

# Аутоматско препознавање аргумената у правним пресудама

Маријета Ракош  
Департман за рачунарство и информатику  
Факултет техничких наука, Универзитет у Новом Саду  
Нови Сад, Србија  
E-mail: [marieta.rakos@gmail.com](mailto:marieta.rakos@gmail.com).

Марија Ћурчић  
Департман за рачунарство и информатику  
Факултет техничких наука, Универзитет у Новом Саду  
Нови Сад, Србија  
E-mail: [marija.curcic07@gmail.com](mailto:marija.curcic07@gmail.com)

**Апстракт** — У овом раду представљена су решења проблема на тему аутоматског препознавања аргумената у правним пресудама писаних на црногорском језику. Пронађени скуп података је преузет са [2] и није лабелиран. Решење се састоји из два главна дела, односно две фазе: детекција реченица које садрже аргумент, и реченица које га не садрже, и другог дела који се бави самим проналажењем делова аргумената у детектованим реченицама. Додатно је било потребно препроцесирати скуп података тако да буде машински читљив и лабелирање реченица на оне које садрже и не садрже аргумент. Како анотирање, односно лабелирање, изискује много времена и присуство стручног лица за одобравање и надгледање анотирања, извршена је делимична анотација, односно анотирано је свега неколико докумената, односно њихових реченица. Над тиме је употребљена техника полунадгледаног учења, коришћењем *BiLSTM* модела, при чему је достигнута тачност од 98%. За потребе детектовања делова аргумената, односно премиса и закључака, детектоване аргументативне реченице су издвојене на делове на основу реза. Извршена је кластеризација добијених делова реченица на нивоу документа. На крају, испитани су модели за наведена два дела, поређени су њихови резултати и продискутоване су грешке које они праве. Модел који класификује реченице евалуиран је помоћу асисасу метрике. Укратко су размотрени радови који се баве сличном темом. Представљени су резултати и наведена усмерења за будући рад и истраживања.

**Keywords**—*machine learning; NLP; argument mining; legal documents; semi-supervised learning;*

**Кључне речи**—*машинско учење; правне пресуде; полунадгледано учење;*

## I. УВОД

Детекција аргумената представља истраживачку област у сфери обраде природног језика (енг. *Natural Language Processing*, скраћено *NLP*). Сврха ове области јесте олакшати разумевање пресуда, убрзати претраге и увести структуру у саме документе.

Издавање аргумената и њихових делова - премиса и закључака, може се извести на више начина. Један од њих јесте употребом XML шема, где су специфицирани одређени тагови који означавају поменуте делове и њихове односе. Оне су развијене са циљем увођења формализма за записивање аргумената у аргументационим системима,

или, пак, како би се омогућила размена документата између различитих софтверских архитектура. Како овај приступ захтева знање експерта и велику количину утрошеног времена, настаје идеја да се за овај проблем употреби машинско учење (енг. *Machine Learning*).

Препознавање аргумената у правним документима захтева тимове стручњака, који би за сваки документ аотирани реченице које садрже аргументе и делове документа који представљају премисе и закључке. Постоји прегршт радова који се баве детекцијом аргумената у текстовима и правним документима на енглеском језику. Ови радови као користе скупове података аотирани од стране експерата у области права. Изазов овог рада јесте извести детекцију аргумената и њихових делова над скупом правних пресуда писаних на црногорском језику, где аотација није извршена.

Задатак овог рада јесте аутоматско препознавање аргумената у правним пресудама на црногорском језику. Приступ предложен у овом раду пролази кроз две главне фазе: детекција реченица које садрже аргумент и детекција делова аргумента на нивоу документа, односно правне пресуде. Због горенаведених проблема аотирања, у првој фази је употребљена техника полунадгледаног учења (енг. *Semi-Supervised Learning*), где је довољна минимална аотација скупа, а у другој фази је извршена кластеризација сегмената текста у кластере. Идеја технике у другој фази јесте утврдити да ли је могуће добити кластере такве да су делови аргумената (премисе, закључци) у засебним кластерима, без аотације скупа података.

Остатак рада је организован на следећи начин. У поглављу II биће представљена решења која се баве сличном темом проналажења аргумената у правним пресудама или каквим другим документима који су производи дебата. У поглављу III биће описан скуп података који је коришћен за израду предложеног решења: начин на који је прикупљен, од чега се састоји, извршене трансформације и сл. У поглављу IV биће описане методологије које су употребљене у овом раду. У поглављу V биће приказани постигнути резултати предложених решења, као и дискусија на тему грешака које модели праве. У последњем поглављу VII биће описана кратка сумаризација целог рада и приказаће се

најважнији закључци. Такође, биће представљена даља истраживања на ову тему.

## II. РЕЛЕВАНТНА ЛИТЕРАТУРА

У овом поглављу ће укратко бити описана решења која су релевантна за тему овог рада. Описаће се проблем или задатак којим се баве, скуп података који користе и постигнута решења и метрике евалуације. Треба имати на уму да не постоје радови који користе истоветан скуп података пресуда на црногорском језику, тако да није могуће у потпуности поредити решења.

### A. A Machine Learning Approach to Argument Mining in Legal Documents

У раду [8] аутори се баве детекцијом аргумената у правним документима употребом машинског учења. Цео процес деле на три целине:

1. идентификовање аргументативних и неаргументативних реченица у правним документима
2. кластеровање компоненти аргумената
3. разликовање компоненти аргумената (премисе и закључци).

Као скуп података користе корпус правних докумената *European Court of Human Rights (ECHR)* [5], аотиран од стране аутора у [6]. Задатак прве целине јесте наћи оптималан алгоритам машинског учења за класификацију реченица у две категорије: оне које садрже аргумент, и оне које га не садрже. Имплементирају *Support Vector Machines (SVM)* и *Random Forest (RF)* моделе. За евалуацију користе *precision*, *recall* и *f-measure* метрике. Закључују да *SVM* модел постиже боље резултате од *RF* модела. Друга целина описана је у наставку рада.

### B. Using Clustering Techniques to Identify Arguments in Legal Documents

Рад [3] представља наставак претходно описаног рада. У овом раду, аутори се баве аутоматским идентификовањем аргумената у правним документима. Задатак обухвата употребу алгорита кластеровања аргументативних реченица како би издвојили аргументе. Како тврде, јавља се проблем да неке аргументативне реченице могу истовремено бити део више аргумената. Тај проблем настоје да реше употребом *Fuzzy c-means* алгоритма кластеровања (*FCM*).

Као скуп података користе корпус правних докумената *European Court of Human Rights (ECHR)* [5], аотиран од стране аутора у [6]. Корпус се састоји од 20 категорија одлука и 22 категорија пресуда које су објављене пре 20. октобра 1999. године. Као репрезентацију текста користе *n-grams*, *word2vec*, *sentence closeness* и комбинацију ова три. Постигнуте резултате евалуирају тако што пореде са постојећим аотацијама, а као мере евалуације користе *precision*, *recall*, *f-measure* и *cluster purity*. Табеларно представљају постигнуте мере за сваку од репрезентација текста и за неке од докумената. Закључују да у већини

случајева *word2vec* представља најбољу опцију, са неколико изузетака где је бољи комбиновани приступ. У просеку постижу *accuracy* од 0.59, *macro f-measure* од 0.497 и *cluster purity* од 0.499.

### C. Unsupervised Argumentation Mining in Student Essay

Аутори рада [4] баве се детекцијом аргумената и њихових компоненти употребом ненадгледаног машинског учења. Идеја рада је истражити да ли је у овој области неопходно аотирати податке. Своје резултате пореде са техникама надгледаног учења које користе аотиране податке. Као скуп података користе корпус од 402 убедљивих студентских есеја, сакупљеног и аотираног за детекцију аргумената од стране аутора у [7]. Представљене методе се састоје од две фазе: детекција реченица које садрже део аргумента (главну тврдњу, тврдњу, премису) и изградњу њихових релација помоћу стабла. Закључују да метод ненадгледаног учења постиже боље резултате од метода надгледаног учења које су испробали.

### D. Neural Semi-supervised Learning for Text Classification Under Large-Scale Pretraining

Рад [11] бави се изучавањем понашања полу-надгледаног учења у области NLP-а, укључујући преобучавање језичких модела великих размера. Представљене су различите стратегије полу-надгледаног обучавања примењене за задатак класификације текста, и њихови резултати. Имплементирани су различити модели за класификацију текста, међу којима је и BiLSTM. Показало се да метод преобучавања у домену има боље перформансе са већим скупом података, док метод заснован на псеудо-ознакама има боље перформансе са мањим скупом података. Коришћењем полунадгледаног учења, постигнути су резултати од 93-94% на малом скупу података и 96% на комплетном скупу података. Показано је да је полунадгледато учење делотворно у присуству претренираних модела великих размера.

### E. Резиме

Радови А, В и С решавају проблем детекције аргумената поделом проблема на више целина: разликовање реченица са аргументом и без њега, а затим детекција компоненти аргумената, што ће бити имплементирано и у овом раду. Рад С предлаже решење које се заснива на ненадгледаном учењу – без аотирања података, што наводи на идеју да се сличан поступак примени и у предложеном решењу. Рад В оправдава приступ кластеровања као решење за детекцију делова аргумената, што ће се на сличан начин искористити и у овом решењу. У раду D представљене су различите технике класификације текста које користе метод полунадгледаног учења.

За разлику од прегледане релевантне литературе, у овом раду ће се користити скуп података који није аотиран и није на енглеском језику, што може смањити перформансе. Такође, како скуп није аотиран, неће бити начина поредити перформансе ненадгледаног и полунадгледаног са надгледаним учењем.

### III. СКУП ПОДАТАКА

У овом поглављу биће детаљно представљен поступак формирања скупа података потребног за израду предложених решења. У првом делу овог поглавља биће више речи о самом преузимању скупа података, од чега се састоји и који је његов обим. У другом делу биће описане трансформације које су извршене за потребе поменутих фаза предложеног решења. У трећем делу биће више речи о препроцесирању текста и припремама за машинско учење.

#### A. Опис скупа података

За потребе развоја предложеног решења коришћен је скуп црногорских правних пресуда. Овај скуп података је прикупљен од стране аутора у [1]. За потребе свог решења прикупљали су пресуде са сајта [2], а у свом раду детаљније описују сам поступак. Пресуде су организоване у фолдерима који одговарају врсти пресуда, а укупно је присутно 43 фолдера. За сваку пресуду постоји *html* и текст фајл, где *html* фајл садржи тагове, док је текст фајл изведен из првог одстрањивањем поменутих тагова. Како тагови не садрже корисне информације о врсти одређеног садржаја на који се односе, коришћени су само текстуални фајлови. За реализацију решења употребљене су пресуде из фолдера “Првостепени кривични предмети”.

#### B. Препроцесирање података

Да би обучавање модела било могуће, неопходно је претходно препроцесирати текстуалне податке. Овиме се из текста одстрањује све што није потребно. У овом раду су извршене следеће обраде текста:

- свођење текста на мала слова
- одстрањивање вишка размака (енг. *white space*)
- одстрањивање звездице \*
- избацивање тачке из следећих скраћеница: бг. чл. таč. кт. да их парсер не би препознао као крај реченице

Прва фаза предложеног решења захтева да се из пресуда издвоје реченице, како би се вршила класификација на реченице које садрже аргумент и оне које га не садрже. Ово је изведено употребом *nlk* библиотеке и њеног модула *sent\_tokenize*. Резултати овог модула биће прокоментарисани у поглављу V. Добијени скуп података састоји се из три колоне:

- реченица
- фајл из кога је издвојена реченица
- лабела.

За потребе полунадгледаног учења, анотирано је 194 реченице, чиме је свакој реченици додељена лабела 1 ако садржи аргумент, или лабела 0, ако га не садржи. Након класификације, реченице које су означене као аргументативне, спојене су у текст на основу назива фајла из ког су претходно биле издвојене. Ово је неопходно за наредну фазу детектовања делова аргумената.

### IV. ПРИСТУП

Предложено решење се састоји од два главна дела, односно две фазе:

- Детекција аргументативних реченица
- Проналажење делова аргумената – премиса и закључака.

У наставку ће свака од фаза бити детаљније описана.

#### A. Детекција аргументативних реченица

Пре него што се скуп података проследи неуронској мрежи, неопходно је да се свака реч у скупу представи као *word embedding* – за ово смо употребили *Word2Vec* алгоритам.

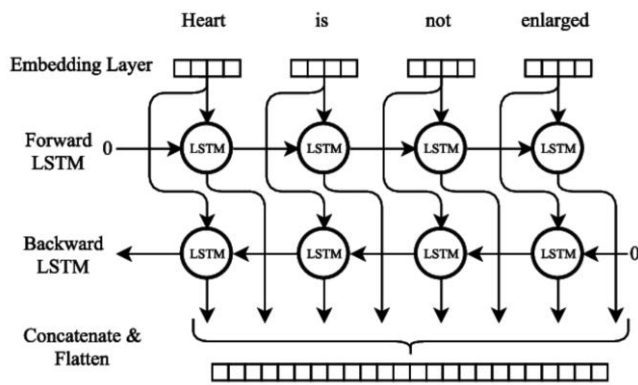
*Word2Vec* [9] је *word embedding* техника представљена 2013. године од стране тима истраживача из компаније *Google*. *Word embedding* је поступак у којем се речи трансформишу у нумеричку репрезентацију у виду векторабојки. Идеја је да се речи представе на начин на који их људи уче и разумеју – узимајући у обзир разне карактеристике речи, контекст у којој се налази, семантички однос са другим речима, итд. На овај начин се омогућава да се речи разумљиве људима прослеђују као улаз класичним моделима машинског учења.

*Word2Vec* функционише тако што конструише векторски простор у којем је свакој јединственој речи додељен један вектор. Идеја је да се забележе и семантичке и синтактичке сличности између речи, на такав начин да се њихов однос може приказати математичким операцијама над векторским репрезентацијама сваке речи.

Конечно, на реченице примењујемо два корака која су неопходна да би скуп могао да се проследи *embedding* слоју наше архитектуре. Прво, речи у реченицама се претварају у нумеричку репрезентацију једноставним енкодирањем на основу свог индекса у целом скупу речи. Након тога, све реченице се проширују на дужину најдуже реченице у скупу додавањем нула на десни крај сваког вектора који представља једну реченицу – овај процес се назива *padding*.

За класификацију реченица у овом раду коришћена је бидирекциона *LSTM* (*Long short-term memory*) архитектура рекурентне неуронске мреже – *BiLSTM*. Рекурентне неуронске мреже (*Reccurent Neural Network, RNN*) представљају класу вештачких неуронских мрежа специјализовану за обраду секвенцијалних података различитих дужина, попут реченица природног језика.

За разлику од *RNN*, *LSTM* у првом слоју уместо перцептрона има *LSTM* јединицу, која има своју рекуренцију и може се сматрати „подмрежом“ [10]. На овај начин *LSTM* омогућава дугу меморију и у пракси је ефикаснија од стандардне *RNN*. Двосмерна *LSTM* (*Bidirectional LSTM, BiLSTM*) мрежа се састоји из две *LSTM*, једна за пропагацију напред и друга уназад (Слика 1).

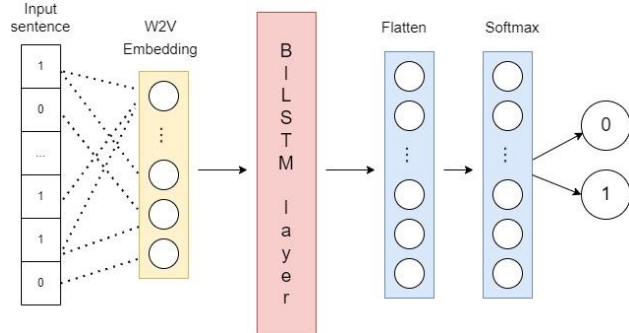


Слика 1 Двосмерна LSTM

Модел *BiLSTM* који је коришћен у овом раду (Слика 2) састоји се из 5 слојева:

1. Улазни *Embedding* слој за уграђивање *Word2Vec* модела
2. Двосмерни *LSTM* слој
3. Нормализациони слој
4. Слој за поравнање (*Flatten*)
5. Излазни слој који користи *Softmax* активациону функцију са 2 класе – једна за аргументативне реченице и друга за неаргументативне

Између нормализационог и *Flatten* слоја примењује се *Dropout* регуларизација.

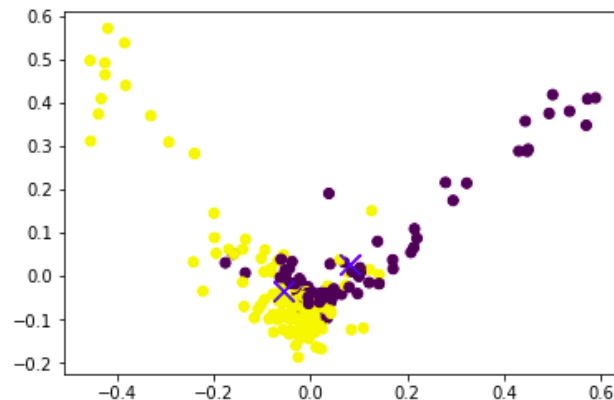


Слика 2 Архитектура *BiLSTM* модела коришћеног у овом раду

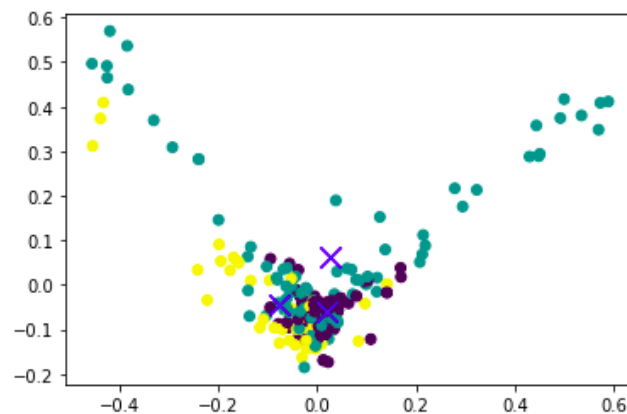
## B. Проналажење делова аргумената

Након класификације употребом полунадгледаног учења, реченице које су класификоване као оне које сарже аргумент (класа 1) издвојене су у нови скуп података који је коришћен за детекцију делова аргумената у реченицама – премиса и закључака. Због недостатка лабела за премисе и закључке у скупу података, одрађена је кластеризација делова реченица у 2 или 3 кластера. Делови реченица су добијени тако што је свака реченица подељена по зарезима. Прва варијанта поделе у два кластера би подразумевала да су добијени делови реченица или премисе, или закључци. Међутим, поделом на три кластера, оставља се простора да неки део реченице није

ни премиса ни закључак. Треба напоменути да је кластеризација вршена на нивоу документа, па тако за један документ добијамо његове потенцијалне премисе и закључке. Приликом кластеризације коришћен је *TfidfVectorizer* и *MiniBatchKMeans* алгоритам из библиотеке *sklearn*. Резултати су редуковани на 2 димензије коришћењем *PCA* алгоритма и графички приказани (Слика 3 и Слика 4).



Слика 3 Кластеризација делова реченица у 2 кластера



Слика 4 Кластеризација делова реченица у 3 кластера

## V. РЕЗУЛТАТИ И ДИСКУСИЈА

У овом поглављу биће описани постигнути резултати, а такође, анализираће се грешке модела.

### A. Фаза детекције аргумената

За обучавање *BiLSTM* модела је коришћен метод полунадгледаног учења. Мали део скупа података је лабелиран ручно и подељен на обучавајући и валидациони скуп у размери 9:1. Модел је обучаван у 50 епоха, при чему је коришћен callback *EarlyStopping* из библиотеке *keras*, ради спречавања *overfitting*-а. Након обучавања, из нелабелираног скупа података се узима 100 секвенци (реченица) и ради предвиђање. Након тога се предвиђени подаци додају у лабелирани скуп и користе за следеће обучавање. Овај процес се понавља у 5 итерација. Резултати након сваке итерације приказани су у табели. На крају је одрађена евалуација модела на валидационом

скупу коришћењем метрика *assurasy* и *loss*. За *assurasy* је достигнура вредност од 0.98 а *loss* је био 0.023.

#### B. Фаза проналажења делова аргумената

У овој фази употребљен је *K-Means* алгоритам кластеровања, који представља ненадгледано учење, односно не учи на основу лабела. Испробан је на неколико примерака фајлова. У Табела 1 и Табела 2 приказани су примери кластеризације делова реченица једног фајла из скупа првостепених кривичних предмета. Овај фајл није ручно лабелиран, већ је производ лабелирања *BiLSTM* модела.

Кластер 0	Кластер 1	Кластер 2
'od oca b. i majke i.', 'stalno nastanjen u ul.', 'suprotno čl 111 zakona o oružju ( sl.', 'sa okvirom u kojem se nalazi 8 metaka', 'na način što ga je reperitirao', 'kaznu zatvora u trajanju od 30 ( trideset ) dana', 'ako osuđeni za vrijeme od 1 (jedne) godine', 'zastupnik javne optužbe',	'negirao je izvršenje krivičnog djela i istakao da je učesnik nor-a i ima urednu dozvolu – oružni list za pištolj marke „zastava“ m-70', 'pokrenuo je parnični postupak za sprečavanje prolaza preko njegove imovine i branio je v. j. da prolazi sa kamionom preko njegove imovine i da vrši deportaciju šume.', 'koja je udaljena 20 - tak metara', 'jer je to moja imovina', 'pa je otišao u kuću', 'već ga je samo upozorio da ne ruši njegovu imovinu i da spor rješavaju sudom.',	'm. s. i oštećenog v. j.', 'neovlašćeno nosio na javnom mjestu pištolj marke „zastava m70“', 'okrivljeni m. b.', 'izašao i isti je reperitirao', 'već je bio okrenut u pravcu zemlje',

Табела 1 Делови реченица у 3 кластера

Кластер 0	Кластер 1
'po pravosnažnosti presude', 'zastupnik javne optužbe', 'u svojoj odbrani datoj tokom postupka', 'pred osnovnim sudom u beranama', 'u vidu sajli i drvenih greda – vrljika', 'uzeo iz kuće pištolj',	'- te mu na osnovu navedenih zakonskih propisa', 'te traži da ga sud ogласи krivim i osudi na kaznu po zakonu.', 'negirao je izvršenje krivičnog djela i istakao da je učesnik nor-a i ima urednu dozvolu – oružni list za pištolj marke „zastava“ m-70', 'pokrenuo je parnični postupak za sprečavanje prolaza preko njegove imovine i branio je v. j. da prolazi sa kamionom preko njegove imovine i da vrši deportaciju šume.', 'predmetnog dana je poslao kamione preduzeća da izvrše utovar grade',

Табела 2 Делови реченица у 2 кластера

Анализирањем добијених кластера, не може се јасно одредити разлика између кластера. У Табела 1 делује да се у кластеру 0 не налазе делови аргумената, док примери из кластера 1 и 2 више личе на делове аргумената. Очекивали смо да ће бити подела на премисе, закључке и остале, а не

може се у потпуности утврдити да тако и ради. Примећујемо и да кластер 1 садржи дуже рченице. У Табела 2 осим што су рченице у кластеру 1 дуже, примећујемо и више примера аргумената у кластеру 1. Очекивали смо да ће кластеровати делове рченица на премисе и закључке, а из приложеног можемо закључити да дели на аргументативне и не-аргументативне делове.

#### C. Анализа грешака

Грешке које модели праве су очекиване и могу се утврдити и објаснити разлози који до њих доводе. Један од разлога је сам скуп података, начин на који *nltk* библиотека дели текст на рченице. Најчешћи знак интерпункције који дели рченице јесте тачка, а библиотека не води рачуна о скраћеницама, иницијалима особа, бројевима и датумима који у себи садрже тачку. Тако добијамо рченице које се састоје од једног или два слова, као и рченице са једним бројем, или се иницијали особе распореде у две рченице, што нема никаквог смисла. У следећем примеру је рченица из скупа података која је неисправно подељена по тачкама које чине датум:

1. k r i v j e : što : u periodu od 09.12.2009. godine do 09.03.2011.godine, nije davao izdržavanje za svoju maloljetnu djecu, sinove m. s., a. s., o. s. i kćerku a. s., u iznosu od 100,00 eura svakomjesečno, na ruke njihovoj zakonskoj zastupnici - majci b. s. ( s. ), iako je na to obavezan pravosnažnom presudom osnovnog suda u beranama p.br ... od 09.
2. 12.
3. 2009. godine, što ukupno iznosi 780,00 eura, - čime je izvršio četiri krivična djela nedavanje izdržavanja iz čl 221 st. 1 krivičnog zakonika, u realnom sticaju.

Као што се из примера може видети, рченица на више места садржи датуме, али нису сви овако подељени. Како би се ово надоместило, испробано је проналажење скраћеница и датума помоћу регуларних израза, при чему су тачке избачене. Међутим, овиме се избаце све тачке из текста пресуде, што није пожељно.

Даље, саме правне пресуде су лоше форматиране: неке рченице су превише дугачке и тешко је обухватити њихово значење, а то додатно отежава анотацију. Ово доводи до тога да дуге рченице често садрже аргумент, док га кратке не садрже, услед чега се модели превише прилагођавају тренинг подацима. Такође, у пресудама се често употребљавају скраћенице, иницијали, што отежава разумевање текста. Грешке модела се могу приписати и недвољно анотираном скупу рченица.

## VI. ЗАКЉУЧАК И БУДУЋИ РАД

Размотрен је проблем проналажења аргумената у правним пресудама. Представљени су и неки сродни радови који нуде решење на сличну тему и предложено је решење које обухвата две фазе: детекцију аргументативних рченица и издвајање делова аргумената, односно премиса и закључака. Укратко је описан скуп података и начин на који је обрађен и аотиран за потребе

прве фазе. Испитано је неколико модела, који су тестирани и евалуирани.

Може се закључити да модел прве фазе даје изузетне резултате. Модел друге фазе даје нешто лошије резултате у смислу да се не могу јасно разлучити кластери, но овакав приступ би се могао унапредити додатним препроцесирањем и пажљивијем извлачењем особина текста.

Поређење перформанси предложеног решења са претходним сродним решењима није изводљиво јер се користе различити скупови података на различитим језицима и различите мере евалуације.

Даља истраживања би укључила додатна препроцесирања и обраду текста, додатно анотирање, као и испитивање других модела и намештање њихових хипер-параметара.

#### РЕФЕРЕНЦЕ

- [1] Trpovski, Igor, Stojanović, Nikola, Orčić, Timotej. "Pronalaženje sličnih presuda upotrebom metoda veštačke inteligencije"

- [2] Sudovi Crne Gore - <http://sudovi.me>
- [3] Poudyal, Prakash, Teresa Gonçalves, and Paulo Quaresma. "Using Clustering Techniques to Identify Arguments in Legal Documents." *ASAIL@ ICAIL*. 2019.
- [4] Persing, Isaac, and Vincent Ng. "Unsupervised argumentation mining in student essays." *Proceedings of the 12th Language Resources and Evaluation Conference*. 2020.
- [5] *European Court of Human Rights* - <https://hudoc.echr.coe.int/eng>
- [6] Mochales-Palau, Raquel, and M. Moens. "Study on sentence relations in the automatic detection of argumentation in legal cases." *Frontiers in Artificial Intelligence and Applications* 165 (2007): 89.
- [7] Stab, Christian, and Iryna Gurevych. "Parsing argumentation structures in persuasive essays." *Computational Linguistics* 43.3 (2017): 619-659.
- [8] Poudyal, Prakash. "A machine learning approach to argument mining in legal documents." *AI Approaches to the Complexity of Legal Systems*. Springer, Cham, 2015. 443-450.
- [9] MIKOLOV, Tomas, et al. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.
- [10] L. Pešut: Rekurentne neuronske mreže, 2019. <https://repozitorij.pmf.unizg.hr/islandora/object/pmf%3A8434/datastream/PDF/view>
- [11] Z. Sun, C. Fan, X. Sun, Y. Meng, F. Wu, J. Li: "Neural Semi-supervised Learning for Text Classification Under Large-Scale Pretraining"