

Оливера Стојановић¹
Библиотека града Београда
oliverakrstic@ymail.com

ПРИМЕНА ВЕШТАЧКЕ ИНТЕЛИГЕНЦИЈЕ У АУТОМАТСКОЈ УДК КЛАСИФИКАЦИЈИ: ПРЕГЛЕД ОБЈАВЉЕНИХ ИСТРАЖИВАЊА

САЖЕТАК: Рад доноси преглед литературе и истраживања, објављених у периоду од 2020. до априла 2025. године, посвећених примени техника машинског учења и обраде природног језика у области аутоматске класификације грађе, са посебним освртом на систем Универзалне децималне класификације (УДК). Циљ рада је да пружи увид у савремене токове који се баве овом темом у библиотекарству, као и да понуди кратак увод у кључне појмове као што су аутоматска класификација, семантички веб, обрада природног језика и машинско учење. У закључку се истиче неопходност развоја локалних ресурса, као и едукација стручног кадра за рад с технологијама вештачке интелигенције, како би овакве или сличне примене биле могуће и одрживе у домаћој пракси.

КЉУЧНЕ РЕЧИ: аутоматска класификација, УДК, семантички веб, вештачка интелигенција, машинско учење, обрада природног језика.

Увод

Библиотечки класификациони системи, као што је познато, служе за систематско организовање грађе у складу са њеном стручном припадношћу, чиме се омогућава лакша, ефикаснија и бржа претрага информација. Посао који обављају стручњаци из области библиотекарства и информатике на пољу каталогизације и класификације грађе у конвенционалним библиотекама је временски

¹ <https://orcid.org/0000-0002-6387-1546>

захтеван и не тако брз процес, а игра виталну улогу у раду сваке библиотеке. Недовољно прецизна или непотпуна каталошка обрада доводи у питање њену употребну вредност, па самим тим и сврху библиотеке као информационог центра. Са појавом дигиталних библиотека, семантичког веба и све већим обимом дигиталних садржаја, јавила се потреба за решењима која би могла да убрзају и олакшају поступак класификације. Савремени развој дигиталних технологија и вештачке интелигенције отвара могућности за модернизацију и усклађивање класификационих шема са захтевима дигиталног окружења.

У том контексту јавља се питање у којој мери и на који начин вештачка интелигенција може допринети ефикаснијој аутоматској класификацији библиотечке грађе. Циљ овог рада јесте да представи актуелне приступе и резултате примене вештачке интелигенције у области аутоматске класификације, са посебним освртом на Универзалну децималну класификацију (УДК), као и да читаоцу приближи кључне појмове неопходне за разумевање теме.

Семантички веб и библиотечки каталози

Семантички веб је почетком 21. века дефинисан као „проширење [тадашњег веба] у којем је информацијама дато добро дефинисано значење, што боље омогућава сарадњу између рачунара и људи”, засновану на децентрализованим моделима представљања и повезивања знања.² Тренд слободног приступа базама података, који подразумева да базе података буду јавне, отворене и доступне свима на интернету, довео је до развоја система „отворених повезаних података” (*Linked Open Data – LOD*). LOD омогућава повезивање различитих скупова података, стварајући тако глобалну мрежу знања. Ново веб окружење захтева од библиотека да своје метаподатке и информације повежу са ресурсима на интернету. Традиционални библиотечки каталози, базирани на машински читљивим (MARC) записима, нису усклађени са принципима семантичког веба. Данас библиотеке раде на трансформацији ових статичних записа у RDF (*Resource Description Framework*), како би им се доделила значења и омогућила боља повезаност са веб-ресурсима. RDF је оквир за описивање веб-ресурса, односно стандард семантичког веба који је развио конзорцијум W3C (*World Wide Web Consortium*).³ Базира се на тзв. „креирању тројки” (субјекат –

Б
И
Б
Л
И
О
Т
Е
К
А
Р

бр.
2,
год.
2025.

² Berners-Lee, Tim, James Hendler, and Ora Lassila, “The Semantic Web.” *Scientific American* 284 no. 5 (2001): 36–37. <https://doi.org/10.1038/scientificamerican0501-34> (преузето 4. 4. 2025).

³ World Wide Web Consortium (W3C) – <https://www.w3.org/> (преузето 4. 4. 2025).

објекат – предикат)⁴ и представља основу семантичког веба, односно фундаментални начин на који се информације представљају и повезују у овом окружењу. Кључна разлика између формата MARC и RDF јесте у томе што је MARC примарно усредсређен на структуру података за каталогизацију, док се RDF фокусира на односе и семантичко значење. „Примена технологија семантичког веба не подстиче нужно промену парадигме у подручју библиотечког пословања, већ нуди механизам за примену нових технолошких могућности које омогућавају размену информација и повезивање колекција са релевантним ресурсима било где на вебу, а корисници имају могућност да са једног места сагледају информације са више аспеката и у више расположивих ресурса”⁵ Иако је у Србији успостављен систем узајамне каталогизације (COBISS), који обухвата и развој нормативних датотека на националном нивоу, њихова примена у пракси и даље је неравномерна.⁶ Нормативне датотеке су есенцијалне за успостављање ауторизованих облика имена, наслова и предмета, што је предуслов за ефикасно креирање семантички обогатених и повезаних података.

Аутоматска класификација и УДК

Развој дигиталних библиотека и дигиталних репозиторијума унутар истраживачких и академских заједница, где метаподаци играју кључну улогу, доводи до потребе за убрзаном класификацијом садржаја који се у њима налази. Ручна класификација је временски напоран задатак и у многим случајевима сматра се непрактичном, будући да се велики број новог материјала свакодневно објављује. „Све већи број е-докумената, као и високи трошкови службе за ручну каталогизацију и класификацију, доводе до појаве аутоматске класификације, односно категоризације докумената (*Automatic Text Classification – ATC*)”⁷ Када се у литератури говори о аутоматској класификацији, она се генерално спомиње у два различита контекста. Први подразумева аутоматску класифика-

⁴ David Beckett, *RDF 1.2 N-Triples*, <https://www.w3.org/TR/rdf12-n-triples/> (преузето 24. 8. 2025).

⁵ Јелена Андоновски, „Мрежа отворених података и језички ресурси у процесу изградње српско-немачког литерарног корпуса: докторска дисертација” (Београд: [Ј. Андоновски], 2019), <https://phaidrabg.bg.ac.rs/o:22874>, (преузето 3. 4. 2025).

⁶ Александра Тртовац и Наташа Дакић, „База CONOR.SR у систему COBISS.SR”, *Infoteka: Journal for Digital Humanities* v. 20, n. 1–2a (feb. 2021): 75–88, https://infoteka.bg.ac.rs/ojs/index.php/Infoteka/article/view/2020.20.1_2.5_sr (преузето 4. 4. 2025).

⁷ Arash Joorabchi and Abdulhussain E. Mahdi, “An Unsupervised Approach to Automatic Classification of Scientific Literature Utilizing Bibliographic Metadata”, *Journal of Information Science* 37 no. 5 (2011): 499–514, <https://doi.org/10.1177/0165551511417785>, (преузето 4. 4. 2025).

цију електронског или дигиталног текста, текста који је доступан у неком електронском формату, а други контекст подразумева аутоматску доделу класификационих бројева у библиотечким системима, при чему се користи електронска верзија класификационе шеме. Аутоматска класификација дигиталног текста „представља значајну истраживачку област, јер тачност постојећих система није савршена и потребна су даља побољшања”.⁸ Сматра се виталним методом за управљање и обраду огромне количине текста у дигиталном облику, који су свеprisутни и континуирано расту.

Када се говори о УДК као о повезаним подацима (*Linked Data - LD*), мисли се такође на две ствари: УДК као изворни податак, тј. сам УДК систем и распоред чувања у његовој матичној бази података (*UDC Master Reference File - MRF*) и УДК бројеви који се примењују у описима ресурса и који се појављују у метаподацима библиографских база података, услугама индексирања библиотечких каталога и библиотечких полица. „Да би служио сврси семантичког веба, потребно је повезати УДК као LD са сервисом који може анализирати и интерпретирати сложене УДК бројеве. Док је употреба УДК бесплатна, објављивање и дистрибуција података је заштићена лиценцом. Објављивање УДК и као LD и као LOD мора бити обезбеђена у оквиру комплексног сервиса, који би омогућио отворен приступ као и приступ кроз различите „paywall”⁹ баријере за различите нивое лиценци”.¹⁰ Дакле, сложеност УДК бројева, која произилази из комбинација основних бројева и употребе помоћних таблица, захтева употребу напредних алата за машинско разумевање и повезивање података. Библиотеке могу бесплатно користити УДК за потребе каталогизације и организације колекција, али комплетна база података (MRF) и њена званична дистрибуција су власништво УДК Конзорцијума (*UDC Consortium - UDCC*) и заштићени су лиценцом. Ово представља ограничење за потпуну интеграцију УДК у окружење LOD. У пракси је доступан само део УДК података, као што је *UDC Summary*,¹¹ који је објављен као LOD, док је приступ комплетној бази условљен лиценцама.

⁸ Emmanouil Ikonomakis, Sotiris Kotsiantis, and V. Tampakas, “Text Classification Using Machine Learning Techniques”, *WSEAS Transactions on Computers* 4 (2005): 966–974, https://www.researchgate.net/publication/228084521_Text_Classification_Using_Machine_Learning_Techniques, (преузето 16. 8. 2025).

⁹ Метод ограничавања приступа онлајн садржају путем плаћања претплате.

¹⁰ Aida Slavic, Ronald Siebes and Andrea Scharnhorst, “Publishing a Knowledge Organization System as Linked Data: The Case of the Universal Decimal Classification”, *ArXiv abs/2205.01395* (2022), <https://doi.org/10.5771/9783956506611-69>, (преузето 27. 3. 2025).

¹¹ Universal Decimal Classification Summary; <https://udcsummary.info/php/index.php> (преузето 28. 8. 2025).

Машинско учење и обрада природног језика у функцији аутоматске класификације

Према једној од дефиниција, **вештачка интелигенција** је „наука која се користи за конструисање интелигенције употребом хардверских и софтверских решења, а инспирисана је начином на који функционишу неурони у мозгу”.¹² Суштински она аутоматизује битне аспекте људске интелигенције, као што су учење, закључивање, решавање проблема, па и одлучивање. Како се у литератури наводи, често се меша са појмом машинско учење (*Machine Learning, ML*), иако је машинско учење само једна од њених подобласти. Према Артуру Лију Самјуелу, који је појам дефинисао још 1959. године, машинско учење омогућава рачунарима да уче без потребе за експлицитним програмирањем.¹³ То у основи значи да машинско учење има за циљ конструисање алгоритама који уче на бази података и искуства. Рачунару се да опис неког објекта и као резултат се очекује класификација тог истог објекта, односно препознавање његовог обрасца. Алгоритми се тренирају на великим скуповима података, са циљем да препознају обрасце и донесу одлуке или предвиђања на основу тих образаца. При томе, треба нагласити битност начина на који рачунар прикупља податке који су неопходни за исправну класификацију, јер „ако су подаци који се достављају вештачкој интелигенцији ограничени, пристрасни или ниског квалитета, онда се нужно добија и пристрасна вештачка интелигенција ограниченог обима”.¹⁴ У оквиру машинског учења разликују се и различити приступи, у зависности од начина на који се алгоритми тренирају над скуповима података, па тако учење може бити надгледано (*Supervised Learning*), ненадгледано (*Unsupervised Learning*) или учење поткрепљивањем (*Reinforcement Learning*).¹⁵ Надгледано учење се заснива на коришћењу означених података где је за сваки улазни пример унапред познат тачан излаз, што омогућава моделу да учи на основу примера са већ дефинисаним одговорима. То у основи значи да се означени подаци користе да „науче” моделе како да ефикасно категоризују текстове. Насупрот томе,

¹² Жолт Нађ, *Основе вештачке интелигенције и машинског учења* (Београд: Компјутер библиотека, 2019), 2.

¹³ Жолт Нађ, *Основе вештачке интелигенције и машинског учења*.

¹⁴ Универзитет Унион Рачунарски факултет, „Шта је машинско учење и шта су интелигентни алгоритми?”, <https://raf.edu.rs/citaliste/najnoviji-it-dogadjaji/sta-je-masinsko-ucenje-i-sta-su-inteligentni-algoritmi> (преузето 3. 4. 2025).

¹⁵ Jeff Heaton, “Review of *Deep Learning*, by Ian Goodfellow, Yoshua Bengio, and Aaron Courville”, *Genetic Programming and Evolvable Machines* 19 (2018): 305–307, <https://doi.org/10.1007/s10710-017-9314-z>, (преузето 24. 8. 2025).

ненадгледано учење примењује се у ситуацијама када су доступни само подаци без пратећих ознака, па модел самостално открива унутрашње структуре и обрасце у тим подацима. Трећи приступ, учење поткрепљивањем, подразумева да модел кроз континуирану интеракцију са окружењем стиче знање путем механизма награђивања и кажњавања, односно добијањем позитивних или негативних повратних информација на основу својих акција.¹⁶ Конкретно „у класификацији текста користе се различите технике машинског учења, које могу бити надгледане, ненадгледане или хибридне“.¹⁷ Хибридни модели су модели који комбинују надгледане и ненадгледане методе како би побољшали перформансе класификације.

Надгледано учење је најчешћа стратегија у класификацији текста, где се означени подаци користе да „науче“ моделе како да ефикасно категоричу текстове. Ова метода дели скупове података на три подскупа: за тренинг, тест и валидацију. Први се користи за изградњу модела; валидациони скуп подешава параметре и процењује перформансе модела током развоја; а тест-скуп је резервисан за коначну процену.¹⁸

Када се говори о обради природног језика (*Natural Language Processing, NLP*) наводи се да је то „област рачунарских наука која развија системе за разумевање природног језика“.¹⁹ Као подобласт вештачке интелигенције, бави се омогућавањем рачунарима да разумеју, интерпретирају, анализирају и стварају природни људски језик. То се постиже комбинацијом лингвистичких модела (за описивање граматике, значење и структуре реченица) и алгоритама машинског учења, како би рачунари могли да из великих количина текстова препознају обрасце и значења која се појављују у језику. Један од најзначајнијих напредака у пољу обраде природног језика представљају велики језички модели (*Large Language Models – LLM*).

Велики језички модели су настали из обраде природног језика, али су без сумње постали један од најреволуционарнијих технолошких напредака у пољу вештачке интелигенције последњих

¹⁶ Jeff Heaton “Review of *Deep Learning*, by Ian Goodfellow, Yoshua Bengio, and Aaron Courville”.

¹⁷ Hesham Allam, Lisa Makubvure, Benjamin Gyamfi, Kwadwo Nyarko Graham, and Kehinde Akinwolere, “Text Classification: How Machine Learning Is Revolutionizing Text Categorization”, *Information* 16 no. 2 (2025): 130, <https://doi.org/10.3390/info16020130>. (преузето 15. 7. 2025).

¹⁸ Hesham Allam, et al. “Text Classification: How Machine Learning Is Revolutionizing Text Categorization”.

¹⁹ Oxford University Press, “Natural-language processing”, in *Oxford Reference*, 2024, <https://www.oxfordreference.com/display/10.1093/oi/authority.20110803100225333>. (преузето, 23.05.2025).

година. Важан увид који су донели јесте да се знање о свету и језицима може стећи кроз задатке моделирања језика великих размера, и на тај начин можемо креирати универзални модел који решава различите проблеме.²⁰

У суштини, LLM-ови су обучени да предвиде следећу највероватнију реч у низу, користећи своје научно знање које произилази из сложености веза које су успоставили између података и образаца током тренирања. У контексту библиотекарства, обрада природног језика је значајна у примени аутоматске класификације и индексирања грађе јер омогућава рачунарима да „разумеју” текстуални садржај докумената. Своју примену налази и у побољшању претраживања, као и у препоручивању литературе.

Преглед објављених истраживања

Истраживања представљена у овом раду објављена су у рецензираним стручним часописима и на академским платформама, а одабрана су она која су у слободном, односно отвореном приступу. Издвојени су радови који се директно баве применом вештачке интелигенције у аутоматској класификацији, са посебним фокусом на употребу УДК, будући да је то главни систем класификације који користе библиотеке у Србији.

Претрага литературе извршена је у априлу 2025. године у базама података Emerald Insight, IEEE Xplore, ScienceDirect, као и путем отворених академских репозиторијума као што су Academia.edu и Phaidra. Ове базе података су коришћене јер су широко познате и често се користе у научној заједници. Као кључне речи коришћене су комбинације: “Universal Decimal Classification”, “automatic classification”, “machine learning”, “artificial intelligence” и “natural language processing”. Временски оквир био је ограничен на период од 2020. до 2025. године, како би се осигурала актуелност обухваћених резултата.

Критеријуми за укључивање били су:

- доступност пуног текста;
- повезаност са темом примене УДК у аутоматској класификацији;
- примена техника вештачке интелигенције;
- рецензирани научни или стручни извор.

²⁰ Tong Xiao and Jingbo Zhu, *Foundations of Large Language Models* (NLP Lab, Northeastern University & NiuTrans Research, 2025), <https://github.com/NiuTrans/NLPBook/tree/main> (преузето 20. 6. 2025).

Из анализе су искључени радови који:

- нису имали директан фокус на УДК, већ на друге класификационе системе;
- нису били доступни у пуном тексту;
- нису се бавили темом директне примене УДК у аутоматској класификацији.

У коначни корпус укључена су три рада, јер су то једини радови који испуњавају све наведене критеријуме претраге. Мала величина корпуса и ограничен број радова доступних у слободном приступу наглашава да је ово поље још у раној фази развоја и да су даља истраживања неопходна. Ово представља и природно ограничење овог прегледа.

У чланку „Automatic classification of older electronic texts into the Universal Decimal Classification – UDC”, објављеном 2021. године у часопису *Journal of Documentation*, Матјаж Крагељ из Народне и универзитетске библиотеке Словеније и Мирјана Кљајић Борштнар са Факултета организационих наука Универзитета у Марибору, нуде модел решења за аутоматску класификацију старих дигитализованих докумената, који је заснован на машинском учењу. Сврха њиховог истраживања била је да осмисле модел способан за аутоматску класификацију дигитализованих текстова, чиме би се библиотекама олакшао процес класификовања старијих публикација без УДК ознаке. Како наводе, „Новији чланци и текстови у дигиталним библиотекама су обично опремљени метаподацима, као што су теме, кључне речи или УДК бројеви, док стари текстови нису. Количина архивских текстова и чланака који су доступни преко дигиталних библиотека огромна је, па се намеће и логичан закључак да их библиотекари не могу самостално обрадити. Процењено је да неколико стотина хиљада текстова који су објављени у 19. и 20. веку не може бити ручно обрађено, нити ће библиотекари икад моћи да самостално сваком чланку доделе УДК број”.²¹ Својим истраживањем желели су да пруже одговоре на два питања (Могу ли се нови научни текстови, који су у потпуности библиографски обрађени и којима су људски стручњаци доделили УДК ознаке користити за изградњу аутоматског модела УДК класификације? Може ли се такав модел класификације, који је изграђен на претходно обрађеним научним текстовима користити за класификацију старих необрађених докумената?)

Б
И
Б
Л
И
О
Т
Е
К
А
Р

бр.
2,
год.
2025.

²¹ Matjaž Kragelj i Mirjana Kljajić Borštnar, „Automatic classification of older electronic texts into the Universal Decimal Classification-UDC”, *Journal of Documentation* 77 (3) (2021): 755–776, <https://www.emerald.com/insight/content/doi/10.1108/JD-06-2020-0092/full/html> (преузето 2. 4. 2025).

Приликом тренирања модела, користили су корпус од 70.000 научних текстова, које су претходно библиотекари у потпуности библиографски обрадили. Тај корпус је послужио за тестирање модела приликом доделе УДК бројева, необрађеном корпусу од 200.000 јединица. Потом су корпус необрађених текстова анализирали коришћењем „кластеризације методом К-средњих вредности” (*K-means clustering*), како би установили усклађеност између УДК група које се појављују и оних које су доделили професионални библиотекари. Кластеризација у основи представља алгоритам машинског учења, који се користи за груписање података на основу њихове сличности.²² У контексту овог рада, како наводе аутори, кластеризација није коришћена за саму финалну УДК класификацију, већ је послужила као прелиминарна анализа, где се желело установити да ли се УДК класификација „природно” поклапа са груписањем које алгоритам предвиђа на основу садржаја текста. Корпус је затим коришћен за изградњу модела класификације, а тако изграђени модел коришћен је за класификацију старијих текстова. Резултати истраживања су показали да је модел машинског учења успео тачно да класификује 80% научних текстова, чиме је потврђена прва хипотеза. Када је модел примењен на старе текстове, библиотекари су проценили 150 насумично одабраних узорака и утврдили да је преко 90% аутоматски додељених УДК бројева било исправно, чиме је потврђена и друга хипотеза.

Као недостатак истраживања наводе да је студија била ограничена недоступношћу старијих дигиталних текстова у смислу заступљености свих УДК група подједнако, већ су неке стручне области доминирале у односу на друге. То се иначе сматра и једним од учесталих изазова када је аутоматска класификација текста у питању. Један од изазова у класификацији текста јесте проблем неуравнотежених података: у корпусу за обуку често постоји велики број докумената који припадају једној категорији, док су друге категорије слабо заступљене. Ово може довести до класификатора који „игноришу” ређе категорије.²³

Као други недостатак истраживања наводе ограничене ресурсе професионалног кадра задуженог за каталогизацију и класификацију докумената, као и то што се модел за УДК класификацију може користити само за словачку и словеначку литературу.

²² K means Clustering – Introduction, <https://www.geeksforgeeks.org/machine-learning/k-means-clustering-introduction/> (преузето 25. 8. 2025).

²³ Emmanouil Ikonomakis, Sotiris Kotsiantis and V. Tampakas, “Text Classification Using Machine Learning Techniques”, *WSEAS Transactions on Computers* 4 (2005): 966–974, https://www.researchgate.net/publication/228084521_Text_Classification_Using_Machine_Learning_Techniques (преузето 24. 7. 2025).

Ипак, препоручена је његова употреба у пракси – по процени библиотекара модел је користан и може да служи као помоћни алат и подршка каталогизаторима у њиховом свакодневном раду, јер значајно смањује време потребно за ручну класификацију. Аутори закључују да ће највероватније управо вештачка интелигенција, односно машинско учење, бити од највеће користи у будућој аутоматској класификацији докумената.

У раду „Automated Subject Identification using the Universal Decimal Classification: The ANN Approach”, објављеном у априлу 2023. године у часопису *Journal of Information and Knowledge*, колеге са Катедре за библиотекарство и информатику Универзитета Северног Бенгала у Бангладешу презентовале су развој аутоматског система за препоруку УДК бројева у полуаутоматској класификацији, који је такође заснован на техници машинског учења. Студија се бавила употребом УДК шеме као „инпута” у систем вештачке интелигенције, а затим је посматрана тачност предвиђања додељених УДК бројева за класификацију текстова. За ово истраживање креиран је корпус од 151 чланка из области библиотекарства и информатике, објављених у часопису *Annals of Library and Information Studies*, у периоду од 2018. до 2022. године. Истраживање је, дакле, имало за циљ да развије систем додељивања комплексних УДК бројева. Разлог због којег су се одлучили да за систем класификације употребе УДК, а не неки други, је тај што сматрају да је „једна од кључних предности Универзалне децималне класификације њена способност да олакша међудисциплинарност претраге”, јер како наводе „хијерархијска структура УДК дозвољава лаку навигацију између стручних области, чинећи тако лакшим проналажење материјала на одређену тему”.²⁴ Осим тога, истичу да је још један од разлога тај што је „УДК вишејезичан, са преводима доступним на преко 30 језика, што га чини заиста глобалним системом класификације”.²⁵ За своје истраживање користили су софтвере BERT (*Bidirectional Encoder Representations from Transformers*) и KNIME (*Konstanz Information Miner*). BERT је софтвер заснован на NLP-у и методама дубоког машинског учења, који је развила компанија Гугл у 2018. Како кажу „оно што издваја BERT од осталих модела NLP јесте способност разумевања контекста речи у реченици. То је двосмерни модел, што значи да може да обрађује реченицу у оба смера, омогућујући тако правилно тумачење контекста. Ова карак-

Б
И
Б
Л
И
О
Т
Е
К
А
Р

²⁴ Aditi Roy and Ghosh Saptarshi, “Automated Subject Identification Using the Universal Decimal Classification: The ANN Approach”, *Journal of Information and Knowledge* 60 (2): 69–76 (2023), <https://doi.org/10.17821/srels/2023/v60i2/170963> (преузето 2. 4. 2025).

²⁵ Aditi Roy et al. “Automated Subject Identification Using the Universal Decimal Classification: The ANN Approach...”

бр.
2.
год.
2025.

теристика га чини посебно корисним за разумевање нијанси језика, као што су сарказам, идиоми и хомоними²⁶. У истраживању, BERT су користили за аутоматску анализу садржаја текста и предикцију класификационих ознака. KNIME је бесплатан софтвер отвореног кода који се користи за анализу и обраду података. Резултати њиховог истраживања показали су да је њихов модел за тестирање конструисан са високом прецизношћу. Међутим, систем предвиђања и додељивања сложених УДК бројева показао је ниску стопу тачности. Као разлог томе наводе да дигитална верзија УДК у RDF форматима (JSON/Turtle) није била доступна, док је UDC MRF из 2011. године био доступан у облику LOD датотеке, па су њу и користили. Закључују да ће, уколико потпуна лиценцирана верзија буде доступна, вероватноћа успешног развоја аутоматске класификационе шеме за практичну употребу бити веома велика.

У раду „A Hybrid Approach to Recommending Universal Decimal Classification Codes for Cataloguing in Slovenian Digital Libraries”, објављеном у јулу 2022. године у часопису *IEEE Access*, аутори Младен Борович, Милан Ојстершек и Дамјан Странд са Факултета електротехнике и рачунарства Универзитета у Марибору, презентују хибридни систем за препоруку УДК бројева у каталогизацији неklasификованих докумената, на корпусу од 114.485 докумената, с циљем да обухвате читаву структуру стручних области заступљених унутар хијерархијског система УДК таблица. Том приликом користили су функцију рангирања „BM25” са класификацијом на основу више приписаних ознака (етикета), које међусобно нису искључиве (*Multi-label classification*). Алгоритам BM25 (*Best Matching 25*), заснован на BERT-у, има функција рангирања, коју користе претраживачи за процену релевантности докумената за дати упит.²⁷ У овом истраживању аутори су функцију BM25 користили за грубу селекцију најсличнијих докумената и њихових ознака према УДК, а затим је BERT дубље анализирао изабране документе, како би се донела прецизнија одлука о додели најпогоднијег УДК броја. Вишеетикетна класификација у машинском учењу представља варијанту проблема класификације, где се свакој инстанци може доделити више ознака, односно „представља процес додељивања скупа унапред дефинисаних ознака непознатом објекту на основу посматрања његових карактеристика. Ради се о приступу надгледаног учења. Класификација је најраспрострањенији при-

²⁶ Aditi Roy et al. “Automated Subject Identification Using the Universal Decimal Classification: The ANN Approach...”

²⁷ LangChain, „BM25”, <https://python.langchain.com/docs/integrations/retrievers/bm25/> (преузето 4. 4. 2025).

ступ анализи података у оквиру надгледаног учења, а машинско учење се у ове сврхе широко примењује већ деценијама”.²⁸ Оваква врста класификације се „користи у ситуацијама када постоје две или више класа, а подаци које желимо да класификујемо могу истовремено припадати ниједној или свим класама”.²⁹ Главни циљ је да модел научи сложене односе између улазних карактеристика и свих релевантних излазних ознака, како би могао да за сваку улазну инстанцу предвиди више њених излазних ознака. За примену у библиотекарству може имати значајан потенцијал, будући да је сусрет са документима који покривају више тема или категорија прилично учестала пракса.

У раду је даље детаљно описано коришћење структуре УДК, текстуалног корпуса докумената, као и само функционисање хибридног система за препоруку. Аутори наводе да је мало покушаја учињено на пољу аутоматске класификације докумената у којима би се користиле неке од познатих класификационих шема, истичући да се велики обим грађе у библиотекама широм света још увек класификује ручно, било због неповерења у аутоматску класификацију, било због њене неадекватне примене. „Иако су учињени неки истраживачки напори ка аутоматској УДК класификацији, већина њих је била ограничена на традиционалне методе машинског учења. Слични покушаји били су направљени и када је у питању аутоматска ДДК класификација. Ово је важно, зато што и ДДК и УДК имају сличну структуру, јер оба система користе децималну класификацију”.³⁰ Својим истраживањем су желели да реше проблем неповерења или слабог поверења у аутоматску класификацију према УДК, као и да представе метод за препоруку УДК бројева који може да функционише као подршка у одлучивању за каталогизаторе.

²⁸ Vaishali S. Tidake and Shirish S. Sane, “Multi-label Classification: A Survey”, *International Journal of Engineering & Technology* 7 no. 4.19 (n.d.): 1045, doi:10.14419/ijet.v7i4.19.28284 (преузето 24. 7. 2025).

²⁹ Geeks for Geeks, “An Introduction to MultiLabel Classification”, <https://www.geeksforgeeks.org/machine-learning/an-introduction-to-multilabel-classification/> (преузето 20. 8. 2025).

³⁰ Mladen Borovic, Milan Ojstersek and Damjan Strnad, “A Hybrid Approach to Recommending Universal Decimal Classification Codes for Cataloguing in Slovenian Digital Libraries”, *IEEE Access* 10 (n.d.): 85595–85605, https://www.academia.edu/11230005A_Hybrid_Approach_to_Recommending_Universal_Decimal_Classification_Codes_for_Cataloguing_in_Slovenian_Digital_Libraries, doi:10.1109/ACCESS.2022.3198706 (преузето 3. 4. 2025).

Табела 1: Приказ анализираних радова

1.	Automatic classification of older electronic texts into the UDC (Kragelj and Kljajić Borštnar, 2021)
Циљ:	Аутоматизовати класификацију дигитализованих старих текстова према УДК, како би се библиотека-рима уштедело време
Корпус:	70.000 обрађених новијих научних радова (за тре-нинг) и 200.000 старих некласификованих текстова на словеначком језику.
Метод:	Машинско учење (надгледани и ненадгледани метод; класификација и кластеровање)
Резултат:	Модел може тачно да препоручи најмање један УДК број у >80% случајева. Потврђено од библиотекара.
Ограничење:	Недостатак означених старих текстова; ограничен број библиотекара за валидацију.
Значај:	Омогућава полуаутоматску класификацију и побољшава претраживање у дигиталним библиотекама. Помаже библиотека-рима приликом класификације старих текстова.

2.	Automated Subject Identification using the Universal Decimal Classification: The ANN Approach (Roy and Ghosh, 2023)
Циљ:	Развој полуаутоматског система за класификацију до-кумената према УДК помоћу модела BERT.
Корпус:	151 чланак из <i>Annals of Library and Information Studies</i> (2018–2022).
Метод:	Ручно састављени УДК бројеви, фино подешавање мо-дела BERT у окружењу KNIME.
Резултат:	Модел показује високу тачност али мању прецизност када је реч о сложеним класама.
Ограничење:	Мали корпус (151 чланак), проблем недоступности пуне УДК LOD базе
Значај:	Омогућава полуаутоматску класификацију у академ-ским библиотекама. Отвара пут ка изградњи аутомат-ске класификације у будућности.

3.	A Hybrid Approach to Recommending Universal Decimal Classification Codes for Cataloguing in Slovenian Digital Libraries (Borovič, Ojsteršek and Strnad, 2022)
Циљ:	Изградити хибридни систем препоруке УДК бројева као подршку библиотекарима у каталогизацији.
Корпус:	114.485 докумената из словеначке базе отвореног приступа (тезе и научни радови).
Метод:	Хибридни систем: рангирање помоћу алгоритма BM25, вишезначни BERT класификатор, филтрирање по најчешћим УДК бројевима.
Резултат:	Хибридни модел је показао побољшане перформансе у односу на појединачне методе
Ограничења:	Проблем недоступности пуне УДК LOD базе
Значај:	Омогућава да више УДК класа буде предложено за један документ, подршка библиотекарима током каталогизације

Дискусија

Резултати прегледа литературе показују да је примењена област машинског учења и обраде природног језика важан корак у модернизацији и унапређењу традиционалних библиотечких пракси. Истраживања описују конкретне технике машинског учења које се користе у процесу аутоматске класификације према систему УДК, при чему се демонстрирају различити приступи, али и проблеми који се морају узети у обзир. Један од приступа у истраживању, који су применили Крагељ и Кљајић Борштар (2021), користи функцију вишезначне (вишеетикетне) класификације у оквиру надгледаног учења. Резултати су показали да ова метода може да омогући релативно високу тачност при додели УДК бројева, међутим, у великој мери је зависна од квалитета и саме репрезентативности скупова података за обуку. То указује на чињеницу да резултати могу бити пристрасни, уколико су скупови улазних података ограничени или непотпуни. С друге стране, техника кластерована као врста ненадгледаног учења, донела је уштеде у погледу времена, јер није захтевала претходно означавање улазних података. Ипак, ризик од губитка значајних семантичких веза између докумената јесте постојао. Методе машинског учења, а посебно модели дубоког учења као што је BERT, показали су велики потенцијал за аутоматску класификацију докумената у оквиру система УДК, што су потврдиле обе студије које су користиле BERT у свом истраживању.

Као предности аутоматизације истичу се убрзано обрађивање грађе, као и доследност у класификацији докумената. Међутим, појавили су се проблеми због ограниченог приступа корпусу и лиценцираним УДК подацима. Посебно је наглашена битност квалитета улазних података, јер успешност система у великој мери зависи од тога. На пример, у студији случаја (*Automated Subject Identification using the Universal Decimal Classification: The ANN Approach*), значајна слабост био је корпус од свега 151 чланка, јер он није био репрезентативан. Додатни проблеми јавили су се код докумената којима је додељено више од једног УДК броја. То су документи који обрађују више тема и самим тим су сложеније природе. Студија случаја која је применила хибридни приступ *gala* је прецизније резултате у том пољу, што се постигло комбиновањем различитих техника. Највећи изазов у УДК класификацији је сама сложеност система. УДК садржи хиљаде класа, а захтева високу прецизност и конзистентност, што умногоме отежава аутоматско класификовање. Истраживања су показала да аутоматски системи могу врло често да генеришу непрецизне ознаке када је потребно повећати детаљност класификације, па је стручни надзор библиотекара и даље неопходан.

Ово указује на чињеницу да NLP технике тренутно имају највећу вредност као подршка, а не као потпуна замена за традиционални, стручни процес класификације. Сва три рада су препознала потребу за системима који ће помагати библиотекаrimа, али их неће потпуно заменити. Сама интеграција машинског учења у библиотечку праксу јесте нешто што је већ започето и као да не оставља простор за питања: Да ли ће се и када ће се то догодити? Кључно питање које би, ипак, требало поставити тиче се квалитета оствареног. Овакве студије могу послужити као корисна основа и оријентир за планирање будућих активности и пилот-пројеката у домаћем контексту, нарочито у сарадњи са Народном библиотеком Србије, Катедром за библиотекарство и информатику, као и одговарајућим техничким факултетима. Као ослонац могу послужити већ постојећи NLP модели за српски језик, као што су BERT³¹, SRBERT³² и SRBERTA,³³ који би се могли прилагодити корпусима у

³¹ Nikola Ljubešić and Davor Lauc, "BERTiC – The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian", *CoRR* abs/2104.09243 (2021), <https://arxiv.org/abs/2104.09243> (преузето 28. 8. 2025).

³² S. Aum and S. Choe, "srBERT: automatic article classification model for systematic review using BERT", *Syst Rev* 10, 285 (2021), <https://doi.org/10.1186/s13643-021-01763-w> (преузето 28. 8. 2025).

³³ Miloš Bogdanović, Jelena Kocić and Leoind Stoimenov, "SRBERTa-A Transformer Language Model for Serbian Cyrillic Legal Texts", *Information* 15 (2024), doi:10.3390/info15020074 (преузето 28. 8. 2025).

систему COBISS. За практичну примену свакако је неопходно изградити локалне језичке корпусе, развити нормативне датотеке и обезбедити едукацију стручног кадра за рад са технологијама вештачке интелигенције.

Закључак

Аутоматска класификација је изазовно истраживачко поље већ неколико деценија. Интересовање за ову област је нагло порасло с напретком и еволуцијом веба, али ће тек с даљим развојем вештачке интелигенције достићи свој врхунац. Многи библиотекари се надају да ће компјутери и слични технолошки алати и технике у будућности моћи самостално да идентификују и класификују документе, односно да сами обављају процес који је до сада захтевао искључиво људску интелигенцију и ангажовање, јер је у његовој основи ментална процедура. Стручна заједница је сагласна у томе да су несумњиво потребна даља истраживања у овој области. Сматра се да би развој ове иновативне технологије могао умногоме да помогне библиотекарима у њиховом свакодневном раду у области библиотечке класификације, уколико би се, наравно, осигурала њена исправна и одговорна примена. Стога је улагање у даљи развој непотпуних националних нормативних датотека и језичких ресурса, као и у обуку стручњака за праћење и коришћење технологија вештачке интелигенције, неопходно за развој заиста поуздане аутоматске класификације као подршке библиотекарима у раду на доступности информација о публикацијама.

Б
И
Б
Л
И
О
Т
Е
К
А
Р

бр.
2.
год.
2025.

Literatura:

1. Allam, Hesham, Lisa Makubvure, Benjamin Gyamfi, Kwadwo Nyarko Graham, and Kehinde Akinwolere. “Text Classification: How Machine Learning Is Revolutionizing Text Categorization”. *Information* 16 no. 2 (2025), <https://doi.org/10.3390/info16020130> (преузето 15. 7. 2025).
2. Andonovski, Jelena. „Mreža otvorenih podataka i jezički resursi u procesu izgradnje srpsko-nemačkog literarnog korpusa: doktorska disertacija”. Beograd: J. Andonovski, 2019. <https://phaidrabg.bg.ac.rs/o:22874> (преузето 3. 4. 2025). (na ćirilici)
3. Aum, S. and S. Choe. “srBERT: Automatic Article Classification Model for Systematic Review Using BERT”. *Syst Rev* 10 (2021): 285. <https://doi.org/10.1186/s13643-021-01763-w>. (преузето 28. 8. 2025).
4. Berners-Lee, Tim, James Hendler and Ora Lassila. “The Semantic Web”. *Scientific American* 284 no. 5 (2001): 34–43. <https://doi.org/10.1038/scientificamerican0501-34>. (преузето 4. 4. 2025).
5. Beckett, David. RDF 1.2 N-Triples. <https://www.w3.org/TR/rdf12-n-triples/> (преузето 24. 8. 2025).
6. Bogdanović, Miloš, Jelena Kocić and Leonid Stoimenov. “SRBerta–A Transformer Language Model for Serbian Cyrillic Legal Texts”. *Information* 15 (2024): DOI:10.3390/info15020074. (преузето 28. 8. 2025).
7. Borovic, Mladen, Ojstersek, Milan and Strnad, Damjan. “A Hybrid Approach to Recommending Universal Decimal Classification Codes for Cataloguing in Slovenian Digital Libraries”. *IEEE Access* 10 (n.d.): 85595–605. https://www.academia.edu/111230005/A_Hybrid_Approach_to_Recommending_Universal_Decimal_Classification_Codes_for_Cataloguing_in_Slovenian_Digital_Libraries, doi:10.1109/ACCESS.2022.3198706 (преузето 3. 4. 2025).
8. GeeksforGeeks. “An Introduction to MultiLabel Classification”, 2025. <https://www.geeksforgeeks.org/machine-learning/an-introduction-to-multilabel-classification/> (преузето 20. 8. 2025).
9. Heaton, Jeff. “Review of *Deep Learning*, by Ian Goodfellow, Yoshua Bengio and Aaron Courville”. *Genetic Programming and Evolvable Machines* 19 (2018): 305–307. <https://doi.org/10.1007/s10710-017-9314-z> (преузето 24. 8. 2025).
10. Ikonomakis, Emmanouil, Sotiris Kotsiantis and V. Tampakas. “Text Classification Using Machine Learning Techniques”. *WSEAS Transactions on Computers* 4 (2005): 966–974, <https://www>.

researchgate.net/publication/228084521_Text_Classification_Using_Machine_Learning_Techniques (преузето 15. 7. 2025).

11. Joorabchi, Arash and Abdulhussain E. Mahdi. "An Unsupervised Approach to Automatic Classification of Scientific Literature Utilizing Bibliographic Metadata". *Journal of Information Science* 37 no. 5 (2011): 499–514. <https://doi.org/10.1177/0165551511417785> (преузето 4. 4. 2025).
12. Kragelj, Matjaž and Mirjana Kljajić Borštnar. "Automatic Classification of Older Electronic Texts into the Universal Decimal Classification-UDC". *Journal of Documentation* 77 no. 3 (2021): 755–76. <https://doi.org/10.1108/JD-06-2020-0092/full/html> (преузето 2. 4. 2025).
13. K. Means Clustering – Introduction, <https://www.geeksforgeeks.org/machine-learning/k-means-clustering-introduction> (преузето 25. 8. 2025).
14. LangChain. "BM25". <https://python.langchain.com/docs/integrations/retrievers/bm25/> (преузето 4. 4. 2025).
15. Ljubesic, Nikola and Davor Lauc. "BERTiC - The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian". *CoRR* abs/2104.09243 (2021). <https://arxiv.org/abs/2104.09243>. (преузето 28. 8. 2025).
16. Nađ, Žolt. *Osnove veštačke inteligencije i mašinskog učenja*. Beograd: Kompjuter biblioteka, 2019. (na ćirilici)
17. Oxford University Press. "Natural-Language Processing". In *Oxford Reference*, 2024. <https://www.oxfordreference.com/display/10.1093/oi/authority.20110803100225333> (преузето 23. 5. 2025).
18. Roy, Aditi and Saptarshi, Ghosh. "Automated Subject Identification Using the Universal Decimal Classification: The ANN Approach". *Journal of Information and Knowledge* 60 (2): 69–76. (2023). <https://doi.org/10.17821/srels/2023/v60i2/170963>. (преузето 2. 4. 2025).
19. Slavic, Aida, Ronald Siebes and Andrea Scharnhorst. "Publishing a Knowledge Organization System as Linked Data: The Case of the Universal Decimal Classification". *ArXiv* 2205 no. 01395 (2022). <https://doi.org/10.5771/9783956506611-69> (преузето 27. 3. 2025).
20. Tidake, Vaishali S. and Shirish S. Sane. „Multi-label Classification: A Survey". *International Journal of Engi-neering & Technology* 7 no. 4.19 (2018). <https://doi.org/10.14419/ijet.v7i4.19.28284>. (преузето 24. 7. 2025).
21. Trtovac Aleksandra i Dakić Nataša, „Baza CONOR.SR u sistemu COBISS.SR". *Infototeca: Journal for Digital Humanities* v. 20, n.

Б
И
Б
Л
И
О
Т
Е
К
А
Р

бр.
2.
год.
2025.

- 1–2a (feb. 2021): 75–88. https://infoteka.bg.ac.rs/ojs/index.php/Infoteka/article/view/2020.20.1_2.5_sr (преузето 4. 4. 2025).
22. Univerzitet Union, Računarski fakultet. „Šta je mašinsko učenje i šta su inteligentni algoritmi?” <https://raf.edu.rs/citaliste/najnoviji-it-dogadjaji/sta-je-masinsko-ucenje-i-sta-su-inteligentni-algoritmi/> (преузето 3. 4. 2025).
23. Xiao, Tong and Jingbo Zhu. “Foundations of Large Language Models”. NLP Lab, Northeastern University & NiuTrans Research, 2025. <https://github.com/NiuTrans/NLPBook/tree/main> (преузето 20. 6. 2025).
24. World Wide Web Consortium (W3C). <https://www.w3.org> (преузето 4. 4. 2025).

Olivera Stojanović
Belgrade City Library
oliverakrstic@ymail.com

APPLICATION OF ARTIFICIAL INTELLIGENCE IN AUTOMATIC UDC CLASSIFICATION: A REVIEW OF PUBLISHED RESEARCH

SUMMARY: The paper provides a review of the literature and research published between 2020 and April 2025, focusing on the application of machine learning and natural language processing techniques in the field of automatic bibliographic classification, with a special emphasis on the Universal Decimal Classification (UDC) system. The aim of the paper is to provide insight into current trends addressing this topic in librarianship, as well as to offer a brief introduction to key concepts such as automatic classification, the Semantic Web, natural language processing, and machine learning. In conclusion, the paper highlights the need for developing local resources and educating professional staff to ensure the possible and sustainable application of similar AI technologies in domestic practice.

KEYWORDS: automatic classification, UDC, Semantic Web, artificial intelligence, machine learning, natural language processing

Примљено: 14. априла 2025.
Прихваћено: 24. септембра 2025.