Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.42:004.738.12
DOI: https://doi.org/10.24867/20BE05Matkovic

АУТОМАТСКА ЕКСТРАКЦИЈА ИНФОРМАЦИЈА ИЗ ОДЛУКА НОВОЗЕЛАНДСКИХ СУДОВА

AUTOMATIC EXTRACTION OF INFORMATION FROM NEW ZEALAND COURT DECISIONS

Јелена Матковић, *Факултет техничких наука, Нови Сад*

Област – ЕЛЕКТРОТЕХНИКА И РАЧУНАРСТВО

Кратак садржај – *У раду је извршена екстракција метаподатака новозеландских судских одлука коришћењем техника обраде природног језика. Предложене су методе машинског и дубоког учења у циљу проналажења прописа релевантних за доношење одлука и извршено поређење добијених резултата. Помоћу екстрахованих референци, пронађени су прецеденти и документи који су последица поступака по правном леку на основу којих су оформљене групе повезаних докумената.*

Кључне речи: *судске одлуке, екстракција података, обрада природног језика, неуронске мреже*

Abstract – *This paper proposes using Natural Language Processing technics for metadata extraction from New Zealand's court decisions. Machine Learning and Deep Learning methods are proposed for obtaining legislations relevant for making these decisions and the quality of obtained results is evaluated. Using extracted references, precedents and subsequent decisions are retrieved and groups of linked legal documents are formed.*

Keywords: *court decisions, data extraction, NLP, neural networks*

1. УВОД

Правосуђе представља сложени систем са судовима организованим на више нивоа надлежности. Количина судских одлука, које представљају исходе судских поступака, свакодневно се увећава.

Анализа донетих одлука је од великог значаја за правну струку, како у правним системима заснованим на англосаксонском праву тако и у правним системима заснованим на континенталном праву. Праћење изречених пресуда од стране судија је предуслов за уједначавање судске праксе, што истовремено захтева упознавање стручне јавности са садржином бројних судских одлука.

Екстракцијом података из правних аката олакшава се проналажење конкретних докумената на основу задатих критеријума. За судске одлуке су од значаја подаци о самом предмету, као што су подаци о странкама, судијама и идентификационим бројевима судског предмета, али и подаци о одлуци као што су датум доношења, референце на примењене прописе и прецеденте и референце на првостепене пресуде у случају поступака по жалби. Постојање ових података у машински читљивом облику би омогућило повезивање судских одлука са другим судским одлукама и прописима релевантним за њихово доношење. Таквом мрежом докумената би се могло унапредити претраживање база судске праксе и проналажење сродних правних аката.

Да би се уз помоћ рачунара подржала анализа правних аката важну улогу имају формат и структура ових докумената. У судским одлукама се користи природан језик, уз поштовање прописа који се односе на ову врсту аката. Осим разлика у погледу околности из судских предмета за које су донете, судске одлуке се могу разликовати и по стилу писања. Тако се у садржини ових аката могу јавити неконзистентности које отежавају њихову аутоматску обраду.

Предмет истраживања овог мастер рада је подршка проналажењу међусобно повезаних законодавних и судских аката који се односе на сродно правно питање. На овај начин би се допринело лакшем прегледању судске праксе како у смислу идентификације примењених прописа и прецедената, тако и праћењу исхода поступака по правним лековима. За решавање овог проблема изабрана је метода обраде природног језика како би се омогућила екстракција података из судских одлука и како би се на основу тих података установиле везе између правних аката. Фокус истраживања је на обради одлука новозеландских судова.

У наредном одељку дат је кратак преглед истраживања која се баве сличним проблемом. Трећи одељак објашњава метод примењен у овом истраживању. Четврти одељак садржи кратак преглед и анализу остварених резултата. Последње поглавље доноси закључак и разматра предности и мане имплементираног решења, уз навођење смерница за његово побољшање.

2. СРОДНА ИСТРАЖИВАЊА

У овом одељку је дат кратак преглед истраживања која се баве проблемом екстракције података из правних аката, односно техникама за обраду природног језика у правним документима.

У [1] аутори су представили систем под називом *History Assistant* који екстрахује информације из суд-

---

НАПОМЕНА:
Овај рад проистекао је из мастер рада чији ментор је био др Марко Марковић, доцент.

ских пресуда, па на основу ових информација проналази сродне, раније донете пресуде. Систем комбинује екстракцију информација, машинско учење и проналажење информација. Обрада природног језика је постигнута са четири нивоа обраде: (1) обрадом стрингова се врши растављање текста на реченице, (2) синтаксном обрадом се групишу делови реченице према њиховој функцији у реченици, (3) семантичком обрадом се екстрахују значења делова реченица и (4) прагматичком обрадом се допуњују непотпуне информације и разрешавају двосмислености. Евалуација система је извршена над 676 случајева и постигнута је прецизност (енг. *precision*) од 54% и одзив (енг. *recall*) od 78%. Системом је подржана једино обрада пресуда америчког правосуђа.

У [2] аутори примећују да се правни акти најчешће објављују у форматима погодним за приказивање и штампање докумената. Ради ефикаснијег коришћења ових података предложен је систем за екстракцију метаподатака који омогућава брже претраживање и филтрирање колекције правних аката. Систем омогућава приступ португалским одлукама у правним поступцима. Из изворних докумената, који су у PDF формату, преузимају се блокови текста и повезују према њиховој позицији на страници. Парсирање текста се врши помоћу скупа правила прилагођених прописаној структури ових правних аката. Правила су базирана на логичким и позиционим односима између делова текста унутар документа. Аутори истичу да би услед недоследности у структури правних аката предложен систем требало унапредити механизмима машинског учења.

Аутори у [3] предлажу метод за детектовање и разрешавање референци у правним актима ЕУ и државама чланицама. Развијен је систем који на бази овог модела утврђује цитиране прописе у холандским правним актима. Систем је предвиђен за потпуно аутоматизовану обраду скупова докумената и не укључује интеракцију са корисником. Тестирање је извршено над пресудама из холандских судова различитих надлежности (по две пресуде из сваког суда) и добијена је прецизност од 98% и одзив од 94%.

Алат *Eyecite* [4] омогућава аутоматско проналажење цитата прецедената у судским одлукама. Прилагођен је препознавању цитираних одлука у америчким пресудама. Алат је написан у програмском језику Python и базиран је на колекцији регуларних израза. Поред алата *Eyecite*, аутори нуде и алате за претпроцесирање којима се отклањају евентуалне грешке у улазним документима. Ипак, овај алат омогућава једино обраду правосудних докумената у САД и није тестиран на пресудама других судова.

У [5] аутори су представили метод којим се из правних аката врши екстракција информација о реферанцираним нормативним актима. Метод је прилагођен референцама у италијанском правосуђу. Подржана је екстракција референци које се односе на делове текста унутар истог правног акта и референци које се односе на делове других правних аката. Метод је евалуиран над 1641 документа и постигао тачност од 93,6%.

3. МЕТОДОЛОГИЈА

У овом одељку је објашњен метод за екстракцију информација из одлука новозеландских судова и детекцију веза између ових аката са законима и другим судским одлукама.

3.1. Преузимање и припрема података

Текстови судских одлука су преузети са портала *NZLII* [6] на којем највећи број објављених докумената потиче из следећих правосудних органа: Апелациони суд (*Court of Appeal*), Високи суд (*High Court*) и Окружни суд (*District Court*). С обзиром на то да се сви документи који потичу из три претходно наведена суда могу преузети у PDF формату, они су преузети са сајта у том формату, а затим даље обрађивани учитавањем и парсирањем докумената. Укупан број преузетих докумената износи: 27293 за Високи суд, 3592 за Окружни суд и 6084 за Апелациони суд.

Документи судских одлука на самом почетку садрже податке о судским предметима на које се односе. Анализом скупа података примећено је да су ти подаци у саставу прве две странице документа. Ови подаци обухватају: датум саслушања, датум доношења одлуке, имена учесника у процесу, идентификациони број судског предмета, идентификациону ознаку у виду неутралног цитата и име једног или више судија.

За издвајање текста из PDF докумената искоришћена је техника оптичког читања карактера (*Optical character recognition - OCR*) уз помоћ које добијени текст остаје доследан распореду одговарајућих текстуалних објеката унутар страница изворних докумената чиме је очувана структура докумената.

Закони су преузети са званичног новозеландског сајта [7] где се могу претражити по абецедном реду или по години објављивања. Текстови закона су преузети у HTML формату и у њима су садржане ознаке чланова, као и назнаке када су одређени делови закона мењани и на који начин, тј. да ли су измењени, додати, премештени или уклоњени. Укупан број преузетих закона износи 1808.

3.2. Екстракција података

За екстракцију метаподатака из докумената употребљени су регуларни изрази због препознатљивог формата ових података и њиховог карактеристичног распореда у документима. Судске одлуке могу садржати идентификационе ознаке судског предмета на који се односе и/или тзв. неутралне цитате. Поред тога ови документи могу да садрже референце ка другим правним актима. Ове референце се односе на прецеденте или законе, као и на ожалбене судске одлуке које се преиспитују у поступку по правном леку. Судске одлуке су цитиране употребом идентификационе ознаке судског предмета или употребом идентификатора у виду неутралног цитата. Како се обе врсте цитата користе и за цитирање прецедената и за цитирање ожалбених одлука отежано је одређивање врсте цитата само на основу његове структуре. Због тога су за анализу цитата упоређују имена странака у

документима. Уколико се пронађе апсолутно поклапање барем једног учесника и са једне и са друге стране, односно у одлуци у којој је пронађена референца и референцираној одлуци, сматра се да се ради о ожалбеној одлуци и одлуци по правном леку.

За екстракцију назива закона релевантних за доношење неке одлуке је искоришћена чињеница да новозеландски закони на крају свог назива најчешће садрже реч '*Act*', као и да се записују великим почетним словима.

Из тог разлога је приликом преузимања текстова закона са званичног новозеландског сајта, формирана листа назива свих закона, па су регуларним изразима издвајани делови реченица који садрже реч '*Act*' којој претходи барем једна реч великог почетног слова. На овај начин добијају се издвојени делови реченица након чега се врши упоређивање са листом назива свих закона и у случају подударања, препознат закон се сматра релевантним за доношење одлуке.

Ипак известан број докумената не садржи ниједно помињање неког прописа. Одређивање потенцијално применљивих закона на правна питања садржана у текстовима одлука реализовано је одређивањем сличности текстова докумената и класификацијом докумената. За одређивање сличности докумената је примењено косинусно растојање и сијамска неуронска мрежа [8]. За класификацију докумената употребљене су неуронска мрежа (*NN*) [9], конволутивна неуронска мрежа (*CNN*) [10] и метод подржавајућих вектора (*SVM*) [11].

3. РЕЗУЛТАТИ И ДИСКУСИЈА

У овом одељку су приказани и дискутовани резултати екстракције података из укупно 36969 преузетих судских одлука.

Табела 1 приказује за сваки од три анализирана новозеландска суда колики је проценат докумената из којих су успешно екстраховане вредности атрибута.

Табела 1 – Проценат докумената из којих су екстраховани атрибути

Атрибут	Високи суд	Окружни суд	Апелациони суд
Назив документа	100%	100%	100%
Неутрални цитат	100%	100%	100%
Идентиф. ознака	99,35%	37,24%	99,22%
Седиште суда	97,51%	97,30%	/
Судија	87,92%	96,49%	/
Веће судија	/	/	92,34%
Датум саслушања	92,09%	89,37%	80%
Датум одлуке	94,57%	97,02%	99,42%
Странка 1	93,81%	98,11%	99,41%
Странка 2	91,72%	99,05%	99,33%
Називи прописа	78,49%	64,09%	77,98%
Референце на одлуке	62,80%	78,55%	68,92%

Може се приметити да је за већину атрибута висок проценат успешно екстрахованих вредности. Треба напоменути да Високи суд и Окружни суд поседују седишта на више локација, за разлику од Апелационог суда који има само једно седиште, те се оно не наводи. Такође, рад у већима је карактеристичан за Апелациони суд, док у Високом суду и Окружном суду одлучују судије појединци.

Укупно је пронађено 11027 докумената који референцирају другу одлуку која има барем једног истог учесника што се може сматрати референцама на ожалбене пресуде. Након груписања докумената по именима учесника оформљено је 4982 групе докумената. Међутим, у извесном броју одлука се као учесник наводи '*The Queen*' и '*The Police*', као и неке друге државне институције, па је отежано утврђивање међусобног односа таквих одлука на основу имена учесника. Ово је посебно изражено у кривичном праву.

С обзиром на то да у неким одлукама није наведен назив ниједног закона, за препознавање прописа релевантних за доношење тих одлука су примењене методе одређивања сличности текста и методе класификације текста.

У табели 2 се могу видети резултати одређивања сличности текста правних аката методом косинусног растојања за три изабрана прописа.

Табела 2 – Резултати одређивања сличности текста косинусним растојањем

Пропис	*Accuracy*	*Precission*	*Recall*	*F1-score*
Land Transport Act 1998	0.95	0.63	0.7	0.67
Fair Trading Act 1986	0.91	0.6	0.58	0.59
Misuse of Drugs Act 1975	0.94	0.59	0.71	0.64

Сијамска неуронска мрежа тренирана је над укупно 40000 парова уводних делова судских одлука у којима су садржани описи чињеничног стања. Сет података је подељен на тренинг и тест податке у односу 5:1. Накнадно је формиран валидациони сет који се састоји од парова судских одлука из Окружног суда и њиме је извршена валидација. У табели 3 приказани су резултати над тест и валидационим сетом података.

Табела 3 – Резултати одређивања сличности текста сијамском неуронском мрежом

Сет података	*Accuracy*	*Precission*	*Recall*	*F1-score*
тест	0.78	0.78	0.7	0.74
валидациони	0.63	0.38	0.45	0.4

Над одабраним прописима су анализирани резултати добијени бинарним класификационим моделима неуронских мрежа и методе подржавајућих вектора. Прегледом уводних делова судских одлука уочено је да се у документима који референцирају поједине законе, текст у већини случајева састоји од веома сличних појмова који умногоме олакшавају класификацију. Насупрот томе, за поједине законе, текст је веома опширан и разнолике је садржине. Последица наведеног јесте да се у зависности од одабира прописа за који се конструише класификациони модел, тачност модела разликује због различитости текста самих докумената. Резултати одређивања релевантности прописа на доношење судске одлука путем класификације докумената приказани су у табели 4. на примеру прописа *'Land Transport Act 1998'*.

Табела 4 – Резултати бинарне класификације за '*Land Transport Act 1998*'

Модел	*Accuracy*	*Precission*	*Recall*	*F1-score*
NN	0.98	0.98	0.99	0.99
CNN	0.99	0.99	0.99	0.99
SVM	0.97	0.97	0.97	0.97

Анализом резултата може се приметити да су методе коришћене за прорачун сличности текстова генерално дале слабије резултате у односу на методе класификације текста.

3. ЗАКЉУЧАК

У раду је предложен начин преузимања и екстраховања података из судских одлука донетих од стране новозеландских судова. Обухваћене су одлуке са три нивоа судске надлежности, а подаци су екстраховани у циљу омогућавања напредне претраге базиране над свим добављеним атрибутима, повезивања одлука које претходе правноснажној пресуди, као и одређивања гране права којој документи припадају.

Употреба регуларних израза за проналажење сложенијих шаблона унутар текста неструктуираних докумената даје задовољавајуће резултате у овом истраживању. Кључну улогу, овакви изрази, имају у екстраховању идентификационих ознака судских предмета и детектовању цитираних прописа.

Разлике у стиловима писања правних аката, односно неконзистентности у организацији и садржини докумената, представљају један од изазова за екстракцију података. Разлике у структури докумената су уочљиве и између докумената насталих у различитим временским раздобљима, што доводи до потребе за креирањем већег броја шаблона за проналажење информација од значаја. Уколико би ови документи били доступни у стандардизованом формату прилагођеном репрезентацији правних аката, олакшао би се процес екстракције података. До усвајања машински читљивих формата докумената у правосуђу, приступ предложен у овом раду могао би послужити као помоћно решење којим би се већ донетим пресудама придруживали метаподаци.

Систем приказан у овом раду би се могао користити у склопу информационих система судске праксе за ефикасније проналажење прецедената и релевантних прописа. Иако је значајан број правних аката новозеландског правног система доступан на Интернету, и даље постоје документи који нису објављени. Ово битно отежава праћење исхода судских поступака по правним лековима. Због тога повезивање одлука о неком случају у хронолошком редоследу може бити онемогућено. Приступ који је коришћен за повезивање одлука на основу имена учесника показао се као недовољно поуздан због тога што се над екстрахованим референцама не може прецизно утврдити када нека одлука представља поступак по правном леку неког другог поступка.

Због различитости у дужини текстова одлука, као и због разлика у стиловима писања, поређење текстова у циљу одређивања примењених прописа показало се као непоуздано решење у односу на бинарну класификацију текста. Међутим, недостатак неопходног броја докумената за тренирање бинарних модела онемогућава текстуалну класификацију за већину екстрахованих прописа.

4. ЛИТЕРАТУРА

[1] P. Jackson, K. Al-Kofani, A. Tyrrell and A. Vachher, "Information extraction from case law and retrieval of prior cases.," Artificial Intelligence, vol. 150, no. 1-2, pp. 239-290, 2003.

[2] B. M. Oliveira, R. V. Guimarães and L. Antunes, Sifting Through Chaos: Extracting Information from Unstructured Legal Opinions. In MIE (pp. 441-445)., 2018, January.

[3] M. van Opijnen, N. Verwer and J. Meijer, "Beyond the experiment: the eXtendable legal link eXtractor," In Workshop on Automated Detection, Extraction and Analysis of Semantic Information in Legal Texts, held in conjunction with the 2015 International Conference on AI and Law (ICAIL)., 2015, June.

[4] J. Cushman, M. Dahl and M. Lissner, "Eyecite: A tool for parsing legal citations," Journal of Open Source Software, 2021.

[5] M. Palmirani, R. Brighi and M. Massini, "Automated extraction of normative references in legal texts," In Proceedings of the 9th international conference on AI and law, pp. 105-106, 2003, June.

[6] New Zealand Legal Information Institute, http://www.nzlii.org/. (приступљено у августу 2022.)

[7] New Zealand Legislation, https://www.legislation.govt.nz/ (приступљено у августу 2022.)

[8] D. Chicco, "Siamese neural networks: an overview," Artificial Neural Networks, Methods in Molecular Biology, vol. 2190, no. 3, pp. 73-94, 2020.

[9] D. Kriesel, "A Brief Introduction to Neural Networks," 2005.

[10] Convolutional Neural Networks for Text, https://lena-voita.github.io/nlp_course/models/convolutional.html (приступљено у августу 2022.)

[11] „A Complete Guide to Support Vector Machines“, https://medium.com/@kushaldps1996/a-complete-guide-to-support-vector-machines-svms-501e71aec19e (приступљено у августу 2022.)

Кратка биографија:

Јелена Матковић рођена је 1996. године у Новом Саду. Основне академске студије је завршила 2020. године на Факултету техничких наука на студијском програму Рачунарство и аутоматика. Мастер рад одбранила је 2022. године на истом факултету, студијски програм Рачунарство и аутоматика - Интелигентни системи.