Класификација текста заснована на српском wordnet-у

УДК 811.163.41'322.2

Јелена Граовац
jgraovac@matf.bg.ac.rs
Универзитет у Београду, Математички факултет, Катедра за рачунарство и информатику

АПСТРАКТ: У овом раду је приказана метода за класификацију текста на српском језику заснована на српском wordnet-у. Метода је вођена хипотезом да се укључивањем морфолошких, синтаксичких и семантичких информација садржаних у лексичким ресурсима може унапредити процес класификације текстова на српском језику, као једном од морфолошки богатијих језика. Коришћен је Ебарт-3 корпус који представља скуп новинских чланака на српском језику подељених у три класе: економија, политика и спорт. Метода користи паметан одабир концепата из српског wordnet-а као представника сваке од класа, а сам одабир се врши на основу вредности уведене мере за тежину која квантификује значај концепта за дату класу. Проблем флексије у српском језику је решен коришћењем морфолошког речника за српски језик. Ради евалуације приказане методе коришћени су микропросечни и макропросечни показатељи – прецизност, одзив и ф-мера. Добијени резултати су показали да се паметним избором концепата добијају бољи резултати него коришћењем свих концепата придружених доменима који одговарају класама, мада су домени дефинисани у wordnet-у, поред осталог, и због његове успешније примене на задатке класификације текста.

КЉУЧНЕ РЕЧИ: класификација докумената, wordnet, српски wordnet, морфолошки речник за српски језик.

ДАТУМ ПРИЈЕМА РАДА: 13. новембар 2013.
ДАТУМ ПРИХВАТАЊА РАДА: 18. март 2014.

1. Увод

У свету у коме живимо, интернет и дигитални запис учинили су да огромне количине сирових података постану доступне широкој јавности. Један амерички менаџер је још давно изјавио: „Рачунари су нам обећали фонтану мудрости, а ово што смо добили је поплава података" (Frawley и др. 1992). Сирови подаци, неадекватно структурирани и различитих формата, садржаја и квалитета су ретко од користи. Неопходно их је припремити, анализирати и на основу тога доћи до информација и знања која на тај начин стичу непроцењиву вредност. Истраживање података (енгл. data mining) представља интердисциплинарно поље информатике које се бави аутоматским или полуаутоматским откривањем знања у подацима. Његов основни задатак је нетривијална екстракција информација из података, и то

информација које су имплицитне, претходно непознате и потенцијално корисне. Један од основних задатака који се решавају у оквиру истраживања података је класификација.

Класификација представља пресликавање података у предефинисани скуп класа које су унапред познате. Подаци се могу класификовати ручно, али то је дуготрајан и скуп процес. Могућности и доступност брзих рачунара довели су до тога да је аутоматско класификовање постало кључни приступ ефикасном организовању и обради велике количине података. Многе технике машинског учења се ефикасно користе за решавање проблема класификације. Неке од њих су: К-најближих суседа, метода подржавајућих вектора, дрво одлучивања, наивна Бајесова метода, неуронске мреже, скривени Марковљеви модели и друго. Сматра се да је преко 80% доступних информација сачувано у текстуалном облику. Већина информација је записана природним језиком, тј. језиком којим људи свакодневно комуницирају.

Српски језик спада у групу морфолошки богатијих језика. Одликује се богатом реченичном структуром и великом слободом у редоследу речи у реченици. Такође, користе се два писма (ћирилица и латиница), правопис је фонолошки заснован и постоји правило о смештању енклитика (облика помоћних глагола и личних заменица које немају акценат) у реченици. Сви проблеми везани за сложеност језика одражавају се на процес класификације текста написаног на том језику (Vitas и Krstev 2005).

У оквиру Групе за језичке технологије на Математичком факултету Универзитета у Београду (Група) већ дужи низ година се развијају лексички ресурси за српски језик (Vitas и др. 2003). Овај рад је подстакло питање како се информације садржане у богатим лексичким ресурсима могу ефикасно искористити ради решавања проблема класификације текста на српском језику. Ова анализа представља унапређење методе за класификацију текста коју смо развили и приказали у раду "Ontology-driven conceptual document classification" (Pavlović-Lažetić и Graovac 2010).

Рад је организован на следећи начин. У оквиру главе 2 приказане су неке опште информације о класификацији текста, о начину представљања документа и начинима редукције скупа атрибута којима је документ представљен. У оквиру главе 3 описани су лексички ресурси коришћени у овом раду: српски wordnet и морфолошки речник за српски језик. У оквиру главе 4 дат је преглед резултата других аутора који се односе на технике класификације текста засноване на wordnet-у. У глави 5 приказан је корпус на српском језику коришћен за класификацију, а у глави 6 мере евалуације које се користе за проверу тачности приказане методе. У глави 7 дат је опис методе и процедура класификације. Уведене су мере за тежину концепта која одређује колико је неки концепт значајан за неку класу и мера припадности документа класи. Глава 8 даје преглед добијених резултата и поређење са приступом заснованим на доменима у wordnet-у. Закључак и даљи рад су приказани у оквиру главе 9.

2. Класификација текста

Класификација докумената може да се дефинише на следећи начин (Sebastiani 2002):

Дефиниција 1 Нека је $C = \{c_1, c_2, ..., c_{|C|}\}$ скуп предефинисаних класа и $D = \{d_1, d_2, ..., d_{|D|}\}$ скуп докумената. Класификација докумената је процес одређивања непознате функције

$$\phi: D \times C \rightarrow \{T, NT\}$$

која може имати логичке вредности T (тачно) и NT (нетачно). Вредност T функције ϕ придружена пару (d_j, c_i) означава да дати документ d_j припада класи c_i, док вредност NT, означава да дати документ d_j не припада класи c_i. Формалније, циљ је пронаћи функцију:

$$\overline{\phi}: D \times C \rightarrow \{T, NT\}$$

која што боље апроксимира непознату функцију ϕ. Ова фунција $\overline{\phi}$ зове се још и класификациони модел, класификатор или класификациона функција. Целокупна класификација се врши само на основу садржаја документа, то јест података добијених из самог документа.

Називи класа нису од значаја и представљају само симболичке ознаке (лабеле) које не утичу на класификацију (Sebastiani 2002).

Поступак класификације се може поделити у две фазе: *учење*, када се на основу скупа за учење креира класификациони модел и *тестирање*, када се на основу скупа за тестирање проверава квалитет направљеног модела. Скуп за учење и скуп за тестирање представљају независне скупове класификованих докумената.

2.1. Представљање документа

Први корак у процесу класификације је представљање документа на начин да га рачунар може разумети. Документ би требало да буде представљен тако да обухвати информације садржане у документу што је могуће боље како би се документи сличног садржаја могли поистоветити а различитог разликовати. Показало се да начин представљања документа има веома велики утицај на рад класификатора, посебно на способност генерализације (Joachims 2002). Модели представљања документа зависе од нивоа на ком се врши анализа текста. Што је виши ниво анализе текста, више се информација добија, али се тиме јако повећава сложеност аутоматског добијања таквих података. Анализа текста може бити извршена на нивоу дела речи (када се добијају морфолошке информације), на нивоу речи (добијају се лексичке информације), на нивоу скупа речи (добијају се синтаксичке информације), на семантичком нивоу (добијају се информације о значењу) и прагаматичком нивоу (добијају се информације о значењу текста с обзиром на контекст). Представљање документа преко н-грама карактера или бајтова један је од начина да се документ прикаже анализом на нивоу дела речи (Graovac 2012; Graovac to appear 2014). Ипак, најпознатија техника представљања документа је на нивоу речи, позната као „врећа речи" (енгл. bag of words) (Lewis and Ringuette 1994), где се документ представља као вектор речи које се појављују у документу. Сваки елемент овог вектора може да садржи информацију о фреквенцији дате речи у документу.

Један од основних проблема који се јављају приликом представљања докумената јесте предимензионисаност, тј. многобројни атрибути којима се документ представља. Два су основна мотива за редукцију скупа атрибута којима се представља неки документ (Joachims 2002): заштита од превише прилагођеног модела (енгл. overfitting) и високодимензионални простор непримењив за многе алгоритме. Постоје два приступа у процесу избора атрибута: *селекција атрибута* (енгл. feature selection) и *екстракција атрибута* (енгл. feature extraction). Код селекције атрибута се на основу одређеног знања бира подскуп почетног скупа атрибута, као „најкорисних" међу њима. У ове методе убрајају се елиминација стоп речи, фреквенција речи – инверзна фреквенција докумената (енгл. term frequency-inverse document frequency, tf-idf), информативност атрибута, узајамна информација, $\chi 2$ тест, снага атрибута и друго. Код екстракције атрибута се од почетног скупа атрибута ствара нов скуп као производ атрибута старог скупа. У ове методе убрајају се свођење на корен речи, лематизација, тезаурус, латентно семантичко индексирање, концептуално индексирање и друго.

3. Лексички ресурси

Лексички ресурси за српски језик се развијају у оквиру Групе већ дуго година, тако да је данас на располагању велики број различитих ресурса, развијених у значајном обиму (Vitas и др. 2003). Поред корпуса српског језика, као и вишејезичних паралелних корпуса, од посебног значаја су српски wordnet и систем морфолошких речника српског језика (Krstev и др. 2004).

3.1. Wordnet

Wordnet, који је данас познат као принстонски wordnet, развили су Џорџ Милер и његов тим са циљем да се користи као једна врста менталног лексикона у оквиру психолингвистичких пројеката (Fellbaum 2010). У оквиру традиционалних речника, лексички појмови су алфабетски уређени и за сваки појам је наведена дефиниција сваког могућег значења.

За разлику од тога, код wordnet-а су све речи којима се може изразити неки појам груписане у скуп синонима (енгл. synset, synonymous set) представљајући тако један концепт. Пројекат EuroWordnet (EWN) (Vossen 1998) дао је пројекту wordnet нову димензију уводећи вишејезичност у семантичку мрежу. Вокабулари седам европских језика су најпре организовани на сличан начин као принстонски wordnet а затим међу собом повезани преко такозваног међујезичког индекса (енгл. Inter-Lingual-Index – ILI). *BalkaNet* (Tufiş и др. 2004) је пројекат чији је циљ био развој поравнатих семантичких мрежа типа wordnet за балканске језике, и то бугарски, грчки, румунски, српски и турски, као и проширење мреже за чешки која је почетно била развијена у оквиру пројекта EWN. Основна сврха BalkanNet пројекта је развој савремених језичких ресурса за балканске језике који би омогућили нов начин приступа информацијама које потичу из балканских језика (Krstev и др. 2004).

3.2. Српски wordnet

Српски wordnet (http://korpus.matf.bg.ac.rs/SrpWN) је лексичко-семантичка мрежа српског језика (Krstev и др. 2004). Структура српског wordnet-а је у основи иста као структура принстонског wordnet-а и организована је преко чворова и релација између тих чворова. Као што је већ речено, ови чворови се у wordnet-у називају синсетови и представљају заправо скупове речи које у неком контексту имају исто значење. Свака реч у синсету представљена је ниском карактера или литералом, за којим следи значење тог конкретног литерала у конкретном синсету. Ово решење се заснива на приступу који се користи у класичним речницима говорног језика, где једној речи одговара више могућих значења, која су на посебан начин обележена. Како у wordnet-у нека реч може имати више значења, то може бити члан више различитих синсетова.

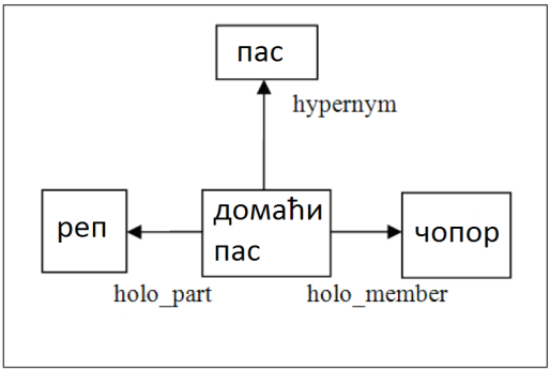
Ова база је подељена у делове према врстама речи, и то према именицама, глаголима, придевима и прилозима. Према стању српског wordnet-а из јануара 2013. године, у табели 1 је представљена расподела синсетова према врсти речи.

Врста речи	Српски wordnet
Именице	14765
Глаголи	2104
Придеви	1380
Прилози	117
Укупно	18366

ТАБЕЛА 1: Расподела синсетова према врсти речи у српском wordnet-у

Именички део базе је организован као хијерархијска мрежа именичких чворова, почевши од општих ка специфичним, која се успоставља на основу постојања релације подређености (енгл. hyponym) и надређености (енгл. hypernym) између појмова које ти чворови представљају. За један појам кажемо да је подређен другом појму ако поседује сва својства која поседује и надређени појам, али има и нека специфична својства. Концепти који нису ни сувише општи ни сувише специфични и који се налазе негде на средини хијерархије означени су као „базични концепти". У раду "Ontology-driven conceptual document classification" (Graovac и Pavlović-Lažetić 2008) уведене су мере продуктивности концепта које одређују колико неки концепт ефективно представља хијерархију којој припада.

Релација подређености и надређености свакако није једина која се успоставља између појмова. Значајне су и релације део–целина (енгл. holo_part) и члан–целина (енгл. holo_member). На пример, синсет {pas:C1x, pseto:1, domacxi pas:1} повезан је релацијом надређености са синсетом {pas:C1}, синсет {rep:2a} повезан је релацијом део–целина са синсетом {pas:C1x, pseto:1, domacxi pas:1} а овај синсет је повезан релацијом члан–целина са синсетом {suopor:1}. На сликама 1 и 2 су приказани идеализовани модел и xml репрезентација српског wordnet-а који илуструју овај пример. Релација антонимије (енгл. near_antonym) успоставља се између именичких синсетова и њоме се повезују (приближно) супротни појмови.



СЛИКА 1: Део српског wordnet-а – идеализовани модел

Друга значајна група релација успостављених између синсетова у wordnet-у јесу оне које повезују појмове који се лексикализују различитим врстама речи. Важна релација која повезује именичке и придевске синсетове јесте *бити у стању – стање нечега* (енгл. be_in_state), а један пример представља синсет {cyistocxa:1} (стање чисте особе; без прљавштине или других нечистоћа) који је повезан са придевским синсетом {cyist:1a} (који је без прљавштине или нечистоће или има навику да буде чист). Релација антонимије је честа међу придевима, па је синсет {cyist:1a} повезан релацијом *скоро супротан* са синсетом {prlxav:1} (који на себи има прљавштину или нечистоћу). Овај синсет је пак у вези са именичким синсетом {prlxavsxtina:1, necyistocxa:1} (стање некога или нечега што није чисто) преко релације *бити у стању – стање нечега*, док је овај синсет опет повезан релацијом *скоро супротан* са синсетом {cyistocxa:1}. Ако овоме придодамо и релације које се успостављају између глаголских синсетова, као што је релација *узрокује–узрокован* (енгл. verb_group) која, на пример, повезује синсетове {uspraviti:1, podignuti:3} (поставити у усправан положај) и {stajati:1a} (бити у усправном положају) јасно је да је у wordnet-у успостављена густа мрежа између чворова (Krstev и др. 2008). У табели 2 је приказана листа свих релација и број њиховог појављивања у српском wordnet-у, према стању из јануара 2013, уређена опадајуће према броју појава.

```
<SYNSET>
      <ID>ENG30-02083346-n</ID>
      <SYNONYM>
            <LITERAL>pas</LITERAL>
      </SYNONYM>
      <DEF>Bilo koji od raznovrsnih
           sisara koji obicyno imaju
           dugu nxusxku i kandye.
      </DEF>
      <POS>n</POS>
</SYNSET>

<SYNSET>
      <ID>ENG30-02084071-n</ID>
      <SYNONYM>
            <LITERAL>pas</LITERAL>
            <LITERAL>pseto</LITERAL>
            <LITERAL>domacxi pas</LITERAL>
      </SYNONYM>
      <DEF>Pripadnik Canis familiaris,
           srodan vuku, pripitomlxen od
           preistorijskog doba;
           postoje mnoge rase.
      </DEF>
      <POS>n</POS>
      <ILR>ENG30-02083346-n
            <TYPE>hypernym</TYPE>
      </ILR>
      <ILR>ENG30-07994941-n
            <TYPE>holo_member</TYPE>
      </ILR>
</SYNSET>
<SYNSET>
      <ID>ENG30-02158846-n</ID>
      <SYNONYM>
            <LITERAL>rep</LITERAL>
      </SYNONYM>
      <DEF>Upadlxivo oznacyen ili oblikovan
           zadnxi deo.</DEF>
      <POS>n</POS>
      <ILR>ENG30-02084071-n
            <TYPE>holo_part</TYPE>
      </ILR>
</SYNSET>

<SYNSET>
      <ID>ENG30-07994941-n</ID>
      <SYNONYM>
            <LITERAL>cyopor</LITERAL>
      </SYNONYM>
      <DEF>Grupa zxivotinxa koje love.</DEF>
      <POS>n</POS>
</SYNSET>
```

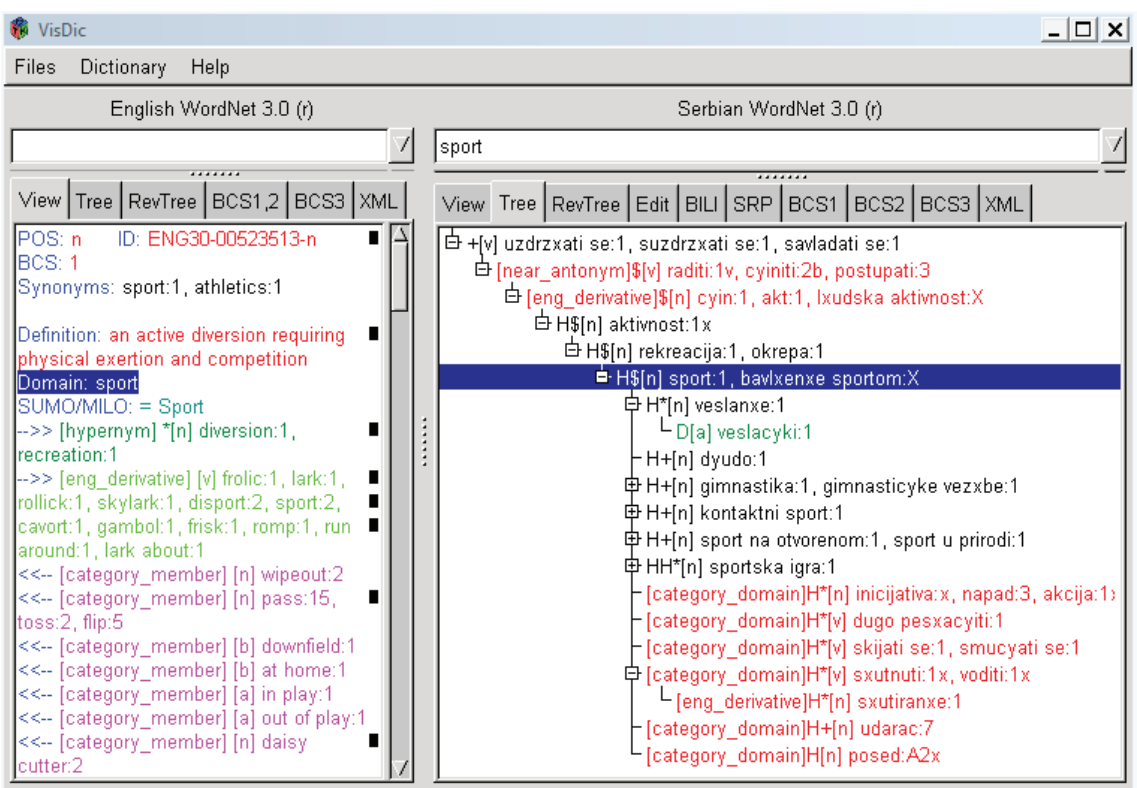
СЛИКА 2: Део српског wordnet-а – xml репрезентација

Релације у српском wordnet-у	Број појава
Hypernym	16886
holo_member	3879
eng_derivative	2926
holo_part	1560
near_antonym	762
category_domain	738
Derived	662
be_in_state	287
similar_to	244
also_see	226
holo_portion	209
verb_group	170
region_domain	82
Subevent	78
Causes	64
derived_gender	38

Релације у српском wordnet-у	Број појава
derived_pos	45
usage_domain	15
Particle	10
derived_vn	2

ТАБЕЛА 2: Расподела релација у српском wordnet-у (јануар 2013)

С обзиром на значај принстонског wordnet-а, његова структура је у више наврата проширивана додатним информацијама ради примењивости у природнојезичким обрадама. Прво проширивање односи се на проширивање принстонског wordnet-а семантичким доменима. Семантички домени представљају природан начин да се успоставе семантичке релације између значења речи које би могле да се успешно користе у разним доменима обраде природних језика (нарочито у решавању проблема класификације). Скоро сваки концепт



СЛИКА 3: Приказ концепта sport у српском и принстонском wordnet-у коришћењем VisDic алата

је анотиран барем једним обележјем домена изабраног из скупа од око двеста хијерархијски организованих домена (Krstev и др. 2008). Српски wordnet још није проширен доменима, али се они лако могу добити из принстонског wordnet-а. Код већине концепата у српском wordnet-у постоји директна веза са одговарајућим концептима у принстонском wordnet-у. Пример за концепт „sport" је приказан на слици 3. На левој страни слике може се видети концепт у принстонском wordnet-у који одговара концепту „sport" у српском wordnet-у (десна страна слике). Том концепту у принстонском wordnet-у придружено је обележје домена „sport" па се сматра да је тај домен придружен и одговарајућем концепту у српском wordnet-у.

Домени који су од значаја у случају класификације докумената класе економија, политика и спорт су:

- Економија: economy (economy, banking, enterprise, money, tax), commerce, industry.
- Политика: politics, anthropology.
- Спорт: sport (sport, badminton, baseball, basketball, football, golf, soccer, tennis, volleyball, skiing, rowing, swimming, diving, athletics, boxing, fishing, hunting, bowling).

Ови домени су изабрани са списка свих домена дефинисаних за принстонски wordnet и доступних на адреси: http://wndomains.fbk.eu/hierarchy.html.

3.3. Морфолошки речник српског језика

Систем морфолошких речника српског језика SrpMD (http://korpus.matf.bg.ac.rs/SrpMD/) (Krstev 2008, Vitas 2003), састоји се од неколико основних делова: DELAS, који представља речник простих речи у основном облику (простих лема), DELAC, речника сложених речи (контингентних ниски простих речи) и DELAF, речника облика простих речи, као и морфолошких граматика које омогућују препознавање „непознате" речи, тј. речи која није препозната на основу постојећих речника. Сваком запису у речнику простих лема (DELAS) придружена је информација о врсти речи и, ако је потребно, код флективне класе, прецизан опис промене речи. Елементима речника DELAS могу се додати морфосинтаксичке, синтаксичке или семантичке информације, као и информације о изговору. Тако је, на пример, придев „devojcyin" у речнику DELAS записан (Obradović i Stanković 2007):

devojcyin, A1+Pos+Ek (1)

што значи да је реч о придеву који припада флективној класи A1, који је присвојан (+Pos), екавског изговора (+Ек). Информације из система речника SrpMD могу се помоћу система Unitex (Sébastien 2002) користити за формулисање комплексних упита за претраживање текстова. На пример, упитом: <A+Pos-Ek> добиће се сви присвојни придеви у тексту који не припадају екавском изговору. Записи у DELAS речнику могу садржати и деривационе релације којима се повезују речи које припадају истом деривационом гнезду. Овај тип информација се раздваја знаком доње повлаке (_). На пример:

devojcyin, A1+Pos+Ek_N=4ka (2)

devojka, N618+Hum+Ek_A=2cyin

Информације које се налазе иза доње повлаке у придеву „devojcyin" (A) повезују га са именицом „devojka" (N) тако што се последња четири карактера „cyin" замене са „ка". У другом реду показано је како се, на сличан начин, именица „devojka" повезује са придевом „devojcyin". Сем тога, морфосинтаксичким информацијама (испред којих стоји знак плус), може се описати и тип деривационе везе између двеју лема. У примеру (2), из ознаке +Pos види се да је придев „devojcyin" присвојни придев именице „devojka". Ове информације из речника DELAS могу се користити за лематизацију текстова на основу произвољно изабране леме из једног деривационог гнезда, помоћу коначних трансдуктора. Такође, могу се користити за добијање свих граматичких облика неке речи.

4. Други радови из ове области

Један од најпознатијих радова о класификацији текстова заснован на wordnet-у је "Text classification using wordnet hypernyms" (Scott and Matwin 1998) где је описана метода примењена на корпусу текстова на енглеском језику. Тај рад је подстакао методу која ће бити приказана у овом раду и која ће бити примењена на корпусу на српском језику. Приказан поступак класификације састоји се од три пролаза кроз корпус. У првом пролазу, свим речима у документу из посматраног корпуса придружује се ознака врсте речи (именице, придеви, глаголи и друго). У другом пролазу, за сваку именицу или глагол, врши се преглед wordnet-а и прави се глобални списак свих синонима и хиперонима сваке од тих речи. Они концепти који се ретко јављају у корпусу искључују се из посматрања, а они преостали формирају скуп особина. Током трећег пролаза, израчунава се густина сваког концепта (дефинисана као однос броја појаве тог концепта и укупног броја речи у документу). У овом раду се дефинише и параметар *h* који представља величину генерализације, тј. колико нивоа навише треба посматрати хипериониме за дати концепт. Припадност некој класи се дефинише коришћењем добијених густина за концепте. Тако је у овом раду дат пример са две класе, Историја и Порез. Правило припадности класи се дефинише преко густине концепта „власништво". Ако је његова густина мала у неком документу, онда тај документ припада класи Историја, а у супротном припада класи Порез. Сматра се да се речи које се односе на реч „власништво", углавном користе у текстовима који се баве темом пореза и врло ретко су то неки историјски документи.

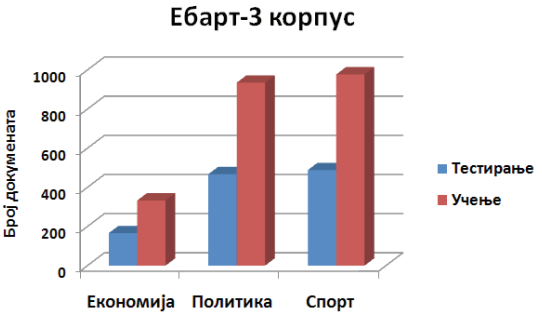
Други познати рад из ове области је "Using WordNet to complement training information in text categorization" (Rodrigues и др. 2000). У оквиру овог рада користи се wordnet за унапређивање метода заснованих на неуронским мрежама и примењује се над Reuters-21578 новинском колекцијом. Рад у коме се такође истражује утицај семантичких информација на задатке класификације текстова и истраживања података је "Text categorization and information retrieval using wordnet senses" (Rosso и др. 2004). У оквиру ових радова, wordnet се користи само зарад добијања синонима неке речи.

Метода класификације, која ће бити приказана у овом раду, представља унапређење методе истих аутора објављене у раду (Pavlović-Lažetić and Graovac 2010). Заснива се на ручно одабраним концептима као представницима класа на основу најфреквентнијих речи у документима сваке од класа. Овим концептима се придружују сви литерали синоними који одговарају концептима (синсетовима) који граде релацију подређености (енгл. hyponym) са тим концептом. Документ за тестирање се придружује оној класи за коју садржи највећи број литерала придружених концептима представницима те класе. Проблем флексије је решен алгоритамски (уз претпоставку да две речи са довољно великим заједничким префиксом представљају исту реч у различитим граматичким облицима) и у ту сврху није коришћен ниједан речник српског језика. Корпус на коме је извршена класификација је Ебарт-5 корпус на српском језику.

5. Корпус

Први корак у процесу класификације јесте прикупљање докумената за корпус и његова подела на скупове за учење и тестирање. У овом раду коришћен је корпус на српском језику назван Ебарт-3 корпус. Овај корпус је подскуп Ебарт корпуса (http://www.arhiv.rs), највеће дигиталне медијске документације у Србији. Актуелни архив је класификован на тематске целине по угледу на уобичајене новинске рубрике: унутрашња политика, спољна политика, друштво, економија, хроника и криминал, култура и забава, спорт, медији, фељтони, писма читалаца. Ебарт-3 корпус, који је коришћен у овом раду, представља насумично одабране чланке из дневних новина *Политика* који припадају рубрикама/класама: спорт,

економија и политика, издатим од 2003. до 2006. године. Сваки чланак/документ може припадати само једној класи. Има укупно 3.366 таквих чланака. Овај корпус подељен је на податке за учење и податке за тестирање у односу 2:1. Слика 4 приказује расподелу овог корпуса по рубрикама/класама, узимајући у обзир поделу на скупове за учење и тестирање. Корпус је доступан у текстуалном и xml формату.



СЛИКА 4: Расподела Ебарт-3 корпуса

6. Мере евалуације

Након завршене фазе учења, веома је важно проценити колико је добро класификатор успео да генерализује проблем на основу података за учење. Може се догодити да је модел превише прилагођен. Овај проблем се јавља када се генерисани класификатор понаша добро на подацима за учење, али подбацује на новим подацима за тестирање. Процес евалуације се састоји у поређењу унапред познате класе са оном коју је предложио класификатор. Тиме се добијају исправно и неисправно класификовани подаци.

Могући исходи кад је у питању бинарна класификација су:

- Стварно позитивни, СП (енгл. true positives, TP)
- Стварно негативни, СН (енгл. true negatives, TN)
- Лажно позитивни, ЛП (енгл. false positives, FP) и
- Лажно негативни, ЛН (енгл. false negatives, FN).

Неке од најпознатијих мера евалуације су прецизност, одзив и ф-мера. Прецизност је проценат примера који су исправно класификовани међу свим примерима који су додељени одређеној класи. Израчунава се по формули (Baeza-Yates и др. 1999):

$$\textbf{Прецизност} = \frac{\textbf{СП}}{\textbf{СП} + \textbf{ЛП}}$$

при чему СП+ЛП представља укупан број примера који су додељени предложеној класи. Одзив (покривање, енгл. recall) оцењује колико је модел успешан у покривању класе, тј. колико примера за тестирање из дате класе (или класа) класификатор може да препозна. Израчунава се по формули (Baeza-Yates и др. 1999):

$$\textbf{Одзив} = \frac{\textbf{СП}}{\textbf{СП} + \textbf{ЛН}}$$

при чему СП+ЛН представља укупан број примера који припадају стварној класи. Ф-мера представља комбинацију прецизности и одзива у једној мери као њихову хармонијску средину (Van Rijsbergen 1979):

$$\textbf{Ф-мера} = \frac{2 * \textbf{Прецизност} * \textbf{Одзив}}{\textbf{Прецизност} + \textbf{Одзив}}$$

Ове мере су дефинисане за случај бинарне класификације (када постоје само две класе). У случају када се класификација врши на више од две класе, потребно је извршити усредњавање ових мера по свим класама. То може да буде урађено на два начина: може се тражити макропросек, где се свакој класи придаје исти значај, и микропросек, где се фаворизују класе које садрже већи број докумената. Код макропросека се најпре израчунава вредност мере за сваку класу појединачно, а затим се врши усредњавање тих вредности по броју класа. Код микропросека се израчунавају вредности за СП, СН, ЛП и ЛН за сваку класу

појединачно. Затим се израчунавају вредности $\overline{СП}$, $\overline{СН}$, $\overline{ЛП}$ и $\overline{ЛН}$ као суме свих СП, СН, ЛП и ЛН за све класе редом. На крају се израчунава вредност мере за добијене сумиране вредности $\overline{СП}$, $\overline{СН}$, $\overline{ЛП}$ и $\overline{ЛН}$.

У овом раду, зарад евалуације добијеног класификационог модела, коришћени су микропросечни и макропросечни показатељи: прецизност, одзив и ф-мера.

7. Процедура класификације

У Ебарт-3 корпусу документи су подељени у три класе: Економија, Политика и Спорт. Нека су у скупу за учење те класе и документи у њима означени на следећи начин: $E = \{d_1, d_2, ..., d_{N_E}\}$, $P = \{d_1, d_2, ..., d_{N_P}\}$ и $S = \{d_1, d_2, ..., d_{N_S}\}$, при чему $N_E = 333$, $N_P = 935$ и $N_S = 977$ у конкретном случају представљају редом број докумената придружених тим класама.

Фаза учења може да се опише кроз следећи низ корака:

За сваку класу из скупа за учење формира се листа речи у основном облику, опадајуће уређена према фреквенцији појављивања у тој класи. Под фреквенцијом појављивања речи у класи подразумева се број појава свих облика те речи, у свим документима скупа за учење те класе. Тако на пример, за реч „sport" сви облици те речи који се могу наћи у морфолошком речнику за српски језик су: „sport", „sporta", „sportu", „sportom", „sportovi", „sportova", „sportovima", „sportove". Сматра се да се реч „sport" појавила 90 пута у класи Спорт ако су се сви облици ове речи појавили укупно 90 пута у свим документима за учење класе Спорт.

Овако добијеним речима се додељује ознака за врсту речи: именица, глагол или нека од преосталих осам врста речи у српском језику (заменица, придев, број, предлог, прилог, узвик, речца или везник).

Из листе се избацују све речи које нису именице или глаголи.

На основу добијене листе речи, за сваку класу се формира општа листа концепата из српског wordnet-а која ће представљати ту класу. Листа се формира тако што се у српском wordnet-у проналазе концепти који ће обухватити речи које су по својој учесталости и специфичности важне за посматрану класу. За неку реч се каже да је обухваћена концептом из српског wordnet-а ако је једнака неком од литерала синонима придружених том концепту или концептима који са њим граде неку од семантичких или лексичких релација. Од семантичких релација узете су у обзир следеће релације: подређен–надређен (енгл. hyponym_hypernym), целина–део (енгл. holo_part), целина–члан (енгл. holo_member), антонимија (енгл. near_antonym) и релација која повезује значење и домен коришћења концепта (енгл. category_domain). Од лексичких релација коришћена је релација извођења (енгл. derived, eng_derivative). Ово су уједно неке од најчешћих релација у српском wordnet-у. Пример за концепт „sport" приказан је на слици 3. Овом концепту су придружени литерали синоними „sport:1" и „bavlxenxe sportom:X". Са слике се види да је концепт „sport:1" надређен концептима „veslanxe:1", „dyudo:1", „gimnastika:1", „kontaktni sport:1", „sport na otvorenom:1" и „sportska igra:1" (гради релацију „hypernym" са њима). Релацију „category_domain" гради са концептима „inicijativa:x", „dugo pesxacyiti:1", „skijati se:1", „sxutnuti:1x", „udarac:7" и „posed:A2x". Концепт „veslanxe:1" гради релацију „derived" са концептом „veslacyki:1", а концепт „sxutnuti:1x" гради релацију „eng_derivative" са концептом „sxutiranxe:1". Дакле, када се каже да је реч обухваћена концептом „sport", то значи да је једнака неком од литерала синонима придружених том концепту или концептима који граде споменуте везе са њим. У обзир се узимају и концепти који посредно граде неку од споменутих веза са датим концептом. На пример, „sport:1" гради везу „hypernym" са концептом „veslanxe:1", а „veslanxe:1" гради везу „derived" са концептом „veslacyki:1". Дакле, и литерали синоними придружени концепту „veslacyki:1" обухваћени су концептом „sport:1".

На овај начин, свакој класи се придружује глобална листа концепата и свих литерала

који су обухваћени тим концептима, евентуално филтрираних по неком домену. Сви ови литерали се додатно филтрирају тако што се из разматрања искључују они литерали који се не појављују ни у једном документу корпуса. Због тога се уводи појам *активних литерала*, то јест литерала који се у неком од својих граматичких облика појављују у барем једном документу у читавом корпусу. Тако ће листу придружену класи чинити концепти, то јест на већ описан начин њима обухваћени активни литерали, евентуално филтрирани по неком домену.

Ради правилног избора концепата који ће чинити листу представника класе, за сваки концепт који је кандидат да буде члан израчунава се тежина која одређује колико је тај концепт (заједно са њему придруженим активним литералима евентуално филтрираним по неком домену) значајан за дату класу. Тежина концепта се израчунава на следећи начин: Нека је дат концепт kS који је кандидат за листу представника класе Спорт. Тежина овог концепта, која представља варијанту *tf-idf* мере (енгл. term frequency-inverse document frequency) израчунава се по следећој формули:

$$težina(k_S) = tf(k_S, S) * idf(k_S)$$

при чему је:

$$tf(k_S, S) = SrednjaGustinaPoLiteralu(k_S, S)$$

$$idf(k_S) = \log(N/\mathrm{df}\, k_S + 0.01)$$

где је N укупан број свих докумената у корпусу Ебарт-3, а $df\, k_S$ је број докумената у корпусу Ебарт-3 у којима се појављује барем један литерал (у неком од својих облика) придружен концепту k_S. Нека је концепту k_S придружено M_S литерала из српског wordnet-a (активних литерала обухваћених тим концептом евентуално филтрираних по неком домену):

$$k_S = \{l_1, l_2, ..., l_{M_S}\}$$

Тада је:

$$tf(k_S, S) = SrednjaGustinaPoLiteralu(k_S, S) = \frac{\sum_{i=1}^{M_S} SrednjaGustina(l_i, S)}{M_S}$$

при чему је:

$$SrednjaGustina(l_i, S) = \frac{\sum_{j=1}^{N_S} Gustina(l_i, d_j)}{N_S}$$

где $Gustina(l_i, d_j)$ представља однос фреквенције или броја појаве литерала l_i у документу d_j и укупног броја речи у том документу. N_S је број докумената који припадају класи Спорт у скупу за учење.

На основу тежине додељене концепту, одређује се да ли ће тај концепт бити члан листе која одређује класу или неће. Уколико је тежина већа од неког унапред задатог броја (на конкретном примеру Ебарт-3 корпуса показало се да је број 3 добра граница за тежину концепта), концепт се придружује листи. У супротном, покушава се са неким другим концептом кандидатом. За сваку класу се бира по десетак концепата који ће бити придружени тој класи. Овим је фаза учења класификатора завршена.

Фаза тестирања се може описати кроз следећи низ корака:

За сваки документ за тестирање и за сваку класу израчунава се мера припадности тог документа тој класи. Мера припадности датог документа датој класи израчунава се као густина појављивања у том документу свих активних литерала, евентуално филтрираних по неком домену, придружених свим концептима из листе представника те класе.

Формално, мера припадности класи може да се дефинише на следећи начин: Нека је класи Економија придружена листа концепата $E = \{k_{E_1}, k_{E_2}, ..., k_{E_n}\}$, класи Политика листа $P = \{k_{P_1}, k_{P_2}, ..., k_{P_n}\}$, а класи Спорт листа $S = \{k_{S_1}, k_{S_2}, ..., k_{S_n}\}$. Мера припадности документа, на пример, класи Спорт се рачуна по следећој формули (слично важи и за друге класе):

$$MeraPripadnosti(\mathrm{d}, \mathrm{Sport}) = \sum_{k_S \in S} \sum_{l \in k_S} Gustina(l, d)$$

при чему је Gustina(l, d) однос фреквенције литерала l у документу d и укупног броја речи у том документу. Под фреквенцијом литерала l у документу d подразумева се број појаве тог литерала и свих његових граматичких облика у том документу.

Документ се придружује оној класи за коју има највећу вредност мере припадности класи.

Коришћени алати: За добијање домена и литерала придружених концепту коришћена је *eXist* XML база података. За преглед wordnet-а коришћен је алат *VisDic*, а за израчунавање тежина концепата и саму процедуру класификације развијен је нов алат под називом *WordnetKlasifikacija* који је примењен у C програмском језику.

8. Резултати

Применом процедуре класификације описане у претходном одељку, у табелама 3, 4 и 5 су редом приказани концепти представници класа Политика, Економија и Спорт. За сваки концепт и за сваку класу приказани су и домени по којима је евентуално извршено филтрирање активних литерала придружених (обухваћених) тим концептима, тежина сваког од тих концепата за класу којој је придружен и број активних литерала евентуално филтрираних по неком домену, придружених тим концептима.

При избору концепата потребно је водити рачуна и о томе да не буде велика разлика у броју активних литерала који се придружују класама како не би дошло до погрешног фаворизовања неке класе. У описаном случају број различитих активних литерала за класу Економија је 66, Политика је 59 и Спорт је 69. Треба нагласити да један литерал може бити придружен више него једном концепту.

За сваки документ за тестирање се израчунава његова мера припадности свакој од класа. Документ се придружује оној класи за коју има највећу вредност мере припадности.

ПОЛИТИКА

А	Б	Ц	Д
izbor:4		63.78	1
partija:1a, stranka:1		30.97	3
parlament:1, skupsxtina:1		28.97	3
medxunarodan:1		28.94	1
politika:1b		22.69	1
ministar:1		18.52	41
poglavar drzxave:1, sxef drzxave:1		12.37	5
rat:1x,ratno stanxe:1		10.66	1
zajednica:1a, drusxtvo:1a	anthropology	8.90	8
politicyko telo:X	politics	10.66	1
svrstavanxe:2, alijansa:1, koalicija:1a	politics	8.90	8
glas:6,glasanxe:1		8.20	11
narod:1 nacija:1	politics	7.12	6
podrsxka:1y, potpora:2a	politics	6.47	6

ТАБЕЛА 3: Листа концепата представника класе Политика. **А** - Концепт (синсет), **Б** - Филтрирано по домену, **Ц** - Тежина, **Д** - Број придружених активних литерала

ЕКОНОМИЈА

А	Б	Ц	Д
banka:2		127.77	1
kredit:3		42.48	1
trzxisxte:2b, berza:x		41.74	2
prodaja:1, prodavanxe:1		19.65	5
industrija:1a, manufaktura:3	enterprise	19.44	5
ustanova:1, institucija:1	economy enterprise	16.80	4
dug:1		10.12	2
poduzecxe:2, preduzecxe:2	enterprise	4.99	26
novcyana jedinica:1	economy	3.53	25

ТАБЕЛА 4: Листа концепата представника класе Економија. **А** - Концепт (синсет), **Б** - Филтрирано по домену, **Ц** - Тежина, **Д** - Број придружених активних литерала

СПОРТ

А	Б	Ц	Д
klub:1b, udruzxenxe:1x		30.23	15
sezona:2, doba:2b		25.75	2
trijumf:2, pobeda:1b		22.44	2
tim:y, kolektiv:2b, ekipa:1a		19.25	2
skor:1, rezultat:2		16.86	3
lopta:2a		16.63	3
takmicenxe:1, nadmetanxe:1, natecanxe:1		8.68	2
igra:y	sport	7.2	1
oprema:1a, deo opreme:X, pribor:1a	sport	5.47	12
takmicyar:1		5.06	13
sport:1, bavlxenxe sportom:X		3.19	27

ТАБЕЛА 5: Листа концепата представника класе Спорт. **А** - Концепт (синсет), **Б** - Филтрирано по домену, **Ц** - Тежина, **Д** - Број придружених активних литерала

Проблем који се при томе може јавити јесте да један документ има максималну вредност мере припадности за више различитих класа. Корпус се међутим карактерише непреклапајућим класама (енгл. single-labeled), односно сваки документ се може придружити тачно једној класи. Оптимистичан приступ овом проблему био би да уколико се класа којој документ заиста припада налази међу класама за које документ постиже максимум мере припадности, сматра се да је документ исправно класификован. Реалистичан приступ је да се сматра да је тај документ и исправно класификован (за класу којој стварно припада) и неисправно класификован (за све остале класе за које постиже максимум а заправо им не припада). Једина разлика између ових приступа је, дакле, у броју лажно позитивних примера. Добијени резултати за оптимистичан приступ приказани су у табели 6, а за реалистичан приступ у табели 7.

А - Прецизност, **Б** - Одзив, **Ц** - Ф-мера

	СП	ЛП	ЛН	А	Б	Ц
С	450	37	38	92.40	92.21	92.31
Е	130	34	36	79.27	78.31	78.79
П	425	45	42	90.43	91.01	90.72
Макропросек				87.37	87.18	87.27
Микропросек				89.65	89.65	89.65

ТАБЕЛА 6: Резултати класификације Ебарт-3 корпуса методом заснованом на wordnet-у – оптимистичан приступ.

	СП	ЛП	ЛН	А	Б	Ц
С	450	66	38	87.21	92.21	89.64
Е	130	103	36	55.79	78.31	65.16
П	425	85	42	83.33	91.01	87.00
Макропросек				75.45	87.18	80.60
Микропросек				78.15	90.00	79.83

ТАБЕЛА 7: Резултати класификације Ебарт-3 корпуса методом заснованом на wordnet-у – реалистичан приступ.

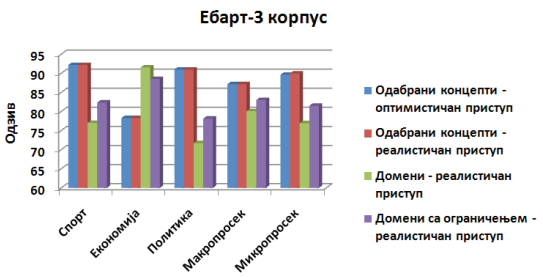
	СП	ЛП	ЛН	А	Б	Ц
С	376	46	112	89.10	77.05	82.64
Е	152	302	14	33.48	91.57	49.03
П	335	85	132	79.76	71.73	75.54
Макропросек				67.45	80.12	69.07
Микропросек				66.59	76.98	66.59

ТАБЕЛА 8: Резултати класификације Ебарт-3 корпуса методом заснованом на доменима у српском wordnet-у.

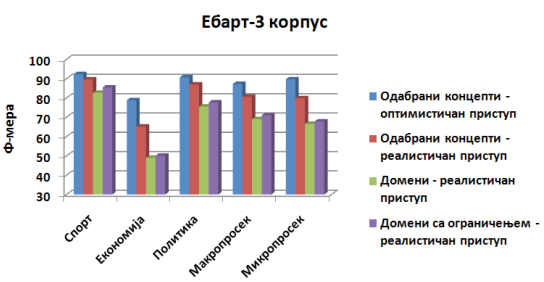
	СП	ЛП	ЛН	А	Б	Ц
С	402	51	86	88.74	82.38	85.44
Е	147	274	19	34.92	88.55	50.09
П	365	108	102	77.17	78.16	77.66
Макропросек				66.94	83.03	71.06
Микропросек				67.85	81.53	67.85

ТАБЕЛА 9: Резултати класификације Ебарт-3 корпуса методом заснованом на доменима у српском wordnet-у, са ограниченим бројем литерала.

СЛИКА 5: Поређење методе засноване на одабраним концептима са методом заснованом на доменима у wordnet-у, у термину прецизности.



СЛИКА 6: Поређење методе засноване на одабраним концептима са методом заснованом на доменима у wordnet-у, у термину одзива



СЛИКА 7: Поређење методе засноване на одабраним концептима са методом заснованом на доменима у wordnet-у, у термину ф-мере

Поређења ради, извршена је класификација корпуса на основу домена придружених концептима. Тако су класама Економија, Политика и Спорт придружени сви концепти, то јест литерали из српског wordnet-а, којима су у принстонском wordnet-у придружена обележја раније набројаних домена који су од значаја за дате класе. На овај начин добијају се

резултати класификације приказани у табели 8 (реалистичан приступ). Број активних литерала у овом случају за класу Економија је 249, за класу Политика је 111, а за класу Спорт је 66. Из табеле може да се закључи да је класа Економија, поред осталог и тиме што јој је додељен већи број активних литерала, погрешно фаворизована (302 документа су јој погрешно додељена). Због тога је извршен још један експеримент где је број активних литерала по класи ограничен на 66 (колико их има у класи Спорт), али оних који припадају концептима са највећом фреквенцијом у посматраној класи. Добијени резултати приказани су у табели 9.

Поређење резултата добијених методом заснованом на одабраним концептима приказаном у овом раду, и методом заснованом на доменима, графички је приказано на сликама 5, 6 и 7 у терминима прецизности, одзива и ф-мере. Приказане су вредности ових мера за сваку класу појединачно, а затим и њихове микропросечне и макропросечне вредности. Из свега овога може да се изведе закључак да се паметним избором концепата методом описаном у овом раду добијају бољи резултати него методом заснованом на доменима из wordnet-а. У случају методе засноване на доменима, бољи резултати се добијају ако се број активних литерала придружених класама ограничи у циљу њиховог уједначавања за све класе.

9. Закључак

Основни циљ овог рада био је да се испита како се морфолошке, синтаксичке и семантичке информације садржане у лексичким ресурсима за српски језик могу ефикасно искористити ради што боље класификације текста. Приказана је метода за класификацију текстуалних докумената на српском језику, заснована на српском wordnet-у. Проблем флексије у српском језику је решен коришћењем морфолошког речника за српски језик. Добијени резултати су показали да се паметним избором концепата из српског wordnet-а као представника класа (заснованим на вредностима тежине коју дати концепти имају за дату класу) добијају бољи резултати него коришћењем свих концепата из домена који одговарају класама, иако је један од основних разлога дефинисања домена у wordnet-у управо боља искоришћеност овог ресурса ради класификације текста.

Резултати добијени приказаном методом били би знатно бољи уколико би корпус чинили дужи документи (који садрже већи број речи) јер би тада било мање случајева да је мера припадности неког документа класи иста за више од једне класе (па би реалистични приступ дао резултате сличније оптимистичком). Такође, ова метода би дала боље резултате када би се применила на корпусу који користи нестандардан и проширен вокабулар.

У оквиру даљег рада, циљ је унапредити ову методу и тестирати је на неком другом богатијем корпусу текстова на српском језику, како по величини докумената тако и по ширини вокабулара који се користи. Иако је ова метода развијена за српски језик, она се може применити и на било који други језик за који су развијени одговарајући ресурси.

Литература

Baeza-Yates, R., B. Ribeiro-Neto, et al. 1999. *Modern information retrieval*, *vol*. *463*. ACM press New York.

Fellbaum, Christiane. 2010. Wordnet: An electronic lexical database. In *Theory and Applications of Ontology: Computer Applications*, eds. Roberto Poli, Michael Healy and Achilles Kameas, 231-243. Dordrecht : Springer.

Frawley, William J, Gregory Piatetsky-Shapiro and Christopher J. Matheus. 1992. Knowledge discovery in databases: An overview. *AI magazine*, 13(3): 57.

Graovac, Jelena and Gordana Pavlović-Lažetić. 2008. Productivity of concepts in Serbian Wordnet. In *Proceedings of the 11th International Multiconference Information Society - IS 2008*, vol. C, eds. Tomaž Erjavec and Jerneja Žganec Gros, 86–91. Ljubljana : Institut "Jožef Stefan".

Graovac, Jelena. 2012. Serbian text categorization using byte level n-grams. In *Proceedings of CloBL 2012: Workshop on Computational Linguistics and Natural Language, 5th Balkan Conference in Informatics, Novi Sad, Serbia, September 16-20, 2012*, eds. Zoran Budimac, Mirjana Ivanović and Miloš Radovanović, 93–97. Novi Sad : Faculty of Sciences, Department of Mathematics and Informatics.

Graovac, Jelena. To appear 2014. A variant of n-gram basedlanguage-independent text categorization. *Intelligent Data Analysis*, 18(4).

Horák, Aleš and Pavel Smrž. 2004. VisDic–wordnet browsing and editing tool. In *Proceedings of the Second International WordNet Conference - GWC 2004, Brno, Czech Republic, January 20 –23*, eds. Petr Sojka et al. 136–141. Brno : Masaryk University.

Joachims, Thorsten. 2002. *Learning to classify text using support vector machines: methods, theory and algorithms*. Berlin : Springer.

Krstev, Cvetana, Gordana Pavlović-Lažetić, Duško Vitas, and Ivan Obradović. 2004. Using textual and lexical resources in developing Serbian wordnet. *Romanian Journal of Information Science and Technology*, 7(1–2): 147–161.

Krstev, Cvetana, Bojana Đorđević, Sanja Antonić, Nevena Ivković-Berček et al. 2008. Kooperativan rad na dogradnji srpskog wordneta. *Infoteka*, 9(1): 57–75.

Krstev, Cvetana, Dusko Vitas, Ranka Stankovic, Ivan Obradovic and Gordana Pavlovic-Lazetic. 2004. Combining Heterogeneous Lexical Resources. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal*, eds. Maria Teresa Lino, 1103-1106. Paris : European Language Resources Association.

Krstev, Cvetana. 2008. *Processing of Serbian: automata, texts and electronic dictionaries*. Belgrade : Faculty of Philology, University of Belgrade.

Lewis, David D. and Marc Ringuette. 1994. A comparison of two learning algorithms for text categorization. In *Third annual symposium on document analysis and information retrieval*, ed. Theo Pavlidis, 1-14. Las Vegas : Information Science Research Institute, University of Nevada

Obradović Ivan and Ranka Stanković. 2007. Integracija heterogenih tekstualnih resursa. U *Zbornik radova međunarodnog simpozijuma Razlike između bosanskog /bošnjačkog, hrvatskog i srpskog jezika*, ed. B. Tošović, 596–616. Graz.

Pavlović-Lažetić, Gordana and Graovac Jelena. 2010. Ontology-driven conceptual document classification. In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval Valencia, Spain, October 25-28, 2010*, eds. Ana L. N. Fred and Joaquim Filipe, 383-386. Valencia : SciTePress.

Rodrıguez, Manuel de Buenaga, Hidalgo, Jose Maria Gomez and Agudo and Belen Diaz. 2000. Using WordNet to complement training information in text categorization. In *Recent Advances in Natural Language Processing II: Selected Papers from RANLP'97*, eds. Nicolas Nicolov and Ruslan Mitkov, 353-364. Amsterdam : John Benjamins Publishing Company.

Rosso, Paolo, Edgardo Ferretti, Daniel Jimenez and Vicente Vidal. 2004. Text categorization and information retrieval using wordnet senses. In *Proceedings of the Second International WordNet Conference - GWC 2004, Brno, Czech Republic, January 20 --23*, eds. Petr Sojka et al. 299-304. Brno : Masaryk University.

Van Rijsbergen. 1979. *Information Retrieval*. London : Butterworths.

Scott, Sam and Stan Matwin. 1998. Text classification using wordnet hypernyms. In *Usage of WordNet in Natural Language Processing Systems: Proceedings of the Workshop ; 16th August 1998*, ed. Sandra Harabagiu, 38–44. Stroudsburg : Association for Computational Linguistics.

Sebastiani, Fabrizio. 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1): 1–47.

Sébastien, Paumier. 2002. *Manuel d'utilisation du logiciel Unitex*. Champs-sur-Marne : Université de Marne-la-Vallée.

Tufis, Dan, Dan Cristea and Sofia Stamou. 2004. Balkanet: Aims, methods, results and perspectives. A general overview. *Romanian Journal of Information science and technology*, 7(1–2): 9–43.

Vitas, Duško, G. Pavlović-Lažetić, Cvetana Krstev, Lj. Popović and I. Obradović.2003. Processing Serbian written texts: an overview of resources and basic tools. In *Workshop on Balkan Language Resources and Tools, 21 Novembar 2003, Thessaloniki*, eds. S. Piperidis and V. Karkaletsis, 97–104. Thessaloniki : Greek Computer Society.

Vitas, Duško and Cvetana Krstev. 2005. Regular derivation and synonymy in an e-dictionary of Serbian. In *Proceedings of 2nd Language & Technology Conference, April 21-23, 2005, Pozna, Poland*, ed. Zygmunt Vetulani, 139-143. Poznań : Wydawnictwo Poznańskie.

Vossen, Piek. 1998. *EuroWordNet: a multilingual database with lexical semantic networks*. Boston : Kluwer Academic.