснованих на лингвистичком и семантичком знању и карактеристикама ентитета, преко статистичких и модела заснованих на правилима и шаблонима, ка дубоким неуронским мрежама и трансформер архитектурама. Рани радови засновани на *BiLSTM* и *CNN* моделима показали су могућност ефикасног моделовања контекстуалних зависности у секвенцама, док је појава *BERT*-а представљала кључну прекретницу, јер је бидирекциони механизам самопажње омогућио коришћење целокупног контекста приликом предвиђања ентитетских ознака.

Трансформер-базирани приступи, посебно *BERT* и његови деривати попут *DistilBERT*-а и *RoBERTa*-е, постали су доминантни у *NER*-у захваљујући богатим контекстуализованим ембединзима и могућности ефикасног финог подешавања [1].

Више истраживања потврђује да трансформер-базирани модели значајно надмашују класичне секвенцијалне приступе у доменским *NER* задацима. Рад [2] показује да *Legal-BERT* јасно премашује *BiLSTM-CRF* на правним текстовима, док рад [3] демонстрира да трансформери у комбинацији са доменски специфичним корпусима дају најбоље резултате. Слично томе, рад [4] потврђује да *BERT-CRF* надмашује више алтернативних модела, што додатно потврђује стабилност и ефикасност трансформерске архитектуре у специјализованим областима.

Посебан значај за јужнословенске језике има студија [5], која утврђује да доменски прилагођен токенизатор побољшава метрике *NER*-а за 4,5–54,6%, што потврђује важност квалитетне токенизације у морфолошки богатим језицима.

Поред архитектуре модела, више радова бави се и структурним аспектима *NER*-а. Истраживање [6] показује да избор шеме анотације (*IO*, *BIO*, *IOBES* итд.) значајно утиче на перформансе модела: једноставни ентитети се боље препознају у *IO* шеми, док сложенији постижу боље резултате у *IOBES* формату.

Рад [1] даје најобухватнији савремени преглед *NER* метода и истиче помак ка интеграцији *LLM* модела, метода заснованих на упитима и техника структурисаног предвиђања попут *CRF*-а.

У оквиру јужнословенских језика, студија [7] систематски упоређује 10 трансформер модела и показује да *bcms-BERTić* и *XLM-R-BERTić* постижу највише *F1* вредности (до 0,942) на *NER* задацима, што потврђује предност мултијезичке *BCMS* обуке.

Рад [8] се директно односи на правосудне текстове на српском језику и показује да фино подешавање *BERTuh* модела на доменски специфичним пресудама даје изузетно високе резултате ($F1 \approx 0{,}96$), чак и у условима ограничених ресурса.

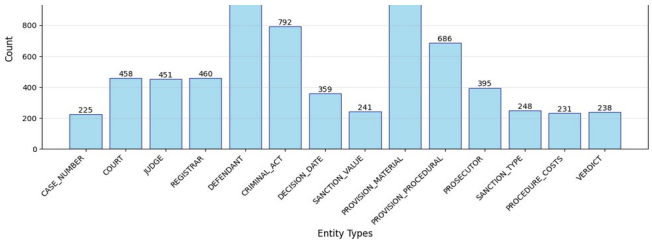
III. Метод

A. 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 судских пресуда (7.201 ентитет). Документи су прикупљени аутоматизованим *scraping*-ом са портала црногорских судова [9], коришћењем *Playwright* библиотеке [10] 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* [11] окружењу, уз дефинисање 14 ентитетских категорија специфичних за правни домен, заснованих на постојећим правним *NER* системима [2], [3]. Обухваћени су следећи ентитети: **COURT** (назив суда), **DECISION_DATE** (датум пресуде), **CASE_NUMBER** (број предмета), **JUDGE** (судија), **REGISTRAR** (записничар), **PROSECUTOR** (тужилац), **DEFENDANT** (окривљени), **CRIMINAL_ACT** (кривично дјело), **VERDICT** (тип пресуде), **SANCTION_TYPE** (врста санкције), **SANCTION_VALUE** (износ/трајање санкције), **PROVISION_MATERIAL** (материјалноправна одредба), **PROVISION_PROCEDURAL** (процесноправна одредба) и **PROCEDURE_COSTS** (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. На слици 1 је приказана расподела класа након извршене анотације, из које се уочава изразит дисбаланс у заступљености појединих ентитетских категорија.



Слика 1 Расподела класа у тренинг скупу

Б. Анотациона шема и BIO формализација

Дефинисање граница ентитета је од кључног значаја за моделовање секвенци. Иако постоје различите шеме анотације, у *NER*-у је најшире усвојена *BIO* шема [1], која експлицитно обележава почетке ентитета. Иако је *IO* шема понекад ефикаснија [6], *BIO* пружа јасније структурне сигнале у сложеним правним текстовима, па је усвојена као основна.

Како *LabelStudio* не подржава *BIO* токенизацију, примењено је ручно поравнавање анотираних сегмената текста са токенима: први токен добија „B-TYPE“, наредни „I-TYPE“, а остали токени „O“.

A. Припрема података за моделе

Дужина пресуда у великом броју случајева премашује ограничење од 512 токена које намећу 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 релевантних структура података у сваком *fold*-у. С обзиром на то да пресуде у корпусу значајно варирају у типовима ентитета које садрже, класична стратификација по фреквенцији класа није била довољна.

Уместо тога, стратификација је заснована на *entity-pattern signature* приступу: сваки документ је мапиран на сигнатуру која представља скуп типова ентитета присутних у том документу. Идентификовано је 16 различитих сигнатура, а документи су распоређени у *fold*-ове тако да сваки *fold* садржи репрезентативан однос честих и ретких структурних образаца. Овакав приступ обезбеђује стабилнију и поузданију процену генерализације модела у односу на стандардне технике.

Имплементација је реализована коришћењем библиотеке *scikit-learn* [12], при чему свака итерација садржи приближно 180 докумената за обуку и око 45 докумената за валидацију.

Г. NER модели

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

Г.1 Трансформер модели

BERTuh, као први велики *BCMS* трансформер, је претрениран у *ELECTRA* оквиру [13], користећи приступ детекције замењених токена (енг. *replaced token detection*, *RTD*) уместо класичног моделовања језика са маскирањем (енг. *masked language modeling*, *MLM*). Овакав приступ омогућава ефикасније учење у нискоресурсним условима, што је посебно важно за црногорски правни домен.

***BERTuh* са тежинама класа** користи стандардни *BERTuh* енкодер, уз модификовану *CrossEntropyLoss* функцију која примењује тежине пропорционалне реткости класа у корпусу. На овај начин се ублажава ефекат израженог дисбаланса у анотираним пресудама, где поједине категорије (нпр. REGISTRAR, CASE_NUMBER) имају знатно мањи број примера од доминантних класа (DEFENDANT, PROVISION_MATERIAL). Тежине класа (енг. *class weights*) побољшавају способност модела да идентификује слабо заступљене ентитете, нарочито у сценаријима где лоша расподела иначе доводи до систематичног занемаривања мањинских класа.

***BERTuh* са фокалним губитком** је трениран помоћу функције фокалног губитка (енг. *focal loss*) , која смањује утицај лако класификованих примера и наглашава оне код којих је модел несигуран. Ова стратегија је посебно корисна за ентитете са великом унутрашњом варијабилношћу (нпр. CRIMINAL_ACT, VERDICT), као и за класе које се јављају у различитим синтаксичким облицима. Функција фокалног губитка не решава само дисбаланс класа, већ побољшава осетљивост модела на појаве „тешких“ примера, што резултује стабилнијим перформансама у правном домену.

XLM-R-BERTuh је мултијезичка варијанта *XLM-RoBERTa* модела, додатно обучена на *BCMS* подацима. Претходна истраживања [7] 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 2.600 правних докумената. Овај поступак омогућава моделовање доменски релевантних дистрибуција речи и побољшава квалитет токенских репрезентација. Као што показује рад [5], *DAPT* доследно унапређује перформансе трансформер модела у задацима *NER*-а специфичним за одређени домен.

Г.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.

Д. Евалуација

Евалуација је вршена на нивоу ентитета, јер делимично препознати ентитети немају практичну вредност у правном домену. За *BIO* секвенце коришћена је библиотека *seqeval* [14], 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. *F1*-мера је главна метрика јер истовремено кажњава и лажне позитиве и лажне негативе, при чему пондерисана варијанта боље одражава реалну дистрибуцију података.

IV. Тренинг процес

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

А. 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.

Б. Заједнички хиперпараметри

За све моделе примењени су идентични хиперпараметри:

- *learning rate*: **3×10^{-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**

В. 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.

Г. 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: више од 90% токена носи ознаку „O“, док су типови као што су DEFENDANT, PROVISION_MATERIAL и CRIMINAL_ACT знатно учесталији од ентитета типа CASE_NUMBER или SANCTION_TYPE.

Д. 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. Овај приступ је значајно умањио ризик да поједини ретки ентитети буду недовољно присутни у валидацији, а истовремено је обезбедио стабилне метрике током свих пет итерација.

Е. Анализа тренинг процеса

BERTuћ је показао стабилну конвергенцију без већих осцилација што указује на његову доследну применљивост у задацима препознавања именованих ентитета у правним текстовима.

BERTuћ са тежинама класа је показао нешто нестабилнију конвергенцију, уз повећан број лажно позитивних предвиђања. Иако је одзив био висок, прецизност је остала осетно нижа, што је указало на ограничен домет класних тежина у овом домену.

BERTuћ са фокалним губитком је обезбедио контролисанији ток тренинга у односу на приступ са тежинама класа. Губитак се брже стабилизовао, а модел је био робуснији на ретким класама, уз мање осцилација током обуке.

BERTuћ-CRF је постигао најстабилније криве губитка. Додавање *CRF* слоја донело је структурну регуларизацију, чиме су значајно умањене грешке на границама ентитета и побољшано целокупно секвенцијално предвиђање.

BERTuћ са *DAPT-MLM* је имао најбржу конвергенцију и најстабилнији валидациони губитак. Доменски адаптивно преттренирање омогућило је моделу да већ у раним епохама демонстрира супериорно разумевање правног језика.

XLM-R-BERTuћ је постигао најбоље укупне резултате. Мултијезичка архитектура и лингвистичка блискост *BCMS* језика омогућили су највећу стабилност између *fold*-ова и највише вредности *F1*-мере.

V. Резултати и дискусија

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

ТАБЕЛА I 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 *BERTuћ* модел представља поуздану референтну основу за *NER* у правним текстовима. Остварени макро-*F1* од 0,76 показује да преттренирани *BCMS* модели добро покривају синтаксичке и семантичке обрасце карактеристичне за правне документе, упркос ограниченој величини корпуса. Модел је постизао најбоље резултате за честе категорије као што су DEFENDANT и PROVISION_MATERIAL, док су најниже вредности забележене за ређе класе као што су SANCTION_VALUE и PROCEDURE_COSTS.

Примена тежина класа је довела до значајног погоршања резултата. Иако је одзив био висок, прецизност је знатно опала, што је резултовало просечним макро-*F1* од само 0,59. Ово указује да инверзно пондерисање класа у правним текстовима није оптимално решење – велике разлике у фреквенцији ентитета доводе до преагресивног „поравнавања“ губитка, што производи велики број лажно позитивних предвиђања.

С друге стране, варијанта *BERTuћ*-а са фокалним губитком се показала знатно стабилнијом. Модел је успео да боље контролише утицај тешких примера, уз макро-*F1* ≈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 је резултовало побољшаном доследношћу секвенцијалних предвиђања. Макро-*F1* ≈ 0,80 указује да структурно моделовање транзиција између ознака елиминише велики број грешака 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 *BERTuћ* 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.

XLM-R-BERTuћ је остварио најбоље резултате – макро-F1 ≈ 0,89, уз најнижу варијансу између *fold*-ова. Мултијезички модел је искористио шире *BCMS* језичко покриће и предобучену структуру *XLM-R* енкодера, што је резултовало бољим разумевањем морфолошких и синтаксичких образаца за све категорије ентитета. Нарочито је био ефикасан у препознавању ентитета који

имају високу разноликост формулација, као што су CRIMINAL_ACT и PROVISION_PROCEDURAL.

Zero-shot приступи су тестирани ради испитивања применљивости модела у условима без анотираних података. *GLiNER* варијанте остварile 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 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.

VI. 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 *BERTuh* и *XLM-R-BERTuh*, 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-BERTuh* (~0,89 макро-*F1*), док *CRF* додатно побољшава конзистентност секвенцијалног означавања. Модел *BERTuh* 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.

References

[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] P. Kalamkar, A. Agarwal, A. Tiwari, S. Gupta, S. Karn, и V. Raghavan, „Named Entity Recognition in Indian court judgments“, 07. Новембар 2022., *arXiv*: arXiv:2211.03442. doi: 10.48550/arXiv.2211.03442.

[4] S.-S. Chen, R.-H. Hwang, C.-Y. Sun, Y.-D. Lin, и T.-W. Pai, „Enhancing Cyber Threat Intelligence with Named Entity Recognition Using BERT-CRF“, у GLOBECOM 2023 - 2023 IEEE Global Communications Conference, Kuala Lumpur, Malaysia: IEEE, Дец. 2023, стр. 7532–7537. doi: 10.1109/GLOBECOM54140.2023.10436853.

[5] 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“, *Applied Sciences*, том 15, изд. 13, стр. 7491, Јули 2025, doi: 10.3390/app15137491.

[6] 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.

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

[8] 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.

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

[10] *Playwright*. [На Интернету]. Available at: https://playwright.dev/

[11] *LabelStudio*. [На Интернету]. Available at: https://labelstud.io/

[12] *scikit-learn*. [На Интернету]. Available at: https://scikit-learn.org

[13] K. Clark, M.-T. Luong, Q. V. Le, и C. D. Manning, „ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators“, 23. Март 2020., *arXiv*: arXiv:2003.10555. doi: 10.48550/arXiv.2003.10555.

[14] *seqeval*. [На Интернету]. Available at: https://huggingface.co/spaces/evaluate-metric/seqeval